THE MODELLING OF ONTARIO IROQUOIAN ARCHAEOLOGICAL SITE PATTERNS: DISTANCE TO THE NEAREST SOURCE OF WATER AND SIZE OF SITE

D.R. Bellhouse, R.J. Pearce, J.H. Peters and L.W. Stitt

Statistical models *are developed* to *describe* the *distance of an Iroquoian* archaeological *site* to its *nearest* source *of water and to de*scribe the size *of an* Iroquoian *archaeological* site. *Models* of this *type provide* the *opportunity* to *factor in any* number *of locational,* environ*mental and cultural variables,* such as *type of water, soil drainage, and cultural* time period. These models *are taken* from *the literature of actuarial* science, medicine *and reliability* theory in engineering. *It* is shown *that* the *distribution* of the distances of Iroquoian sites in southern Ontario to *the nearest* source of *water follows an exponentially* decreasing *pattern away* from *the water.* The factors *which* influence *average distance* to *water in* this *distribution are water type and* soil *drainage. All other environmental and cultural variables are* non-significant in *this model for the data under study. The distribution* of the size *of Iroquoian* sites is *also* exponential in shape, *although it follows a slightly different statistical distribution and* is *affected by* other *variables.*

INTRODUCTION AND RATIONALE

A number of studies have addressed prehistoric land use and the relationship between the location of archaeological sites and natural features (e.g., Burgar 1990; Campbell and Campbell 1992; Hasenstab 1991; Heidenreich 1971; Konrad 1973). A slightly different approach was taken by Neumann (1992) who examined the empirical frequency of site occurrences as a function of the distance to water sources in the South Branch of the Potomac River drainage basin. In the present study mathematical models taken from actuarial science, medicine and reliability theory are used to describe and study the distribution of sites from the nearest water source and the dependence of this distribution on natural features and on the cultural time period of the site. Similar models are applied to the size of the site. The data for this study were collected as part of a contract with Ontario Hydro, which employs a multi-stage process to select the location of its major transmission line corridors and associated facilities such as transformer stations. Archaeological resource concerns are addressed at all stages of this process (Peters 1986:19). Over the past several years, staff of the Land Use and Environmental Planning Department of Ontario Hydro, with input from the Ontario Ministry of Culture, Tourism and Recreation, and archaeological consultants, have reached the conclusion that one valuable avenue of planning and research is to attempt to define those areas which have the highest potential for containing as yet unknown archaeological sites. Initially, the derivation of a model for determining areas of high archaeological potential was based on the collection and analysis of available locational, environmental and cultural data for all known archaeological sites within a certain geographical area. Environmental data included topography, physiography, soils, distance to water and water type. The cultural variables collected were cultural affiliation, time period and site type. The selected geographical area was related to Ontario Hydro's proposal to construct new 500 kilovolt transmission lines. The proposed lines were to run from the Bruce Generating Station on Lake Huron, east to the Barrie area, south to the Milton area, and south to the Nanticoke Generating Station, and/or west to the London area (Mayer et al. 1985; Pearce 1989; Pearce and Pihl 1983).

These early studies found that 80 percent of all known archaeological sites within this broad area were located within a certain distance to the nearest source of water, specif-

ically within 200 to 250 m if the water source were a lake and within 150 m for all other water sources. At the same time, other locational, environmental and cultural data were found to be too variable to be included in this predictive model. The current work builds on these past studies.

Part of Ontario Hydro's multi-stage process for siting its transmission corridors is a general assessment of archaeological potential for use in environmental assessment documentation prior to route evaluation. This assessment includes the plotting of known archaeological sites and areas of high archaeological potential at a scale of 1:50,000. In the subsequent route centrelining stage, known archaeological sites and areas of high archaeological potential are mapped at a larger scale, often 1:4,000 composite maps compiled from detailed aerial photographs, topographic maps and windshield surveys. In these initial stages, it is theoretically possible for Ontario Hydro to revise the centreline of the proposed transmission corridor, or to shift the location of specific facilities, to avoid known archaeological sites or areas of high archaeological potential. The model for determining areas of high archaeological potential is also used in the latter stage to specify those areas which will require an archaeological field survey. In subsequent stages, Ontario Hydro contracts a licensed archaeologist or archaeological consulting firm to carry out field surveys. This is primarily limited to those areas which had earlier been defined as having a high archaeological potential, but includes field checks or investigations of known sites. However, Ontario Hydro also requires that the archaeological survey work include a field-based assessment of archaeological potential, which may lead to a decision to survey additional areas. These areas may include lands around springs, or dry sandy knolls which might be more than 150 m from a source of water and not visible on the 1:4,000 scale maps. Ontario Hydro also requires that archaeologists keep detailed field notes and compile specific field data on environmental conditions. These data may be used as input for planning future corridors and for refining the model of archaeological potential. The process that Ontario Hydro has used has been adapted by archaeologists for use in other projects, including large subdivision developments and other types of linear corridors such as pipelines and highways.

In the present study, refinements to the existing model not only incorporated the distance-to-water and water-type variables, but also included several environmental and cultural variables. A related model was developed to predict and explain site size. The study was based on a specific and limited set of data, namely all known Iroquoian settlement sites within three geographic areas in southern Ontario identified prior to 1990. The study excludes find spots, burials and ossuaries.

DATA COLLECTION AND DESCRIPTION

Data collection for this study was carried out by the London Museum of Archaeology. Data were collected on a total of 191 Iroquoian archaeological sites registered with the Ontario Ministry of Culture, Tourism and Recreation. The sites had been located within three specific areas in southern Ontario: (1) the central Thames River drainage in and around the City of London, (2) the central Bronte/Oakville Creek drainage around the Town of Milton, and (3) the central to southern Rouge River/Duffins Creek drainage in the Pickering area east of Toronto. The study was limited to these areas since the data for sites are more complete than for other areas and the authors are more familiar with the region. There were two reasons for including only Iroquoian sites in the study: (1) more data are available for these sites than other types of sites, and (2) Iroquoian sites are often the largest in size and the costliest to mitigate.

The study was limited to those variables for which data were available and reliable (Table 1). For example, some researchers (Robertson and Robertson 1978; Weston 1981) have based predictive models, in part, on a measurement of the difference in elevation of a site from its nearest perennial water source. At present, such data do not exist in Ontario. More recent studies (Neumann 1992) have looked at distance to water sources as a function of certain physiographic variables such as water flow, elevation and meander position of the water source.

