Chapter 6

Example using Generalized Linear Models

Introduction

This chapter will present an example change detection study utilizing Generalized Linear Models (GLMs). Before we begin modeling, there is an overview of the method. We then describe our sampling procedure, presenting the analysis and logic used to select a sampling design and describing how the sample was collected. We then present some initial results from the sample data. Here we test the response variable for spatial autocorrelation and discuss the implication of these tests on our subsequent modeling. We then regress the binary response of change/no-change on various functions of the radiance values. We use a stepwise regression to select the most influential variables out of 28 possibilities. The stepwise regression is done for three different link functions: the logit link, the probit link and complementary log-log link (Agresti, 1990). We find the GLMs useful in selecting the most significant function of radiance values to use for predicting change areas. The resulting models use a combination of variables that are not intuitively obvious and would not generally be included using traditional method.

Overview

Before we begin with a detailed description of each step in our example, we will describe the general procedure for using GLMs for satellite-based change detection. First, we need to collect a sample of "ground truth" or reference data. The sample is based on a set of particular ground locations. For each location we will determine if the land cover has

changed by interpreting the sample areas on the air photos. We do this from the digital air photos (the DOQs) for 1994 and from the hardcopy air photos for 1988. For each point in the sample we therefore have the land cover type for each time. For each point, if there is a different land cover type for the two time periods then the response for that point is a 1. If there is no change, the sample point will be assigned a response of 0. The next step is to relate the 0 and 1 responses from the sample points to the radiance values from the satellite data. So, for each point in the sample we obtain the satellite radiance values for that area. From this, for each sample point, we have a paired set:

{Change determination from reference data :

Satellite radiance values from corresponding area}

Similar to a standard regression procedure, for every point in the sample we have one response (the "Y" value) and a set of possible explanatory variable (the "X" values). Using the LOGISTIC procedure in SAS we can determine which function of the radiance values does the best at discriminating between the 0's and 1's – the unchanged and changed sample areas. The objective is to find a function of radiance values that can be used to predict where changes have occurred. Once an appropriate function is found, this function will serve as a model from which to predict the probability of change for the entire satellite image. That is, we use the sample points, for which we have DOQ/air photo reference data coupled with the radiance values, to develop a model to estimate the probability of change for every pixel in the image.

Initial autocorrelation analysis

Recall that, for GLMs, the data making up the random component are assumed to be independent. The implication of this assumption is that we want our change/no-change data to be independent. The sampling procedure used to collect the data should be set up to increase the chance of the data being independent. We addressed this concern by doing preliminary investigation on the spatial nature of the reflectance values. This was based on

our belief that the random, change/no-change, component is tied to the reflectance values and the spatial nature of the reflectance values can guide toward a sampling procedure that will result in an independent change/no-change response variable.

The radiance values for adjacent pixels are *not* independent. Adjacent pixels represent adjacent areas on the ground. These areas can share soil type, hydrological regimes, elevation, and slope, as well as any other factors that affect plant communities (Barbour, Burk, and Pitts, 1987). Also, adjacent pixels can share zoning districts, policy regulations, and social demographics. Through the characteristics shared by adjacent pixels comes the increase chance that adjacent pixels also share land cover type and, thus, similar spectral radiance values. Let us call this "ground-related" correlation.

As we discussed chapter 3, the reflection from most land cover is somewhat diffuse and scatters into the surrounding area (Lillesand and Kiefer, 1994, Section 1.4). Also, the atmosphere interferes and scatters the ground reflectance. This implies that the reflectance from the area represented by one pixel contributes -- through its own scattered reflectance and the atmospheric scattering -- to the observed at-sensor reflectance for adjacent pixels. Let us call this "reflectance-based" correlation. Together, these two types of spatial correlation cause autocorrelation between adjacent and/or nearby pixels.

One of the most common techniques used to assess spatial autocorrelation is the semi-variogram. This is a plot that shows the relationship between autocorrelation and the distance between points. As discussed intuitively above, we can expect two nearby pixels to be more correlated than two pixels that are far apart. So, in a plot that shows autocorrelation against distance, we would expect the autocorrelation to become less as the distance becomes larger. An empirical variogram plot has *"lag distance"* along the x-axis and the *"average squared difference between points at that distance"* along the y-axis. The lag distance is how far two points are apart. The values for the y-axis are defined by the following equation (Isaaks and Srivastava, 1989, p. 142):

$$Variogram\,Value = \frac{1}{2N(h)} \sum_{(i,j)|h_{ij} \approx h} (v_i - v_j)^2 \qquad \textbf{(eq. 6.1)}$$

In this equation, ***h*** represents "lag distance" and the summation is for all pairs that are "h" units apart. With the y-axis defined as such, lower values represent stronger autocorrelation. That is, lower values come from smaller differences, smaller differences imply more similar data, and more similar data implies stronger correlation. In many situations the empirical values will start low and increase until they reach a stable plateau. This plateau is referred to as the "sill" and the distance at which the sill is met is referred to as the "range" (Isaaks and Srivastava, 1989, p. 143). We will use the term "autocorrelation structure" to refer to the whole of the range and sill parameters as well as the general shape of the variogram. Figure 6.1 shows a hypothetical variogram model.

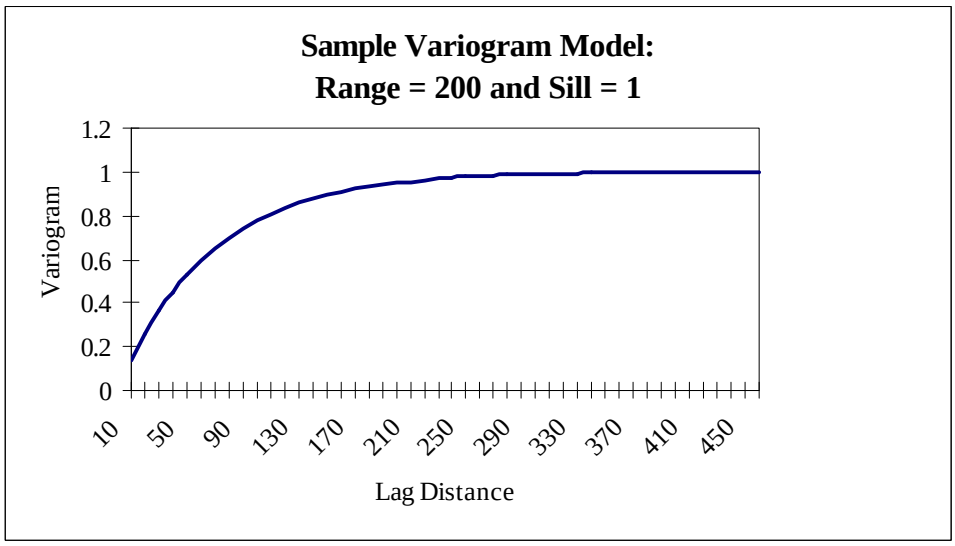


Figure 6.1: Hypothetical variogram model

The variogram has been used to assess satellite image data (Jupp *et al*., 1988; Woodcock *et al*., 1988 and 1988b). Rossi *et al*. (1994) used variograms of TM reflectance values in a Kriging application to interpolate reflectance values in shadow areas. (Note, Kriging is a geostatistical weighted average interpolation technique where the weights are determined by the variogram model, see Isaaks and Srivstava, 1989.) Lacaze *et al*. (1994) analyzed variograms of vegetation indices derived from SPOT (*System Pour l'Observation de la Terre*) and TM satellite data to examine landscape patterns. To assess the effect of pixel

size on autocorrelation structures, McGwire *et al*. (1993) used variograms from vegetation indices derived from TM data while Atkinson (1993) uses variograms from airborne multispectral scanner (MSS) data. Congalton (1988) used variogram values to assess patterns in classification error for different types of landscape. While each of these studies contains complex, in depth analysis, taken together they can be summarized by the following points:

- digital image data do exhibit spatial autocorrelation
- autocorrelation structures differ based on spatial resolution (pixel size) and landscape pattern
- spatial correlation can affect the information derived from the image

The first point implies the need for us to consider the spatial structure of our image data. The second point has led us to consider different areas from each of our two subset images for autocorrelation analysis. The third point implies the need for us to consider the autocorrelation structure in our sample design and model development.

We collected data from three transects in both the North/South and East/West direction for each of the two time periods for each of the two study areas. This gave 6 transects per image. This was done for each study area, for each time period, resulting in 24 variograms. Each variogram is presented in Appendix D. There are three main conclusions we derive from our interpretation of the empirical variogram plots. These are presented in the following three subsections.

Different spatial structures between Raleigh and coastal imagery

Within the Raleigh image, for both 1988 and 1994, the spatial autocorrelation of radiance values has a range of around 1000 meters. For the coastal scenes, for some bands, the autocorrelation appears to continue along the entire transect. This is particularly true for the infrared bands. This may be due to the different landscape patterns that are causing the differences in autocorrelation structure for the two areas. The Raleigh area landscape is more heterogeneous with fewer large contiguous areas than found in the coastal area. The large contiguous water and forested areas within the coastal scene may be the cause

of the long ranging autocorrelation for this area. There is a patch of haze in the 1994 coastal scene and one might suspect this to cause some long-range correlation. However we see a similar structure for the 1988 scene, which was free of haze. With this, we believe the differences to be due to differences in the landscape.

Limited range on the reflectance-based correlation

For both areas, the major incline on the empirical variograms is within the first few hundred meters. We believe this represents the distance needed to overcome the reflectance-based correlation and believe the reflectance-base correlation dies off after the first several hundred meters.

