CHAPTER FOUR

RESULTS AND ANALYSIS

Evaluation of the Logistic Regression Model

Stepwise logistic regression screens the available list of independent variables to select only those that it deems "important" in describing the dependent variable. Once the model has been built and predictions produced, it is necessary to determine how effective that model is at predicting the dependent variable. This is referred to as *goodness-of-fit*, and this section of Chapter 4 will deal with this issue.

The F ratio and R^2 tests are commonly used to determine the significance (goodness-of-fit) of the more familiar linear regression model. In some cases, random sampling variation in the data can produce an improvement in prediction by using the regression equation even when the independent variables are unrelated to the dependent variable. The multivariate F ratio test is used to ascertain whether a reduction in error of prediction is caused by these sampling variations, or whether there is truly a relationship between the independent variables and the dependent variable. The coefficient of determination, or R^2, measures the proportion by which use of the regression equation reduces the error of prediction. In other words, it determines whether the relationship is substantial enough to be significant.

To analyze goodness-of-fit for the logistic regression model, close parallels to the F ratio and R^2 tests were used. In logistic regression the G_M statistic is analogous to the F test in linear regression (Menard 1995, Hosmer and Lemeshow 1989[1]). As mentioned in Chapter 2, the log-likelihood ratio statistic was used for selecting parameters in the logistic regression model. The SPSS statistical package presents not the log-likelihood itself but the log-likelihood multiplied by –2 (SPSS Inc. 1998). Output from SPSS denotes log-likelihood multiplied by -2 as "-2 Log Likelihood" (Appendix One). By multiplying the log-likelihood by -2 it approximates a χ^2 (chi-square) distribution (Menard 1995). Larger values of -2 log likelihood indicate worse prediction of the dependent variable.

Before independent variables were entered into the logistic regression model, the -2 log likelihood for the model with only the intercept constant (α) was given. This intercept-only -2 log likelihood is designated D_0 (Denoted as L_0 by Hosmer and Lemeshow) to indicate that none (zero) of the independent variables were included in the equation. In SPSS, this value is located at the beginning of the output and is labeled

[1] Hosmer and Lemeshow denote G_M as G.

"Initial Log Likelihood Function -2 Log Likelihood" (Appendix One). D_0 is analogous to the total sum of squares (SST) in linear regression analysis (Menard 1995).

At each additional step in the logistic regression procedure a new -2 log likelihood value was determined. This -2 log likelihood statistic was produced using only those independent variables included at that step and the intercept. This statistic is referred to as D_M^n (n denoting the step number in the logistic regression procedure) or the deviation χ^2 for the full model. D_M^n is analogous to the error sum of squares (SSE) in linear regression analysis and indicates how poorly the model fits with the independent variables in the equation (Menard 1995).

Taking the difference between D_0 and D_M, that is (D_0 - D_M), gives the model G_M. In SPSS G_M is labeled as "Model Chi-Square" (Appendix One). G_M is not only analogous to the multivariate F test, but also to the regression sum of squares (SSR) (i.e. SSR = SST - SSE) (Menard 1995). G_M provides a test of the null hypothesis that $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ for the logistic regression model. If G_m is statistically significant, then the null hypothesis can be rejected, and it can be concluded that the independent variables contribute to better predictions.

Table 4.1 holds values of G_M (model chi-square) at consecutive steps in the logistic regression procedure. At each step in the logistic procedure, but more importantly in the last step (6), G_M was determined to be highly significant and the null hypothesis (H_0) was rejected. Therefore, it was determined that information on the independent variables allowed for better prediction of Union fort/battery locations than could be made without their inclusion. Notice that G_M decreased slightly at each of the consecutive steps. This is a factor of utilizing backward stepwise logistic regression (i.e. independent variables are removed from the equation at consecutive steps).

Table 4.1

The -2 log likelihood, and G_M values for the logistic regression model using the Wald backward stepwise method.

