

# **An Exploratory Data Analysis Approach to Artifact Density Correlation**

**Brian D. Jones, UMASS Amherst**

Abstract: Understanding the spatial distribution of artifacts across a site underlies the interpretation of past human activity. Another aspect of spatial analysis is to assess the relationship between different types or classes of artifacts or ecofacts across a site. Such assessments may, for example, be aimed at correlating the density distribution of different lithic raw materials to provide evidence for the contemporaneity of knapping activity, or at better understanding relationships between the disposal patterns of ceramics and food refuse. While complex methods are available for the calculation of three-dimensional spatial correlation, most require specialized statistical or GIS software. This paper proposes a relatively simple exploratory data analysis approach to establishing a measure of three-dimensional spatial correlation between classes of artifacts that can be calculated with any spreadsheet program.

## **Introduction**

Attempting to establish a degree of correlation between data in three dimensions (e.g. x, y, z) can be a daunting task (Kintigh 1990). Recently, GIS software is able to make such analyses more practical for archaeologists using a grid-based approach. Here, however, I would like to introduce a relatively simple measure of three-dimensional correlation that can be calculated based on field data quickly with any spreadsheet program.

The initial assumption of the model is that data exists in three columns expressing a x, y, and z values, or an easting, northing and count per unit area. Typically such data is generated from field notes or a query of one's inventoried data. In most cases, this information is likely to express raw artifact counts of a particular class of finds from excavated units of like size, such as the number of chert flakes or creamware sherds in every square meter of excavation. Such data may be quickly plotted in programs such as Surfer to provide a visual impression of the distribution of artifact density across the site. Such visual comparisons between similar artifact

types, such as creamware and pearlware sherds can be used to develop inferences about chronologically relevant issues of site formation processes, for example, whether the artifact classes in question were disposed of contemporaneously or not.

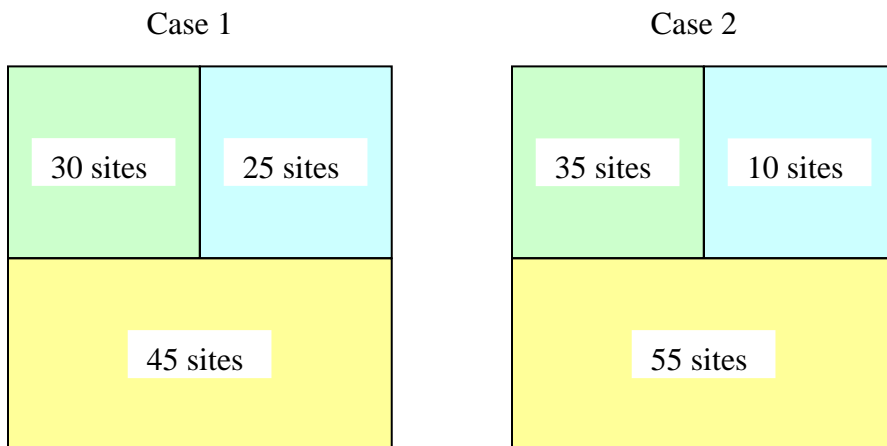
While the visual expression of such three-dimensional data is often compelling enough to develop inferences on which to base a narrative of site formation processes, it is desirable to be able to support such visual representations and narratives with statements of statistical significance. The method described here provides one means of doing this in a relatively straightforward manner based on the use of the t-statistic with proportional comparisons between populations. I first developed this method to assess the distribution of sites against environmental variables measured across a large study region in Connecticut (Forrest et al. 2006). The test was used originally used to determine if the number of sites within an environmentally defined region reflected a random or nonrandom distribution. In the latter case, higher or lower than expected site density could be assumed to reflect some human response to the environmental variable in question. It recently occurred to me that a similar approach could be used to assess the distribution of artifact density at the site level. I will first describe the initial statistical method and then move to its use for assessing artifact density correlation at the intra-site level.

### **The Student t-Statistic and Proportional Data**

Descriptive statistics are commonly used to discuss patterns in site location criteria. For example, it might be said that 60% of sites are located on well-drained soils. This may be an important observation, but the statement cannot address whether this observation is significant.