Collecting a systematic random sample

Although variogram structures are different for the two areas, each does have a generally concave shape (similar to the hypothetical variogram in figure 6.1). Haining (1990) uses several references (Ripley, 1981; Cochran, 1977; Quenouille, 1949) to conclude that variogram structure can help determine the optimal sampling scheme. Given that there is autocorrelation present and the general concave form of all of the empirical variograms, the recommendations of Haining (1990, section 5.3.2) are to conduct a systematic sample. These recommendation are based on several studies which show that systematic sample will outperform simple random sampling as long as there is not strong periodicity in the data. Because we do not see any strong wave-like pattern in our empirical variograms (Appendix D), we felt safe in conducting a systematic sample.

In addition, we believed that the systematic sample was our best way to do a truly random sample such that no two points were closer than a given distance. Although the sample grid is systematic it will be randomly placed over the image and so each location on the ground is just a likely to be included in the sample (Cochran, 1977, Chapter 8). By dividing the area for which we have reference data into a 20 by 13 grid we will obtain 260 points and insure a distance greater than 800 meters between each pair of sample points.

We believe the distance between sample points is far enough to go beyond the "reflectance-based" autocorrelation. A less dense grid may have decreased the chances of autocorrelation in our response variable. However, we wanted to collect a sample size sufficient enough to yield reliable modeling. Previous studies show that the percentage of change areas can range from 5% to 15% (Khorram, in press). With this, a 260-point sample should yield from 13 to 39 change areas. This sample should be enough to avoid have a sparse set of locations that have changed (McCullagh and Nelder, 1989, Section 4.4.5). Our logic for choosing a systematic sample follows very closely the logic of Pereira and Itami (1991) in their use of logistic regression to model Red Squirrel habitat.

Prospective vs. Retrospective Sampling

We decided to sample randomly over the study area with no predetermined amount of samples falling into the "change" category. This sampling method is referred to as "prospective" sampling (Agresti, 1990, p.13; Cox and Snell, 1989, section 4.2). An alternative to this would have been to sample a fixed amount of changed points and a fixed amount of unchanged points, then observe the radiance values for each of these points. The alternative method is referred to as "retrospective" sampling (Agresti, 1990, p 12; Cox and Snell, 1989, section 4.3). Retrospective sampling can be an attractive method of sampling when the change of one of the binary response variables in very low. That is, a prospective sample may need to be prohibitively large in order to find only a few observations of the rare response.

We chose to use prospective sampling for two reasons. The first is that is allows you to obtain an estimate of the proportion of areas that have changed. Secondly, for certain link functions the estimates derived from a GLM require a prospective sample. For a binary response, the logit function can be applied to data collected prospectively or retrospectively. However the probit and complementary log-log link function can only be applied to data collected prospectively (SAS, 1995; Agresti, 1990, section 4.2.2). This is due to the nature of the link function. The probit and complementary log-log links require

prospective sample to obtain an estimate of the marginal proportions of the response variable (Cox and Snell, 1989, section 4.3; McCullagh and Nelder, 1989, Section 4.3.3). Since one of our goals is to explore each of these three different link functions, we needed to sample prospectively. In a case where the chance of change is, say, less then 1%, the cost of limiting the modeling to the logit link function would probably be worth the gains in sampling efficiency. But note that you would still not obtain an estimate of the proportion of areas that have changed. Several existing studies have used retrospective sampling to gather data used in a logistic regression (Narumalani *et al*, 1997; Pereira and Itami, 1991). However, these studies were limited to using the logit link function.

Sample Data Collection

This section describes the sampling grid used to determine the location of each sample point, how the reference data were collected at each sample point, and then how the radiance values were collected for each sample point.

The sampling grid

Based on the conclusions from the autocorrelation of the radiance values, as stated previously, for each study area we collected 260 points from a randomly placed 20 column by 13 row grid. The grids were randomly place by selecting random numbers between 10 and 50 from a random number table (Steel and Torrie, 1980, Table A.1). The random numbers were used to move to the East and South of the upper left corner of the reference data. Starting from this point, a uniform grid, 20 x 13, was set to cover the remaining reference data area. This was done using the "Coordinate Calculator" within Imagine™. This module will place a grid over an image by dividing the image into the number of intervals specified by the user. This produces a list of X and Y coordinates. We converted this list into an ArcINFO™ point coverage. (This was done by saving the coordinate list into database file, which we imported into ArcView™. From ArcView™ we generated a "shape" file. Then, in ArcINFO we converted this into a point coverage

using the “shapearc” command.) Figure 6.2 displays a schematic of the sampling grid and its random placement.

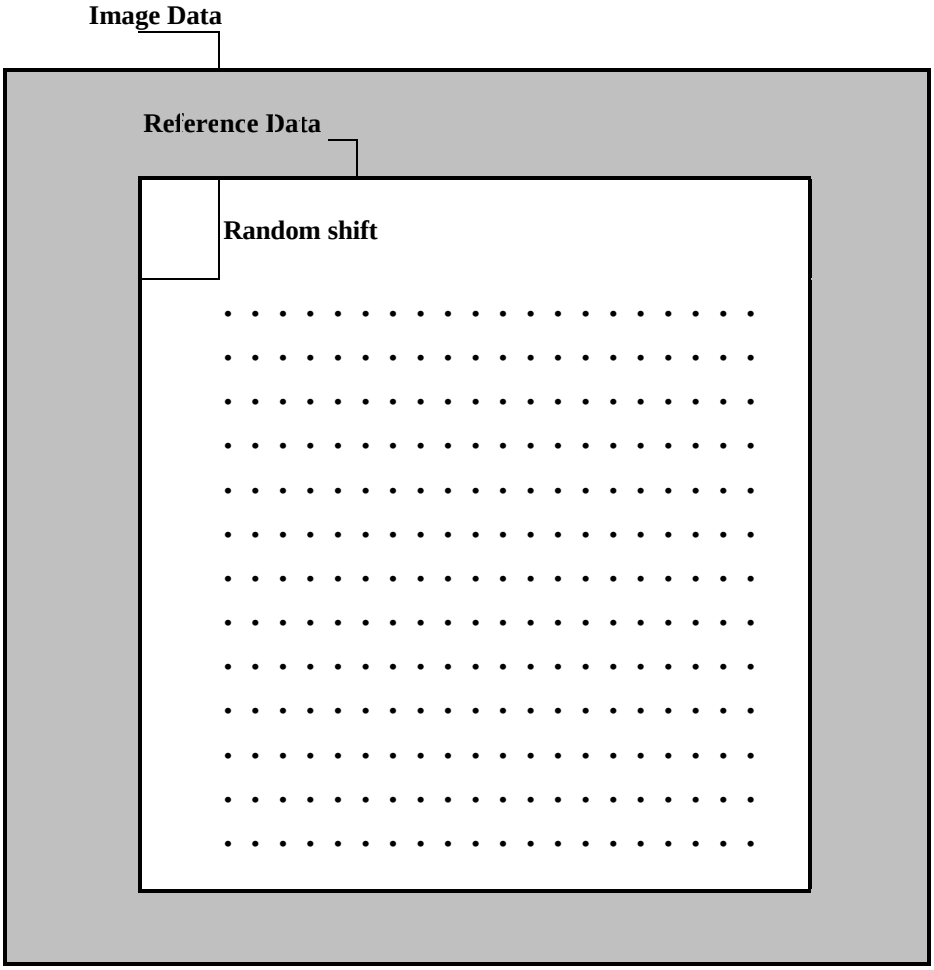


Figure 6.2 Schematic of sampling grid

Collecting the reference data

As noted in the "Materials" chapter (Chapter 2) the DOQs are geocorrected digital data. So it was possible to view these data within Imagine™ using a digital file containing the coordinates of the sample points. Once located on the DOQs, the sample points were interpreted to determine the land cover for the 1994 time period. Then each point was visually located on a 1988 photo to determine land cover for that point at that time. The alternative to visually locating points on the 1988 photos is to plot the points on a paper map, such as a USGS 7.5 minute topographic map. Then use a Stereo Zoom Transfer

Scope™ (Avery and Berlin, 1992, p.87) to locate those points on the photo (see Bauer *et al.*, 1994, and Khorram, 1995, for two studies which utilized this approach). Cook and Pinder (1996) have shown the error associated with using USGS 7.5 minute topographic maps as compared to using the Global Positioning System (GPS) to located points for image rectification. While there may be some error introduced by visually locating the points on the hardcopy air photos, we eliminate the steps of using the paper map and the Stereo Zoom Transfer Scope™. In addition, the one-meter resolution of the DOQ's gave a clear location of sample points, which were easily located on the hardcopy photos. DOQs are a relatively new product and we believe they offer exciting opportunities as geocorrected, digital, one-meter resolution reference data.

For each sample point an area of approximately 1 acre centered on the sample point was interpreted. This area corresponds roughly to the size of a two-by-two square of pixels. That is, for our reference data we consider the minimum mapping unit to be one acre. In determining the land cover at each point we used the following classification scheme:

Table 6.1: Land cover classification for reference data

Land Cover Classification
Deciduous Woody Land
Evergreen Woody Land
Pine Plantation (for the coastal area)
Mixed Woody Land
Developed Land
Cultivated Land
Grassland
Bare Land
Emergent Wetland (for the coastal area)

This classification scheme is based the C-CAP protocol classification scheme (Dobson *et al.*, 1995, Chapter 2). It is simple enough to allow easy classification on the reference data yet detailed enough to determine the changes of interest. That is, for the Raleigh scene this scheme will allow us to determine which points have moved from a vegetated area to urban development and for the coastal area we can determine if significant forest

management (such as clearing, planting, or growth) has occurred. The classification is an intermediate step to get to the ultimate binary response variable. If the land cover classification is different for the two time periods then that sample point is assigned a 1. If the land cover class is the same for each time period then that sample point is assigned a zero. For each area we have a zero or one assigned to each of the 260 points in the sample.