Step	Degrees of Freedom	-2 Log Likelihood	Model Chi-Square (G_M)	Significance (p)
Initial	---	277.25887	---	---
1	11	139.596	137.663	0.00
2	10	139.622	137.637	0.00
3	9	139.812	137.447	0.00
4	8	140.101	137.157	0.00
5	7	140.560	136.698	0.00
6	6	142.390	134.869	0.00

Analogous to R^2 (SSR/SST) for linear regression is the R^2_L statistic for logistic regression. R^2_L indicates how much the inclusion of the independent variables in the model reduces the badness-of-fit D_0 chi-square statistic (Menard 1995). R^2_L varies between 0 (independent variables are useless in prediction of dependent variable) and 1 (independent variables in model predict the dependent variable perfectly), and is calculated by dividing the Model Chi-Square (G_M) by the Initial Log Likelihood Function -2 Log Likelihood (D_0) (Equation 4.1).

$$R^2_L = G_M / (D_0) = G_M / (G_M + D_M) \qquad \textbf{4.1}$$

R^2_L values for the stepwise logistic regression model are found in Table 4.2. These values illustrate a moderately strong association between Union fort/battery location and the independent variables. Converting the models final R^2_L ratio (step 6) into a percentage, it can be said that 48.64% of the variance of the dependent variable was accounted for by the logistic regression equation. This value is substantive enough to consider the equation significant.

Table 4.2

The -2 log likelihood and R^2_L values for the logistic regression model using the Wald backward stepwise method.

Step	-2 Log Likelihood	Model Chi-Square (G_M)	R^2_L
Initial	277.25887 (D_0)	---	---
1	139.596 (D_M^1)	137.663	.4965
2	139.622 (D_M^2)	137.637	.4964
3	139.812 (D_M^3)	137.447	.4957
4	140.101 (D_M^4)	137.157	.4947
5	140.560 (D_M^5)	136.698	.4930
6	142.390 (D_M^6)	134.869	.4864

Interpreting the Logistic Regression Coefficients

One should note that β coefficients can be positive or negative. A positive coefficient indicates that an *increase* in the corresponding variable is associated with a greater likelihood of site presence. Conversely, a negative coefficient indicates that a *decrease* in the corresponding variable is associated with a greater likelihood of site presence. The independent variables: CFORT_DIST (distance from Confederate fort/battery), CON_VIS (visibility from main Confederate works), ELEV38 (elevation), and WATER_DIST (distance from water) all had positive β coefficients. This implies that; as one moves further away from Confederate fort/batteries, as the visibility increases, as the elevation increases, and as one move further from a water source, the probability of Union Civil War fort/battery presence increases. The independent variables: CMAIN_DIST (distance from main Confederate works) and RAIL_DIST (distance from railroad) have negative β coefficients. This implies that as one moves closer to the main Confederate line or closer to a railroad line that the probability of Union Civil War fort/battery presence increases.

It is interesting to note that as the distance from Confederate forts/batteries increases so does the probability of Union fort/battery presence. Conversely, as the distance from the main Confederate works decreases the probability of Union fort/battery presence increases. I theorize that this phenomenon is the result of heavy artillery (e.g. cannons, mortars, etc.) placements within the Confederate forts and batteries as opposed to the light artillery (e.g. riffles) placed along the connecting trenches. Confederate heavy artillery placements are capable of inflicting heavy damage and large causalities at longer ranges than possible by light artillery. When construction of Union forts and batteries, and the lines that connected them was done under enemy fire, *gabions* were used. A gabion is a round basket of cylindrical form, without a bottom, that is made of wood. These gabions were placed between the enemy and the soldiers, and filled with earth as they constructed trenches. Earthwork construction was also performed under the cover of darkness for concealment.

The independent variable representing distance from water is also puzzling. I hypothesized that as the distance from a water source decreased the probability of Union fort/battery presence would increase, when the opposite was true. My hypothesis was based on the assumption that an army would need water for drinking purposes, to clean weapons, or for swabbing artillery bores. Large water bodies could also serve as anchor points for earthwork construction or as natural barriers to enemies. However, the distance from water variable was overwhelmed by elevation, since the two variables are inversely proportional to one another in the study area. Therefore my original hypothesis concerning elevation and distance from water contradicted each other.

