CHAPTER THREE

METHODOLOGY

Model Variables

A major concern in the initial building stage of the predictive probability model was selection of meaningful independent variables. Although a stepwise logistic regression technique was utilized, it was still a concern not to "bog-down" the algorithm with erroneous variables. Independent variables were required to meet two criteria. First, there must be belief that variation within the variable influenced the non-random placement of Union Civil War fortifications[1] within the landscape. Second, these same variables must be represented, to some extent, on available maps (i.e., modern or historic, digital or analog). Nine independent variables were selected for inclusion in the logistic regression procedure, and are discussed in the following subsections. These variables were categorized as either environmental or social.

Selection of the model's dependent variable was straightforward. The objective was to predict locations of Union Civil War fortifications within the study area; therefore, the presence or absence of these structures defined the dichotomy of the dependent variable.

Environmental Independent Variables

Five environmental independent variables were selected for the predictive model. These variables were directly associated with the natural environment of the historic landscape.

Elevation. In Mahan's (1852) "A Complete Treatise on Field Fortification", he stressed taking advantage of commanding heights in planning fortification placement. "…By giving a commanding view of the surrounding ground, increase both the range and effects of fire arms; whilst they [commanding heights], at the same time, serve to screen the troops behind them … (Mahan 1852, 140)." From this, it was hypothesized that most Union Civil War fortification sites would be located in regions of high elevation in relation to the surrounding landscape.

Visibility. As mentioned in Mahan's (1852) previous statement, "commanding views" were a valued landscape characteristic. Such well-placed fortifications would

[1] For this study, the term "Union fortifications" is used to refer to Union forts or batteries only.

allow close observation of Confederate troop and artillery movements, thus allowing for advance warning of military actions. Locations that exhibited extensive viewsheds were considered prime military positions for fortifications.

Visibility from Union fortifications was not directly measured. It was assumed that visibility from Confederate positions would be directly proportional to that of the Union fortifications. Therefore, two independent variables were contrived from visibility. One variable represented visibility from the main Confederate lines[2], while the second depicted the visibility from only Confederate fortifications[3].

Slope. Slope is the amount or degree of deviation from a horizontal surface. Steep slopes were of concern in fortification placement. "Very steep slopes will not admit of a defense with artillery, because the gun cannot be fired under a greater depression than one-sixth, and unless the shot take effect the enemy will be inspirited to advance, confiding in the safety of his position" (Mahan 1852,120). Simply put, steep slopes would not permit defense with artillery; therefore locations with ground descending in a gentle slope from the front of the fort were preferred.

Aspect. Aspect is simply the compass direction of the steepest downhill slope. Variation in aspect was not mentioned in historic text as having any deterministic effect on Union fortification site presence. However, because of aspects' close connection with the first three environmental variables, it was suspected that a *suppressor effect* relationship might exist between itself and elevation, visibility, or slope, therefore it was included in the model.

Distance from water. In some instances watercourses were great enough in size to impede an enemy's progress. Where this wasn't the case dams were sometimes built, inundating the surrounding area to protect a location. Stream banks could also be used to the enemy's advantage, therefore it was pertinent to occupy or vigilantly guard such locations (Mahan 1852). These same watercourses also supplied water for everyday tasks such as cleaning weapons, swabbing artillery tubes, drinking, cooking, washing, and bathing. From this, it was believed that Union Civil War fortifications would most likely be located near water.

Social Independent Variables

Four social independent variables were selected for the predictive model. These variables related to the man-made aspects of the historic landscape.

---

[2] Here the term "main Confederate lines" is used to refer to only Confederate forts, batteries, and the continued lines or lines with intervals that connected them.
[3] The term "Confederate fortifications" is used to refer to Confederate forts or batteries only.

Distance from Confederate earthworks. For obvious reasons, the location of Confederate earthworks was also included as a variable in the model. The Union objective was to take control of Petersburg, and in turn, end the supply of Richmond and the Army of Northern Virginia. Since Confederate earthworks lay between the Union army and its objective, the location of these earthworks directly influenced the placement of Union fortifications. Union fortifications should be close enough to the Confederate positions so that their heavy artillery would have effect, but not close enough that the enemy's medium and small range weapons would be accurate (Mahan 1852).