Collecting the radiance values

The Landsat TM image data collected for each point consists of 36 different values for each time period. This resulted from collecting the six original band radiance values and the six "bands" from the tassel cap transformation. (The tassel-cap transformation was discussed in Chapter 2 when it was used for our atmospheric normalization procedure.) These 12 values were calculated for each pixel on which the sample point fell, as well as from the average of a three-by-three set of pixels centered on the sample point and the average of a five-by five pixel area centered on the sample point. The 36 data values taken from each sample point is depicted in figure 6.3.

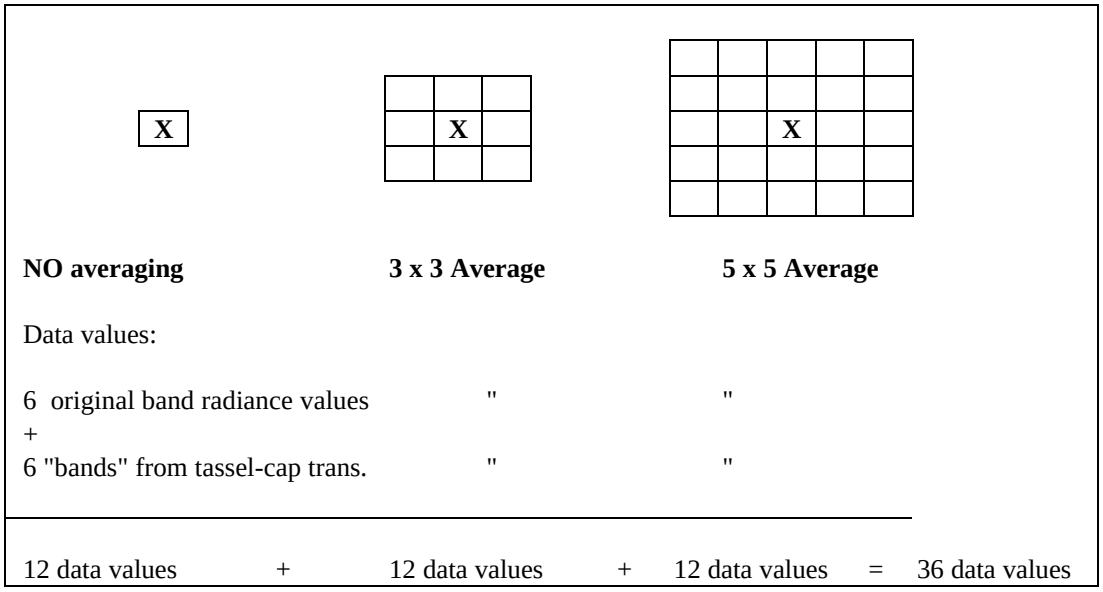


Figure 6.3: Diagram depicting the 36 image values collected for each point

The method used to collect the image data was somewhat involved. Recall that we have a digital (ArcINFO™) point coverage for the sample points. Using a macro created by Casson Stalling at the Computer Graphic Center, we can query an Arc "Grid" to select the grid value for each point in the point coverage. This implied the need to export the image data as grids using the Export module in Imagine™. Once a grid was created for each of the six bands of radiance values and for each band of the tassel-cap transformation, the macro was run to collect the image data for each of the 260 points. We then used the ArcGRID™ "focalmean" operator to filter each grid to produce a grid where each pixel contains the average value for a given neighborhood. This was done for a 3x3 neighborhood and a 5x5 neighborhood. By using the macro we were able to query each of these grids to generate a database containing the 36 different radiance values from the grids. For each area we then had a database with 36 columns and 260 rows. The rows correspond to each point in the sample.

Initial Sample Results

This section will present the result of the reference data collection and test for autocorrelation in the reference data. The following section will present the modeling that will link the response variable to the radiance values.

Reference data results

The change/no-change (black-white) data are presented in figure 6.4 for the coastal area and figure 6.5 for the Raleigh area. Note that in these figures each cell represents a sample point and is not to scale since each point represents an area of roughly 30 x 30 meters while each point is approximately 800 meters apart. However, these figures do show the distribution of change areas over the sample. Perhaps the most straightforward piece of information from the sample is the percentage of sample points that have changed. From our sample of the coastal area we found that 38 out of the 260 sample points, or 14.6%, had experienced changes in land cover. For the Raleigh area 22 out to of the 260 sample points, or 8.5% had experienced change. Most of these changes were

clearly visible on the air photos. For the coastal area, the changes were mainly either growth or cutting of pine forests. Either young pine trees had started to grow on clear land or forested areas had been cleared. The changes in the Raleigh area were also apparent, yet somewhat more diverse. An area in the Northwest has been damned which resulted in a small lake. Some areas had been bare land and experienced vegetation growth while some areas had been vegetated in 1988 and cleared and converted into developed or grassy areas sometime before 1994. Again, the classification scheme was selected so that the land cover type was easily determined from the air photos. Although there may be some errors in the interpretation, we will assume the reference data to be accurate. Justification for this is due to high resolution of the DOQs and air photos and the rather simple classification scheme.

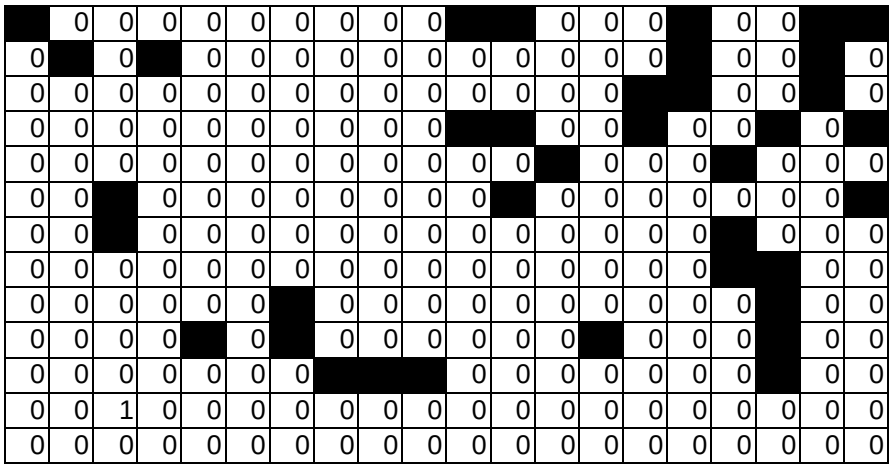


Figure 6.4: distribution of change areas from the coastal area

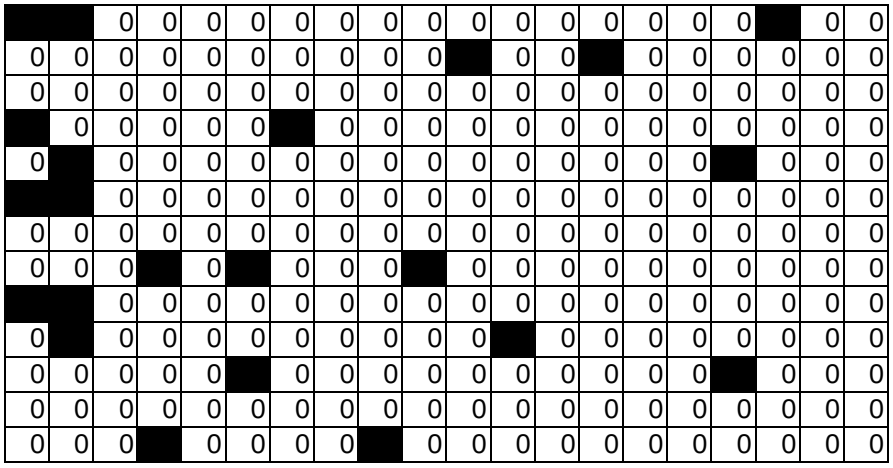


Figure 6.5: distribution of change areas from the Raleigh area

Testing independence for the response variable

As discussed in chapter 5, GLMs assume that the response variable is independent. To test this assumption, we conducted a "join-count" test to see if there was spatial correlation in the observed changed areas. We conducted a "black/black" or "BB" test as described by Upton and Fingleton (1985), which will test if neighboring locations are more likely to display the same color. For our data a "black" cell is one which has changed and a white cell is one that has not changed (as depicted in figures 6.4 and 6.5). We considered a "join" to be two black cells adjacent side-by-side or up and down. Upton and Fingleton (1985) refer to this as a "rooks" definition for adjacency, making the analogy to the game of chess. The BB test will determine if there are more BB joins than expected by a random arrangement of black and white cells. We will use a normal approximation, or a "Z" Score, to assess the significance of the results. More BB joins result in higher "Z" scores. Larger "Z" scores will have lower p-values and low p-values indicate a small chance of observing so many joins if the data are randomly distributed. That is, small p-values indicate there are more BB joins than one would expect by chance. So, for small p-values we reject the null hypothesis that the data are randomly distributed and assume that there is some autocorrelation in the data. For the Raleigh area, since there is a relatively small number of BB joins, we also assess the significance using the Poisson approximation (Upton and Fingleton, 1985, p. 163). The p-value for the Poisson approximations has the same interpretation as the p-value from the normal approximation. The results of the BB join-count test are given in table 6.2.