The variables representing elevation and visibility from main Confederate works are closely related. Topographic characteristics such as these were extremely important to proper positioning of field works. Obviously, locations with a commanding view of the surrounding area would have a high probability for Union fort/battery presence. As visibilities or elevations increase, so in turn would the likelihood of site presence.

Locations of railroad lines were also significant to the presence of Union fort/battery locations in the landscape. These railroads were lifelines for the Confederate army, transporting goods and war supplies to and from the Confederate capital of Richmond. There is no wonder that as the distance to railroad lines decreases, the probability of Union fort/battery presence increased.

Unstandardized Coefficients

The logistic regression coefficient can be interpreted as the change in the dependent variable, logit(*Y*), associated with a one-unit change in the independent (Mennard 1995). However, unlike linear regression, P(*Y*=1) is not a linear function of the independent variables, the slope of the curve varies depending on the value of the independent variables (Mennard 1995). The equation for the relationship between Union Civil War fort/battery locations (USCWFBL) and the predictors was:

$$\text{Logit(USCWFBL)} = 1 / (1 + (\mathit{Exp}\ (- (.2704 + .0076(\text{cfort\_dist}) - .0099(\text{cmain\_dist}) + .0076(\text{con\_vis}) + .0500\ (\text{elev38}) - .0019(\text{rail\_dist}) + .0028(\text{water\_dist}))))) \qquad \mathbf{4.2}$$

Predictions for individual cases (i.e., pixels/cells) were obtained by entering the values for the variables in the formula for the specific case.

Standardized Coefficients

When independent variables are measured at different scales or in different units, β coefficients must be standardized to compare the strength of the relationship between the dependent variable and the many independent variables. By standardizing the coefficients, the independent variables can be compared directly to determine which has the largest magnitude on the dependent variable.

Unlike linear regression, the calculation of standardized coefficients is not as straightforward. In logistic regression it is not the value of *Y*, but the probability that *Y* has one or the other of its possible values. The dependent variable is logit(*Y*) not *Y* in logistic regression, so the calculation of means or direct calculation of standard deviations is not possible (Mennard 1995). Standard deviations are calculated using the predicted values of logit(*Y*) and the explained variance, R^2. The following equation was used to estimate standardized logistic regression coefficients:

$$b^*_{YX} = (b_{YX})(s_X)/\mathrm{SQRT}(s^2_{\text{predicted (logit}(Y))}/R^2) = (b_{YX})(s_X)(R)/s_{\text{predicted logit}(Y))} \qquad \textbf{4.3}$$

Where b^*_{YX} is the standardized coefficient, b_{YX} is the unstandardized coefficient, s_X is the standard deviation of the independent variable *X*, $s^2_{\text{predicted (logit}(Y))}$ is the variance of predicted logit(*Y*), $s_{\text{predicted (logit}(Y))}$ is the standard deviation of predicted logit(*Y*), R^2 is the coefficient of determination.

Using Equation 4.3, standardized coefficients were calculated for each independent variable in the final equation (Table 4.3).

Table 4.3

Unstandardized and standardized logistic regression coefficients

Variable	Unstandardized	Standardized
cmain_dist	-.0099	-2.497
cfort_dist	.0076	1.962
rail_dist	-.0019	-.2293
elev38	.0500	.1190
water_dist	.0028	.0917
con_vis	.0076	.0771

Interpretation of these standardized coefficients is quite straightforward. A one standard deviation increase in the independent variable (*X*) produces a b^* standard deviation change in logit(*Y*). For example, a one standard deviation increase in cfort_dist is associated with an increase of 1.962 standard deviations in logit(USCWFBL). Standardized coefficients listed in Table 4.3 are sorted in descending order, with cmain_dist having the greatest influence over the dependent variable and con_vis with the

least influence. It should be noted that when predictor variables are correlated, as they are to some extent in this analysis, adding a new variable to the model can drastically alter the "importance ranking" because new interrelationships between the predictor variables may be introduced (Kvamme 1985).

