

Re-thinking Accuracy and Precision in Predictive Modeling

by

Thomas G. Whitley
Brockington and Associates, Inc
email: tomwhitley@brockington.org

A Paper Prepared For:
Computer Applications in Archaeology Conference
Prato, Italy, April 13-18, 2004

Abstract

Testing archaeological predictive models has almost always relied upon evaluating the percentage of sites “captured” versus the percentage of area defined as “high” potential. This is known as the “gain” statistic. Fundamentally inherent in correlative models and the gain statistic, though, is the assumption that measuring the deviation from randomness is the best method to evaluate the accuracy and precision of a model. This paper will show, in contrast, that the locations of archaeological sites are always dependent upon the location of the previous instance of settlement and therefore can act only like time-series dependent phenomena, never like random points. This calls for a fundamentally different means of testing models which can account for spatial autocorrelation.

Introduction

In this presentation I am not going to belabor the argument that correlative (what might also be called inductive or regression-based) predictive models are inadequate on both theoretical and practical grounds. I have done so before (e.g. Whitley 2001; 2003a), and not wanting to beat a not-so-dead-yet horse, I will pass on the opportunity to do so again.

Instead, I would like to focus on an issue which has a direct bearing on the way in which predictive models are employed, but does not directly relate to the theoretical nature of their construction. There is an indirect link, as I shall shortly demonstrate, but the focus of my discussion will be on how we test predictive models, not build them. This relates to two terms first discussed by Kvamme (1988), which have distinctive uses and well-defined meanings in predictive modeling; namely *accuracy* and *precision*.

Accuracy, as defined in what we might loosely refer to as “the field of predictive modeling,” means how well does the model (regardless of the nature of its

development) capture the sites used to test it? Put more simply, do most of the sites (either known prior to the development and implimentation of the model, or located with a testing strategy afterward) fall within areas of modeled high probability? Accurate predictive models should theoretically capture a high proportion of archaeological sites.

Precision, also known as specificity (cf. van Leusen et al. 2002), does not refer to the capture of sites, but the reduction of space into useful categories of probability. In other words, highly accurate models are of no use if they refer to all areas as high potential. Precise models limit high potential areas in some way to make them useful for focusing survey strategies or research designs. Remember that the purpose of a predictive model is to find areas likely to produce archaeological sites for whatever reason (e.g. land management, or explaining settlement location patterning).

In essence, accuracy is intended to refer to the confidence one can have in the predictive ability of the formula, while precision is supposed to be a measure of its utility. Because a successful model should have both high

accuracy and good precision, Kvamme (1988:329) developed a means to combine these two attributes and measure them. This is known as the gain statistic. The formula for evaluating model gain is expressed as:

$$1 - (\% \text{ High Potential Area} / \% \text{ Known Sites Captured})$$

It is my argument, however, that the gain statistic, although currently uniformly accepted and employed almost ubiquitously as a measure of predictive model success, is founded upon a faulty premise. It should also not be taken at face value, nor assumed to be objective. Instead, we should examine the way we verify predictive models and base our confidence in them on their explanatory power. This entails thinking of accuracy and precision in a different light.

The Chimera of Randomness

When we initially produce a probability surface (through whatever technique) our first instinct is to want to objectively evaluate how well the model works. As trained scientists and statisticians, we immediately verbalize this as “do the patterns appear to be significant?” In other words, do our observations significantly differ from our expectations. But where do the expectations come from? The typical answer is from an assumption that if all things were equal, archaeological sites would be distributed randomly across the study area.

Using that assumption then, we can extrapolate for the study area a random distribution of points and measure how different the observed sites are from that random distribution. Depending on how the model is constructed, there are numerous ways of attaching a statistical significance value to that difference. This, in fact, is at the root of how correlative models are created. In the realm of model testing, though, the main point is to find a significantly high proportion of sites with high probability values.

That seems to make sense, but does it really? Can it be said that (all things being equal) archaeological sites would “behave” like random throws of a pair of dice? Would sites under those conditions have an equal potential to fall anywhere in the environment for each occurrence? Or, are sites inherently autocorrelated with each other?