It is important to distinguish two distance-to-water measurements used in this study. One of the measurements recorded was the distance of a site from its nearest source of water as

Table 1. Study Variables and Their Values

Variable Name	*Levels*	*N1**	*Dist.*	*N2*	*Size*
Cultural Affiliation	Early (Glen Meyer, Pickering)	44	130.3	38	0.58
	Middle (Uren, Middleport)	29	96.3	28	1.28
	Late (Neutral, Huron)	71	81.9	65	1.01
	Historic (Neutral, Seneca)	24	156.3	20	1.38
	Total	168	-	151	-
Site Type	Cabin	30	79.5	30	0.28
	Camp	50	125.9	47	0.51
	Hamlet	21	143.7	20	0.57
	Village	79	97.2	74	1.68
	Total	180	-	171	-
Water Type	Primary	10	73.5	9	0.83
	Secondary	14	98.1	13	1.20
	Tertiary	138	117.2	120	0.93
	Spring	16	22.1	16	1.23
	Marsh/Lake	4	250.0	**4**	1.10
	Swamp/Bog/Pond	9	111.7	9	0.96
	Total	191	-	171	-
Physiography	Drumlin/Kame Moraine/ Till Moraine/Moraine	27	92.4	71	1.12
	Till Plain	75	90.9	23	0.91
	Escarpment/Limestone Plain	29	128.8	27	1.15
	Spillway/Sand Plain	60	126.4	50	0.72
	Total	191	-	171	-
Soil Drainage	Poor/Imperfect	42	93.4	36	1.36
	Fair/Good	94	93.5	86	0.93
	Well/ Very Good	52	131.3	47	0.77
	Total	188	-	169	-
Topography	Drumlin/Esker	8	164.1	8	0.98
	Flat/Gentle Slope/Rolling	33	137.4	27	0.98
	Knoll/Ridge/Plateau/Terrace	149	99.1	135	0.98
	Total	190	-	170	-
Study Area	Central Thames River drainage (1)	91	112.5	81	0.28
	Bronte and Oakville Creek drainage (2)	51	122.1	45	0.51
	Rouge River and Duffins Creek drainage (3)	49	85.1	45	0.57
	Total	191	-	171	-

**N1* = the number of cases for which measurements were made on distances; *Dist* = the distance to the nearest source of water in metres using the best distance; *N2* = the number of cases for which measurements were made on the size of site; *Size* = size of site in hectares.

plotted on a 1:50,000 or 1:250,000 scale topographic map. A second measurement (obtained for 72 percent of the sites) was the actual distance of a site from its nearest source of water as measured or observed in the field. In all instances, the actual distance as measured in the field was less than or equal to the distance as measured on the topographic map. In some cases, the actual distance was significantly less than the mapped distance. This was especially true if the water source was a spring, since springs seldom appear on topographic maps. For example, a site may have a mapped distance of 150 m from a tertiary watercourse, but an observed field distance only 25 m from a spring. This difference is clearly reflected in the mean distances for each of the two variables. The mean distance based on all the 191 measurements taken from maps was 133.4 m, while the mean distance based on the 137 cases in which the distance was obtained in the field was 84.1 m. For this study we created a new distance variable which combined the two measurements. The new variable was the distance obtained in the field if known (137 of 191 cases); otherwise mapped distance was used. The mean value of this new variable, denoted throughout the article as the best known distance, was 109 m.

It was possible to collect data on several different specific soil types for most sites in the sample. These were taken from county soils maps or site record forms, or were inferred from field notes and personal observations. These soil types (e.g., Cashel Clay, Berrien Sandy Loam, London Loam, etc.) were found to be too variable and to have too few cases per specific soil type in the sample to allow for reasonable statistical analysis. Hence, the soil types were merged into broader categories: clay, clay loam, sandy loam, and sand. An analysis of soil types showed no significant differences among the four broad categories. Other data on soils included the drainage characteristics of the soil. Three statistically significant groupings of the drainage variable were found: (1) poor/imperfect drainage, (2) fair/good drainage, and (3) well drained/very good drainage. Since data on these categories were available for most of the sites, the soils data used in this study were comprised only of the drainage characteristics of the soil.

An initial statistical analysis was performed on all variables based on the data values as compiled. After these analyses and considerable discussion, several of the variables were re-coded. The re-coding was justified on the basis of observations on the initial statistical analysis and for various logical reasons. For example, the variable describing physiography initially had nine values; these were later collapsed into four categories based on the fact that different physiographic zones were formed by similar geological processes (Table 1). A complete description of, and rationale for, the collapsing of categories is contained in a report to Ontario Hydro (Pearce et al. 1989). The report contains a complete tabulation of both the raw data and the re-coded data. One of the variables collected was the area of a site in hectares. With the inclusion of this size variable, statistical settlement models could be developed not only for distance from the water source, of primary interest to Ontario Hydro, but also for size of the site. A number of important questions could now be asked and answered: Is site size related to physiography or to the goodness of the soil drainage? If so, how?

MODELS FOR DISTANCE TO WATER AND SITE SIZE

Models which describe the distribution of the distance of sites to the nearest source of water and the distribution of the sizes of sites can be useful for both explanation and prediction. In the explanatory mode, it would be useful to know the effect on distance to water and site size of changing cultural and physical attributes associated with the sites. It would also be of interest to know the shape of the distribution since this provides one descriptive measure of the settlement process (e.g., the exponentially-decreasing shape of the distribution in Figures 1, 2 and 3). In the predictive mode, it would be of interest to chart the areas around water sources in which sites are likely to be found.

Some general ideas of the type of models which may fit the distance and size variables can be obtained from histograms of data collected on these variables. Figures 1 and 2 show the distribution of distances for all 191 sites in the study. In Figure 1, only the mapped distances to the nearest source of water are shown. In Figure 2, the 137 distances obtained in the field are substituted for their mapped

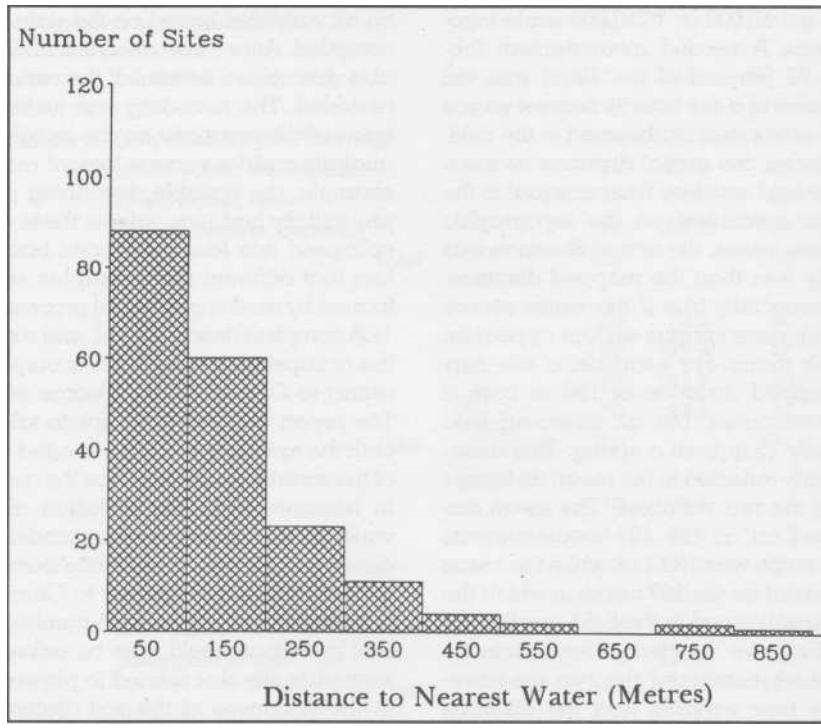

Figure 1. Frequency Distribution of the Mapped Distances in Metres to the Nearest Source of Water. The distance measurement is based on the mapped distance.