Predictive Efficiency

Goodness-of-fit measurements determined that the independent variables contributed significantly to prediction of Union fort and battery locations, and that the logistic regression equation accounted for a considerable percentage of the dependent variable's variance. In addition to these statistics, the accurate prediction of group membership in the model was also of interest. To assess model accuracy, SPSS created two classification tables for each step in the logistic regression model. One table displayed the observed and predicted values of site and non-site locations for the randomly "selected" cases (i.e., stratified random sample of 100 from each class), while the second presented the observed and predicted values of site and non-site locations for the "unselected" cases (51,782) in the model (Appendix One). Figure 4.1 displays a classification table for each step in the logistic regression procedure for only the "selected" cases. Each table summarizes the observed and predicted values to interpolate predictive efficiency/accuracy for each step in the model.

TABLE A: Initial Step

	Predicted		
Observed	*Non-Site*	*Site*	Percent Correct
Non-Site	100	0	100.00%
Site	100	0	.00%
		Overall	50.00%

TABLE B: Step 1

	Predicted		
Observed	*Non-Site*	*Site*	Percent Correct
Non-Site	88	12	88.00%
Site	16	84	84.00%
		Overall	86.00%

TABLE C: Step 2

	Predicted		
Observed	*Non-Site*	*Site*	Percent Correct
Non-Site	88	12	88.00%
Site	16	84	84.00%
		Overall	86.00%

TABLE D: Step 3

	Predicted		
Observed	*Non-Site*	*Site*	Percent Correct
Non-Site	88	12	88.00%
Site	16	84	84.00%
		Overall	86.00%

TABLE E: Step 4

	Predicted		
Observed	*Non-Site*	*Site*	Percent Correct
Non-Site	90	10	90.00%
Site	17	83	83.00%
		Overall	86.50%

TABLE F: Step 5

	Predicted		
Observed	*Non-Site*	*Site*	Percent Correct
Non-Site	87	13	87.00%
Site	18	82	82.00%
		Overall	84.50%

TABLE G: Step 6

	Predicted		
Observed	*Non-Site*	*Site*	Percent Correct
Non-Site	89	11	89.00%
Site	13	87	87.00%
		Overall	88.00%

Figure 4.1. Classification tables for the stepwise logistic regression procedure (Selected samples)

Table G of Figure 4.1 is the classification table for the final step in the logistic regression model. At this final step (6), the regression equation correctly classified 87.00% of non-sites and 89.00% of sites for a total predictive efficiency of 88.00%. From this number, 88.00%, it can initially be inferred that the logistic regression equation efficiently predicted the presence or absence of Civil War Union fortifications for Michler map 38.

It is important to recognize that any level of "accuracy" in predicting sites can be achieved. This is accomplished by accepting a "trade-off": exchanging increased accuracy for predicting sites in return for decreased accuracy in predicting non-sites. This trade-off is controlled by the cut-point value that is selected for the model, 0.62 in this case. If the cut-point value was increased, then a smaller percentage of sites would have been predicted correctly. In turn, a greater percentage of non-sites would be predicted correctly. However, if one were to decrease the cut-point value to less than 0.62, then the opposite would present itself (i.e., increasing "site" percent correct while decreasing "non-site" percent correct). By selecting a cut-point of 0, a model would correctly classify all site locations. In this case every non-site would also be classified as a site, and the model would be useless.

Figure 4.2 displays the SPSS plot of observed groups and predicted probabilities for the logistic regression model of Michler Map 38. This figure also illustrates that the accuracy of prediction for the model is high.

```
      32 +                                                          +
         I                                                          I
         I                                                          I
F        In                                                         I
R     24 +n                                                         +
E        In                                                         I
Q        In                                                         I
U        In                                                         I
E     16 +n                                                         +
N        In                                                         I
C        In                                                         I
Y        Inn                                              ss s s    I
       8 +nns  s                                          ss s ss   +
         Innn  s     n                                 sssssssss s I
         Innn sn n   n     s                        s  sssssssssssss I
         Innnnnnnnnnn nnnnn nnnnnn n    nnnnnnnnsn sssnnnssnnssssssss I
Predicted --------------+--------------+--------------+---------------
  Prob:   0            .25            .5             .75             1
  Group:  nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnsssssssssssssssssssssssss

          Predicted Probability is of Membership for site
          The Cut Value is .62
          Symbols: n – non-site
                   s - site
          Each Symbol Represents 2 Cases.
```