And, if they are, what does that mean for how we should test probability models? First, let’s look at what archaeological sites actually represent.

First-Order Autocorrelation

As we all know and understand quite well, archaeological sites represent the remnants of human behavior. This means that the primary mechanism for the distribution of archaeological sites (regardless of the factors which promote or inhibit use of some area) is the movement of people across the landscape. So, let’s assume a spatial manifold (or study area) of 40 to 50 kilometers in size. Furthermore, let’s assume that this manifold is absolutely frictionless; meaning there are no impediments to movement in any direction.

If we begin with someone, or a group of people, placed in the exact middle of this manifold, there are no physical nor cultural factors which will push or pull them in any specific direction. The only limitation on the placement of their next archaeological site, is how far they can travel before they deposit additional archaeological materials. Putting aside for the moment the nature of what their activities may be (which greatly determine what materials they are leaving behind, and how frequently), let’s assume that artifacts are deposited at regular time intervals (we can call them standard time units).

If those people travel at a speed of 1 km per standard time unit, then the limits on the placement of the next archaeological occurrence must be within a zone of between 0 to 1km from the point of origin. The next archaeological occurrence must, as well, fall within 0 to 1km of the previous location (not the first). All of the subsequent site locations must fall within that same buffer zone of each previous occurrence. Figure 1 extrapolates this process for 500 such simulated archaeological occurrences.

The resulting distribution of “sites” does not resemble randomness. The reason why is that each one is not an independent occurrence. They are, in essence, time-series autocorrelated. Each one is dependent not upon the extent of the manifold, but the location of the previous site and the maximum distance which can be spanned between occurrences. The distribution illustrated here is in fact, a random walk. Direction is chosen randomly but merely a

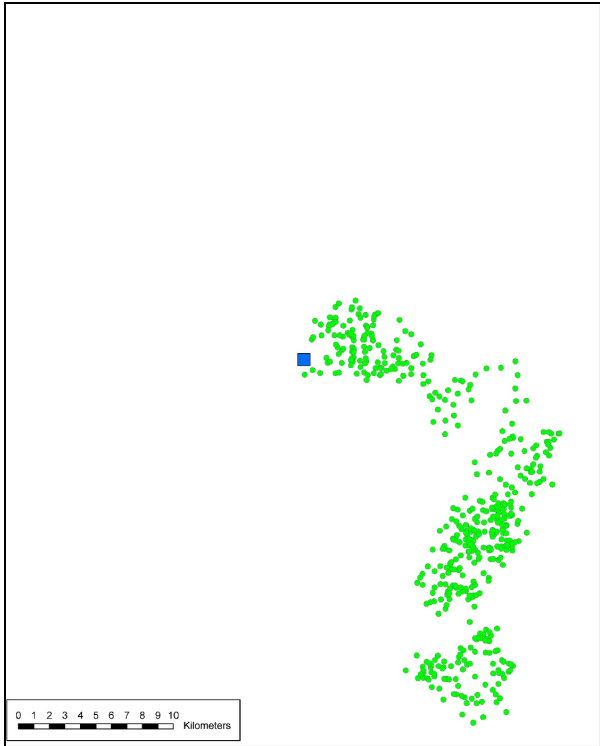


Figure 1. Random walk (1km buffer).

limitation on possible distance has eliminated the potential for an appearance of randomness.

What happens though when you choose a larger distance buffer? Figure 2 illustrates the same sequence of random direction selection with a buffer of 2km for 500 archaeological occurrences. As you can see by the result, the spread of points increases, but a large number of the points falls outside the manifold entirely (outside the boundary of the figure). Increasing the buffer to 5km (Figure 3) adds additional spread within the 40 to 50km manifold (and reduces much more the number of sites which fall within it). Even further increasing the buffer to 10km (Figure 4) adds dramatically to the spread, but still does not simulate the appearance of randomness in the few sites which can be found within the manifold.