Table 6.2: Results of join count test for spatial independence

	Coastal Area	**Ralegh Area**
# of rows	13	13
# of columns	20	20
# in sample	260	260
# of black cells	38	22
# of white cells	222	238
Expected value		
of BB joins	10.16813187	3.341164241
Variance		
of BB joins	7.558979908	2.829386507
Observed BB	**18**	**5**
Normal Approx:	2.6667546	0.688931034
p-value	**0.0038294**	**0.245433261**
poisson	**not applicable**	**0.122126912**
p-value		

In this table we see that the number of BB joins observed for the coastal area result in a very small p-value and, so, the number of BB joins is larger than would be expected by chance. With such a small p-value we would reject the hypothesis of a random distribution of change areas and conclude there is some autocorrelation in the change response variable. The implication of this on our modeling is the need to be cautious with our interpretation of the variance terms and the related p-values that are used in model selection. As stated in Chapter 5, we can proceed with using GLMs on the data set. Once we have decided on a model we will do further autocorrelation analysis on the residuals from that model. This analysis is presented in the Modeling section of this chapter.

For the Raleigh area the p-values from both the normal and Poisson approximation are not small enough to reject the hypothesis of random distribution and so we will accept that there is not autocorrelation in the change response for the Raleigh area. So for the modeling in the Raleigh areas we can proceed with standard GLM analysis.

Modeling

It is obvious by now that we will be using GLMs to model the data. We will use the logit, probit, and complementary log-log links. For each of these link functions we will do stepwise model fitting to determine which of the 28 input variables can be used to explain the "change/no-change" response data. We will first describe the 28 explanatory variables and then present the modeling results.

The explanatory variables are summarized in the following table.

Tabel 6.3: Explanatory variables used in the stepwise model fitting

Variable Number	**Abreviated Notation**	**Decription**
1	dif1	difference in band radiance values for band one
...through...	...	...through...
6	dif6	difference in band radiance values for band six
7	dif-tas1	difference in tassel-cap band one
...through...	...	...through...
12	dif-tas6	difference in tassel-cap band six
13	absdif1	absolute difference of radiance values, band one
...through...	...	...through...
18	absdif6	absolute difference of radiance values, band six
19	absdif-tas1	absolute difference of tassel-cap band one
...through...	...	...through...
24	absdif-tas6	absolute difference of tassel-cap band six
25	vector	square root of the sum of the squared radiance difference*
26	vector-tas	square root of the sum of the squared tassel-cap differences
27	difNDVI	difference in NDVI*
28	absdifNDVI	absolute difference in NDVI

* equations provided in document text

Variables 1 through 6 represent simple differences in radiance values. Each difference is the 1988 data subtracted from the 1994 data. Variables 7 through 12 are also a simple difference but for the different bands of the tassel-cap transformation (Crist and Cicone, 1984; Crist and Kauth, 1986). The next 12 variables, numbers 13 through 24, represent the respective absolute values of the simple differences. The "vector" variable is a

Euclidean distance measure calculated with the equation

$$\textbf{\textit{vector}} = \sqrt{\sum_{i=1}^{6} (x_{i,94} - x_{i,88})^2} \; ; \qquad \textbf{(eq. 6.2)}$$

in which $x_{i,\,date}$ represents the radiance value for band ***i*** at time period ***date***.

The "vector-tas" variable is calculated from a similar equation but instead of radiance values we use the different bands of the tassel-cap transformation.

The "difNDVI" variable represents a simple difference in the Normalized Difference Vegetation Index (Jensen, 1996, p. 182). This variable is calculated with the equation

$$\textbf{\textit{difNDVI}} = \left(\frac{x_{4,94} - x_{3,94}}{x_{4,94} + x_{3,94}} \right) - \left(\frac{x_{4,88} - x_{3,88}}{x_{4,88} + x_{3,88}} \right) \qquad \textbf{(eq. 6.3)}$$

again, $x_{i,\,date}$ represents the radiance value for band ***i*** at time period ***date***.

We have used relatively simple functions for our explanatory variables. This is because our main objective is to describe how GLMs can be used for change detection modeling. Future work could be directed at using more complex variables in GLM-enhanced change detection studies. However simple, these variables have been used in recent studies (see Chapter 4). Additional reasoning for using this particular set of variables is based on the conclusions of Coppin and Bauer (1996, p. 229) in their review of change detection methods for monitoring forest ecosystems. They conclude “vegetation indices are more strongly related to changes in the scene then the response of a single band”. Conversely, Green *et al*. (1994) found differences in band 7 to be more suitable for monitoring changes in forested areas than a difference in a vegetation index. (The index was a ratio of band 3 over band 4.) It seems difficult to draw a general conclusion on which specific band or vegetation index is best for satellite-based change detection. Different studies produce different results. This is because different scenes and different landscapes have different

characteristics. For each of the 28 explanatory variables listed in table 6.3 there are reasons to believe that each could be used to detect change. Differences in the individual bands are essentially differences in “color”; where color goes beyond the visible spectrum and into the infrared bands. Differences in the first band of the tassel-cap transformation can be interpreted as differences in brightness. Differences in the second tassel-cap band can be thought of as differences in greenness. Differences in the third band are attributed to different moisture conditions (Jensen, 1996, pp. 179 - 182; Crist and Cicone, 1984; Crist and Kauth, 1986). The fourth band is related to haze (Lavreau, 1991). However, the fifth and sixth bands have not been associated with any particular scene characteristics (Crist and Cicone, 1984). So, it is somewhat difficult to interpret differences in these bands. The absolute value of the differences makes more sense for the simple “change/no-change” variable in that these variables do not indicate directional changes -- only the magnitude of the change. However, the simple difference could be important because it may be that only one of the extremes (either positive or negative) is related to change. Each of the “vector” variables is a combined difference over all of the bands. The normalized difference vegetation index (NDVI) is a popular measure for vegetation and differences in NDVI can be interpreted as differences in healthy vegetation (Jensen, 1996, pp. 179 - 182; Lillesand and Kiefer, 1994, p. 5-6). All of these variables could justifiably be used to detect change. One of the major benefits of using GLMs is to provide the analysts with a quantitative means of choosing among these variables. By using differences in single bands as well as differences in single tassel-cap bands and also considering the "vector", "vector-tas" and "difNDVI" indices we can use the GLMs to test which is more strongly related with the observed changes.

The models we ran were all done within the SAS™ system using the LOGISTIC procedure. Within this procedure you can specify each of the three link function discussed previously: logit, probit and complementary log-log (note however that SAS™ refers to the probit link with the name “normit”). Within the procedure you can also specify that the model be fit with a stepwise model selection method (SAS, 1995). Stepwise model building starts with the most significant variable and continues to build the model one

variable at a time, choosing the most significant variable of those not yet included in the model. At each step in the model building all included variables are considered and those which are not significant are removed (Neter *et al.*, 1989, pp. 453 - 458). The analyst chooses the level of significance for adding and removing variables. We decided on a stepwise procedure with the level of entry set to alpha = .1 and used the default level for removal at alpha = .05. This combination allows variables that are significant at the 10% level to be brought into the model but limits the variables in the final model to those with 5% significance levels.

We will present tables that summarize the results of the stepwise modeling. We will compare the Akaike Information Criterion (AIC), Schwarz Criterion (SW), the "-2 Log Likelihood" statistic (-2 Log L) and its p-value, the Score statistic and it p-value, and the "concordance" measure. The AIC and SC are useful for comparing different models (SAS, 1995). Better fitting models will have lower AIC and SC statistics. The -2 Log L and Score statistic are used to test the null hypothesis that all regression coefficients are zero (SAS, 1995). So, low p-values for these statistics imply significant explanatory variables. The concordance figure is a measure of the association between the predicted probabilities and the observed responses. It is a percentage and bounded by 100%, which would be perfect agreement. High concordance values indicate a better predictive capacity for the model. Each table presents all of these statistics. The results presented in the tables are used to guide the selection of the best model for each study area.

Modeling results for the Raleigh Area

This section begins with tables containing the values of the criteria described above. There are three tables: one for the original (or unfiltered) data, one for the 3x3 filter, and one for the 5x5 filter. Within each table we present the results from the three different link functions. These tables represent the information used to select a particular model for the Raleigh area. After these tables there are some additional details related to the selected model and an analysis of the residuals for that model.

Table 6.4: GLM output for the Raleigh area with unfiltered data

Raleigh Area	Link Functions:		
	Logit	Probit	Comp. Log - Log
Significant variables	vector-tas, absdif-tas5, difNDVI	Vector-tas, absdif-tas5, difNDVI	vector-tas, absdif-tas5, difNDVI
AIC	128.765	128.762	129.116
SC	143.007	143.004	143.359
-2 Log L	120.765	120.762	121.116
(p-value)	(0.0001)	(0.0001)	(0.0001)
Score	43.368	43.368	43.368
(p-value)	(0.0001)	(0.0001)	(0.0001)
Concordant	78.00%	78.20%	77.50%

Table 6.5: GLM output for the Raleigh area with 3x3 filtered data

Raleigh Area with 3x3 filter	Link Functions:		
	Logit	Probit	Comp. Log - Log
Significant variables	vector-tas, dif-tas5	vector-tas, dif-tas5	vector-tas, dif-tas5
AIC	123.907	124.9	122.961
SC	134.589	135.582	133.643
-2 Log L	117.907	118.9	116.961
(p-value)	(0.0001)	(0.0001)	(0.0001)
Score	51.948	51.948	51.948
(p-value)	(0.0001)	(0.0001)	(0.0001)
Concordant	72.20%	72.70%	72.50%

Table 6.6: GLM output for the Raleigh area with 5x5 filtered data

Raleigh Area with 5x5 filter	Link Functions:		
	Logit	Probit	Comp. Log - Log
Significant variables	vector-tas, dif-tas5	vector-tas, dif-tas5	vector-tas, dif-tas5
AIC	119.615	119.936	118.863
SC	130.297	130.618	129.545
-2 Log L	113.615	113.936	112.863
(p-value)	(0.0001)	(0.0001)	(0.0001)
Score	58.58	58.58	58.58
(p-value)	(0.0001)	(0.0001)	(0.0001)
Concordant	76.50%	77.20%	76.60%