From this variable two independent variables were produced for the predictive model. One variable represented the distance from the main Confederate lines, while the second depicted the distance from Confederate fortifications only.

Distance from structures. Structures were also a prominent feature distributed throughout the landscape. Structures were sometimes utilized to conceal movements, shelter troops, or act as a defensive location (Mahan 1852).

Distance from railroads. Railroads were probably the most valuable resource during the siege at Petersburg. Whomever controlled these railroad lines controlled the movement of supplies, and as time would tell, the duration of the siege. It was hypothesized that Union fortifications would be located in close proximity to railroads.

Distance from roads. It was assumed that roads played a similar role in the siege as railroads. Close proximity to a road network allowed more efficient movement of supplies and communications to and from the fortifications.

Creation of GIS Map Layers

The above mentioned independent variables were represented in two source maps. These maps, although one digital and the other analog, both depicted variance of the predictors (i.e., independent variables) within the study area. To facilitate analysis within the GIS, these data sets were converted into preliminary map layers. Preliminary map layers are digital representations of the variables from the associated source maps. This conversion process involved digitizing (e.g., scanning, manual, or GPS) and various other processing steps (e.g., importing, geo-referencing) to extract variables into their own individual map layers. Preliminary map layers were further processed to produce secondary map layers. These secondary layers were coded in such a way that they directly represented the variation of the independent variable within the landscape (Fig 3.1).