But how do we know that sites are limited in this way? Is it not likely that many activities are tied only to a few specific sites, which themselves may fall much further apart; such as a base camp-activity area type settlement strategy? Does that sort of process result in greater randomness? Figure 5 illustrates just such an example. Each of the base camps (shown as purple squares) is linked

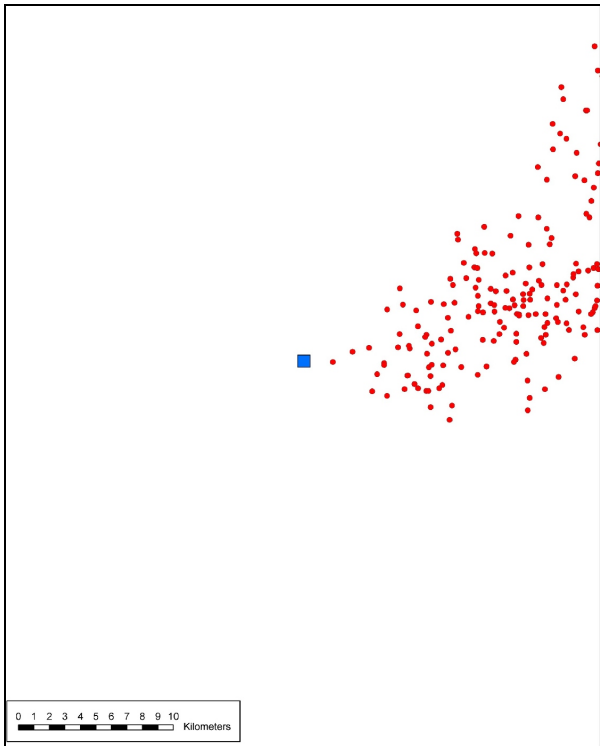


Figure 2. Random walk (2km buffer).

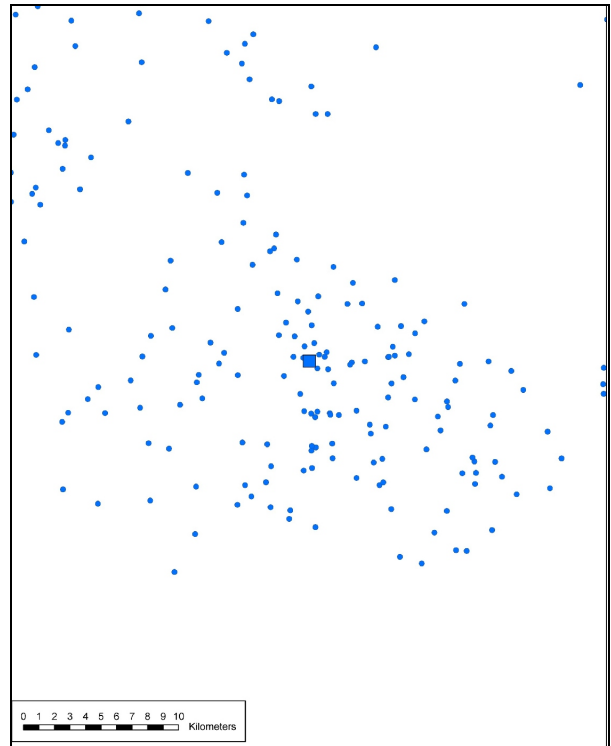


Figure 3. Random walk (5km buffer).

