PRACTICAL QUANTITATIVE LITHIC USE-WEAR ANALYSIS USING MULTIPLE CLASSIFIERS

Nathan E. Stevens[a], Douglas R. Harro[b], and Alan Hicklin[c]

[published version: Journal of Archaeological Science 37 (2010) 2671-2678]

[available online at: http://www.elsevier.com/locate/jas]

[a][corresponding author] Department of Anthropology, University of California, Davis, One Shields Avenue, Davis, CA 95616 USA (nestevens@ucdavis.edu) Telephone: (805) 710-3486, Fax (530) 752-8885

[b]Applied Earthworks, Inc. 515 E. Ocean Ave, Suite G, Lompoc, CA 93436 USA

[c]Spectral Imaging Facility, NEAT ORU, Chemistry Rm 11, University of California, Davis, One Shields Avenue, Davis, CA 95616 USA

Although use-wear analysis of prehistoric stone tools using conventional microscopy has proven useful to archaeologists interested in tool function, critics have questioned the reliability and repeatability of the method. The research presented here shows it is possible to quantitatively discriminate between various contact materials (e.g., wood, antler) using laser scanning confocal microscopy in conjunction with conventional edge damage data. Experiments with replica and prehistoric tools suggest the quantitative method presented here provides valid functional inferences and is flexible enough to accommodate other relevant sources of data on tool function.

Keywords: lithics, use-wear, microwear, laser scanning confocal microscopy, classification

1. Introduction

For several decades, archaeologists have had an uneasy relationship with lithic use-wear analysis. While the data it provides are useful, the subjectivity of the method, which often relies more on experience and expertise than explicit criteria, has left many archaeologists wary. This problem has been addressed in two ways: 1) blind tests to explore accuracy and interobserver error (Bamforth, 1988; Newcomer, et al., 1986; Odell and Odell-Vereecken, 1980), and 2) attempts at quantification (Evans and Donahue, 2008; Gonzalez-Urquijo and Ibanez-Estevez, 2003; reviewed in Grace, 1996; Keeley, 1980; Kimball, et al., 1995; Stemp and Stemp, 2001; Stemp and Stemp, 2003; Van den Dries, 1998). Similar issues and solutions have developed in parallel among researchers studying dental microwear in primates (Scott, et al., 2006).

Laser scanning confocal microscopy (LSCM) is a promising method for use-wear analysis because it produces 3-dimensional point data that can be presented either as a high-resolution image (Figure 1), or as quantitative data (Evans and Donahue, 2008). Making use-wear analysis quantitative rather than qualitative will not only make the process of producing functional inferences more explicit, it will also provide a basis for further methodological improvements. Tests using both human analysts and quantitative classification (i.e., "machine learning" in the language of artificial intelligence, see Alpaydin, 2004) can be used to explore the relevant variables affecting the accuracy of the method including experience (Bamforth, et al., 1990), materials (Bradley and Clayton, 1987; Lerner, et al., 2007), use duration (Bamforth, 1988; Goodale, et al., 2010), and post-depositional processes (Burroni, et al., 2002; Evans and Donahue, 2005; Levi Sala, 1986).

If different contact materials have distinctive quantitative signatures, then it should be possible to represent use-wear data in the form of probability statements that report not only how a tool was used, but also provide information about the certainty of the attribution, something that has not been possible using traditional methods. Additionally, if the results of human and machine use-wear experts converge for both replica and archaeological datasets, then many previous use-wear studies using conventional methods can be validated.

2. Background

Blind tests typically involve analysts trading replica tools to test their ability to identify: 1) which tools or tool edges were used, 2) the tool motions employed, and 3) the contact materials on which tools were used. While such tests have generally produced satisfactory results, some high-profile published tests with poor results (e.g., Newcomer, et al., 1986) led many archaeologists to doubt the efficacy of use-wear analysis.

A common misunderstanding of use-wear studies is that there is one "correct" answer and that anything less is not useful. In fact, the specificity of functional data needed depends on the archaeological question. In some cases, it may be enough to know simply which items are tools and which are waste flakes. In other cases, knowing whether tools were used on hard or soft substances may suffice. In still other cases, it may be desirable to know specifically the numbers of tools used to cut meat, or scrape wood.

Independent of the desired specificity of functional data are a number of factors largely beyond the control of the analyst that affect the quality of the data including raw material type, post-depositional alteration, available equipment, and others. The problem is that it is difficult to convey information about how certain an analyst is in a particular identification. It is also unclear when this lack of certainty should lead to less specific attributions (e.g., "indeterminate hard material") or an admission by the analyst that the contact material is unknown. This problem is analogous to that of faunal analysts deciding when to identify material to the species, genus, or higher taxonomic levels owing to fragmentation, expertise, and other factors (e.g., Gobalet, 2001). These issues can create real differences in how use-wear data are produced by different analysts and interpreted by the archaeological community.