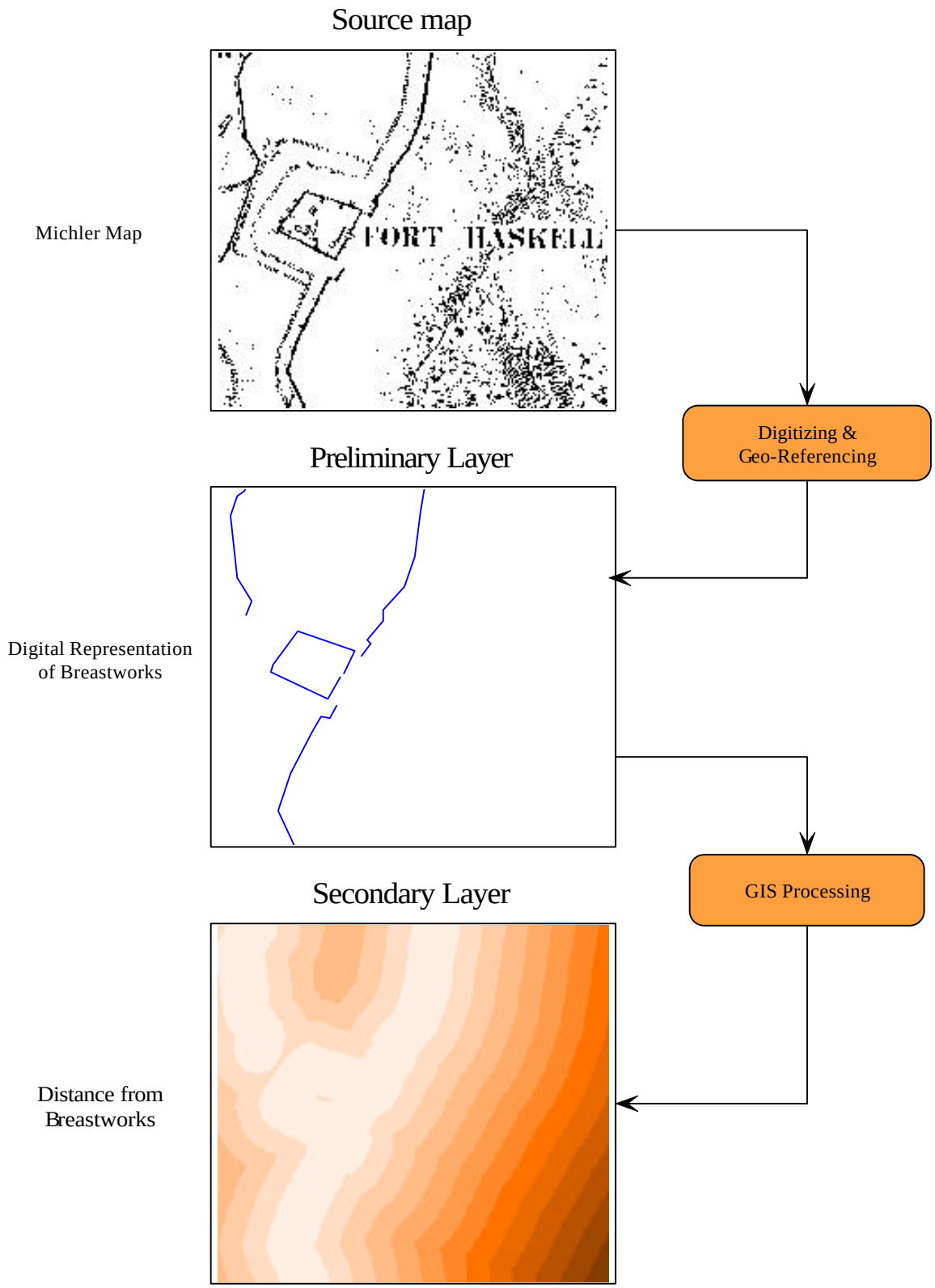

Figure 3.1. Generalized flowchart of predictive map layer development.

Source Maps

Information regarding archaeological or the historic characteristics of a landscape is often the most difficult to compile. Fortunately for this study, the historic Michler-Weyss map sheets were available. As mentioned in the Introduction, these eight maps portray the military landscape of the area surrounding Petersburg after the 1864-65 siege. Each map is at a scale of 1:7950 and together they represent approximately 148 square miles. The following features are illustrated in each map: Civil War earthen fortifications, abatis, roads, railroads, structures, vegetative cover, and land use (Figure 3.2).

Figure 3.2. Enlarged section of Michler map number 38.

Environmental information, excluding water resources, which were gathered from the Michler maps, was collected from 7.5 minute Digital Elevation Models (DEM). DEMs are produced by the United States Geological Survey (USGS) and consist of a sampled array of elevations for a number of ground positions at regularly spaced intervals. This source data, already in digital raster format, corresponds to the USGS 1:24,000 scale topologic quadrangle maps. Nine DEMs were required to cover the area corresponding to the “Michler Maps” (Fig. 3.3).

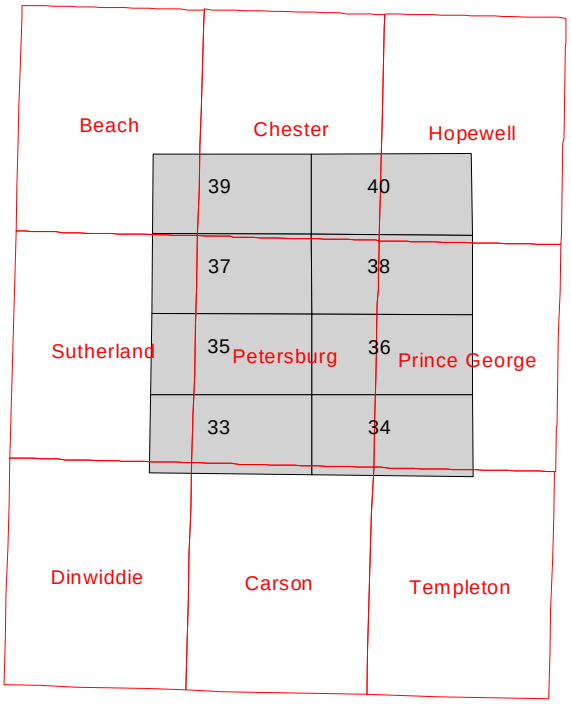

Figure 3.3. Extent of the USGS DEMs (outlined in red) in relation to the eight Michler Maps (gray areas).