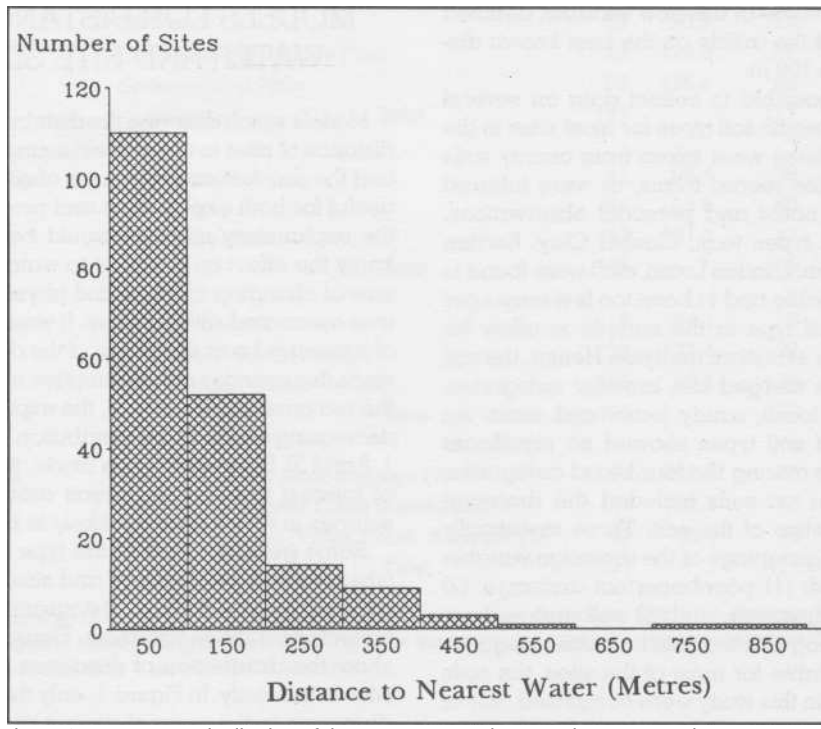

Figure 2. Frequency Distribution of the Best Known Distances in Metres to the Nearest Source of Water. The distance measurement is based on the actual distance.

distances (i.e. the best distance measurement is used). Figures 1 and 2 suggest that mapped distances tend to be longer than distances obtained in the field. Figure 3 shows the distribution of the site areas for 171 of the 191 sites in the study. The distributions in Figures 1, 2 and 3 have two important similarities: both variables — distance and size — are non negative so that both histograms begin at zero, and the number of cases decreases dramatically as distance or size increases.

The type of statistical models which fit data similar to those shown in Figures 1, 2 and 3 are known as survival distribution models in the medical and actuarial science literature, or lifetime distribution models in the probabilistic reliability theory literature of engineering (Lawless 1982; Nelson 1982). In this literature, the distributions are all functions of time, although the models can be adapted to distributions which are functions of distance or size. The mathematical form of these models is often given, in the medical and actuarial literature, in terms of the survival distribution function. The survival distribution function at a given age describes the fraction of individuals expected to be alive at that age, or to survive to that age, from an initial pool of individuals followed from birth. In reliability theory the survival distribution is called the reliability function. If lifetimes of electronic parts are tested, the reliability function at a given time describes the fraction of parts that are expected to be operating, or to be reliable, at that time. In the archaeological context this nomenclature has no meaning, and the term "settlement function" is adopted instead. The distance settlement function at a particular distance describes the fraction of sites expected to be further from their water source than the given distance. Likewise, the size settlement function at a particular size describes the fraction of sites expected to be larger than the given size.

Let "y" denote either the distance of a site from its nearest water source or the size of the site. The settlement function, as a function of y, will be denoted by S(y). The survival distribution function or reliability function (the counterpart of the settlement function in the medical, actuarial or engineering literature) is often expressed in the form

$$S(y) = \exp\{-H(y)\},\ y > 0,$$

where exp denotes the natural base, or 2.7182818 to seven decimal places. In the medical literature $H(y) = -\log[S(y)]$ is called the cumulative hazard function; the first derivative of H(y) with respect toy, denoted by h(y), is called the hazard rate function or force of mortality. In the context of the settlement data, h(y) will be known as the force of settlement. The force of settlement may be viewed as a mathematical summary of the natural forces which limit habitation far from water or which limit the size of settlement. For distance data, the interpretation of h(y) is that it is the instantaneous rate of decrease in the number of sites at distance y from the source relative to the fraction of sites that are at least y metres from their nearest source of water. The force of settlement for size is the instantaneous rate of decrease in the number of sites at size y relative to the fraction of sites that are at least y hectares in area.

Two statistical models that will be used for the settlement data are the exponential and Weibull distributions (Lawless 1982:14-19). In the exponential distribution $H(y) = \forall y$, a straight line through the origin, where $\forall$ is a constant but unknown number greater than zero. The force of settlement $h(y) = \forall$. The constant $\forall$ is a parameter of the distribution. The mean and standard deviation of the exponential distribution are the same, $1/\forall$ (Lawless 1982:14). The Weibull distribution is similar to the exponential distribution. The difference is one additional parameter, denoted by k, also a constant greater than zero. The function H(y) for the Weibull distribution is $H(y) = (\forall y)^k$, so that the force of settlement under this model is $h(y) = \forall k(\forall y)^{k-1}$. Under this model, as y increases, the force of settlement increases geometrically. The mean and standard deviation of this distribution are fairly complicated functions of $\forall$ and k (Lawless 1982:16), but have the property that, for a fixed value of k, the mean is proportional to the standard deviation. Note that for the Weibull model

$$\log[H(y)] = \log\{-\log[S(y)]\} = k\log(\forall) + k\log(y),$$

a straight line in log(y).

Both the exponential and Weibull models can be extended to include regression (Lawless 1982:273-274). With this extension, the dependence of distance or size on other variables can be tested and the size of the de-