For example, if 60% of sites were found on well-drained soils, but these soils made up 65% of the landscape area, can it be said that the nature of well-drained soils played a significant role in site selection criteria? In this case it is likely that the relationship is not significant at all, and is simply due to chance.

The method of analysis described here works with such a test. The question is whether the number of sites found in association with a given environmental variable, such as soil drainage class, can be explained as a result of chance association, or whether the number is so large or so small that it is statistically unlikely that the association is a result of chance. The test can be explained easily with a simple diagram. Below, two study areas are represented by boxes of equal size. For simplicity, we will assume each square represents a 10x10 km area, or 100 square km. The upper halves of each box have been divided into two equal areas of 25 square km each, while the lower half represents an area of 50 square km. A number of sites have been noted for each subsection of the study area.



Given a random distribution of sites, one would anticipate about 50% of the sites to occur in the large (yellow) zone of each box, and 25% in each of the other smaller blue and green zones. As the number of sites in each zone varies from the amount estimated by a random distribution, it becomes increasingly unlikely that chance alone played a role in the observed distribution. We can determine the probability that an observed number of sites occurs in a certain zone of known size given the proportion of sites in that zone, the proportionate area of the zone, and the sample size.

In the Case 1, 30% and 25% of the sites occur in the smaller zones, and 45% of the sites occur in the larger zone. The table below explains the calculation to determine the significance of the observed pattern based on the calculation of a p-value (the probability that chance alone accounts for the observed distribution of sites). The number of sites in each case will be assumed to be 100.

Table 1: Calculation of Spatial Correlation, Case 1

Proportion of sites	Proportion of area	<i>s</i>	<i>SE</i>	<i>t</i>	<i>p</i>
0.45	0.50	0.4975	0.0497	-1.0050	0.3173
0.25	0.25	0.4330	0.0433	0.0000	1.0000
0.30	0.25	0.4583	0.0458	1.0911	0.2779

An approximation of the standard deviation (*s*) of the distribution is calculated as follows:

$$s = \sqrt{p(1 - p)}$$

where *p* is the proportion of sites in a zone (Drennan 1996: 140). The standard error of the mean (*SE*) is calculated by the function:

$$SE = \frac{s}{\sqrt{n}} \text{ or simply } SE = \sqrt{\frac{p(1-p)}{n}}$$

where  $s$  is the standard deviation of the sample, and  $n$  is the complete sample size (e.g. 100 sites). SE is a measure of the expected range of variation of the mean itself, rather than of the population (Drennan 1996: 108). The t-score is a measure of the number of standard errors an observed value falls above or below an expected value. It is valuable because it allows one to determine whether significance is positive or negative, and is a gauge of the degree of deviation from the expected value. In this case, our expected value is the proportion of the area itself. The value of  $t$  is calculated as follows:

$$t = \frac{(o - e)}{SE}$$

where  $o$  is the observed proportion of sites and  $e$  is the expected proportion based on area (e.g. Drennan 1996: 160). Finally, a p-value can be calculated based on  $t$  and the sample size, using the Student t-distribution statistical table. In this case, Microsoft Excel was used to calculate p using the TDIST() function. This function works as follows: TDIST( $SE$ ,  $df$ , 2), where  $df$  is the degrees of freedom ( $n-1$ ) and the final value 2 is used for a two-tailed Student t-distribution test of significance. (It should be noted that in cases where artifact proportions, and therefore SE are equal to zero, an average t can be calculated based on the average value of SE calculated from both the observed and expected proportions. In most cases this average does not vary greatly from that based on either individual proportion.)

In Case 1 above, p-values from all three zones fall well above 0.05, indicating that the observed distribution is readily explained by chance alone. In Case 2, however, some of the p-values are lower:

Table 2: Calculation of Spatial Correlation, Case 2

percent sites	percent area	<i>s</i>	<i>SE</i>	<i>t</i>	<i>p</i>
0.55	0.50	0.4975	0.0497	1.0050	0.3173
0.35	0.25	0.4770	0.0477	2.0966	0.0386
0.10	0.25	0.3000	0.0300	-5.0000	0.0000

The p-values indicate that sites are about as common as expected due to chance in the large yellow zone ( $t= 1.00$ ,  $p=0.3173$ ). However, sites are significantly positively correlated with the smaller green zone ( $t= 2.1$ ,  $p=0.0386$ ), and are strongly negatively correlated with the small blue zone ( $t=-5.0$ ,  $p<0.0001$ ). This relatively simple method of calculating significance between spatial variables forms the backbone of the following analyses.