Figure 4.2. Observed groups and predicted probabilities for the final step in the logistic regression model

The final classification (prediction) table in Figure 4.1, Table G, only measures agreement using the diagonal elements of the matrix. These accuracy figures do not consider errors of commission and errors of omission. Therefore, these results of accuracy are overly optimistic since they do not take into account the chance probability of being correct. To account for errors of commission and errors of omission, Cohen's Kappa coefficient of agreement was used (Cohen 1960 and Congalton and Green 1998). Cohen's Kappa coefficient measures the relationship of beyond chance agreement to expected disagreement using all cells of the matrix, not just the diagonal (Rosenfield and Fitzpatrick-Lins 1986). The Kappa coefficient is commonly referred to as "k-hat" as is calculated using the following formula:

$$\text{“k-hat”} = \hat{k} = \frac{\theta_1 - \theta_2}{1 - \theta_2}$$

$$\text{where, } \theta_1 = \sum_{i=1}^{r} x_{ii} / N ,$$

$$\theta_2 = \sum_{i=1}^{r} x_{i+} \; x_{+i} / N^2,$$

x_{ii} = diagonal subtotal for row i, column i,
x_{i+} = row subtotal for row i and
x_{+i} = column subtotal for column i

4.4

A value of 0.0 for Kappa indicates a situation where obtained agreement equals chance agreement. Positive values of Kappa occur from greater than chance agreement, where a maximum value of 1.0 indicates perfect agreement. Negative values of Kappa are from less than chance agreement.

With values from Figure 4.1, Table G, a “k-hat” coefficient of 0.76 was produced for Michler Map 38. Often Kappa is presented as a percentage, so it the can be said that the probability model performed 76% greater than chance agreement. This value is less optimistic than the accuracy value (i.e., 88.00%) calculated from the final error matrix in Figure 4.1. This difference is not surprising when one considers the large number of non-sites in proportion to sites. Therefore, optimistic percentages in Figure 4.1 are the result of a high probability of agreement for non-site locations.

Confidence intervals around “k-hat” can be computed using sample variance and the fact that the “k-hat” statistic is asymptotically normally distributed. This fact provides a means to evaluate the significance of the “k-hat” statistic. Equation 4.5 is the approximate large sample variance for Kappa, and Equation 4.6 is used to calculate significance (i.e., Z score):

$$\widehat{\mathrm{var}}(k) = \frac{1}{N}\left[\frac{\theta_1(1-\theta_1)}{(1-\theta_2)^2} + \frac{2(1-\theta_1)(2\theta_1\theta_2-\theta_3)}{(1-\theta_2)^3} + \frac{(1-\theta_1)^2(\theta_4-4\theta_2^2)}{(1-\theta_2)^4}\right]$$

Where $\theta_3 = \sum_{i=1}^{r} x_{ii}(x_{i+} + x_{+i}) / N^2$,

$\theta_4 = \sum_{\substack{i=1 \\ j=1}}^{r} x_{ij}(x_{j+} + x_{+i})^2 / N^3$,

x_{ij} = observed cell value at row i, column j and
x_{j+} = column subtotal for column j

4.5

$$Z = \frac{\hat{k}}{\mathrm{SQRT}(\widehat{\mathrm{var}}(k))}$$

Where SQRT = square root of ()

4.6

The Kappa coefficient for Michler Map 38 had a variance of 0.002 and a Z score of 16.54. At the 95 percent confidence level, this Kappa value is significant since it exceeds 1.96. This indicates that error matrix for Michler Map 38 is significant.

Importantly, the predictive probability model eliminated approximately 85.07% of the study area as not likely to contain sites. Within this same region only about 14.04% of sites occurred, while 99.96% of the eliminated region did not contain sites. About 85.95% of all sites occurred in approximately 14.93% of the study region predicted to be "favorable" for sites. The probability of a site in the region favorable to site locations was 0.0134 (104 / (104 + 7657)), nearly 36 times more likely than the probability of a site in the region "unfavorable" to site location was 0.00038 (17 / (17 + 44204)). Figure 4.3 and 4.4 are predictive surfaces produced by the GIS to depict model efficiency.