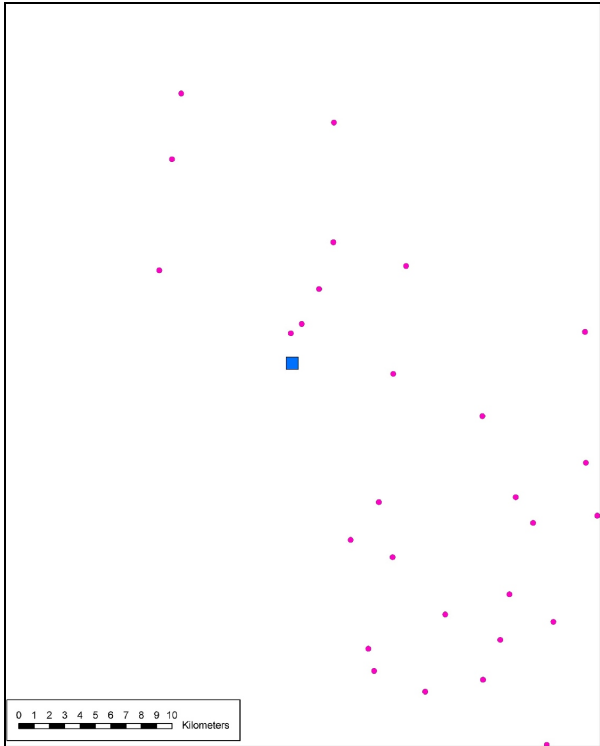


Figure 4. Random walk (10km buffer).

to a series of 10 activity locales (black triangles) by a buffer limit of 5km. Each base camp is also linked to the next by a buffer limit of 25km. This results in the appearance of clusters of associated sites (much like you would find in an archaeological setting), but once again, they do not suggest randomness.

Now, of course it is not likely that archaeological sites can be considered to be deposited in regular intervals, however large those intervals may be. But that doesn't change the nature of their creation as probabilistically dependent upon being some distance and direction from the previous occurrence, or a base camp. That probability can be considered inversely proportional to distance. In other words, given the location of a single site, the probability is very high that the next site will be quite close, and diminishes as you get further and further away; this is the very definition of spatial autocorrelation.

Even with all of the examples combined, the manifold is not a random distribution (Figure 6). And if we expand the manifold to be 400 to 500km across, we can see the distribution of the results of all random walks in

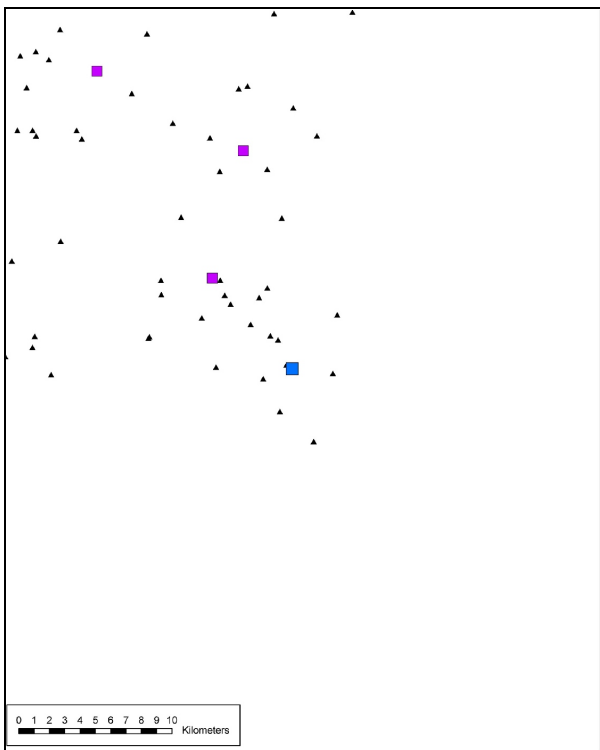


Figure 5. Random walk (5km and 10km buffers - base camp routine).