The results for each link and for the unfiltered and filtered data are similar. All models show significant p-values, indicating that all models are significant. We see that the unfiltered and the 5x5 filter have higher concordance values and lower AIC and SC statistics. So, these two do somewhat better than the 3x3 filtered data. We see that in all models the "vector-tas" variable is significant. We also find the difference in the tassel cap band 5 to be a significant variable. It is somewhat curious that this variable shows up in each model. In their paper on the TM tasseled-cap transformation, Crist and Cicone (1984, p. 262) state that there may be no clear association between physical processes and the fifth and sixth bands of the tassel-cap transformation. We believe this variable is accounting for some atmospheric or sensor differences that were not accounted for in our normalization procedure. For the unfiltered data we also have the difference in NDVI appearing as a significant variable. Both this and the "vector-tas" variable appearing as significant are consistent with Coppin and Bauer's conclusions discussed above. The exciting point is that the GLM procedures indicates that the additional variable, the difference in tassel-cap band five, is also helpful in modeling the probability of change. Because the summary statistics for the unfiltered and 5x5 filter are close, we decided to use the simpler 5x5-filter model that contains only two explanatory variables. There does

not appear to be any practical difference between the three link-functions. Because it is more commonly used, we decided to use the logit link function. So, for the Raleigh area we will use the 5x5 filtered data with a logit link function using the "vector-tas" and "dif-tas-5" variables to model change. Figure 6.6 shows the sample data plotted in the "vector-tas" and "dif-tas-5" two dimensional space. The unchanged sample points are shown as white diamonds and changed areas are shown as solid squares. This plot shows how the changed and unchanged areas are distributed with respect to the combined vector difference and the band five difference in the tassel-cap transformed data.

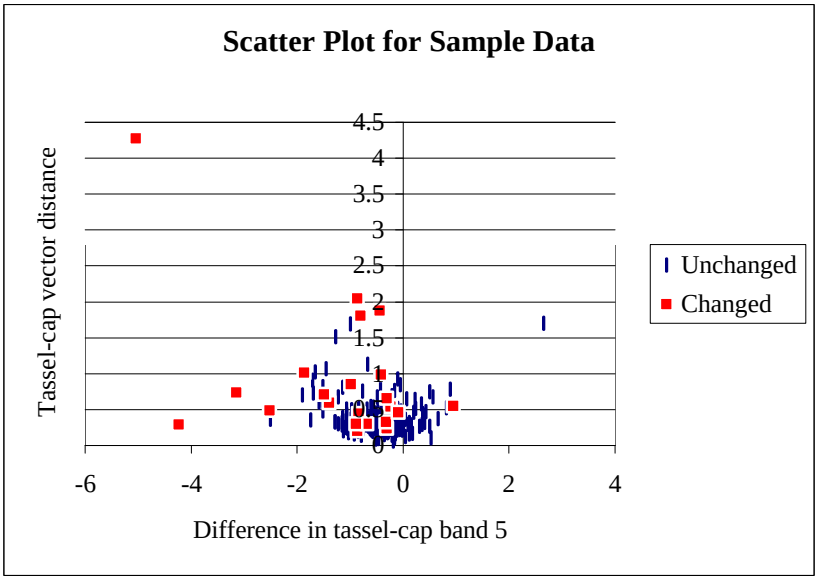


Figure 6.6: Distribution of sample data in two dimensional space

Figure 6.6 shows the empirical data in the two dimensions corresponding to the two variables that the GLMs indicated as the most significant indicators of change. From this data the GLM procedure produces a model in this two dimensional space. The fitted model follows the equation:

$$Prob\ of\ change = \frac{e^{-4.2954 - 0.9294 \times dif\text{-}tas\text{-}band5 + 0.1551 \times vector\text{-}tas}}{1 + e^{-4.2954 - 0.9294 \times dif\text{-}tas\text{-}band5 + 0.1551 \times vector\text{-}tas}} \qquad \textbf{(eq. 6.4)}$$

The surface is presented in figure 6.7 and the parameters, standard errors, Wald Chi-Square and its associated p-value are shown in table 6.7. The Wald Chi-Square statistic

and its p-value provide a measure of a variable's significance; with smaller p-values indicating higher significance (SAS, 1995, p. 22; SAS, 1989, 1075)

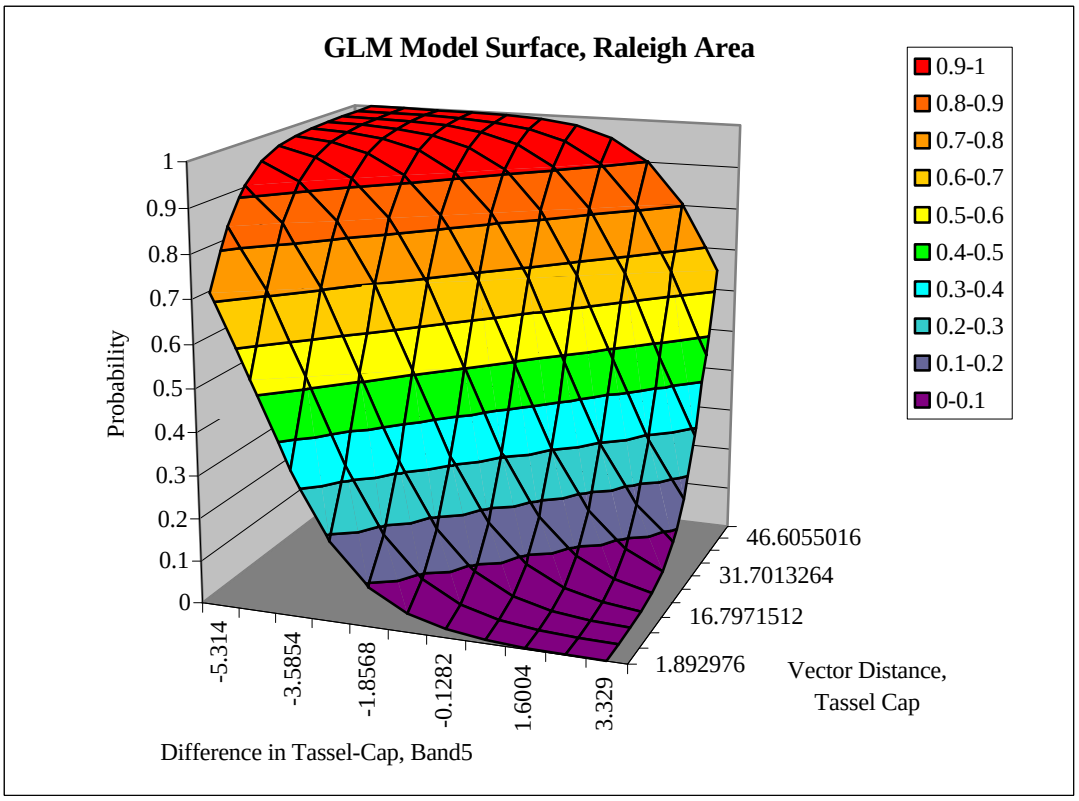


Figure 6.7: Logistic regression for the Raleigh area

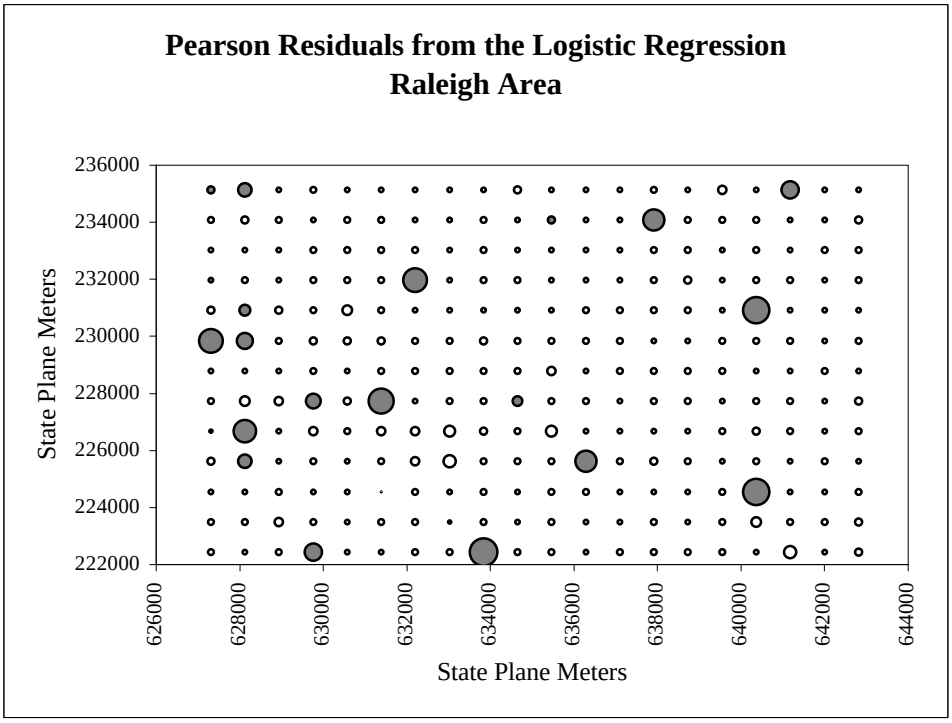
Table 6.7: Logistic regression parameters

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	p-value
intercept	-4.2954	0.5032	72.8745	0.0001
vector-tas	-0.9294	0.3223	8.3178	0.0039
dif-tas-5	0.1551	0.425	13.2998	0.0003

Since the “vector-tas” variable is an Euclidean distance variable, we would expect the coefficient to be positive since large “vector-tas” values indicate non-directional changes in the tassel-cap bands. The negative coefficient on the “dif-tas-5 variable indicates that smaller values of this variable are associated with the change areas. So, since the variable

is 94 values minus 88 values, smaller values in tassel-cap band 5 in 1994 are indicative of change. Both variables are highly significant.