Blind tests of both the "high-power" and "low-power" use-wear approaches (see Odell and Odell-Vereecken, 1980) have shown that accuracy varies predictably with level of

specificity (Bamforth, 1988). Accuracy is quite high (ca. 70-90%) if only the presence or absence of use wear is examined. Accuracy is lower (ca. 60-75%) when discriminating between general contact material classes (e.g., hard vs. soft), and lower still (ca. 20-70%) for specific contact materials (e.g., antler, meat). Low power methods, in particular, have reduced success at discriminating specific contact materials but do well at discriminating general material classes. (Odell and Odell-Vereecken, 1980).

Conventional use-wear studies have shown that identification of specific contact materials can be improved by combining more than one type of data such as polish appearance at high magnification in addition to edge damage at low magnification (Bamforth, 1988; Keeley, 1980). This approach was explicitly implemented by Grace (1989) and Van Den Dries (1998), who used a combination of polish and edge damage variables obtained by conventional microscopy in the construction of expert systems designed to identify tool functions. More recently, studies using sophisticated instrumentation including atomic force microscopy (Kimball, et al., 1995), laser profilometry (Stemp and Stemp, 2001), and LSCM (Evans and Donahue, 2008) have demonstrated quantitative differences between different polish classes, but have used only single descriptors (e.g., Rq, or root mean square roughness) and have not incorporated other sources of data on tool function such as edge damage. Our aim was to combine quantitative data on use-wear polishes (acquired at magnification equivalent to 1000X, using the LSCM) with qualitative data on edge damage (observed at 20X-100X, using a stereomicroscope) to arrive at a multivariate approach to classification incorporating the best attributes of each.

This was accomplished by following the lead of researchers working in the fields of artificial intelligence, pattern recognition, and machine learning, who have advocated an approach to classification where the results of multiple classifiers are combined (e.g., discriminant analysis or decision trees), each with particular strengths, weaknesses, and input data, yielding a more reliable classification (Alexandre, 2001; Ho, et al., 1994; Lam, 2000; Rahman and Fairhurst, 2003).

3. Methods

3.1 Analysis Procedure

The identification of unknown tool polishes was treated as a classification problem employing a *training set* of known classes, and a *test set* of unknown classes. In the field of machine learning, this is known as *supervised learning* (Clarke, et al., 2009; Liu, 2007). The training set "trains" a classifier (e.g., a discriminant function) using correctly classified objects. Subsequently, the test set is introduced to the classifier which assigns class membership based on the parameters of the training set. In this case, each data point is a tool or tool edge and the classes are contact materials (e.g., wood, antler).

The general analysis procedure consisted of collecting two types of data for each tool or tool edge: (1) quantitative polish data acquired using the LSCM, and (2) qualitative edge damage data acquired using a conventional stereomicroscope. Training set data from each was then input into a relevant classifier (i.e., discriminant analysis or decision tree).

Subsequently, test set data of each type was input into the classifiers and probabilities of class membership were calculated. Then, probabilities from each classification method were combined and the class with the highest resulting probability was assigned. A graphical representation of the data acquisition and classification process is presented in Figure 2.

3.2 Training and Test Data Sets

The polish data training set comprised 36 replica tools consisting of unmodified flakes used to work five different materials: antler, wood, soft plants, dry hide, and meat. Six tools were used to work each contact material for 30 minutes and an additional six tools were left unmodified. All replica tools for the polish data training set were of Monterey chert, a high-quality toolstone common along the California coast. Contact materials were chosen to encompass the range of variability present in materials worked by prehistoric peoples in the region. Antler and bone are generally agreed to be indistinguishable (Vaughan, 1985), so polish from scraping mule deer (*Odocoileus hemionus*) antler was used as a proxy for both. Wood polish was produced by scraping seasoned manzanita (*Arctostaphylos* sp.), a common California hardwood used ethnographically for digging sticks, projectile tips, and other implements. Soft plant polish was produced by slicing tule (*Scirpus californicus*) stalks. Dry hide polish was produced by scraping the membrane side of naturally tanned undyed cowhide. Meat polish was

produced by cutting raw pork shoulder while avoiding contact with bone or the cutting board. Because edge damage data were easier and less expensive to acquire, they were also more plentiful. The edge damage data training set included all of the replica tools from the polish data training set and an additional 12 replica tools, for a total of 8 tools per contact material.

All replica tools were cleaned with a soft brush in soap and water then soaked in 5% HCL solution for 1 hour. Archaeological tools were cleaned with soap and water, but no acid bath was used to preserve tool surfaces for future residue studies. While it is possible this difference in cleaning techniques affected the outcome of the experiment, the effects of many factors affecting prehistoric tools, such as post-depositional processes, are also largely unknown. The effects of all such factors become part of the error in classification, something which could be theoretically studied empirically and controlled, but which is beyond the scope of this study. All replica and archaeological tools were also cleaned of finger oils and other extraneous substances by swabbing tool edges with alcohol prior to imaging.