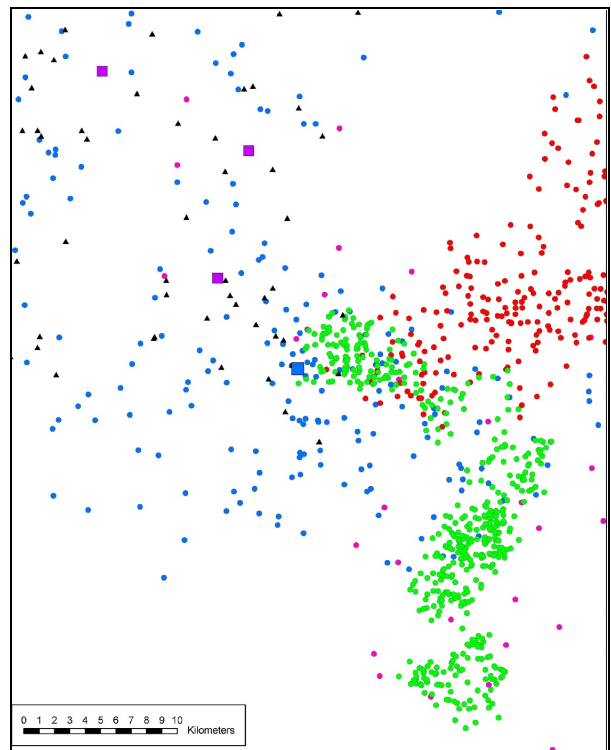


Figure 6. All random walks (1, 2, 5, 10, and 5+10km buffers).

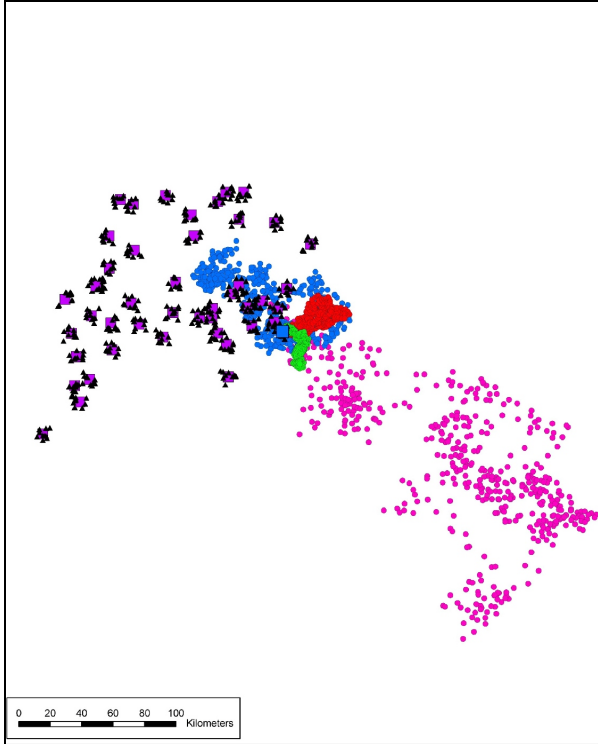


Figure 7. All random walks (as in Figure 6, but with a 400 x 500km manifold).

their entirety (Figure 7). This totals more than 2500 possible archaeological sites and it illustrates an important notion. The appearance of randomness depends on the scale at which the data is being viewed; even in cases where there is assumed to be no restrictions on where people are likely to have traveled.

To even approximate the appearance of randomness, you have to assume a travel buffer much larger than the size of the manifold. A palimpsest of thousands of years of settlement in a frictionless surface in which the travel buffer is much larger than the manifold of observation could eventually resemble a random distribution. But in that situation, the appearance of randomness does not equate with a random distribution. The distribution of sites in such a situation actually represents many different applications of site selection and cannot be considered part of a *single* predictive model. Each archaeological occurrence is still linked to the previous, or to a base camp, in such a way that I believe any test against randomness for archaeological sites is inherently untenable.

Second-Order Autocorrelation

Now, the autocorrelation between sites themselves is not the only autocorrelation inherent in models of site location. There is the additional caveat that we do not live on a frictionless surface, and neither did people in the past. Mitigators of movement come in several forms; those which are specifically chosen as variables upon which spatial decisions are made, and those which are unconsciously employed at all times and can be considered as second-order autocorrelates with site location.

For example, imagine that we have the same 40 to 50km manifold as in the previous examples. Only this time, we have a topographically variable surface. We can still employ a standard time unit buffer between sites, or different site types. But rather than assuming the cost of travel is equal in every direction, we can employ a cost-surface to create the buffers. Figure 8 illustrates such a cost surface based strictly on terrain slope.