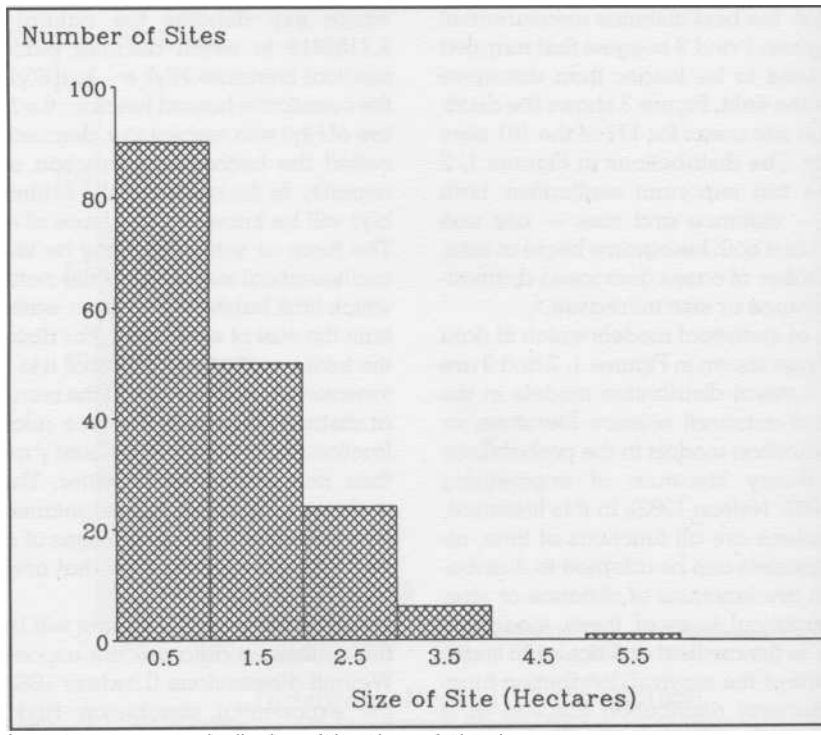

Figure 3. Frequency Distribution of the Sizes of Sites in Hectares

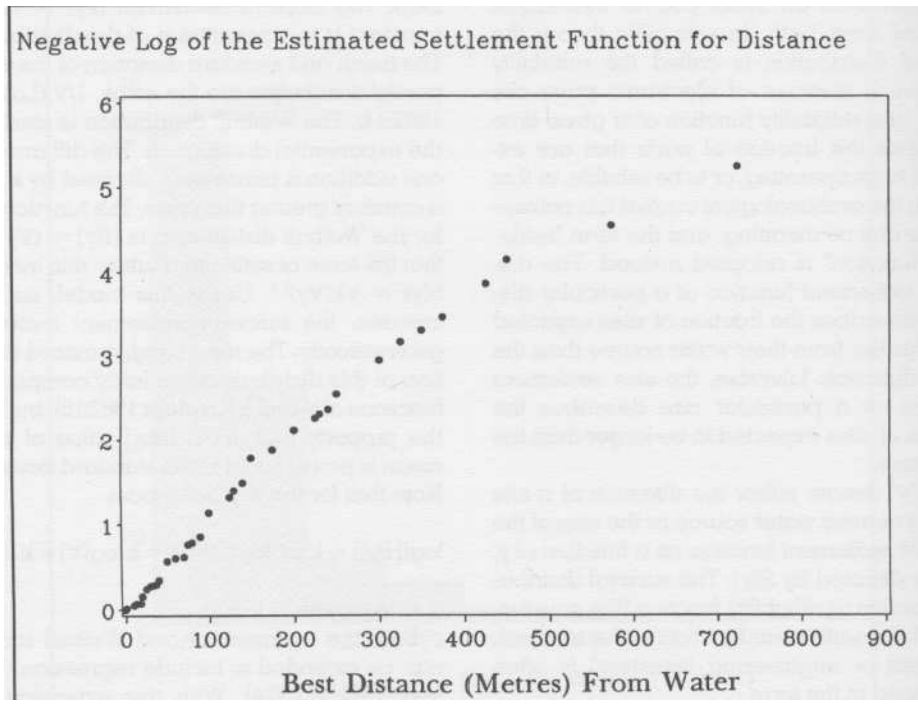

Figure 4. Graphical Test for the Presence of an Exponential Probability Distribution for Distances using the Best Known Distances to the Nearest Source of Water

pendence can be estimated. A regression model can be incorporated into the exponential and Weibull models by replacing $1/\forall$ in either model by the natural exponent e = 2.7182818 to the power of the regression model mean. Suppose, for example, that it is of interest to test the dependence of y, distance or area, on some other variable, say x. In simple linear regression it is assumed that the average value of y is $\forall + \exists x$ (Neter et al. 1990:32). Then the settlement function is given by

$S(y) = \exp\{-y/\exp(\forall + \exists x)\}$,

if the distribution of y is exponential, and by

$S(y) = \exp[\{-y/\exp(\forall + \exists x)\}^k]$,

if the distribution of y is Weibull. Extension of these models to multiple regression, and to regression with indicator variables or analysis of variance models, is straightforward; simply add more regression terms after $\forall + \exists x$.

MODEL IDENTIFICATION

There is a graphical procedure that may be used to distinguish whether the settlement data follow an exponential, a Weibull, or perhaps another distribution. The technique, called hazard plotting or Nelson plotting (Elandt-Johnson and Johnson 1980:196-200; Nelson 1982:131-146), uses the functional relationships between log[S(y)] and H(y) to identify the form of H(y). Suppose data y1, ..., y_n have been collected. From the data an estimate, call it ES(y), of the settlement function may be obtained. ES(y), for a given distance or area y, is that fraction of the observations y_i, ..., y_n which are greater that the given y. If the plot of $-\log[ES(y_i)]$ against y_i for i = 1, ..., n shows an approximate straight line through the origin, then the exponential distribution is appropriate. If the plot of $\log\{-\log[ES(y_i)]\}$ against y_i shows an approximate straight line, then the Weibull distribution is the appropriate model.

Figures 4, 5 and 6 show plots of $-\log[ES(y_i)]$ against y_i, where the y_i are the 191 distances to the nearest source of water. In Figure 4, the distance measurements that are used are the best known distances. Although the plot of points goes through the origin, the points do not follow a straight line. The reason for this is apparent in the plots found in Figures 5 and 6 which show straight lines through the origin, but with different slopes. The plot in Figure 5 is based on mapped distances only and the plot in Figure 6 is based on distances obtained in the field. The results of these plots indicate that the distance to the nearest source of water is exponentially distributed, but the mean of the measurement depends on how the measurement was taken, in the field or from a map. Since the mean may also depend on other variables or factors, an exponential model was adopted with the mean depending on how the measurement was taken or on various environmental and cultural factors.

Figures 7 and 8 show plots of $-\log[ES(y_i)]$ against y_i and $\log\{-\log[ES(y_i)]\}$ against y_i respectively, where the y_i are the 171 known site size in hectares. The plot in Figure 8 appears to fall along a straight line better that the plot in Figure 7. The systematic deviations from the line in Figure 8 are due probably to the differences in the average size for various types of sites. Consequently, it was assumed that the size of a site followed a Weibull distribution with the mean of the distribution depending on various environmental or cultural factors.

DATA ANALYSIS

Once the underlying probability distribution has been decided for distance or size, it is necessary to discover which environmental and cultural factors influence the mean distance or size. Since the data on the environmental and cultural factors were all obtained at the nominal level of measurement, one-way and multi-way analysis of variance techniques are appropriate (Montgomery 1991:43-58, 189-214, 236-240; Neter et al. 1990: 528-550, 761-772, 846-848) to find which of these factors are significant. Additionally, any cross classification of the factors in the current data yield an unequal number of cases per cell. Consequently, the analysis of variance procedures will involve techniques for unbalanced data. The statistical procedure GLM (generalized linear models) in the SAS computer package, described in the SAS User's Guide (SAS Institute 1985:433-506), was used to analyze the data.

The problem with using the usual analysis of variance techniques directly on the data collected is that two of the assumptions required to make valid use of these techniques are

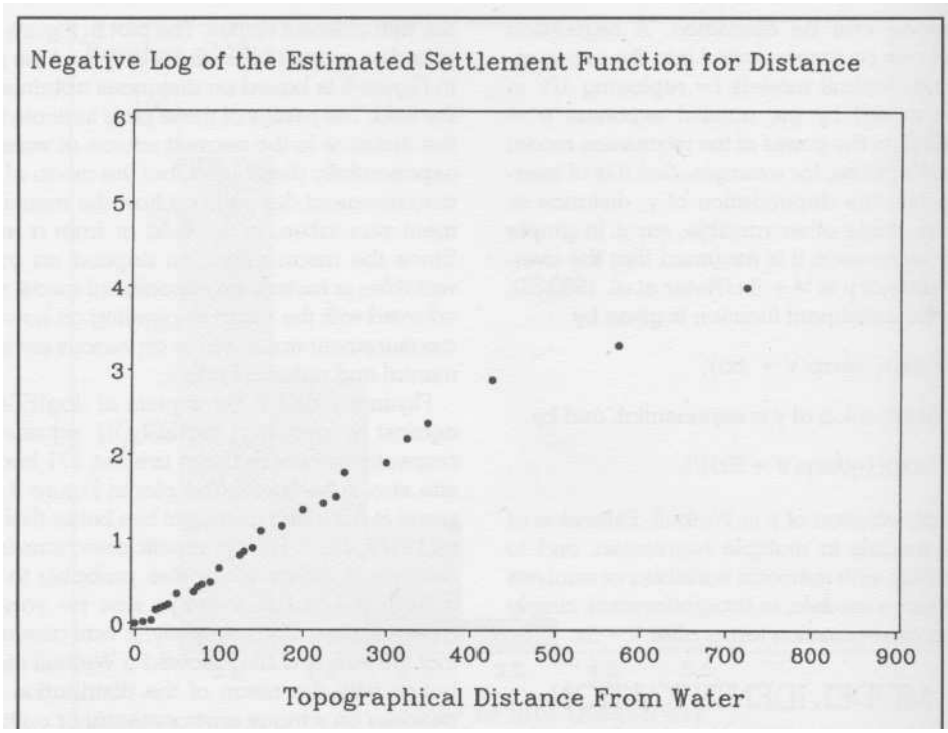

Figure 5. Graphical Test for the Presence of an Exponential Probability Distribution for Distances using Mapped Distances Only