Two separate test sets were employed. First, a group of replica tools (the "replica test set") including a subset of the training sample was examined using both conventional and quantitative methodologies. The aim of this test was to evaluate the performance of the quantitative classification procedure. A total of 20 replica tools was examined by two analysts using a stereomicroscope with fiber optic illuminator under magnifications ranging from 20X to 140X. Analyst "A" was the primary author (N.S.) and analyst "B" was Nicholas Hanten, both of UC Davis. The test was a blind test in that each analyst made 10 tools which were traded with the other analyst who was unaware of their uses. Because most of the blind test replica tools were also used for training classifiers, these samples were iteratively held out from the training set when calculating probabilities and assigning replica class membership so as not to bias the classification procedure.

The second test set (the "archaeological test set") comprised 16 prehistoric Monterey chert tools from CA-SBA-246, an Early Holocene (ca. 9200-8800 calBP) archaeological site on the central California coast (Lebow, et al., 2001). One edge of each of these tools was imaged except for tool 1616-1, of which two edges were imaged. Additionally, the unused interior portions of three tools were imaged, for a total of 20 test cases. A previous "low-power" use-wear analysis had been performed by the second author, permitting a comparison of the analyst calls to those produced by the quantitative analysis.

3.3 Polish Data

Quantitative polish data were generated by imaging tool edges with the LSCM and then processing the images to extract three-dimensional surface characterizations. Imaging of samples was performed at the NEAT ORU Spectral Imaging Facility, University of California, Davis, using an Olympus FluoView FV1000 laser scanning confocal microscope with 405 nm laser diode and 40X (NA 0.6) objective. Each image had a field of view of 158 μm^2 and was acquired at a digital resolution of 0.155 μm/pixel with the confocal aperture set at 100 μm and the z-step interval at 1 μm. While the FV1000 is generally used for fluorescence microscopy, for this application, it was maximized to acquire reflected light in the region of the laser wavelength (400-410 nm).

All image processing and surface characterization was performed using ImageJ software (Abramoff, et al., 2004) and relevant plugin modules. Each tool image stack was first transformed into a single grayscale topographic image with pixel intensity corresponding to height using the TopoJ plugin (Hovis, 2009). Fifteen 10 um^2 areas within each image were sampled and descriptive statistics characterizing tool surfaces were calculated using the plugins SurfChar (Chinga, et al., 2007) and FracLac (Karperien, et al., 2008).

After confocal image data were converted to quantitative data, further statistical analyses were performed in JMP (SAS Institute, 2007). Principal components analysis, bivariate comparisons, and stepwise discriminant analysis were used to identify correlated variables that would be redundant for classification. SurfChar and FracLac both calculate a variety of statistics, many of which are useful (e.g., Evans and Donahue, 2008 successfully use Rq). Ultimately, the choice of specific variables used for classification is somewhat arbitrary as many combinations of three or more produce satisfactory results. Adding too many variables, however, can result in "overfitting" the data, resulting in classifications that are not widely applicable outside the training sample.

The three variables chosen were a compromise between avoiding multiple correlated variables, avoiding overfitting, and ability to discriminate material classes. Of 11 variables initially considered for classification using discriminant analysis, three were found to be particularly useful for discriminating different contact materials: *Mean resultant vector* (MRV), *surface area* (SA), and *fractal dimension* (Df). A canonical biplot with all 11 variables (Figure 3)

shows that many variables are highly correlated and that those with the most influence on the discriminant analysis are MRV, SA, and Df. Rq is also useful for discrimination but is highly correlated with SA (R^2 = .96), so only the later was used.

Mean resultant vector is a measure of the central tendency of facet orientation angles (Chinga, et al., 2007; Curray, 1956; compare to anistropy in Scott, et al., 2005). The measure is scaled to a range of 0 to 1 so that a score of 0 would signify a random distribution of orientations (e.g., a natural, unused chert surface) while a score of 1 would signify all facets oriented in the same direction (e.g., due to abrasive wear). Surface area is a measure of surface complexity in that a more convoluted surface will have a larger surface area while a flat or polished surface will have a smaller surface area. Fractal dimension is a measure of how patterned details change with scale (see also Stemp and Stemp, 2001; Stemp and Stemp, 2003; Ungar, et al., 2003). Its value for classification likely relates to the observation that certain aspects of use-wear are apparent at smaller scales while others require a large-scale overview (Vaughan, 1985).

Polish data were classified by employing a discriminant analysis of the above-specified variables (MRV, SA, and Df) of the training set. An examination of the canonical plot (Figure 4) shows that polish from different contact materials plots in different regions, a necessary condition for proper discrimination. However, it is also apparent that some polish types overlap considerably; in particular, antler/plants and wood/hide. This suggests that if classification relied only on polish variables, tools used on antler would often be mistakenly classified as used on plants (and likewise with hide and wood). In both of these ambiguous cases, one material is hard (antler and wood), and one is soft (plants and hide), meaning if some other classifier could differentiate between hard and soft material classes, then classification accuracy could be improved. This was the role of edge damage data.