Preliminary Coverages

Each of the eight historic source maps was digitally scanned from a copy obtained from the National Archives in Washington, D.C. These images were then geo-referenced to the Universal Transverse Mercator (UTM zone 18) projection and North American Datum 1927 (NAD27). Geo-referencing was accomplished by utilizing ground control points produced with a Global Positioning System (GPS) and then rubber sheeting the digital (scanned) image to those known locations. From these digital images, landscape features (e.g., structures, roads, and vegetation cover) were digitally extracted to produce individual GIS map layers (i.e. points, lines, and areas) using a process known as on-screen-digitizing. Each of these GIS map layers was also attributed. The process performed above (conversion of source map data into preliminary coverages) was performed by David Lowe and Bonnie Burns of the Cultural Resources GIS facility, Heritage Preservation Services Division of the National Park Service.

Overall accuracy of the eight Michler maps was difficult to determine. Military cartographers made some errors, while error was also introduced by the physical quality of the source maps. However, under these less than ideal circumstances Lowe estimated that an average accuracy of about forty meters was accomplished. Edges of the eight adjoining maps do not line up perfectly; this was due to the rubber sheeting process used to geo-reference the images.

All nine USGS DEMs were imported into the GIS to form individual grid (raster) layers. This grid-cell data structure is commonly used to represent digitally continuous data in a GIS. In this data structure, each individual pixel, or cell in the regularly spaced grid represents a land polygon in the corresponding study area. The resolution of the grid layer (i.e., pixel size) describes the size of each land polygon. The 7.5 minute USGS DEMs have a cell size of 30 by 30 meters, therefore each cell within the grid represents a 30 square meter land polygon in the study area. In general, the vertical accuracy of a 7.5-minute DEM is equal to or better than 15 meters (USGS 1998).

Adjacent grid layers were appended together to form one continuous layer of elevation data for the study area. This mosaic process was accomplished by utilizing a weighted average method to calculate values of cells in the overlapping areas, producing a single grid layer with a relatively smooth transition between overlapping areas of the neighboring grids. The Hermite Cubic proximity analysis algorithm was applied for this process and can be describe by the following formula (eq. 3.1):

$$H_3(s) = 1 - 3s^2 + 2s^3 \qquad \textbf{(3.1)}$$

Where, s is the normalized distance (ranging values from 0 to 1) of the width of the overlapping area.

Once combined into a single grid coverage, a focal function was applied to the mosaic layer. This focal function was used to fill gaps of no data within the grid. The following formula (eq. 3.2) describes the function:

Output Grid = CON (ISNULL (no-data grid), FOCALMEAN (no-data grid, RECTANGLE, 4, 4), no-data grid) **(3.2)**

Where, CON performs a conditional if/else situation of ISNULL on a cell by cell basis within an analysis window of RECTANGLE 4x4 on the “no-data grid”. If a cell value of no data is found within the analysis window, then a FOCALMEAN function (mean of cell values within the analysis window) is applied to assign a value to the gap cell.

Secondary Coverages

Using GIS processing procedures, multiple secondary coverages were produced from the above mentioned preliminary map layers. These secondary coverages were utilized to furnish the dependent and independent variable(s) for the predictive probability model.

The dependent variable in the predictive model was the presence or absence of Union Civil War forts or batteries within each grid cell. From the earthworks line coverage, only those works attributed as either forts or batteries were selected. From this selected set, a new line coverage was created. This line coverage was then converted from the vector data model to the raster data model, with fort and battery locations given a value of 1 (site) and every other location given a value of 0 (nonsite). Cell resolution was set to correspond to the 7.5 minute USGS DEMs (30x30 meters) (Figure 3.4).

Figure 3.4. Dependent variable grid of Michler map 38

Two additional layers were produced from the original elevation grid (Fig. 3.5) to represent slope (Fig. 3.6) and aspect (Fig. 3.7) of the region. Slope is the degree change in elevation from cell to cell (land parcel to land parcel). The GIS slope function produced a grid coverage with values that range from 0 to 17 degrees. To identify the slope direction, aspect was used. Aspect identifies the down-slope direction of the maximum rate of change in elevation from cell to cell. The GIS aspect function provided a grid with values that ranged from 0 to 360 degrees measured clockwise from north. Flat areas were assigned a cells value of -1.