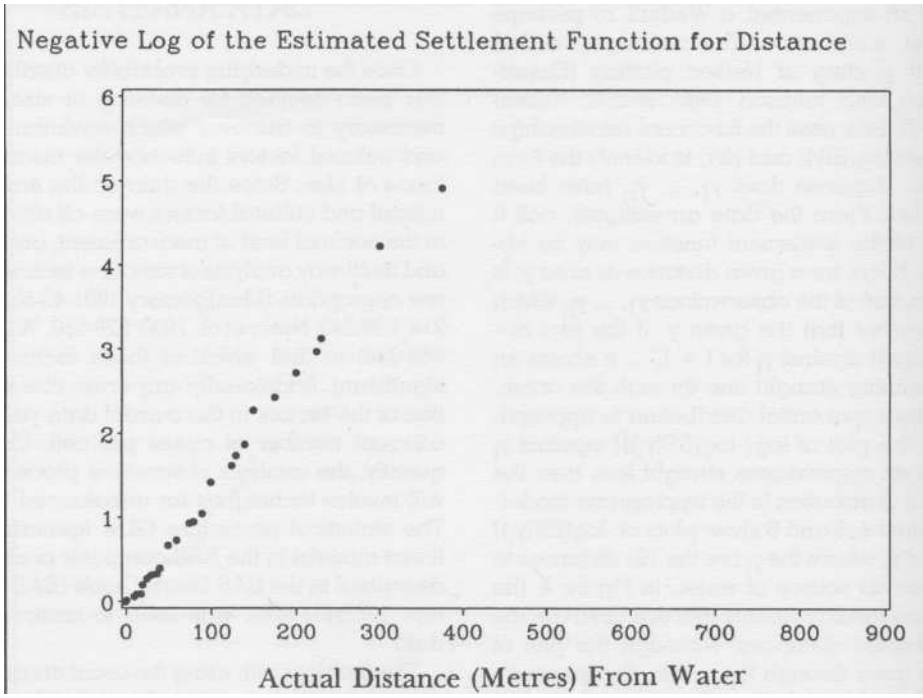

Figure 6. Graphical Test for the Presence of an Exponential Probability Distribution for Distances using Only the Distances Determined in the Field

violated. The first is that the underlying distributions are not normal, but follow exponential and Weibull distributions instead. Secondly, the variances of the data, either the distance to water or the size of site, are not constant over all the variables collected. From the Weibull and exponential models, it is known that the variance varies with the mean; specifically, the square root of the variance, or standard deviation, is proportional to the mean. Equality of variances over all variables can be achieved by transforming the data and then carrying out the analysis on the transformed data. When the standard deviation is proportional to the mean, the appropriate procedure is to perform analysis of variance techniques on the natural logarithm of the data (Montgomery 1991:103-105; Neter et al. 1990:620-622). Fortunately, for the data examined here, this transformation also yields an approximate normal distribution on the transformed values.

A further aspect of the data which complicates the analysis is that there is a large number of variables or factors, both environmental and cultural, for which the significance of each factor is to be tested. Some of these factors are strongly correlated with one another, a situation that may lead to the problem of multicollinearity (Neter et al. 1990:300-304). If an omnibus model is used blindly, the presence of multicollinearity could lead one to conclude that a given variable is insignificant when in fact it is significant. Hence, the approach that was selected was similar to stepwise regression (Neter et al. 1990:453-458). A one-way analysis of variance was run using each factor separately. This was followed by a two-way analysis using each pair of significant factors from the previous analysis. This was continued to three-way and higher as necessary.

Each variable, or factor, has several attributes or levels associated with it. For example, the factor site type" has four levels, namely sites which are cabins, camps, hamlets or villages. If a factor with two or more levels is found to be "significant" in the analysis of variance, one concludes that at least one of the levels of the factor has a different effect on the average than the other levels. The test does not indicate which factor levels are different. For example, in the following analysis the factor site type" is found to be "insignificant" when the analysis is done on the distance to the nearest source of water, and "significant" when the analysis is done on the size of the site. For the variable on distance to the nearest water source, it may be concluded that each site type has the same average distance to the water source. For the size of the site, it may be concluded that at least one of the site types - cabin, camp, hamlet and village - is different from the rest.

In order to find where the differences occur, the Bonferroni t-test, among several others, may be used (Dunn and Clark 1974:80-81). The Bonferroni test works well for comparing pairs of a small number of factors. It is also available in the package program SAS. At the heart of this testing procedure is the calculation of confidence intervals for the difference in the means of two factor levels. These confidence intervals are calculated for all pairs of factor levels. From the point of view of data analysis, factor levels which have the same mean are then grouped together and treated the same for analysis. This is not to say that the grouped levels of a factor would be considered identical. What is said only is that grouped levels of a factor have the same average size.

After the number of factors and factor levels has been reduced to a reasonable size using the usual analysis of variance techniques on the logarithms of size and distance variables, then settlement functions, based on the exponential or Weibull distributions can easily be estimated using the SAS procedure LIFEREG.

From Figures 1 and 2, or from Figures 4, 5 and 6, it is apparent that the mean distance to water depends on whether the distance measurement was taken in the field or from a map. Not all field distances were obtained in this study. Field distances are usually more accurate and reliable than distances taken from a map. Maps may not record the presence of springs — one of the possible sources of water. In view of this, the distance variable used is the best distance, the field distance if it is available, or the mapped distance if the distance measurement in the field is not available. Consequently, any analysis of variance models or any settlement functions, will include the nature of the distance measurement, whether taken in the field or from a map, as a factor in the model.