### **A Measure of Three-Dimensional Intra-Site Artifact Correlation**

A statistical assessment of the strength of correlation between two distributions of artifacts expressed in three-dimensions is possible based on the above method. The main difference in the test of correlation is the starting assumption, or null hypothesis. In the case above, the null hypothesis was that site location reflected a random distribution. When p-values fell below about 5% this hypothesis was rejected, and the assumption was made that other, probably cultural factors relating to site location choices were at work: that is, the observed distribution of sites could not be explained well by chance alone. In the assessment of the distribution of artifacts across a site the null hypothesis is the opposite. The assumption we would like to test is that two artifact classes share a distribution because similar forces resulted in their deposition. This might be the case if a single knapper worked two types of raw material at close intervals in time, or if certain types of waste were thrown out the same house door. In this

case, low p-values indicate a weak correlation, and high p-values reflect the strength of correlation. As in the above examples, correlation may be negative or positive as expressed by the t-score. The strengths of correlation of each unit of excavation are calculated first, but the overall assessment of spatial correlation across the site is determined by the average strength of these individual values. This has the added advantage of permitting a calculation of the standard deviation of the range of individual correlations to provide a further measure of the strength of the observed relationship.

### Test Case Studies of the Measure of Three-Dimensional Correlation

Before moving to case studies based on archaeological field data I will provide some small-scale test cases to better explain the method. These will be simple artifact distributions in four by four arbitrary grids. The first example will provide a simple case where the first, or “observed”, distribution represents values equal to one half of the second, or “expected” distribution. Which distribution will act as the observed and which the expected is arbitrary.

Case 3: Distribution 1 equals one-half Distribution 2 (rounded down), average correlation = 0.978 +/- 0.067

**Distribution 1**

N4	5	40	35	10
N3	10	45	45	15
N2	10	40	35	10
N1	5	15	10	2
	E1	E2	E3	E4

**Distribution 2**

N4	10	80	70	20
N3	20	90	90	30
N2	20	80	70	20
N1	10	30	20	5
	E1	E2	E3	E4

The following table summarizes the calculations made to determine the average correlation. Note that The Standard Error values and associated t-scores are very low because

the proportions are nearly identical in each case. The deviation from a perfect correlation is caused by the number of artifacts in unit N1E4 which has been rounded down to the integer value 2. This slight degree of variation results in subtle differences between all proportions of Distribution 1 compared to Distribution 2. Generally, p-values are quite high and indicate a very strong correlation (the probability that the forces producing the two distributions were the same), as expected. Only Unit N1E4 varies from this. It's t-score of -0.352 results in a p-value of 72.5% because the 2 artifacts recovered were less than the expected 2.5. The average correlation of the two distributions is 97.8 +/- 6.7%, indicating a very strong spatial relationship with little variance. This can be interpreted as a roughly 98% probability that the two distributions resulted from the same past event. Were we permitted to have 2.5 artifacts in unit N1E4, the average correlation would be 100% +/- 0.0.