Figure 4.3. Probability surface for Union Civil War fortifications (Michler map 38).

Figure 4.4. Map illustrating predicted site and non-site locations in relation to the 0.62 cut-point value

Descriptive statistics were produced to compare probabilities for site and non-site locations in the model (Table 4.4). This data indicates that there is a significant difference between site and non-site locations.

Table 4.4

Statistics of predictive surface from Michler map 38

	Number	Mean	Stand Deviation	25th Percentile	75th Percentile	75th Percentile	t-test t/p
Site	121	0.7667	0.2492	0.7617	0.8508	0.9197	20.438 / .000
Non-site	51861	0.2411	0.2827	0.0000	0.1100	0.4200	

Frequency distributions of the 121 sites and 51,861 non-sites in the model were plotted in Figure 4.5. Remember that in this model a cut-point probability of 0.62 was selected to discriminate between sites and non-sites. In an ideal model, where there is 100 percent accuracy, all of the sites would be plotted to the right of the cut-point value, while all of the non-sites would be plotted to the left. As one can plainly see this is not the case, however, I believe this model does an excellent job of discrimination.

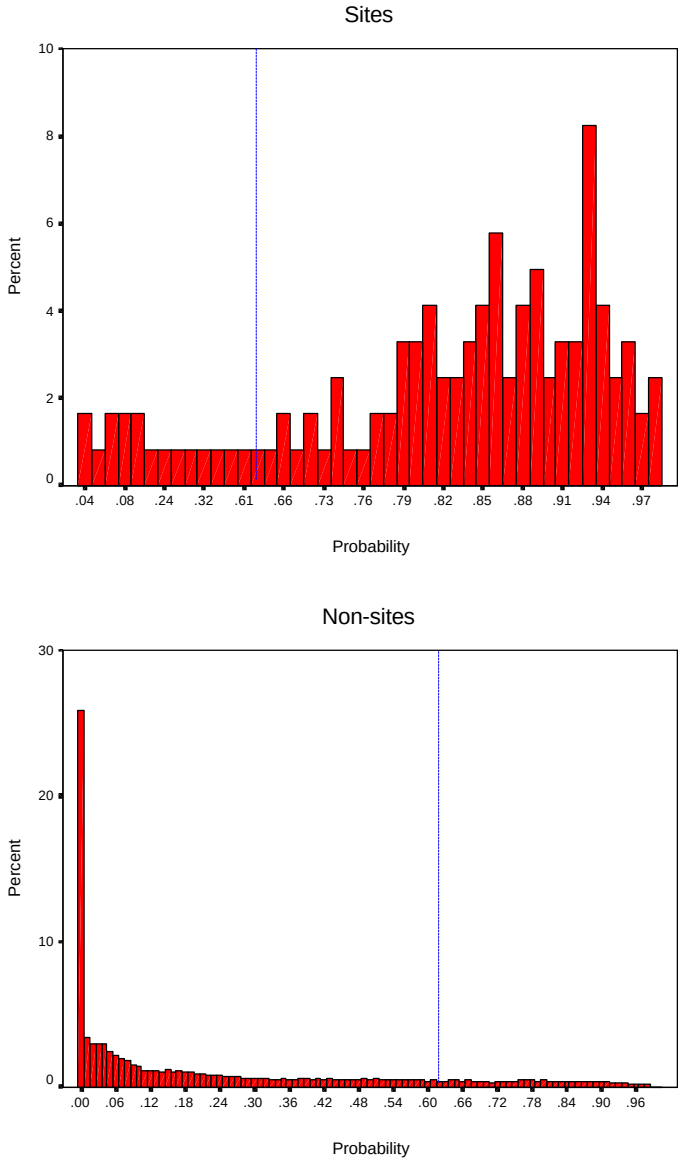

Figure 4.5. Frequency distributions of sites and non-sites along the site-presence probability gradient with cut-point value (0.62) highlighted in blue.