The data analysis on the distance to the nearest water source was begun by running all two-way analyses of variance with one of the factors being whether or not the distance

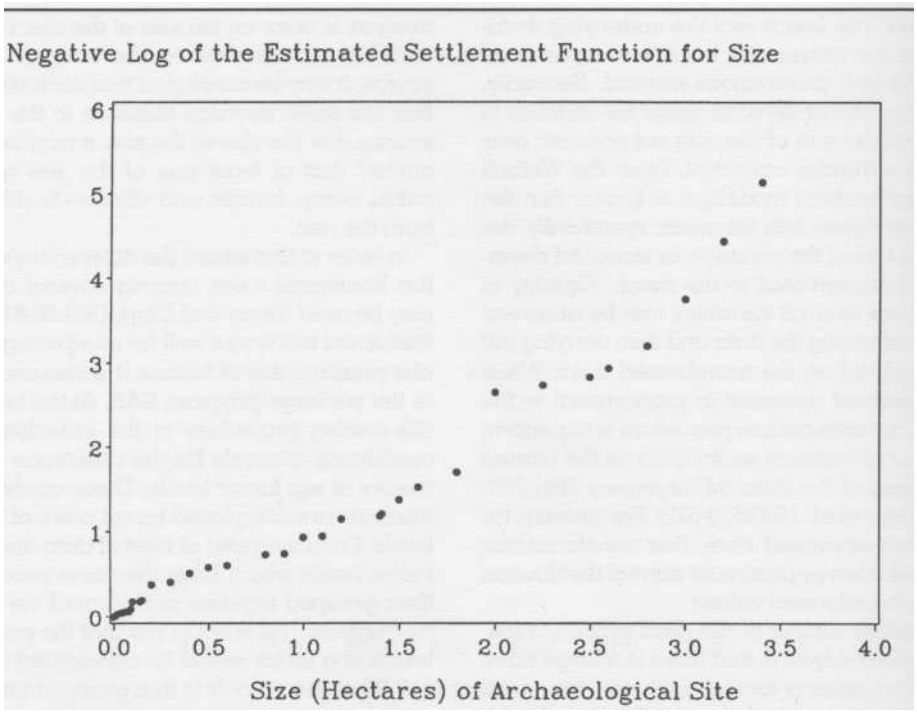

Figure 7. Graphical Test for the Presence of an Exponential Probability Distribution for the Size of Sites

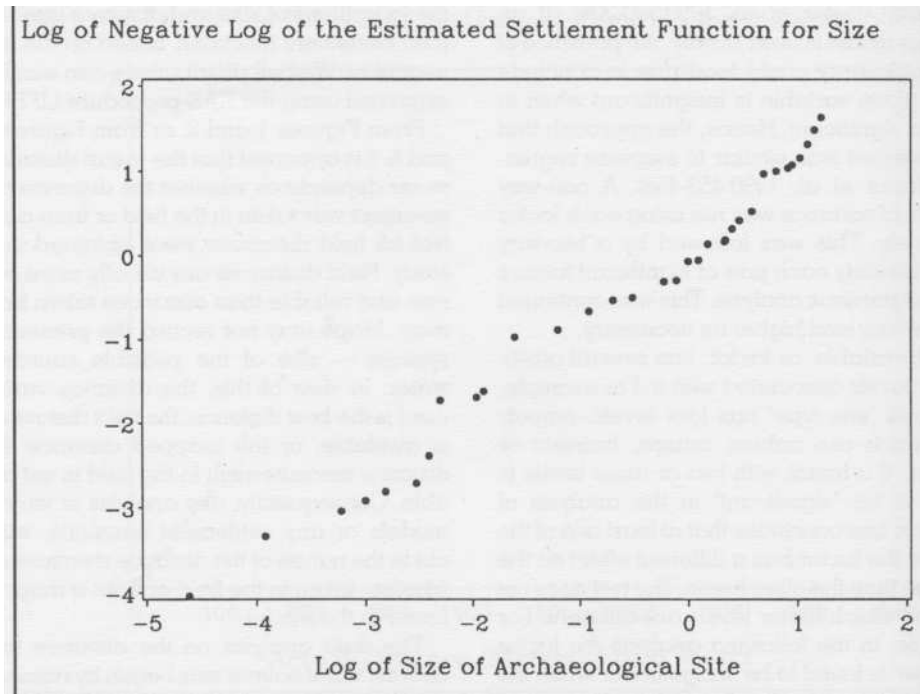

Figure 8. Graphical Test for the Presence of a Weibull Probability Distribution for the Size of Sites

measurement was taken in the field. The factors "cultural affiliation", "site type", "physiography" and "topography" were all found to be not significant in their respective two-way analyses, even at a 10 percent level of significance. Consequently, these factors were dropped from any further analysis of the distance to the nearest water source. On the other hand, the factors "study area", "water type" and "soil drainage" were all statistically significant, even at a 1 percent level of significance. The Bonferroni tests on "study area" indicated that the average distance was the same in Areas 1 and 2, while the average in Area 3 was much less. With regard to "water type", the Bonferroni tests showed that the average distance was the same to all water sources, except springs. The distance to a spring, when it was the nearest water source, was much lower, on average, than the average for all other sources of water. The Bonferroni tests on "soil drainage" showed that the average distance to a site on land described as very good or well drained was much higher than sites on land with poorer drainage (land described as having poor, imperfect, fair, or good drainage). In summary, it was found after the initial set of two-way analyses, that the average distance of a site to the nearest source of water could depend on: (1) the distance measurement that was used, in the field or from a map, (2) whether or not the study was carried out in Area 3, (3) whether or not the nearest source of water was a spring, and (4) whether or not the drainage in the soil could be classified as well drained.

On further examination of the data, it was found that the two factors "study area" and "soil drainage" may be confounded (i.e., they may both explain the same phenomenon). A cross tabulation of the two factors showed that there were no sites from Area 3 which were on very good to well drained land. Since the average distance to water is high for sites with the best drainage, it should be expected that sites in Area 3 (which contains no cases of this type) exhibit a lower average distance than the other two study areas. Looking at the problem from another perspective, since the average distance to water in Area 3 is lower than that of the other two areas, the average distance to water on better drained land (which appears only in Areas 1 and 2) should be higher than the average for less well-drained land.

The next step in the analysis was to combine each of the significant effects into a single multi-way analysis of variance which included interaction terms. Two approaches were used for this step. The first is what is called the Type I sum of squares approach (SAS Institute 1985:83-84). As new factors are entered into the model, their significance is tested while controlling for factors which have previously been included in the model. The testing procedure is sequential in nature and parallels the forward selection procedure for building linear regression models (see Seber [1977:380] for a brief discussion of this methodology). In the second approach, Type III sums of squares were used to test for the significance of each effect. Here, tests of significance for any effect are made while controlling for all other factors considered for the model (SAS Institute 1985:87-90).

To account for the widest possible outcomes in the data, we included terms in the model known as interactions. The concept of interaction is defined by Hicks: "When a change in one factor produces a different change in the response variable [in this case distance to the nearest water source] at one level of another factor than at other levels of this factor, there is an interaction between the two factors" (Hicks 1993:124; emphasis in original). For example, consider "water type" and "soil drainage" as two variables in the model. There are two levels of each factor: whether or not the source of water is a spring and whether or not the soil is well-drained. Interaction is present if the change in the average distance to water when going from well-drained to poorly-drained soil (when the water source is a spring) is different from the change in the average distance to water when going from well-drained to poorly-drained soil (when the water source is not a spring).

In the Type I sums of squares approach, the significant factors from the first analysis were entered in the order of what was thought to be their decreasing relative importance in terms of their effect on the mean distance to water. The order under consideration was: whether or not the distance measurement was taken in the field, "beater type", "soil drainage", an interaction term between "water type" and "soil drainage", "study area", and an interaction term between "study area" and "water type." The levels of each factor were the reduced set

of levels determined by the Bonferroni t-tests. The Type I analysis was carried out by examining the method of distance measurement, then by examining "water type" after removing the effect of the method of measurement, then by examining "soil drainage" after removing the effect of the method of measurement and "water type", and so on. With this approach, it was found that all factors and interactions were significant at a one percent level of significance, except the "soil drainage" by "water type" interaction which was not significant even at a 10 percent level of significance.

When the Type III analysis was used, the same conclusions were reached, although the factor "soil drainage" (when controlling for all other factors in the model) was no longer significant even at a 10 percent level of significance. In view of the confounding between factors "soil drainage" and "study area", the effect of the former factor is masked by that of the latter.