3.4 Edge Damage Data

Lithic use-wear analysts have proposed a variety of methods for describing edge damage ranging from simple presence/absence of attributes (e.g., flake scars) to detailed quantification of flake removal types and locations (Akoshima, 1987; Bird, et al., 2007; Grace, 1989; Tringham, et al., 1974). Types and locations of edge damage have previously been shown to provide information about the types of materials worked as well as the tool motions employed although analysts concentrating on polish characteristics have tended to use edge damage as

supplementary data for assigning tool function (Odell and Odell-Vereecken, 1980). Given that the aim of this study was to concentrate on variables that aid in assigning contact materials and not tool motions, a limited number of attributes that have been previously shown to help in discriminating contact material types (see Grace, 1989; Keeley, 1980; Odell and Odell-Vereecken, 1980; Vaughan, 1985) were recorded using a stereomicroscope at magnifications from 25X to 100X. These included the presence or absence of snap fractures, microflakes, step fractures, and the degree of edge rounding.

This list of edge damage variables was reduced to two simple presence/absence values for use in decision tree analysis: step fractures and edge rounding. Use-wear analysts have repeatedly shown that step fractures are indicative of working hard materials (Akoshima, 1987; Grace, 1989) while edge rounding is an attribute of hide working (Vaughan, 1985). The decision tree (Figure 5) confirms these findings, suggesting the presence of step fractures easily distinguishes hard (antler, wood) from soft contact materials (meat, hide, and plants) and that the presence of edge rounding on tools without step fractures is a good indicator of hide working.

3.5 Classifier Combination

As detailed above (sections 3.1-3.4), two types of data (polish data and edge damage data) and two classifiers (discriminant analysis and decision tree) were used in the analysis. When new data of unknown class assignment (i.e., test set data) are input into either classifier, probabilities of assignment to each contact material class are output. This is represented by:

$$p_{ij} = p_j(c_i|x)$$

or, p_{ij} is the probability that an object with measurable characteristics x belongs to class i, using classifier j.

Classifier combination proceeded according to the method outlined by Alexandre (2001) where the probabilities of class assignment obtained by each classifier (i.e., discriminant analysis and decision tree) are averaged for each class, or,

$$\frac{1}{N}\sum_{j=1}^{N} p_{ij} = p_i$$

where p_i is the average over classifiers 1, 2, … N of the probabilities for class i. In this case, there are two classifiers and six possible classes. The decision rule used to assign class membership is simply: assign the observation x to the class i^* with the highest average probability p_i.

4. Results

4.1 Replica Test Set Results

The conventional analysts correctly identified specific contact materials in 65% of the cases (Table 1), better than average for similar "low-power" blind tests (e.g., Odell and Odell-Vereecken, 1980). This is probably due to the limited number of possible contact materials and restricted tool motions employed. By comparison, quantitative classification performed nearly as well (Tables 1 and 2), correctly assigning tools to specific contact materials in 60% of the cases. The probability of obtaining this result by chance is less than .01 (cumulative binomial probability n=20, p=.17; $b(x \geq 12)$=0.000015 where n=number of trials, p=probability of success in a single trial, b=binomial probability, and x= total number of successes).

As expected, the combination of two classifiers (i.e., discriminant analysis and decision tree) performed better (60%) than either classifier in isolation (each at 40% correct). It is interesting to note that correct assignments were twice as common when associated probabilities were above 0.5 (n=8 correct, n=4 incorrect), whereas incorrect answers were equally common above and below the 0.5 level. Of the incorrect answers with probabilities below 0.5, however, all but one were correct at the level of general material class (i.e., "hard," "soft," or "unused"). Of the misclassified artifacts, in all but one case, the second highest probability is the correct answer (see Table 2).

4.2 Archaeological Test Set Results

In consideration of the results of the replica test set, for prehistoric tools, probabilities of less than 0.5 were reported only at the level of general material class (Table 2). Specific contact materials associated with each probability are also provided in the table for comparison.

Given that the conventional use-wear analysis of the CA-SBA-246 collection was a "low-power" analysis conducted with a stereomicroscope at magnifications between 20X and 90X,

specific contact materials were often collapsed into more general material classes (Lebow, et al., 2001). This fact is likely to have contributed to the good agreement (60%) between the conventional analyst and the quantitative classification. Nevertheless, the correspondence is still better than would be expected by chance and is surprisingly reliable for an initial test (Krippendorff's Alpha = 0.51 (Krippendorff, 2004)). Overall, maximum probabilities were lower (mean = 0.47) when compared to the blind test using replica tools (mean = 0.54).

5. Discussion and Conclusions

Overall, the results of this study suggest quantitative use-wear analysis of stone tools is a promising analytical technique. The method presented here performed nearly as well as conventional analysts and much better than would be expected by chance at the most difficult task of use-wear analysts, that of identifying specific contact materials. Perhaps more important is the fact that use-wear assignments are given in the form of explicit probability statements, making it easier to evaluate their accuracy.

The replica test set results show that quantitative classification of use-wear using multiple classifiers is nearly as effective as conventional analysts at identifying specific contact materials given the materials and protocols of this initial comparison. Furthermore, the replica results confirm that new methods of surface characterization such as LSCM can be augmented by the addition of conventional edge damage data. The fact that correct classifications are more likely to have higher associated probabilities suggests the probability statements produced by the classification procedure provide useful information that can be used to fine-tune the analysis. It is also interesting that the conventional analysts and the quantitative classification did not necessarily make the same mistakes, suggesting it may be possible to isolate which types of use-wear are best suited to human analysts and which can be better identified by quantitative methods.