Figure 6.8 shows a map of the Pearson residuals for the model in equation 6.4. Pearson residuals are useful in identifying observations that are not well explained by the model (SAS, 1990, p. 1093). The larger circles represent larger residuals and, therefore, sample points that are not well explained by the model. The white circles represent negative residuals while the gray circles represent positive residuals.



Gray circles represent positive residuals
White circles represent negative residuals

Figure 6.8: Distribution and magnitude of Pearson residuals for the coastal area logistic model

The Pearson residuals are a function of the observed minus fitted values. Because "changed" areas are labeled as 1 and due to the nature of the logistic regression, we know that all changed areas will have positive residuals and all unchanged areas will have negative residuals. In figure 6.8 we see that all of the larger circles are from positive

residuals and, thus, are from changed areas. We see that the residuals for changed areas are much larger than those for the unchanged areas. The large pixels from changed areas are those points that have changed but have little difference in radiance values from the two images.

In order to explore the residuals further and to investigate the possibility of modeling particular types of change, those areas that have changed were further broken down based on whether the change was an increase or decrease in vegetation. This break down was based on the initial classification of the reference data (see Table 6.1). Figure 6.9 displays the residuals for four groups: group "0" represents the unchanged pixels, group "1" represents areas that experienced increases in vegetation, group "2" represents areas that experienced decreases in vegetation, and group "3" represents areas that have changed but can not be readily classified as more or less vegetated. An example from group 3 would be a change from bare to developed land.

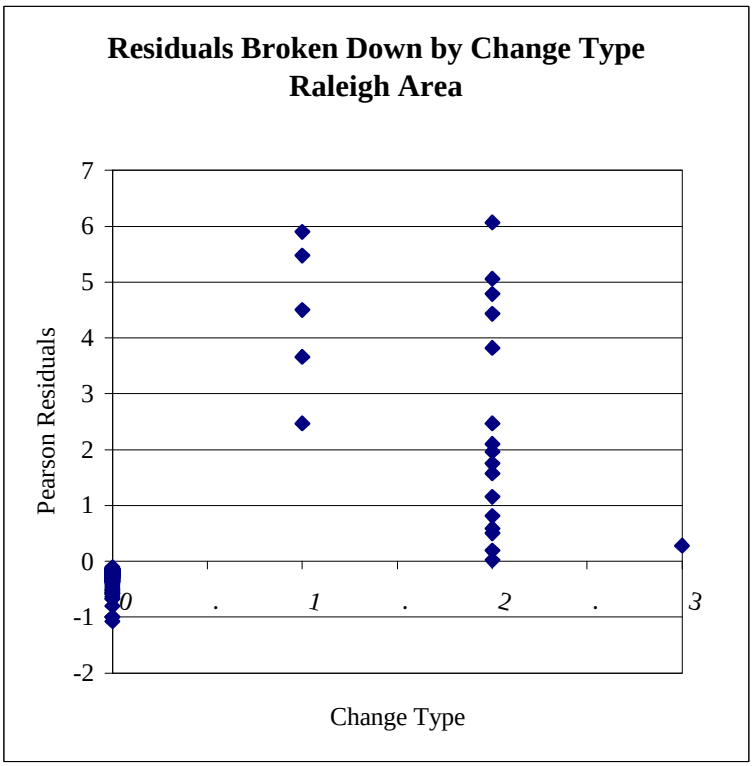


Figure 6.9: Pearson residuals broken down by change type

Concentrating on groups one and two, figure 6.9 shows that there are generally larger residuals for those change areas becoming more vegetated (i.e., group one has a higher average). This implies that the model does better at predicting decreases in vegetation (group two). This indicates that either separate models or a multinomial response could be used to further model the observed changes. Many typical studies classify change as a simple binary variable. Figure 6.9 shows how GLMs used *on a binary response* can be used to investigate possibilities for further modeling particular types of change. For the Raleigh area, further research could be directed at either a multinomial change response or a separate model on a more particular binary response such as "1" for areas which have decreased in vegetation and "0" for all other areas. This is left as further research and we will now proceed to the modeling results from the coastal area.

Modeling results for the coastal Area

Again we will present the tables with the values of the criteria used to evaluate the different models, then present the details for the selected model and investigate the residual from that model.

Table 6.8: GLM output for the coastal area with unfiltered data

Coastal Area	Link Functions:		
	Logit	Probit	Comp. Log - Log
Significant variables	dif4, absdif-tas3, vector	dif4, absdif-tas3, vector	absdif-3, absdif-tas3, diftas-5
AIC	134.202	132.658	141.505
SC	148.445	146.901	152.187
-2 Log L	126.202	124.658	135.505
(p-value)	(0.0001)	(0.0001)	(0.0001)
Score	110.529	110.529	103.263
(p-value)	(0.0001)	(0.0001)	(0.0001)
Concordant	91.20%	91.10%	88.50%

Table 6.9: GLM output for the coastal area with 3x3 filtered data

Coastal Area with 3x3 filter	Link Functions:		
	Logit	Probit	Comp. Log - Log
Significant variables	Absdif-tas3, absdif3, absdif6	absdif-tas3, absdif3, absdif6	absdif-tas3, absdif3, absdif6, vector
AIC	124.85	123.347	124.339
SC	139.093	137.59	142.143
-2 Log L	116.85	115.347	114.339
(p-value)	(0.0001)	(0.0001)	(0.0001)
Score	116.921	116.921	118.337
(p-value)	(0.0001)	(0.0001)	(0.0001)
Concordant	91.90%	92.40%	90.60%

Table 6.10: GLM output for the coastal area with 5x5 filtered data

Coastal Area with 5x5 filter	Link Functions:		
	Logit	Probit	Comp. Log - Log
Significant variables	Absdif-tas3, absdif-tas5, dif-tas5	absdif-tas3, absdif-tas5, dif-tas5	absdif-tas3, absdif-tas5, dif-tas5
AIC	124.011	122.529	125.852
SC	138.254	136.772	140.094
-2 Log L	116.011	114.529	117.852
(p-value)	(0.0001)	(0.0001)	(0.0001)
Score	126.065	126.065	126.065
(p-value)	(0.0001)	(0.0001)	(0.0001)
Concordant	91.60%	91.70%	91.30%

Among these tables we do not see results as consistent as those from the Raleigh area. That is, we do not see the same variables included in all the models. In looking at the image data, we felt that the patch of haze and the large water area in the southeast portion of the image may be affecting the model. To check how much these points were

influencing the model, we deleted points that fell into the haze area and those in the estuarine waters to the southeast. These results are presented in the following three tables.

Table 6.11: GLM output for the coastal area with unfiltered data, haze and water removed

Coastal Area (haze and water removed)	Link Functions:		
	Logit	Probit	Comp. Log - Log
Significant variables	dif4, absdiftas3, vector	dif-4, absdiftas3, vector	absdif-tas3, absdif-3
AIC	138.835	129.974	137.849
SC	145.702	143.709	148.15
-2 Log L	134.835	121.974	131.849
(p-value)	(0.0001)	(0.0001)	(0.0001)
Score	86.275	95.638	88.93
(p-value)	(0.0001)	(0.0001)	(0.0001)
Concordant	90.50%	90.40%	87.90%

Table 6.12: GLM output for the coastal area with 3x3 filtered data, haze and water removed

Coastal Area with 3x3 filter (haze and water removed)	Link Functions:		
	Logit	Probit	Comp. Log - Log
Significant variables	absdif-tas3, absdif3, absdif6	absdif-tas3, absdif3, absdif6	absdif-tas3, absdif3, absdif6, vector
AIC	120.638	119.529	123.147
SC	134.372	133.264	136.882
-2 Log L	112.638	111.529	115.147
(p-value)	(0.0001)	(0.0001)	(0.0001)
Score	101.047	101.047	101.047
(p-value)	(0.0001)	(0.0001)	(0.0001)
Concordant	91.30%	91.80%	90.40%

Table 6.13: GLM output for the coastal area with 5x5 filtered data, haze and water removed

Coastal Area with 5x5 filter (haze and water removed)	Link Functions:		
	Logit	Probit	Comp. Log - Log
Significant variables	absdif-tas3, absdif-tas5, dif-tas5	absdif-tas3, absdif-tas5, dif-tas5	absdif-tas3, absdif-tas5, dif-tas5
AIC	122.038	120.854	123.426
SC	135.772	134.589	137.161
-2 Log L	114.038	112.854	115.426
(p-value)	(0.0001)	(0.0001)	(0.0001)
Score	108.828	108.828	108.828
(p-value)	(0.0001)	(0.0001)	(0.0001)
Concordant	90.70%	90.80%	90.50%

We see that these results are similar to the initial analysis. However, since we are trying to monitor changes in land cover, we believe it is appropriate to remove those sample points that are in estuarine or ocean water in locations that have essentially no chance of changing. We believe these sample points are causing a slight increase in the "concordance" figures presented in the first three tables for the coastal area. Also, we believe it is appropriate to mask out haze areas, as haze in one image will produce differences in radiance values not related to changes in land cover. So, in selecting a model we will use the data set with the haze and water areas removed.