Figure 3.5. Elevation grid of Michler map 38

Figure 3.6. Slope grid of Michler map 38

Figure 3.7. Aspect grid of Michler map 38

The original elevation grid, along with the vector layer representing Confederate earthworks, was used to produce two visibility analysis grids. One grid represented the visibility from Confederate forts and batteries (Fig. 3.8) while the other represented the visibility from all main Confederate works (Fig. 3.9). Using the elevation grid, the GIS visibility function built an output grid that recorded the number of times each grid cell could be seen by the Confederate positions (i.e., earthworks). Simply put, the larger the value in the grid cell, the more visible the location.

Figure 3.8. Grid of visibility from Confederate fortifications

Figure 3.9. Grid of visibility from main Confederate earthworks

The remaining independent variables were produced using a distance function and historic map layers. This function measured the true Euclidean distance, rather than the cell distance, from the source cell to all surrounding cells in the grid. Distance grids were produced from the structures point coverage (Fig. 3.10), Confederate forts and batteries (Fig. 3.11), main Confederate works (Fig. 3.12), railroads (Fig. 3.13), roads (Fig. 3.14), and water (Fig. 3.15) line coverages.

Figure 3.10. Grid of distance from historic structures

Figure 3.11. Grid of distance from Confederate fortifications

Figure 3.12. Grid of distance from main Confederate earthworks

Figure 3.13. Grid of distance from railroad lines

Figure 3.14. Grid of distance from roads

Figure 3.15. Grid of distance from water

Site and Nonsite Sampling

The area encompassed by Michler map number 38 was selected for which to build the predictive probability model for the study area. This map covers approximately 18.5 sq. miles and is digitally represented by 163 rows and 327 columns (30 meter cell size) in a raster coverage. Of the 53,301 cells, 1319 contained no data and were eliminated from the study. Of the remaining 51,982 cells, 121 were sites (Union forts or batteries) and 51,861 were non-sites.

Using the dependent variable grid (Figure 3.4), in which site cells had a value of 1 and non-site cell a value of 0, a sample was taken at each of the 51,982 cells from each of the independent variable grids (Figures 3.5 – 3.15). This procedure produced an ASCII text file containing a row of data for each of the grid cells (i.e., 52,103 rows). Each row of data contained a value for the dependent variable (0 or 1), the x and y coordinate for the cell, and a value for each of the independent variables at that cell location (Table 3.1). This ASCII data file was then imported into a statistical analysis package.

Table 3.1
Example of ASCII text file produced from sample

Cell	Dep	X	y	Elevation	Railroad Distance	Fort Visibility	…
1	0	289830.05	4125510.01	21	1194	319	…
2	1	289770.04	4125480.01	8	1237	437	…
…	…	…	…	…	…	…	…

As mentioned in Warren, et al. (1987), sample size problems are common in predictive models, since non-sites are generally more common than sites. One way to overcome this problem is to extract a random sample of sites and non-sites for analysis so their relative frequencies are about equal. Another method is to retain the biased group size in the analysis and adjust the Y-intercept constant (α) after the model has been calibrated. Group size has no effect on the regression coefficients (β), only on the intercept term (Warren 1987). To solve the problem of unbalanced group size a random sample of 100 was extracted from both site and non-site locations in the ASCII data file.

This random sample was then utilized as input for the stepwise logistic regression algorithm.

Wald Backward Elimination Stepwise Logistic Regression

Before running the logistic procedure, criteria for entry into and removal from the model were selected. The criteria value for entry into the model is denoted by p_E, while removal from the model is denoted by p_R. In backwards stepwise logistic regression the value for p_E should exceed the value of p_R to guard against allowing the procedure to enter and then remove the same variable at successive steps. To prevent failure to find a relationship when one exists; the default setting of 0.05 was relaxed. As suggested by Menard (1995) and Hosmer and Lemeshow (1989) a value of 0.20 was selected for p_E and a value of 0.15 for p_R. This range was much less stringent, and while it may have increased the risk of finding a relationship that didn't exist, it would have also decreased the risk of not finding a relationship did exist (Menard 1995).

A cutpoint probability (*c*) was also selected. The most common value for *c* is 0.5 (Hosmer and Lemeshow 1989). However, to determine the proper *c* value for this model, figure 3.16 was used. This figure graphs the percent of correct predictions for sites, non-sites and a total of both (sites/non-sites) at different *c* values. A *c* value of 0.62 was revealed from the intersection of these three lines. Each estimated probability in the model was compared to this value for *c*. If the estimated probability exceeded *c*, then the derived variable was equal to 1 (site). If the estimated probability fell below *c*, then the derived variable was equal to 0 (non-site).