Accuracy of Fortification Predictive Surface for Independent Data

Internal measures of accuracy of predictive models often produce a high estimate of performance. However, these optimistic estimates are biased because there is a lack of independence among observations. Data used to develop the model was also used to test the model. To realistically evaluate the predictive probability model, a "test" predictive

surface was produced using independent observations and the coefficients (i.e., α and β) produced from the predictive probability model (Figure 4.6 and Figure 4.7).

Figure 4.6. "Test" predictive probability surface of Union fort/battery locations.

Figure 4.7. Map illustrating predicted site and non-site locations for the "test" surface in relation to the 0.62 cut-point value

Table 4.5 is the classification table for the "test" predictive probability surface. From the available 191 sites and 53,929 non-sites in this "test" area, a stratified random sample was taken (100 sites, 100 non-sites) to produce this accuracy table. This model eliminated approximately 86.29% of the study area as not likely to contain sites. Within this same region about 89.00% of sites occurred, while 99.64% of the eliminated region did not contain sites. About 11.00% of all sites occurred in approximately 13.71% of the study region predicted to be "favorable" for sites. The probability of a site in the region favorable to site locations was 0.0028 (21 / (21 + 7398), which was less than the probability of a site in the region "unfavorable" to site location (0.0036 = (170 / 170 + 46531)).

Table 4.5
Classification table for the "test" predictive surface

	Predicted		
Observed	*Non-Site*	*Site*	Percent Correct
Non-Site	86	14	86.00%
Site	88	12	12.00%
		Overall	49.00%

Unlike Michler Map 38, the classification table for the independent "test" surface illustrated a low degree of overall accuracy. This error matrix was used to produce a "k-hat" value of -0.02, a variance of 0.014, and a Z score of –0.17. These values indicate that the predictive probability model fared poorly and is not significant for this "test" area.

Frequency distributions of the 191 sites and 53,929 non-sites in the test model were plotted in Figure 4.8. Like the previous model, a cut-point probability of 0.62 was selected to discriminate between sites and non-sites. Remember, in an ideal model, in

which there is 100 percent accuracy, all of the sites would be plotted to the right of the cut-point value, while all of the non-sites would be plotted to the left. Notice that the "test" predictive surface performed poorly in delineating sites from non-sites.