Table 3: Calculation of the Average Spatial Correlation Value of Case 3

Unit	Dist. 1	Dist 2	Prop. 1	Prop. 2	s	SE	t	P
N4E1	5	10	1.51%	1.50%	0.122	0.007	0.003	0.997
N3E1	10	20	3.01%	3.01%	0.171	0.009	0.005	0.996
N2E1	10	20	3.01%	3.01%	0.171	0.009	0.005	0.996
N1E1	5	10	1.51%	1.50%	0.122	0.007	0.003	0.997
N4E2	40	80	12.05%	12.03%	0.326	0.018	0.010	0.992
N3E2	45	90	13.55%	13.53%	0.342	0.019	0.011	0.991
N2E2	40	80	12.05%	12.03%	0.326	0.018	0.010	0.992
N1E2	15	30	4.52%	4.51%	0.208	0.011	0.006	0.995
N4E3	35	70	10.54%	10.53%	0.307	0.017	0.009	0.993
N3E3	45	90	13.55%	13.53%	0.342	0.019	0.011	0.991
N2E3	35	70	10.54%	10.53%	0.307	0.017	0.009	0.993
N1E3	10	20	3.01%	3.01%	0.171	0.009	0.005	0.996
N4E4	10	20	3.01%	3.01%	0.171	0.009	0.005	0.996
N3E4	15	30	4.52%	4.51%	0.208	0.011	0.006	0.995
N2E4	10	20	3.01%	3.01%	0.171	0.009	0.005	0.996
N1E4	2	5	0.60%	0.75%	0.077	0.004	-0.352	0.725
<i>Sum</i>	332	665	1.00	1.00				
							<b>average</b>	<b>0.9777</b>
							<b>std dev</b>	<b>0.0674</b>

Case 4: Distributions 1 and 2 are random, average correlation = 0.1218 +/- 0.2406

**Distribution 1**

N4	36	49	44	3
N3	92	76	40	30
N2	71	76	62	9
N1	25	32	59	43
	E1	E2	E3	E4

**Distribution 2**

N4	58	59	83	74
N3	3	56	53	50
N2	75	84	48	96
N1	95	18	13	88
	E1	E2	E3	E4

The results of the calculation of average correlation between the random artifact distributions in Case 4 is  $p = 12.18\% \pm 24.06$ . In other words there is only about a 12% chance that the distributions reflect the same deposition process, and the high standard deviation indicates that we should have very little confidence in even that low estimate. In fact, ten of the sixteen individual p-values fall below the 5% margin, strongly suggesting that chance alone accounts for the observed distribution.

Case 5: Mirrored Distributions, average correlation = 0.0135 +/- 0.023

**Distribution 1**

N4	20	10	5	2
N3	40	30	20	10
N2	60	80	40	20
N1	40	20	10	5
	E1	E2	E3	E4

**Distribution 2**

N4	2	5	10	20
N3	10	20	30	40
N2	20	40	80	60
N1	5	10	20	40
	E1	E2	E3	E4

In the final sample case, the two distributions of artifacts are mirrored across the site area to reflect an example of truly divergent artifact distributions. In this case, the calculation of average correlation results in a value of  $p = 0.0135 \pm 0.02$ . This value falls well below the traditional 5% threshold for significance, which in this case indicates a clear rejection of the null hypothesis that the two distributions reflect the same event.

## **Archaeological Case Studies**

Archaeological artifact distributions are more complex than the small-scale cases presented above. Artifacts seldom express unimodal decay from a single central concentration but instead are clustered unevenly across a site area (Kintigh and Ammerman 1982; Kintigh 1990). Such irregular clustering is an expression of the episodic artifact discard that typifies the palimpsest nature of most archaeological sites. This suggests that strong correlations, as explored in case 3 above represent the exception rather than the rule. In fact, my search for an example of strong artifact correlation, whether horizontally or vertically expressed proved fruitless. Instead, I will present just two case studies that express the complexity of actual archaeological scenarios and will discuss some caveats and limitations of the method in the conclusions.

### **Historic ceramic distributions at Site 72-66**

Site 72-66 represents a small Native American house structure occupied on the Mashantucket Pequot Reservation between about 1780 and 1800. The site assemblage is dominated by coarse red earthenware and creamware sherds (n=4013 and 4395 respectively). The artifacts were recovered from 353 square meters of excavation of the house and surrounding area in 1996. The ceramic artifact distribution appears to primarily reflect discard out a door located in the southeast wall of the house structure (Figure 1). Creamware sherds (green) and earthenware (red) express rather distinct distributions. Figure 1 has been simplified for clarity to express only artifact densities above 20 sherds per square meter. The measure of correlation expressed by this distribution is 14.3% +/- 25.9%, suggesting the distributions are not likely