The probability of site placement still diminishes with distance, but that distance reflects the actual cost of

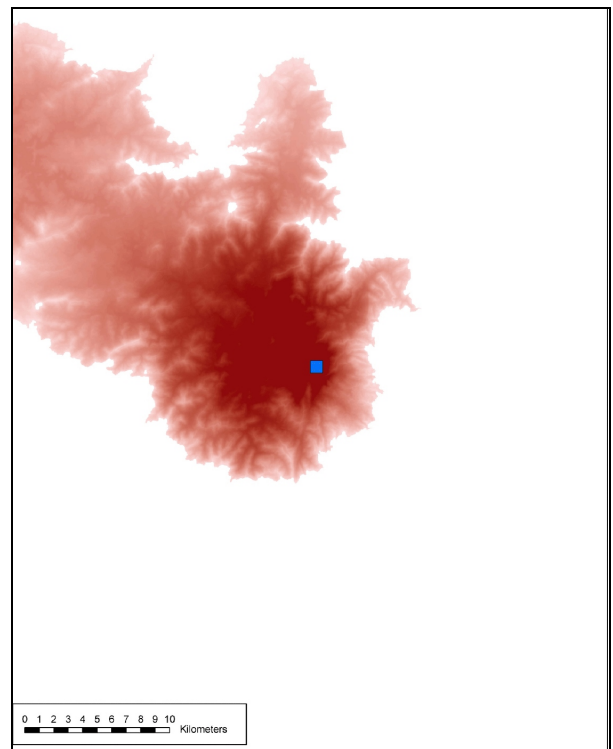


Figure 8. Cost distance based buffer (1000 cost units over variable terrain).

travel across the landscape. Importantly, such a surface is not strictly a measure of the negative costs which increase friction, but may include benefits which reduce it as well. These can be conceptualized as attractors (spatial variables which reduce friction) and repulsors (ones which increase it).

In this way there are some variables which elevate the cost of travel to such a degree that the probability of any site occurring in high cost areas is very minimal, *even when there has been no explicit strategy to employ them*. This is best exemplified by two very important variables in predictive modeling; terrain slope and distance to water. Whenever these two variables are employed in a predictive model they typically work quite well as predictors of generic site location, but *only* in areas of high relief or at least moderate aridity.

Can we argue that extracting such variables as these in a predictive model is realistically a measure of our understanding of settlement strategies? I believe that *de facto* variables which were not explicitly cognized as attractors or repulsors to settlement (or other activity) should not be considered as a means to create confidence in a predictive model.

In other words, I am not particularly impressed when the accuracy of some model is based primarily on the use of slope and distance to water. It should be a given that people were averse to living on steep hillsides and far from a water source. There is no evidence to show that they cognized such variables and used them as the primary motivators for settlement.

But what does it matter, if the precision of the model is good, and therefore it has high utility? I would argue that in this instance high accuracy and good precision is likely to instill a false sense of security that significant archaeological resources will always occur in high probability areas. And secondly, that low probability areas are of the least concern for archaeological purposes.

In fact, if a resource were to occur in a low potential area when it has not been predicted, that should clearly raise a red flag that perhaps such a site is unusual and consequently more significant than if it had been found in a high potential zone. When evaluated in a test against randomness, however, such a site would be automatically

thrown out as an outlier. The second-order autocorrelative nature of some *de facto* variables has over-ridden the interpretative potential of the sites themselves.

The Effect on Accuracy and Precision

What does this mean for significance testing of predictive accuracy and precision? Putting aside for a moment the problems with building models on false assumptions of randomness, we need to rethink how we derive our expected site distributions. Should we use random walk models (such as those illustrated in the foregoing figures) to test our predictive surfaces? I do not believe so. The nature of random walks is such that you would need to be able to extract specific sequences of archaeological occurrences in order to evaluate a model's accuracy. We know that it is extremely difficult, if not impossible, to do so. Instead we need to recognize that we already have a viable dataset for testing predictive models that can be classified by behavioral characteristics rather than specific temporal sequences.

To illustrate this concept, let's go back to the frictionless manifold for a moment. The main problem with comparing known sites against random points (or randomly expected numbers of sites by probability area, even if actual points are never used) is that the built-in autocorrelations cannot be accounted for. The reason is that a random distribution has no autocorrelations, yet as I have shown here, all archaeological sites inherently have them on two different levels. Thus a comparison between a pattern of sites (or a probability surface) and an actual random distribution is like comparing apples and oranges.