The interesting result for the coastal area is how several of the individual original bands are significant variables. This differs from the results from the Raleigh area where the combined “vector-tas” variable was consistently the most significant variable. This shows the power of using GLM. From the Raleigh area we might assume that the “vector-tas” variable is the best indicator of change and use this variable for the coastal area. However, we see from the stepwise regression results that this variable is not included in any of the models.

Now, considering that there is autocorrelation present in the response variable, we need to exercise caution when interpreting the results from the stepwise procedure. The criterion used in a stepwise procedure is based on the significance of the variables as they are considered for inclusion into the model. Since the autocorrelation can influence the variance estimate, it can also influence the p-values from which the significance is inferred. Even if there is no autocorrelation in the response variable, the stepwise procedure is mainly an exploratory technique (SAS, 1995, p. 51) used to find a subset of significant variables and then consider these variables more closely -- considering both the statistical significance and the physical interpretation of the variables. With this, and the autocorrelation present in the response variable, we used the model results represented in tables 6.11 through 6.13 mainly to guide a more careful consideration of a subset of the original 28 possible variables.

Seeing that the absolute difference in tassel-cap band 3 is significant in all of the models we felt that this should be included in our change model. Recall that band 3 of the tassel-cap transformation is related to differences in moisture. The significance of this variable makes sense in light of the precipitation data (see appendix C). We saw that in each of the four days prior to the 1994 image acquisition there was from 1/3" to 1" of rainfall. Since changes in the response variable, as determined from the air-photo reference data, were due to land cover and not moisture conditions, we believe the differences in tassel-cap band 3 do not so much predict changes but adjust for the moisture related variability in the data. The observed differences are primarily in either cutting or planting/growing of evergreen forests. This could result in differences in band 3, 4, 5 or 6 in the original image data and/or differences in band 2 of the tassel-cap transformation.

We tried fitting a model with the absolute difference in tassel-cap three and absolute differences in the original bands 3, 4, 5, and 6 and tassel-cap band 2. The most significant model, which also had the highest concordance of 92.6%, was from the model using the 3x3 filtered data with only the original band 3 and tassel-cap band 3. The differences in band three are due to differences in red reflectance. By viewing a gray scale image with

only band 3 we found that most vegetated areas have a low red reflectance while bare soil or developed areas have a high red reflectance. This implies that the extremes of the absolute value of band three are due to vegetation vs. non-vegetated areas. Since the band three variable is highly significant (p-value < .001) and since the inclusion of this variable has a physical interpretation we decided on a model that included the absolute difference in band 3 and the absolute difference in tassel-cap band 3. Figure 6.10 shows the sample data plotted in the "abs-dif-3" and "abs-dif-tas3" two dimensional space. The unchanged sample points are shown as white diamonds and changed areas are shown as solid squares. This plot shows how the changed and unchanged areas are distributed with respect to the difference in tassel-cap band 3 and the original band 3.

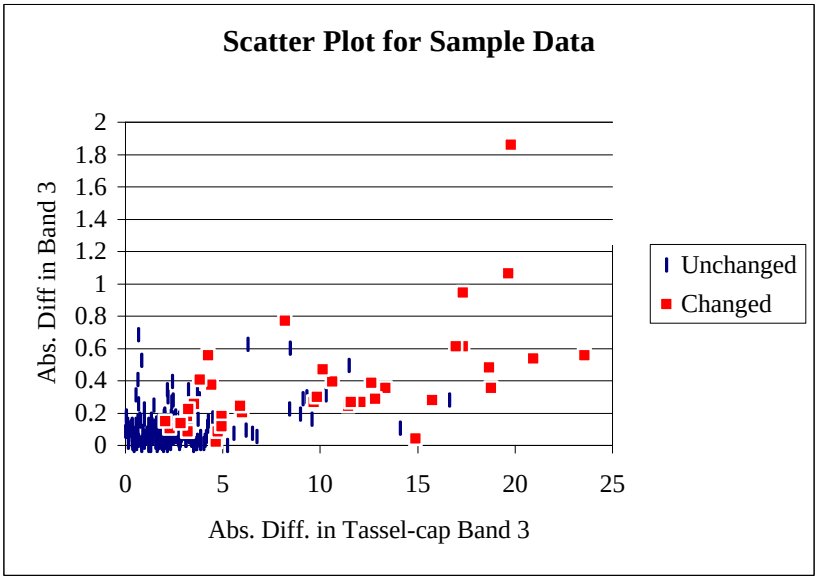


Figure 6.10: Distribution ofcoastal sample data in two dimensional space

Figure 6.10 shows the empirical data in the two dimensions corresponding to the two variables that the GLMs indicated were significant indicators of change. From this data the GLM procedure produces a model in this two-dimensional space. The fitted model follows the equation:

$$Prob\,of\,Change = \frac{e^{-3.9320+0.3012\times abs-dif\,-tas3+4.1207\times abs-dif\,3}}{1+e^{-3.9320+0.3012\times abs-dif\,-tas3+4.1207\times abs-dif\,3}} \qquad \textbf{(eq. 6.5)}$$

The surface is presented in figure 6.11 and the parameters, standard errors, Wald Chi-square and its associated p-value are shown in table 6.14.

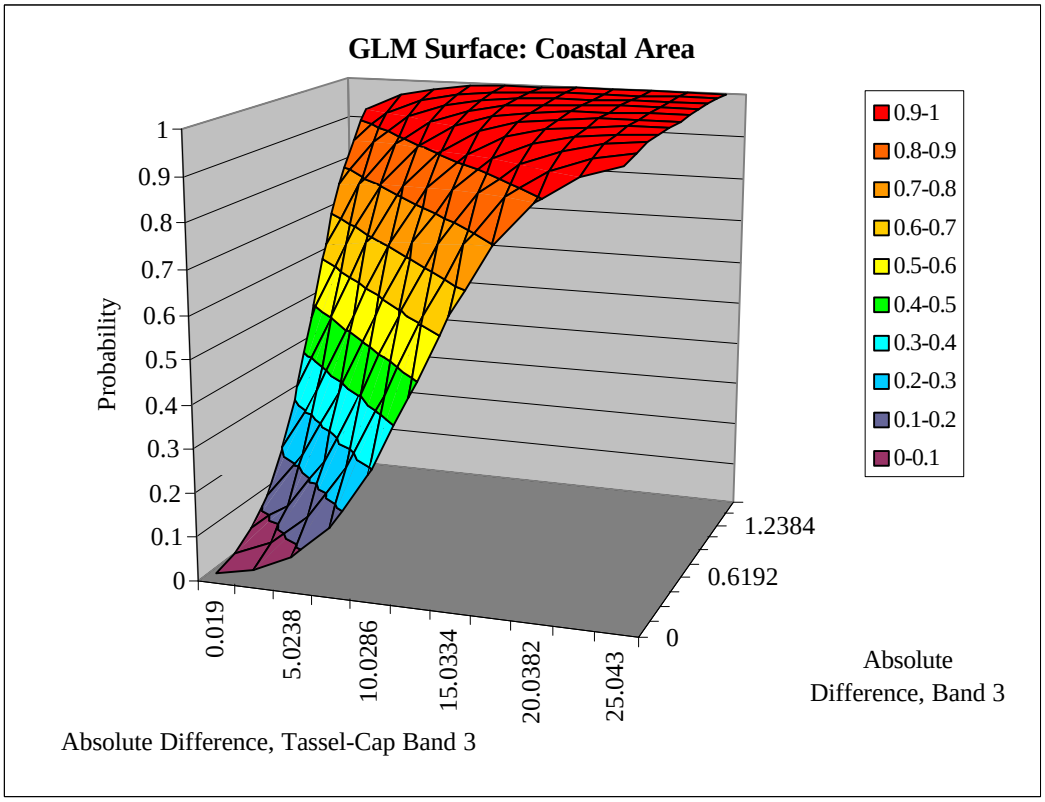


Figure 6.11: Logistic regression model for the coastal area

Table 6.14: Logistic regression parameter results

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	p-value
Intercept	-3.9320	0.4446	78.2053	0.0001
abs-dif-3	0.3012	0.0623	23.4117	0.0001
abs-dif-tas-3	4.1207	1.5567	7.0069	0.0081

Both the "abs-dif-3" and the "abs-dif-tas3" coefficients are positive. This means that larger differences in these two variables are indicative of change. It is good to see the p-value of the "abs-dif-3" is more significant than that of "abs-dif-tas3" since we believe

band 3 is explaining the actual changes and the tassel-cap band 3 is explaining some of the variation caused by differences in moisture conditions.

Now that we have a selected model we can test for autocorrelation in the residuals. The variance estimate is derived from the residuals, so if the residuals show no spatial autocorrelation, we can trust the variance estimate and inference based on the standard output for the model.

Figures 6.12 and Figure 6.13 show the empirical variograms in the North/South and East/West direction. Both figures show essentially flat lines. This indicates that there is no autocorrelation present in the residuals. With this we can assume our variance estimates and the inference derived from the model are appropriate. (Had there been autocorrelation present in the residuals the empirical variogram can be used to estimate a true variance/covariance between the sample points and this could be used to estimate the robust variance estimator given in Diggle *et al.*, 1995, equation A.6.1.)