related to the same episodes of artifact discard at the site. The results compare favorably with the random distribution pattern in case 4. While there is little doubt that the artifacts were discarded by occupants of the house, the poor correlation indicates that the discard of these two ceramic types rarely occurred at the same time. In fact, the house was likely occupied for twenty years, and the observed pattern probably reflects scores of episodes of dumping. The nature of these discard events appears to have varied greatly. In one case, over 400 sherds of earthenware, probably representing the crushed remnants of a single vessel, were discarded in the southwestern portion of the site. This episode alone accounts for over 10% of the earthenware sherds recovered and has resulted in a highly skewed distribution pattern. Interestingly, removing this anomalous case increases the correlation by only 1%.

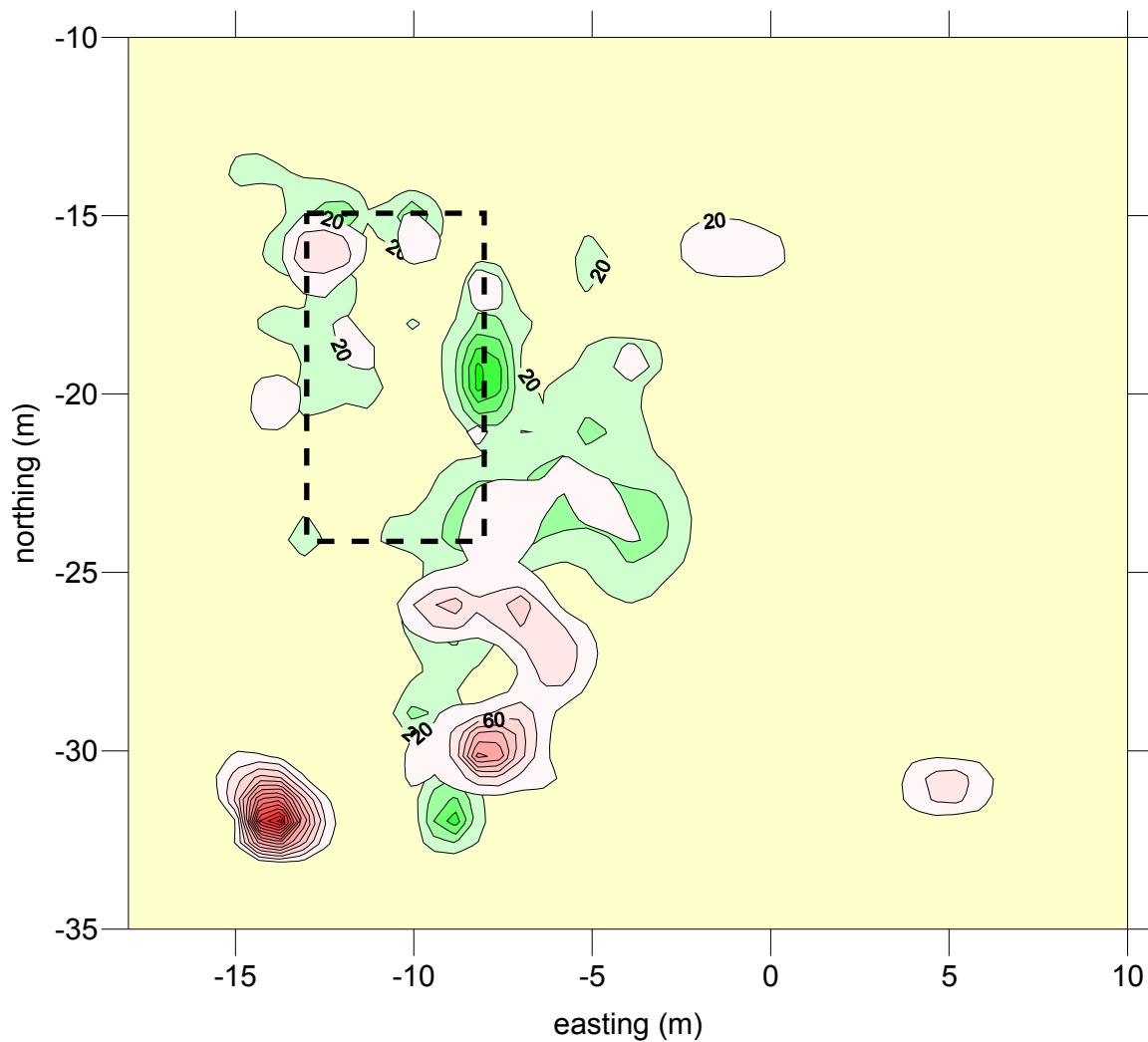


Figure 1: Distributions of coarse red earthenware (red, n=4013) and creamware (green, n=4395) at site 72-66, Mashantucket Pequot Reservation. Average correlation = 0.143 +/- 0.259, suggesting the distributions are not likely related to the same episodes of artifact discard at the site. The dashed line represents the approximate house location.

A second test case was examined in the hopes of exploring a stronger correlation between two variables. In this case the distributions of creamware recovered from the upper (0-10cm below ground surface) and lower (10-20cm) topsoil zones were assessed. The assumption was that the creamware sherds were deposited on the surface over a relatively short period of time during site use. The distributions recorded approximately 200 years later during excavation were expected to reflect relatively homogenous post-depositional forces that resulted in gradual artifact burial across the site. Instead, the measured average correlation between this class of artifacts from the upper and lower topsoil horizons was low:  $p = 19.86\% \pm 29.21\%$  (Figure 2). This value compares favorably to simulated random distributions suggesting little direct association between the distributions in the upper and lower topsoil horizons.

There are two primary explanations for this observation. The first is that over the twenty year period of occupation, artifacts were unevenly incorporated into the topsoil as they were discarded. More likely, however, post-depositional processes occurring across the site over the last 200 years were quite heterogeneous, resulting in uneven rates of artifact movement through the soil over time and across space. Burrowing animals, tree-throws and land-clearing are the most likely forces behind the remnant vertical artifact distributions observed.