Based on the analysis given thus far, two settlement functions in the form of statistical models can be proposed to describe the distribution of sites to the nearest source of water. Both statistical models are of the exponential form

$$S(y) = \exp\{-y/\mu\},$$

where μ, the mean, can be written as the natural base e to a regression expression in the form of the multi-way analysis of variance. The parameters which comprise μ in either model were estimated using the LIFEREG procedure with the exponential option in SAS. The first statistical model is

$$S(y) = \exp[-y/\exp\{\forall + \exists I + (J + *K + yJK\}], \quad (1)$$

where y is the distance in metres to nearest water source and where ∀, ∃, (, * and y are parameters estimated from the data. Also in model (1), the term I = 1 if the measurement to the nearest water source was obtained in the field and 0 if from a map; the term J = 1 if the nearest water source is a spring and 0 otherwise; and the term K = 1 if the study area is Area 3 and 0 if the area is 1 or 2. In the second statistical model, the factor for "study area" is removed and replaced by "soil drainage". The model is

$$S(y) = \exp[-y/\exp\{ 8+ 0I + BJ + <L\}], \quad (2)$$

where 8, **0, B** and < are parameters estimated from the data. The terms I and J are as in model (1), and the term L = 1 if the soil drainage can be described as well or very good drained soil and is 0 otherwise. Although the factor "soil drainage" was not significant in the preliminary analysis, because its effect was masked by the factor "study area", model (2) may be preferable to model (1). Model (2) depends only on physical factors while one term in model (1), "study area", was an arbitrary geographical division.

Either model (1) or (2) may be used to predict archaeological potential around various sources of water. These models may be used to answer one of two questions: how far from a water source should one go to ensure that a certain percentage of the sites are within that distance; given a certain distance from a water source, what fraction of sites are that particular distance or closer to their respective sources of water?

The parameter estimates for both (1) and (2) were obtained using SAS. The estimates of the parameters in (1) are: 5.360 for ∀, -0.750 for ∃, - 1.388 for (, -0.592 for * and -1.315 for y. Similarly, the estimates of the parameters in (2) are: 4.967 for 8, -0.535 for 0, -1.370 for B and 0.285 for <. In model (1), since the estimate for 3, for example, is negative, the average distance to water is smaller when I = 1 than when I = 0 (i.e., when the measurement is taken in the field rather than on a map). In model (2) the estimate of < is positive. From this it may be interpreted that the average distance to water increases when the soil drainage is very good. In both (1) and (2), the estimates of the parameters associated with the variable J, the indicator showing whether or not the water source is a spring, are large negative values. This may be interpreted to mean that a site is much closer to water when the water source is a spring than with any other water source.

There is much more work that could be done in statistical model building to explain the variation in the distance of a site to its nearest source of water. The current factors used in the models thus far account for less that 40 percent of the total variation in the data. To increase the percentage of explained variation, more data on other factors should be collected. It may also help to increase the sample

size beyond the number (191) available at the time the study was initiated. Some evidence of other possible interaction terms was found between the "cultural affiliation" and "water type" variables, between the "cultural affiliation" and "topography" variables and between the "physiography" and "topography" variables. The evidence for interaction was based on very small sample sizes in many of the cells, with very large sample sizes in the remaining cells. In the present study it would be inappropriate to include these terms in the statistical model.

SIZE OF SITE

The analysis of the size of site variable followed along lines similar to the distance to the nearest water source variable. Using analysis of variance techniques on the logarithms of the sizes of the sites, factors were eliminated which had no significant effect on the average site size. Bonferroni tests were then used to group factor levels which had the same effect on the site size. Finally, the settlement function for size was obtained using the LIFEREG procedure in SAS with the Weibull option for the model. Only the settlement function for the explanatory model on size is given here. The statistical model, obtained through the data analysis is given by

$$S(y) = \exp[-(y/\exp\{\forall + \exists I + (J + {}^{*}K + yL + NJM\})^{k}], \quad (3)$$

where ∀, ∃, (, * and y model parameters different from those in model (1), and where N and k are also model parameters. Also in model (3): I = 1 if the cultural affiliation is "early" and is 0 otherwise; J = 1 if the site type is a cabin and is 0 otherwise; K = 1 if the site type is a camp and is 0 otherwise; L = 1 is the site type is a hamlet and is 0 otherwise; and M = 1 if the study is area 3 and is 0 otherwise. The parameter estimates for model (3) are: 0.700 for ∀, -0.704 for ∃, -5.304 for (, -1.023 for *, -1.073 for y, 3.166 for N and 1.452 for k.

SUMMARY OF RESULTS

The distance of a site to the nearest source of water in metres may be adequately modelled by an exponential distribution with the mean distance in the model expressed as a function of various other factors. Two different models, given by expressions (1) and (2), for the modelled mean distance to water sources were supported by the data collected. In both models, the mean depended upon whether the distance measurement was taken from a map or in the field, and upon whether or not the nearest water source was a spring. One form for the mean distance depended upon whether or not the study area was near Pickering, Ontario (Study Area 3), while in the other form the mean depended on the goodness of the soil drainage. As noted in Table 1, there is substantial variation in the mean distances to various types of water. However, in the statistical tests, only the mean distance to springs was significantly different from the rest. This may be attributed to the fact that there is substantial variability in distances to the water source within each type of water source as well as between each type.

As stated earlier, models (1) or (2) may be used to answer the questions of how far one should go from a water source to ensure that a certain percentage of the sites are within that distance, and (for a given distance from a water source) what fraction of sites are expected to be within that distance. For the former question the percentage is specified before the distance is determined, and for the latter question the distance is specified before percentage is determined. Suppose, in the first question, that the source of water is a spring, that the soil drainage is not very good and that we wish to capture 90 percent of Iroquoian sites. The answer is the solution for y in the equation

$$1 - 0.9 = \exp[-y/\exp\{4.967 - 0.535(1) - 1.379(1) + 0.285(0)\}],$$

or y = 49 m. In the second question, suppose that the source of water is a tertiary stream and that the soil drainage is very good. For all areas with these given characteristics, the fraction of sites within 120 m of the water source is given by 1 - S(120), or

$$1 - \exp[-120/\exp\{4.967 - 0.535(1) - 1.379(0) + 0.285(1)\}],$$

which reduces to 66 percent.

For the purposes of comparison, various solutions to the first question are given in Table 2 using model (1) and in Table 3 using model

Table 2.

Solution for y, the distance to the nearest water source in metres, at various levels of archaeological potential for statistical model (1) based on the data collected

$$S(y) = \exp[-y/\exp\{5.36 + 0.751 - 1.388J - 0.592K - 1.315JK)]$$

Model Values			Level of potential in percent		
1	J	K	70	80	90
0	0	0	**256**	342	**490**
0	0	1	142	189	271
0	1	0	64	85	122
0	1	1	10	13	18
1	0	0	121	162	231
1	0	1	67	90	128
1	1	0	30	40	58
1	1	1	5	6	9

1= 1 if the distance measurement was obtained in the field, 0 otherwise
J = 1 if the water source is a spring, 0 otherwise
K = 1 if the study area is Area 3, 0 otherwise

Table 3.