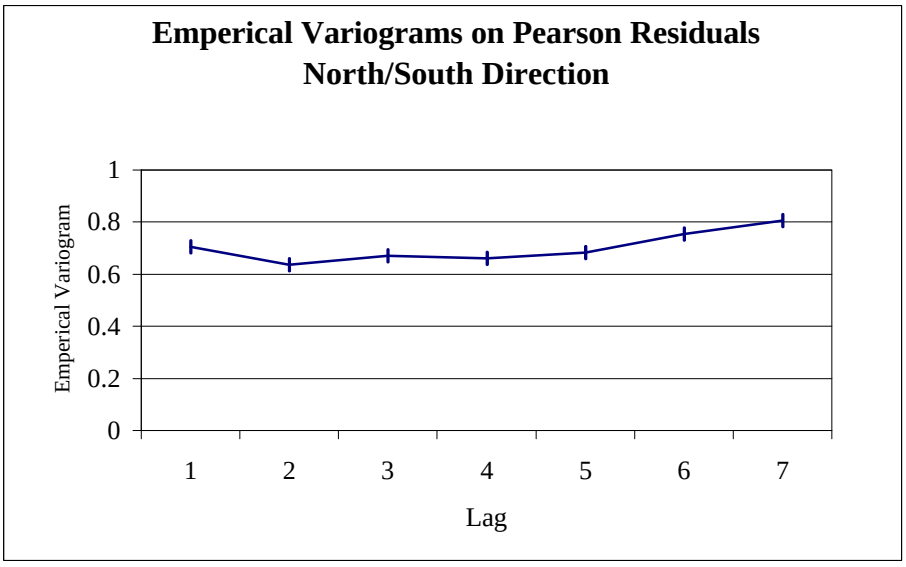


Figure 6.12: Empirical variogram of Pearson residuals for the coastal model, East/West direction

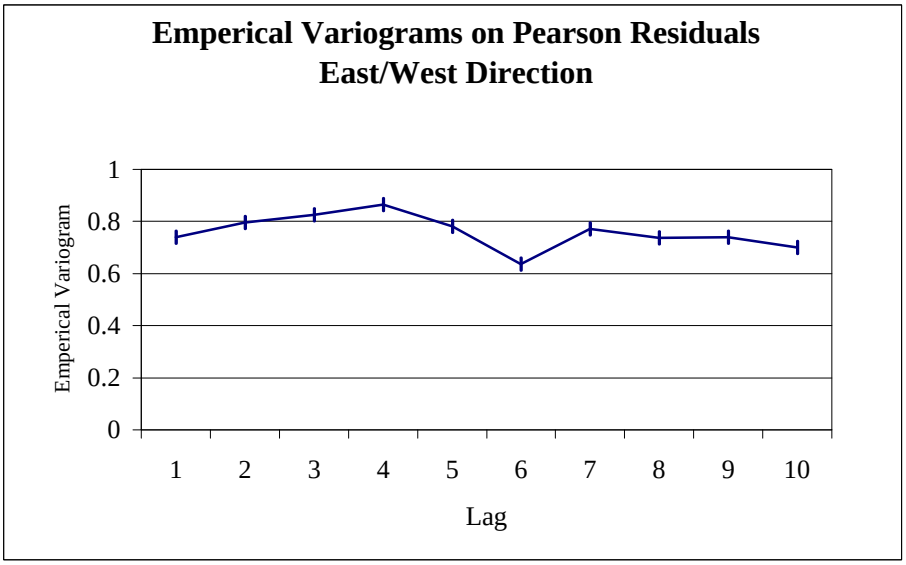
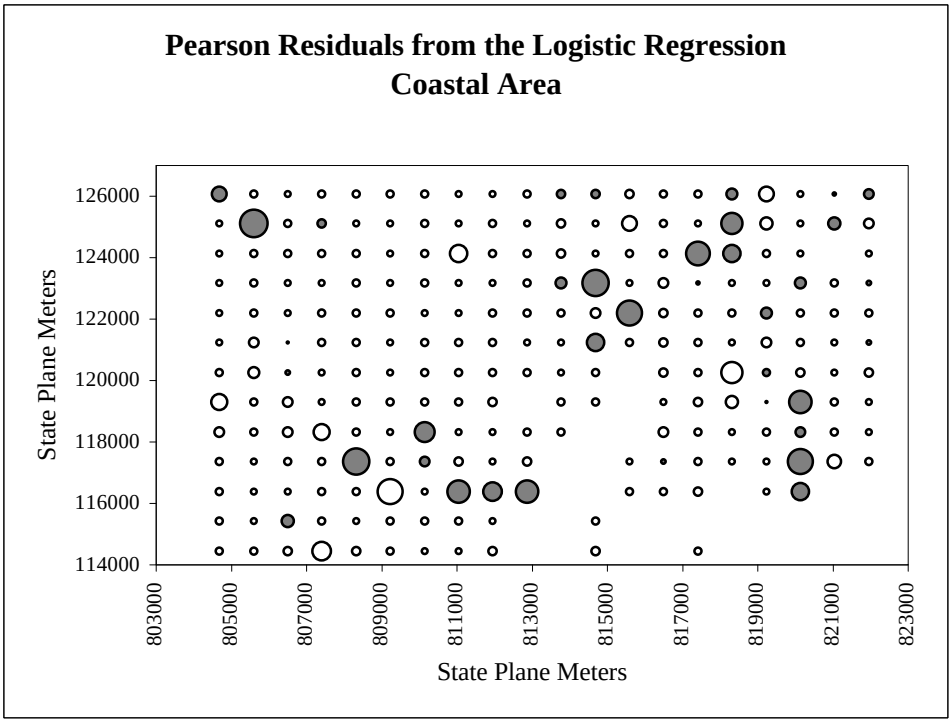


Figure 6.13: Empirical variogram of Pearson residuals for the coastal model, North/South direction

Seeing that the residuals show no autocorrelation, we will go on to investigate the residuals in a fashion similar to the analysis for the Raleigh area. Figure 6.14 shows a map of the Pearson residuals for the model in equation 6.5. Looking at the locations with larger circles, as with the Raleigh residuals (Figure 6.8), we see that they are mainly the gray circles, indicating positive residuals, which are from the changed areas. Note that the missing locations are the haze or water areas excluded from the modeling.



Gray circles represent positive residuals
White circles represent negative residuals

Figure 6.14: Distribution and magnitude of Pearson residuals for the coastal area logistic model

Similar to the analysis of the residuals from the Raleigh area, the residuals are broken down by the particular type of change. This is shown in figure 6.15. Here, the group numbering follows the same meaning as for the Raleigh area. Again we see a higher average value for group 1 -- those changes resulting from an increase in vegetation. Here too the model does not predict these types of change as well. This leads us to the same conclusion that further modeling of either a multinomial change variable or a more particular binary response might be appropriate for the coastal area.

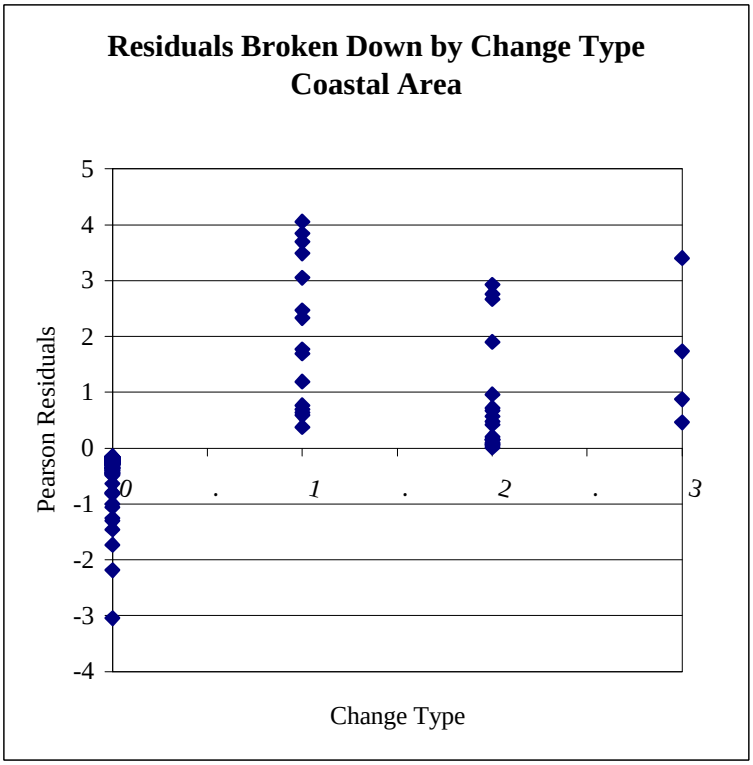


Figure 6.15: Pearson residuals broken down by change type

Conclusions from the Example

Through this example we can see how GLMs provide a statistically sound procedure to select a change detection algorithm. We decided to model change in the Raleigh area with the logit link function, using the 5x5 filtered data, with the vector distance from the tassel-cap bands and the difference in tassel-cap band 5. We decided to model change in the coastal area with the 3x3 filtered data using the absolute difference in band 3 and the absolute difference in tassel-cap band 3. For both study areas there was little difference among the three link functions and the filtered data did somewhat better.

From this Chapter we have two strong arguments for utilizing GLM in change detection. Different results reported in recent literature imply that there is not one single "best" change detection method. Likewise, for our two study areas we find different significant

variable. So, the first argument is that GLMs can be used to select among different change detection variables. The second argument is how GLMs can be used to incorporate different combinations of variables. For both the Raleigh and coastal area we use two variables in our change detection model. The GLM allows us to use more than one variable and assigns a coefficient and significance to each variable. Then we can analyze the residual from the chosen model to test for autocorrelation and to direct possible further analysis on the change areas. GLMs help select a change detection model and then provide information for the chosen model.

In the following two chapters we will continue with this example, showing how the results of the GLM can be used to enhance satellite based change detection. In the next chapter, for both the coastal and Raleigh areas, we will compare the models we have selected to a standard change detection model. We will explore how the GLM leads to more accurate modeling and how the modeling can guide a more informed selection of the change/no-change threshold. This is followed by a chapter in which we apply the change detection models to our study areas. In that chapter we present change detection products that utilize GLMs by using the probability of change as a way to present a more meaningful change detection analysis.