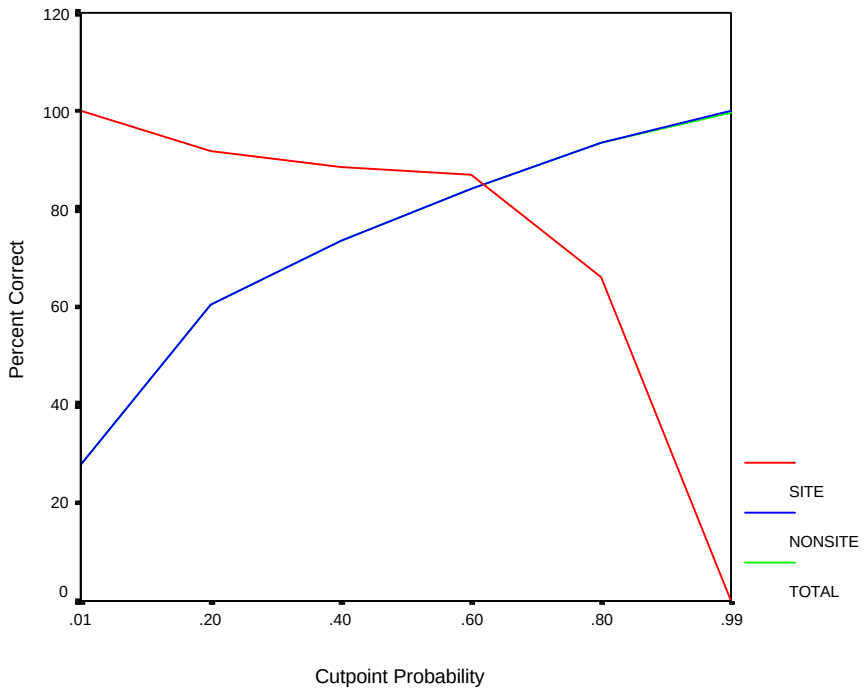

Figure 3.16. Graph of internal accuracy of the logistic regression model using different cutpoint (*c*) levels to determine the optimum value. Note that the plotted green line lies almost directly underneath the plotted blue line.

The initial step of backward stepwise logistic regression involved adding only the intercept constant (α) to the model. No independent variables were included at this step; therefore it was referred to as the "intercept only model".

At step one all eleven independent variables, plus the intercept, were added to the model (Table 3.2). Since all variables were included at this step, it was referred as the "full model". Also at this step, the program selected the variable bldg_dist for removal from the model at step two. This variable had the largest *p*-value (.8707) (Table 3.3). Since there were additional variables with *p*-values greater than 0.15 (p_R), the program proceeded to step two.

Table 3.2
Proposed independent variables for predictive model.

Variable	File Name	Measurement Interval	Scale
Aspect	asp38	1 degree	Ratio
Slope	slp38	1 degree	Ratio
Elevation	elev38	1 m	Ratio
Distance from water	water_dist	1 mm	Ratio
Distance from structures	bldg_dist	1 mm	Ratio
Distance from roads	road_dist	1 mm	Ratio
Distance from railroads	rail_dist	1 mm	Ratio
Distance from main Confederate earthworks	cmain_dist	1 mm	Ratio
Distance from Confederate forts or batteries	cfort_dist	1 mm	Ratio
Visibility from main Confederate earthworks	con_vis	1 unit of visibility	Ratio
Visibility from Confederate forts or batteries	fort_vis	1 unit of visibility	Ratio

Table 3.3
P-values for independent variables at successive steps in the backward elimination logistic regression model.