The archaeological test set results are more difficult to evaluate because the true uses of the tools are unknown. The fact that the correspondence between the classifications of the conventional analyst and the quantitative method is greater than would be expected by chance, however, suggests human and machine analysts may “see” similar patterns in the data despite different methods of acquisition and processing. The fact that maximum probabilities were lower when compared to the blind test using replica tools, suggests additional sources of variability,

such as postdepositional alteration, use of tools on multiple substances, or a wider variety of tool use intensity, duration, or contact materials, are present among the prehistoric tools that were not captured by the training set.

The use of LSCM to produce 3-dimensional surface data is certainly an advance in imaging and identification of use-wear polishes, but it should not be forgotten that additional sources of data obtained through more conventional means are still a valuable component for understanding tool function. In this case, if LSCM data alone were used to classify artifacts, many misidentifications would have resulted due to overlapping characteristics of certain polishes such as antler and plants (see Figure 4 and Table 1). The addition of edge damage data improved classification performance from 40% correct to 60% correct (see Table 1).

As instruments like LSCM become more available to the archaeological community, the temptation will be to replace existing traditional (i.e., user-generated) analytical methods with automated ones. However, a good case can be made for incorporating many types of information in making use-wear identifications, rather than relying on the appearance of polish alone. This was a lesson learned previously by use-wear analysts using conventional microscopy (Bamforth, 1988). The use of multiple classifiers allows for the incorporation of a variety of quantitative data obtained through a diversity of methods and, potentially, can maximize the contribution of each method. Given certain conditions, two (or more) classifiers should work better than one because each classifier uses different input data and each operates in different regions of the feature space. In other words, each classifier will tend to be wrong in different ways, so that when combined, the best attributes of each method can be emphasized (Cunningham, et al., 2008).

Although it is possible to imagine using a similar methodology to produce a completely machine-driven classification procedure, that was not the goal of this study. Instead, we see a real opportunity to integrate data generated by new imaging technologies such as LSCM as well as data generated by conventional microscopic examination and an understanding of fracture mechanics; in other words, human expertise. While many additional variables can easily be recorded and more elaborate classification procedures implemented, the plausibility of any answer must still be evaluated by an experienced lithic analyst who can incorporate information such as site context that is difficult to quantify.

Quantitative classification could also be incorporated as part of the larger analysis procedure according to the availability and expertise of human analysts. Certain features that require little training to identify (e.g., presence/absence of step fractures) could be cataloged by less experienced analysts while machine classification could be used to perform the more mundane identifications while, at the same time, identifying those artifacts that are best examined and interpreted by a human expert.

For this initial test, we limited the scope of potential use-wear identifications to a specific set of contact materials and also attempted to include the smallest number of relevant variables and use simple classification procedures. Future applications of quantitative methods to use-wear analysis could incorporate more variables and likely result in more accurate classifications and the ability to discern how a tool was used (e.g., tool motions) in addition to reporting possible contact materials. Truly machine-driven automatic classification of lithic tools is also a future possibility, but for the time being, human experts are an integral part of the process.

Acknowledgements:

Jelmer Eerkens helped with design and implementation of the project. Nicholas Hanten's hard work was crucial in the production of replicas and the implementation of the blind test. Special thanks to Denise Jurich for sharing replica specimens and expertise. Statistical guidance was graciously provided by Mark Grote and Ian Robertson. Brian Codding, Jelmer Eerkens, and Mark Grote read a previous draft and improved it greatly. Funding for this work came from a UC Davis Institute of Governmental Affairs Dissertation Research Award, a UC Davis and Humanities Graduate Research Award, and funds from the UC Davis Department of Anthropology.

References Cited

Abramoff, M.D., Magelhaes, P.J., Ram, S.J., 2004. Image Processing with ImageJ, Biophotonics International 11, 36-42.

Akoshima, K., 1987. Microflaking Quantification, in: Sieveking, G., Newcomer, M. (Eds.), The Human Uses of Flint and Chert, Cambridge University Press, Cambridge, pp. 71-79.

Alexandre, L.A., Campilho, A.C., Kamel, M.,, 2001. On Combining Classifiers using Sum and Product Rules, Pattern Recognition Letters 22, 1283-1289.

Alpaydin, E., 2004. Introduction to Machine Learning, MIT Press, Cambridge, MA.

Bamforth, D.B., 1988. Investigating microwear polishes with blind tests: The institute results in context, Journal of Archaeological Science 15, 11-23.

Bamforth, D.B., Burns, G.R., Woodman, C., 1990. Ambiguous use traces and blind test results: New data, Journal of Archaeological Science 17, 413-430.

Bird, C., Minichillo, T., Marean, C.W., 2007. Edge damage distribution at the assemblage level on Middle Stone Age lithics: an image-based GIS approach, Journal of Archaeological Science 34, 771-780.