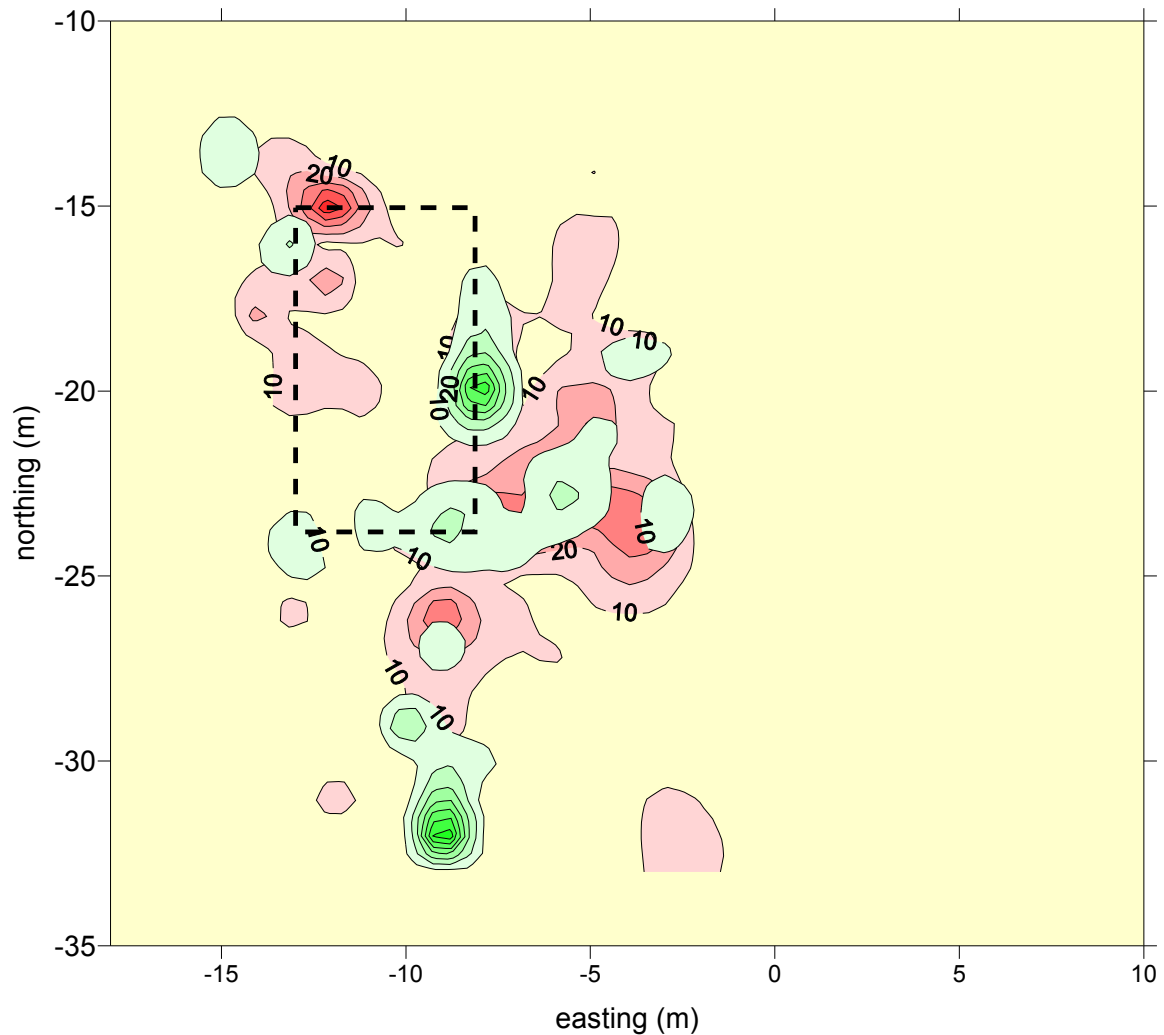


Figure 2: Density of creamware sherds ( $>10/m^2$ ) at site 72-66 within the topsoil horizon: 0-10cm below surface (green) and 10-20cm below surface (red). Although a stronger correlation was expected, the average p-value of the correlation between these two contexts is only 19.86%  $\pm$  29.21%. This is comparable to values produced by random distributions of similar dimensions. The density plot supports the statistical conclusion that the two distributions are poorly correlated. The dashed line represents the approximate house location.

## Conclusions and Caveats

The simple artificial test cases above (cases 3 – 5) indicate that the method can provide an excellent indicator of correlation in three dimensions between two artifact distributions. Real world application suggests that archaeological data are seldom as neatly structured, and that post-depositional processes can have a significant impact on data correlation that might have once existed. The method is very sensitive to variation from expected distributions, and the average p-value will plummet when a portion of the site expresses essentially random patterning. Units of scale are therefore important to bear in mind. An examination of p-values from individual units of analysis is possible, and may provide insight regarding specific locations of high correlation. However, spurious correlations will occur between random data sets, so caution is warranted in this approach.