Step #	bldg_ dist	slp38	fort_ vis	asp38	road_ dist	Con_ Vis	water _dist	elev3 8	cfort_ dist	rail_ dist	cmain _dist
1	.8707	.6561	.6205	.5338	.2261	.4277	.0100	.0037	.0071	.0047	.0006
2	*	.6631	.6282	.5391	.1422	.4319	.0076	.0037	.0070	.0045	.0006
3		*	.5950	.4228	.1490	.4103	.0084	.0038	.0075	.0045	.0006
4			*	.4975	.1675	.1158	.0082	.0042	.0084	.0021	.0006
5				*	.1795	.0643	.0095	.0042	.0078	.0023	.0006
6					*	.0961	.0178	.0048	.0041	.0032	.0004

* variable in column removed from equation

At step two all independent variables except bldg_dist, plus the intercept, remained in the model. Next, the program selected the variable slp38 for removal from the model at step three. This variable had the largest *p*-value (.6631) at this step (Table 3.3). Since there were additional variables with *p*-values greater than 0.15 (p_R), the program proceeded to step three.

At step three all independent variables except bldg_dist and slp38, plus the intercept, remained in the model. Next, the program selected the variable fort_vis for removal from the model at step four. This variable had the largest *p*-value (.5950) at this step (Table 3.3). Since there were additional variables with *p*-values greater than 0.15 (p_R), the program proceeded to step four.

At step four all independent variables except bldg_dist, slp38 and fort_vis, plus the intercept, remained in the model. Here, the program selected the variable asp38 for removal from the model at step five. This variable had the largest *p*-value (.4975) at this step (Table 3.3). Since there were additional variables with *p*-values greater than 0.15 (p_R), the program proceeded to step five.

At step five all independent variables except bldg_dist, slp38, fort_vis and asp38, plus the intercept, remained in the model. Here, the program selected the variable road_dist for removal from the model at step six. This variable had the largest *p*-value (.1795) at this step (Table 3.3). Since there were additional variables with *p*-values greater than 0.15 (p_R), the program proceeded to step six.

At step six the variables con_vis, water_dist, elev38, cfort_dist, rail_dist and cmain_dist, plus the intercept, remained in the model. Here the maximum *p*-value to remove was 0.0961 for con_vis. This *p*-value was less than 0.15 (p_R), so the variable remained in the model, and the program stopped. The remaining independent variables and their β coefficients are given in Table 3.4. Full output from the Wald backward elimination logistic regression procedure is available in Appendix One.

Table 3.4
Independent variables remaining in the final model

Variable	β	P-value
Con_vis	.0076	.0961
Water_dist	.0028	.0178
Elev38	.0500	.0048
Cfort_dist	.0076	.0041
Rail_dist	-.0019	.0032
Cmain_dist	-.0099	.0004
α (constant)	.2704	.6902

Fortification Predictive Surface

Using the β coefficients from the remaining independent variables (Table 3.4) and the α coefficient, a predictive surface was produced using the GIS. The following formula was implemented as a function in GRID:

$$\text{Predictive Surface} = \frac{1}{1 + Exp(-(\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i))} \qquad \textbf{3.3}$$

Where *Exp* is a function that raises the number *e* exponentially to the power of the value enclosed in parentheses. The number *e*, Euler's number, is the irrational number whose natural logarithm is 1 (ln(1) = 2.71828…). The α coefficient denotes the intercept constant for the model, while the βs are used to represent the coefficients for the independent variables. Independent variables for the corresponding β coefficients (in this case the grid coverage of the variable) are denoted by *X*s. Further discussion of this predictive surface will be covered in Chapter Four: Results and Analysis.

Fortification Predictive Surface for Independent Data

Landscape characteristics (variables) of Michler Map 38 were used to produce α and β coefficients for the predictive probability model. In theory, these coefficients could produce efficient/accurate probability models of Union Civil War fort/battery sites in locations that were withheld from the initial model development phase. However, these α and β coefficients are biased towards the development area (i.e., Michler Map 38), and therefore overly optimistic. This bias occurs because cases used to conduct an internal test of accuracy were not independent of the cases used to develop the model. To realistically test the validity and utility of these coefficients for predictions in areas other than map 38; an additional predictive surface was produced. Using Formula 3.3, α and β coefficients produced from Michler map 38 (model development area), and an independent study area (Figure 3.17), an additional predictive surface was created to analyze the validity/utility of the predictive probability model. Further discussion of this predictive surface will be covered in Chapter Four: Results and Analysis.

Figure 3.17 Extent of independent "test" area.