Bradley, R., Clayton, C., 1987. The influence of flint microstructure on the formation of microwear polishes, in: Sieveking, G., Newcomer, M. (Eds.), The Human Uses of Flint and Chert, Cambridge University Press, Cambridge, pp. 81-89.

Burroni, D., Donahue, R.E., Pollard, A.M., Mussi, M., 2002. The Surface Alteration Features of Flint Artefacts as a Record of Environmental Processes, Journal of Archaeological Science 29, 1277-1287.

Chinga, G., Johnsen, P.O., Dougherty, R., Berli, E.L., Walter, J., 2007. Quantification of the 3D microstructure of SC surfaces, Journal of Microscopy 227, 254-265.

Clarke, B., Fokoué, E., Zhang, H.H., 2009. Supervised Learning: Partition Methods, Principles and Theory for Data Mining and Machine Learning, Springer, New York, pp. 231-306.

Cunningham, P., Cord, M., Delany, S.J., 2008. Supervised Learning, Machine Learning Techniques for Multimedia, Springer, Berlin, pp. 21-49.

Curray, J.R., 1956. The Analysis of Two-Dimensional Orientation Data, The Journal of Geology 64, 117-131.

Evans, A.A., Donahue, R.E., 2005. The Elemental Chemistry of Lithic Microwear: An Experiment, Journal of Archaeological Science 32, 1733-1740.

Evans, A.A., Donahue, R.E., 2008. Laser Scanning Confocal Microscopy: A Potential Technique for the Study of Lithic Microwear, Journal of Archaeological Science 35, 2223-2230.

Gobalet, K.W., 2001. A Critique of Faunal Analysis; Inconsistency among Experts in Blind Tests, Journal of Archaeological Science 28, 377-386.

Gonzalez-Urquijo, J.E., Ibanez-Estevez, J.J., 2003. The Quantification of Use-Wear Polish Using Image Analysis. First Results, Journal of Archaeological Science 30, 481-489.

Goodale, N., Otis, H., Andrefsky Jr, W., Kuijt, I., Finlayson, B., Bart, K., 2010. Sickle blade life-history and the transition to agriculture: an early Neolithic case study from Southwest Asia, Journal of Archaeological Science 37, 1192-1201.

Grace, R., 1989. Interpreting the Function of Stone Tools: The quantification and computerisation of microwear analysis., B.A.R. international series 474, Oxford.

Grace, R., 1996. Use-Wear Analysis: The State of the Art, Archaeometry 38, 209-229.

Ho, T.K., Hull, J.J., Srihari, S.N., 1994. Decision combination in multiple classifier systems, Pattern Analysis and Machine Intelligence, IEEE Transactions on 16, 66-75.

Hovis, D., 2009. Personal Communication, Case Western Reserve University.

Karperien, A.L., Jelinek, H.F., Buchan, A.M., 2008. Box-Counting Analysis of Microglia Form in Schizophrenia, Alzheimer's Disease, and Affective Disorder, Fractals 16, 103-107.

Keeley, L., 1980. Experimental Determination of Stone Tool Uses: A Microwear Analysis, The University of Chicago Press, Chicago.

Kimball, L.R., Kimball, J.F., Allen, P.E., 1995. Microwear polishes as viewed through the atomic force microscope, Lithic Technology 20, 6-28.

Krippendorff, K., 2004. Content Analysis: An Introduction to Its Methodology, Sage, Thousand Oaks, CA.

Lam, L., 2000. Classifier Combinations: Implementations and Theoretical Issues, in: Kittler, J., Roli, F. (Eds.), MCS 2000, LNCS 1857, Springer-Verlag, Berlin, pp. 77-86.

Lebow, C.G., Harro, D.R., McKim, R.L., Denardo, C., 2001. Archaeological Excavations at CA-SBA-246, An Early Holocene Site on Vandenberg Air Force Base, Santa Barbara County, California., Applied EarthWorks, Inc., Fresno, California, for Tetra Tech, Inc., Santa Barbara, California. Submitted to 30 CES/CEV, Vandenberg Air Force Base, California.

Lerner, H., Du, X., Costopoulos, A., Ostoja-Starzewski, M., 2007. Lithic raw material physical properties and use-wear accrual, Journal of Archaeological Science 34, 711-722.

Levi Sala, I., 1986. Use wear and post-depositional surface modification: A word of caution, Journal of Archaeological Science 13, 229-244.

Liu, B., 2007. Supervised Learning, Web Data Mining, Springer, Berlin, pp. 55-116.

Newcomer, M., Grace, R., Unger-Hamilton, R., 1986. Evaluating microwear analysis with blind tests, Journal of Archaeological Science 13, 203-218.

Odell, G.H., Odell-Vereecken, F., 1980. Verifying the Reliability of Lithic Use-Wear Assessments by 'Blind Tests': The Low-Power Approach, Journal of Field Archaeology 7, 87-120.

Rahman, A.F.R., Fairhurst, M.C., 2003. Multiple Classifier Decision Combination Strategies for Character Recognition: A Review, International Journal on Document Analysis and Recognition 5, 166-194.

SAS Institute, 2007. JMP, version 7, Cary, NC: SAS Institute, Inc.