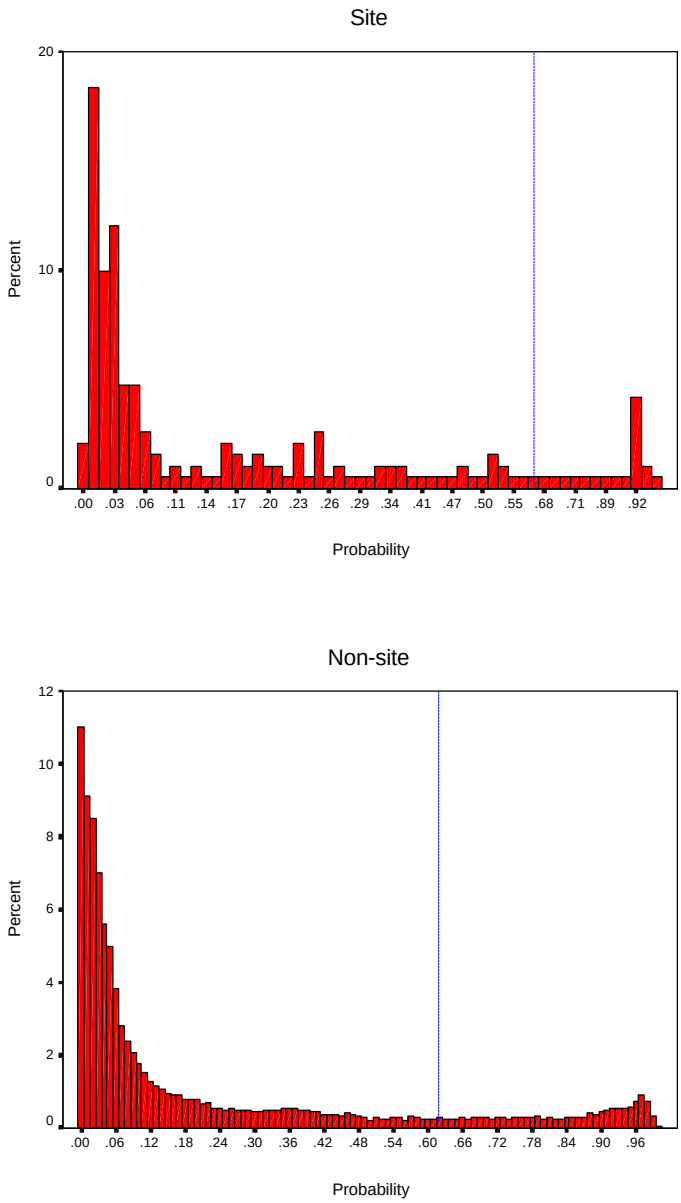

Figure 4.8. Frequency distributions of sites and non-sites along the site-presence probability gradient for the test model with cut-point value (0.62) highlighted in blue.

The statistics and probability maps above clearly illustrate that the probability model faired poorly when applied to an independent "test" area that was withheld from the initial building phase. No further analysis is required to depict this fact. However, the question must be asked: Why did the probability model fair so poorly when applied to an independent "test" area, when it predicted fortification sites accurately in Michler Map 38? It is my belief that two factors played a role in the failure of this model: One, as mentioned earlier, α and β coefficients produced by the logistic regression procedure were biased towards the development area (i.e., Michler Map 38), and therefore overly optimistic. Two, differences in battle strategies between the two areas (i.e., Michler Map 38 and the independent "test" area) affected the weight and importance of the independent variables.

Factor one, that is the event that α and β coefficients were biased towards the development area, is relatively self-explanatory. However, to further elaborate on factor two, that is differences in battle strategies between the two areas, Figure 4.9 will be used.

Figure 4.9. Extent of Michler Map 38 and the "test" coverage in relation to one another

In viewing this illustration, notice how the distance between Union and Confederate lines increases greatly the further southwest they stretch. This phenomenon alone had a direct effect on two, possibly three independent variables. Here I am referring to "Distance from Main Confederate Lines", "Distance from Confederate Fortifications", and possibly "Visibility from Main Confederate Lines". As evident from Figure 4.6, this change, in addition to Factor two, had a detrimental effect on the logistic regression formula and its' ability to predict Union fortification sites.

In the early stages of the siege on Petersburg, the prime aim of the Union army was to take Petersburg by direct attack (This strategy is depicted in Michher Map 38). However, Grant's initial directive proved futile since storming Confederate earthworks by any means was too formidable a task. To keep Lee off balance and under pressure, Grant directed surprise attacks north of the James River on the Richmond front, while continually extending the Union lines west from Petersburg. He hoped to stretch the Confederate lines to the breaking point, encircle Petersburg, and take control of all the railroads and roads supplying it.

The August 18th, 1864 attack on the Weldon Railroad at Reams Station was the first move in this new tactic (This strategy is depicted in the independent "test" area). Although unsuccessful in taking the railroad at Reams Station, it came under Union control at Globe Tavern. With the Weldon railroad in Union hands, that left only the Southside Railroad that connected Petersburg with the rest of the south, and the Richmond and Danville Railroad the only direct connection to Richmond. There were

many other subsequent engagements that followed (e.g., Peeble's Farm, September 29, 1864; Boydton Plank Road, October 27, 1864; Hatcher's Run, February 5, 1865; Five Forks, March 29, 1865) the same directive, that is, take control of all railroads and roads feeding Petersburg. It is this variation in strategy that changed the layout of Union fortifications in the landscape around Petersburg, and had a detrimental effect on the logistic regression formula and its' ability to predict these sites.

Looking at it another way, the predictive surface produced for the independent "test" area (Figures 4.6 and 4.7) highlights areas that would be considered "good" sites for a frontal assault on the City of Petersburg. That is, if the Union Army hadn't modified its' siege tactics then the fortifications would most likely be located in the 0.62 and above probability range as illustrated in Figure 4.7.