With practice, the method is relatively simple to apply to the types of data archaeologists are used to looking at: sums of counts of particular classes of artifacts from units of excavation. It works well as an exploratory data analysis method, that is one intended to quickly assess the presence of correlation or lack thereof. In both cases it should instigate hypothesis development with the intent of exploring alternative explanations of the observed patterning, or lack thereof. In the archaeological test cases above it certainly inspired a new look at some old data and raised additional questions about formation processes that occurred both during and after site use.

## References Cited

Drennan, Robert D.

1996 *Statistics for Archaeologists: A Common Sense Approach*. Plenum Press: New York.

Forrest, Daniel T., Michael S. Raber, Brian D. Jones and Robert M. Thorson

2006 *Archaeological and Historical Resource Study, Adriaen's Landing Project, Hartford, Connecticut*. Prepared for the Connecticut Office of Policy and Management. Archaeological and Historical Services, Inc.: Storrs, CT

Kintigh, Keith W.

1990 Intrasite Spatial Analysis: A Commentary on Major Methods. In *Mathematics and Information Science in Archaeology: A Flexible Framework*, edited by Albertus Voorrips. *Studies in Modern Archaeology* 3: 165-200. Holo, Bonn.

Kintigh, Keith W., and Albert J. Ammerman

1982 Heuristic approaches to spatial analysis in archaeology. *American Antiquity* 47(1): 31-63.