Scott, R.S., Ungar, P.S., Bergstrom, T.S., Brown, C.A., Grine, F.E., Teaford, M.F., Walker, A., 2005. Dental microwear texture analysis shows within-species diet variability in fossil hominins, Nature 436, 693-695.

Scott, R.S., Ungar, P.S., Bergstrom, T.S., Brown, C.A., Childs, B.E., Teaford, M.F., Walker, A., 2006. Dental microwear texture analysis: technical considerations, Journal of Human Evolution 51, 339-349.

Stemp, W.J., Stemp, M., 2001. UBM Laser Profilometry and Lithic Use-Wear Analysis: A Variable Length Scale Investigation of Surface Topography, Journal of Archaeological Science 28, 81-88.

Stemp, W.J., Stemp, M., 2003. Documenting Stages of Polish Development on Experimental Stone Tools: Surface Characterization by Fractal Geometry Using UBM Laser Profilometry, Journal of Archaeological Science 30, 287-296.

Tringham, R., Cooper, G., Odell, G., Voytek, B., Whitman, A., 1974. Experimentation in the Formation of Edge Damage: A New Approach to Lithic Analysis, Journal of Field Archaeology 1, 171-196.

Ungar, P.S., Brown, C.A., Bergstrom, T.S., Walker, A., 2003. Quantification of Dental Microwear by Tandem Scanning Confocal Microscopy and Scale-Sensitive Fractal Analyses, Scanning 25, 185-193.

Van den Dries, M.H., 1998. Archaeology and the application of artificial intelligence; case studies on use-wear analysis of prehistoric flint tools, Archaeological Studies Leiden University I., Leiden.

Vaughan, P.C., 1985. Use-wear Analysis of Flaked Stone Tools, University of Arizona Press, Tucson.

Figures

Fig. 1. LSCM images of tool edges. a: unused replica, b: replica used for scraping wood, c: replica used for cutting soft plants, d: prehistoric tool (981-3) classified as used on hard material (wood), e: 3D projection of replica tool depicted in c.

Fig. 2. Flowchart illustrating quantitative classification process. Probabilities from each classifier are marked P_{ij}, $P_{ij'}$ …etc. C = combiner rule (e.g., arithmetic mean), P_i = combined probabilities, D = decision rule (e.g., assign to class with largest P_i).

Fig. 3. Discriminant analysis canonical plot of polish data training set using all 11 variables considered. Rays represent the direction of the variables in the canonical space. Circles represent 95% confidence limits. Rq: root mean square deviation, Rku: kurtosis, Rsk: skewness, FPO: mean polar facet orientation, MFOV: variation of the polar facet orientation, FAD: direction of azimuthal facets, MRV: mean resultant vector, SA: surface area (Chinga, et al., 2007), Df: fractal dimension, LAC: lacunarity, PLAC: prefactor lacunarity (Karperien, et al., 2008).

Fig. 4. Discriminant analysis canonical plot of polish data training set using MRV, SA, and Df. Circles represent 95% confidence limits.

Fig. 5. Decision tree of edge damage training set using presence/absence of step fractures and edge rounding to partition hard materials from soft materials and hide from other soft materials.

Tables

1. Replica test set results. Results in bold are correctly classified artifacts. The “Analyst” column reports blind test results using conventional microscopy; “a” = Analyst A, “b” = Analyst B. Quantitative classifications are reported in the following three columns with the last column containing maximum probabilities of combined classification.

2. Replica test set results with combined probabilities to all contact materials. Boxed numbers in bold are maximum probabilities. “Pred Material” column reports combined quantitative classification with contact materials in bold matching actual uses.

3. Archaeological test set results. “Pred Material” column reports combined quantitative classification. Boxed numbers in bold are maximum probabilities while contact materials in bold are consistent with those of the conventional analyst. *1616-1a is considered plausible, but is not included in calculations of the success of this method.