Solution for y, the distance to the nearest water source in metres, at various levels of archaeological potential for statistical model (2) based on the data collected

$$S(y) = \exp[-y/\exp\{4.967 + 0.535I - 1.37J - 0.285L)]$$

Model Values			Level of potential in percent		
1	J	L	70	80	90
0	0	0	173	231	331
0	0	1	230	307	440
0	1	0	**44**	58	84
0	1	1	58	78	112
1	0	0	101	135	194
1	0	1	135	180	258
1	1	0	26	34	**49**
1	1	1	**34**	**46**	**65**

1= 1 if the distance measurement was obtained in the field, 0 otherwise
J = 1 if the water source is a spring, 0 otherwise
L = 1 if the soil drainage is describe as well/very good, 0 otherwise

(2). Both tables show, under various scenarios (or values of I, J, K, and L), the distances away from water that are necessary to capture 70, 80, and 90 percent of the Iroquoian archaeological sites. Calculations such as those appearing in Tables 2 and 3 can be used to draw likelihood bands of archaeological potential around sources of water.

A reasonable statistical model for the size of a site in hectares is the Weibull distribution. The data support a model for the mean size depending on the cultural variables, "cultural affiliation" and "site type." Within the site type there are differences in the mean size for each

of cabins, camps, hamlets and villages. The sizes for site types follow a logical progression from cabin (0.28 ha) to camp (0.51 ha) to hamlet (0.57) to village (1.68). The mean sizes for various cultural periods also follow a logical progression from Early (0.58 ha) to Middle (1.28 ha) to Late (1.01 ha) to Historic (1.38), a pattern that has been previously identified for Ontario Iroquoians (Dodd 1984:280). With the current data, however, there was sufficient variability in site size within each of the cultural periods, so that it could only be concluded that the Early cultural period showed a significantly different site size from the rest.

RECOMMENDATIONS FOR FURTHER RESEARCH

With the appropriate collection of data, further work could be done in statistical model building. For example, using the models for the distance to the nearest water source, bands could be drawn around water sources in which a desired percentage of the sites are expected to be found. These bands would show only the expected percentage, rather than the spatial distribution of sites within the bands. A further step would be to formulate and test models providing insight into factors which affect the spatial distribution of sites within areas of high archaeological potential. One might proceed by overlaying a grid system between the bands drawn around water sources. By using logistic or log-linear models (Fienberg 1977:84 - 86), the effect of various factors on the probability of finding a site within a grid-square can be examined. These models could be used to predict smaller areas of high archaeological potential near water or to characterize the terrain near water in which sites are most likely to be found.

Acknowledgments. Financial support for this work was made possible by a contract from Ontario Hydro. The authors would like to thank the reviewers and editor for many helpful comments which have improved the paper.

REFERENCES CITED

Burgar, R. W. C.
1990 An Archaeological *Master Plan for* the *Metropolitan Toronto and* Region *Conservation* Authority. The Metropolitan Toronto and Region Conservation Authority, Toronto.

Campbell, C., and I. D. Campbell
1992 Pre-Contact Settlement Pattern in Southern Ontario: Simulation Model for Maize-Based Village Horticulture. *Ontario* Archaeology 53: 3 - 24.

Dodd, C.
1984 Ontario Iroquois Tradition Longhouses. Archaeological Survey of Canada Mercury Series Paper 124, National Museum of Man, National Museums of Canada, Ottawa.

Dunn, O. J., and V. A. Clark
1974 Applied Statistics: *Analysis of Variance* and Regression. Wiley, New York.

Elandt-Johnson, R. C., and N. L. Johnson
1980 *Survival Models and Data Analysis.* Wiley, New York.

Fienberg, S. E.
1977 The Analysis of Cross-Classified Categorical *Data.* MIT Press, Cambridge.

Hasenstab, R.
1991 Wetlands as a Critical Variable in Predictive Modeling of Prehistoric Site Locations: A Case Study From the Passaic River Basin. *Man in the* Northeast 42: 39 - 61.

Heidenreich, C. E.
1971 Huronia. McClelland and Stewart, Toronto.

Hicks, C. R.
1993 *Fundamental* Concepts in the Design *of Experiments.* Saunder College Publishing, New York.

Konrad, V. A.
1973 The Archeological Resources of the Metropolitan Planning Area: Inventory and Prospect. Department of Geography, York University, North York, Ontario.

Lawless, J. F.
1982 *Statistical Models* and Methods *of* Lifetime *Data.* Wiley, New York.

Mayer, R., R. Pihl, and D. Poulton
1985 *A Review of Archaeological* Re-sources for *Ontario Hydro's* South-*west* Study: *Supply* to London and *Brantford Area.* Ms. on file, Landuse and Environmental Planning Department, Ontario Hydro, Toronto.

Montgomery, D. C.
1991 *Design and* Analysis *of Experiments.* 3rd ed. Wiley, New York.

Nelson, W.
1982 Applied Life *Data Analysis.* Wiley, New York.

Neter, J., W. Wasserman, and M. H. Kutner
1990 *Applied* Linear Statistical *Models.* 3rd ed. Irwin, Homewood.

Neumann, T. W.
1992 The Physiographic Variables Associated with Prehistoric Site Location in the Upper Potomac River Basin, West Virginia. Archaeology *of Eastern North America* 20: 81 - 124.

Pearce, R. J., and Pihl, R
1983 A Review of Archaeological Re-sources *for Ontario Hydro's* South-*west* Study. Submitted to Ontario Hydro. Ms. on file, Landuse and Environmental Planning Department, Ontario Hydro, Toronto.

Pearce, R. J.
1989 A Review of Ontario *Hydro's Approach* to Prehistoric Archaeological Resource Assessment in Southern *Ontario.* Ms. on file, Landuse and Environmental Planning Department, Ontario Hydro, Toronto.

Pearce, R. J., D. R. Bellhouse, and L. W. Stitt
1989 An Assessment *and* Refinement of Prehistoric Archaeological Site Potential *Models* for Southwestern *Ontario.* Ms. on file, Landuse and Environmental Planning Department, Ontario Hydro, Toronto.

Peters, J.
1986 Transmission Line Planning and Archaeological Resources: A Model of Archaeological Potential for Southwestern Ontario. In *Archaeological Consulting in Ontario:* Papers of the London *Conference* 1985, edited by W. A. Fox, pp. 19 - 27. Occasional Publication of the London Chapter, Ontario Archaeological Society 2. London, Ontario.

Robertson, B. P., and L. B. Robertson
1978 The Generation of Locational Models in an Inductive Framework. In Conservation *Archaeology in* the *North*east: *Toward a Research* Orientation, edited by A. E. Spiess, pp. 27 - 36. Peabody Museum of Archaeology and Ethnology Bulletin 3. Harvard University, Cambridge, Massachusetts.

SAS Institute
1985 SAS *User's* Guide: Statistics, Version 5 *Edition.* SAS Institute, Cary, North Carolina.

Seber, G. A. F.
1977 *Linear* Regression *Analysis.* Wiley, New York.

Weston, B. R.
1981 Upper Illinois River Unit. In Predictive *Models* in Illinois Archaeology: Report Summaries, edited by M. K. Brown, pp. 21 - 32. Illinois Department of Conservation, Division of Historic Sites, State of Illinois, Chicago.

D.R. Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada N6A 5B7
R.J. Pearce, London Museum of Archaeology, Lawson-Jury Building, 1600 Attawandaron Road, London, Ontario, Canada N6G 3M6
J.H. Peters, Land Use and Environmental Planning Department, Ontario Hydro, 700 University Avenue, Toronto, Ontario, Canada M5G 1X6
L.W. Stitt, Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada N6A 5C 1