Instead we need to consider comparing a modeled probability surface against the known dataset *not* an expected distribution of sites. This means that we first need to develop an explanatory model; one which addresses the causality of site locations (cf. Whitley 2003b). Such a model may or may not be based on a previously known set of archaeological occurrences, but it should reflect distinctive behavioral characteristics and not lump sites together merely because they are from the same time period or have similar artifact contents.

Causal models build hypotheses about why certain areas were explicitly and cognitively chosen for a particular

behavior. The spatial components are then broken down into manageable quantitative surfaces and combined to develop a probability surface which reflects those hypotheses. When compared with the archaeological occurrences which we believe represent those behaviors then we have a much more robust testing strategy, because theoretically every one of those sites should fall within land units which have a high probability value for that behavior. If even a single one falls in a low potential zone, it suggests that there is something wrong with the model (or conversely that there is something wrong with the designation of that site as representing the targeted behavior).

Ultimately, our confidence in the predictive model (i.e. its accuracy) would be based on the statistical relevance of the known archaeological sites (cf. Salmon 1971; 1998) not their statistical propensity. This means we have the ability to build models which address behaviors that are infrequent, and the size of the dataset need not be large to do so.

Instead, the model could be verified with just a few archaeological occurrences of the modeled behavior, or even the absence of such sites in low potential zones. Naturally, an increased abundance of sites can generate higher confidence, but with a test against randomness that abundance is required from the beginning and the rejection of a faulty hypothesis is not as robust.

The model's precision is much more dependent upon the scale at which it is employed. Just like the traditional notion of precision, if the study area is too small, precision is likely to be greatly reduced. More importantly however, we need to regard predictive precision as a concept intimately tied to causality as well. An explanatory model may still be of great utility even if it is quite imprecise (as long as it is also very accurate). This is not the case for a model which may be accurate but has no explanatory power (such as a correlative one).

Conclusions

My intent with this discussion has been primarily to introduce the idea that we need to reconsider our notions of what makes an archaeological predictive model successful. Our failure to thoroughly integrate the very

theoretical ideas so often debated in archaeology with the methods of predictive modeling have allowed us to latch onto faulty assumptions and rely on an inadequate and non-robust testing strategy. By rethinking how we envision accuracy and precision in a predictive setting, we begin to recognize the drawbacks and inadequacies of how we create the models in the first place.

References Cited

- Kvamme, Ken L.
1988 Development and Testing of Quantitative Models. In *Quantifying the Present and Predicting the Past: Theory, Method, and Application of Archaeological Predictive Modeling*. W.J. Judge and L. Sebastian (eds), pp. 325-428. U.S. Government Printing Office, Washington, DC.
- Salmon, Wesley C.
1971 *Statistical Explanation and Statistical Relevance*. University of Pittsburgh Press, Pittsburgh, PA.
1998 *Causality and Explanation*. Oxford University Press, Oxford.
- van Leusen, Martijn, Jos Deeben, Daan Hallewas, Paul Zoetbrood, Hans Kamermans, and Philip Verhagen
2003 *Predictive Modelling for Archaeological Heritage Management in the Netherlands*. Baseline Report for the NWO (Humanities Section) of the BBO.
- Whitley, Thomas G
2001 *Using GIS to Model Potential Site Areas at the Charleston Naval Weapons Station, South Carolina: An Alternate Approach to Inferential Predictive Modeling*. Paper presented at the Conference: "GIS and Archaeological Predictive Modeling: Large-Scale Approaches to Establish a Baseline for Site Location Models", Argonne National Laboratory, Argonne, Illinois, March 21-24, 2001.
2003a *Causality and Cross-Purposes in Archaeological Predictive Modeling*. Paper Prepared for the Computer Applications in Archaeology 2003 Conference, Vienna, Austria, April 8-12, 2003.

2003b *A Brief Outline of Causality-Based Cognitive Archaeological Probabilistic Modeling*. Position Paper prepared for the Symposium on Predictive Modeling and Archaeological Heritage Management, Amersfoort, The Netherlands, May 22-23, 2003