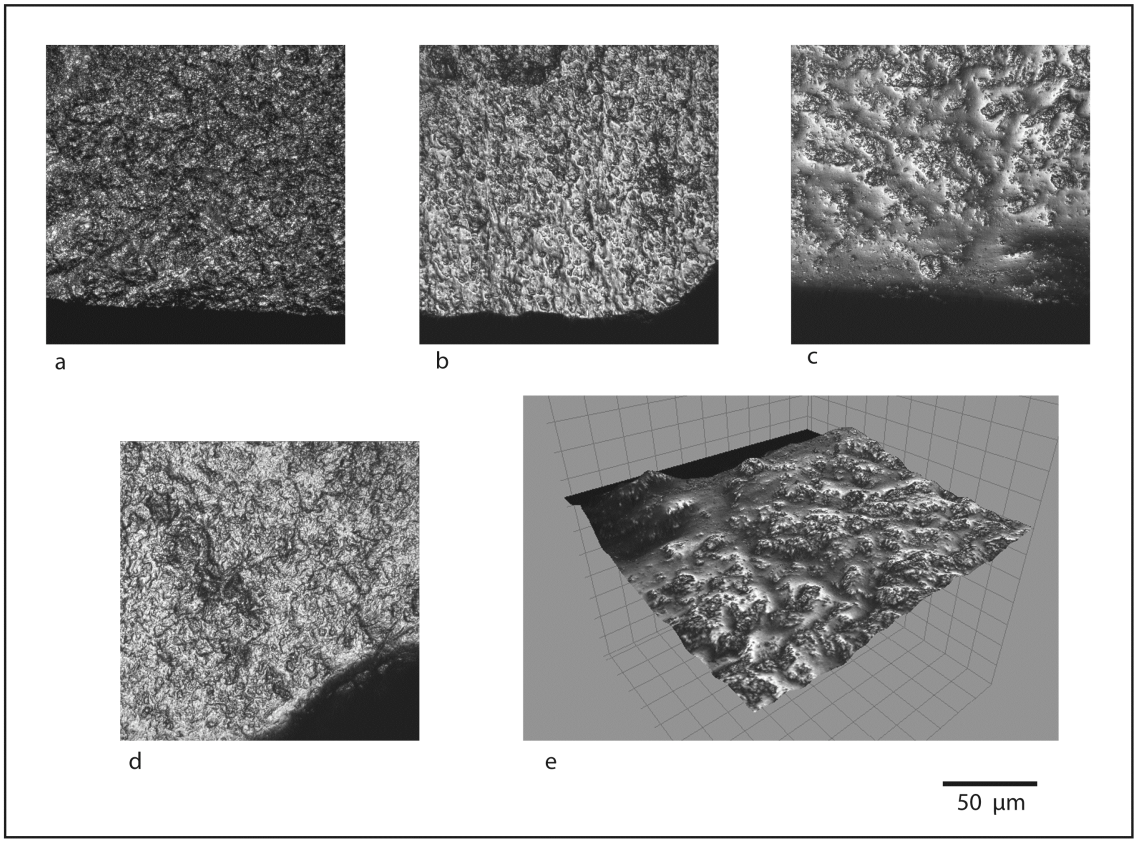

Figure 1.

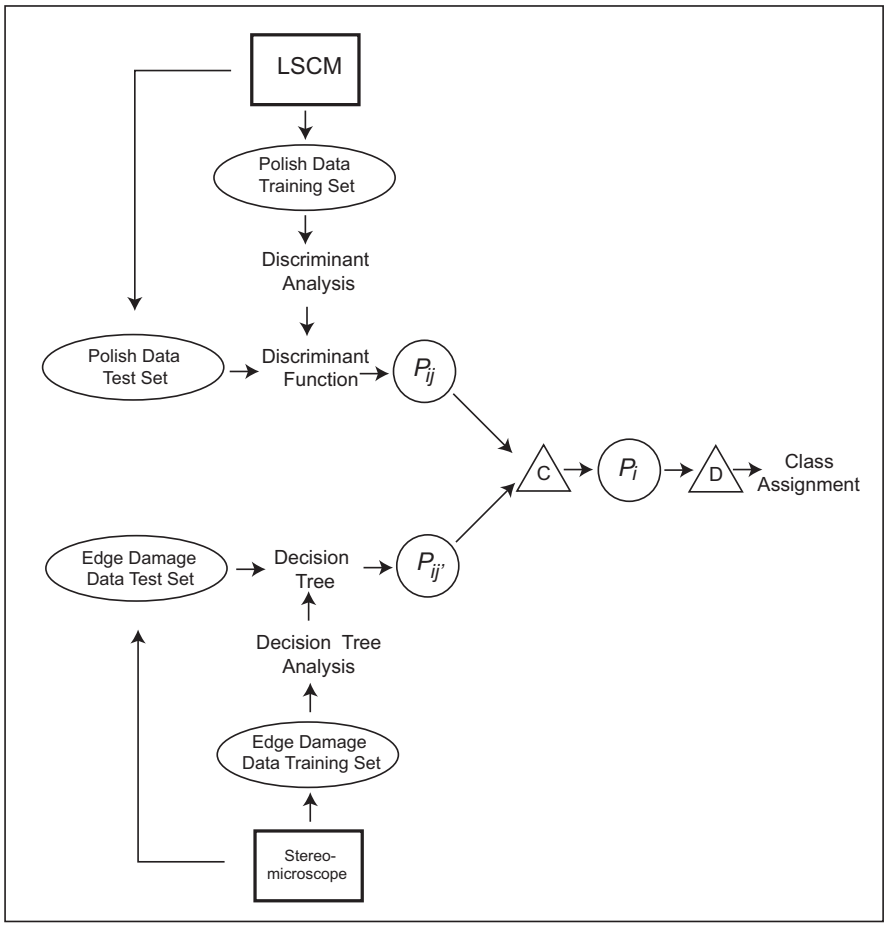
LSCM
Polish Data Training Set
Discriminant Analysis
Polish Data Test Set
Discriminant Function
P_{ij}
C
P_i
D
Class Assignment
Edge Damage Data Test Set
Decision Tree
$P_{ij'}$
Decision Tree Analysis
Edge Damage Data Training Set
Stereo-microscope

Figure 2.

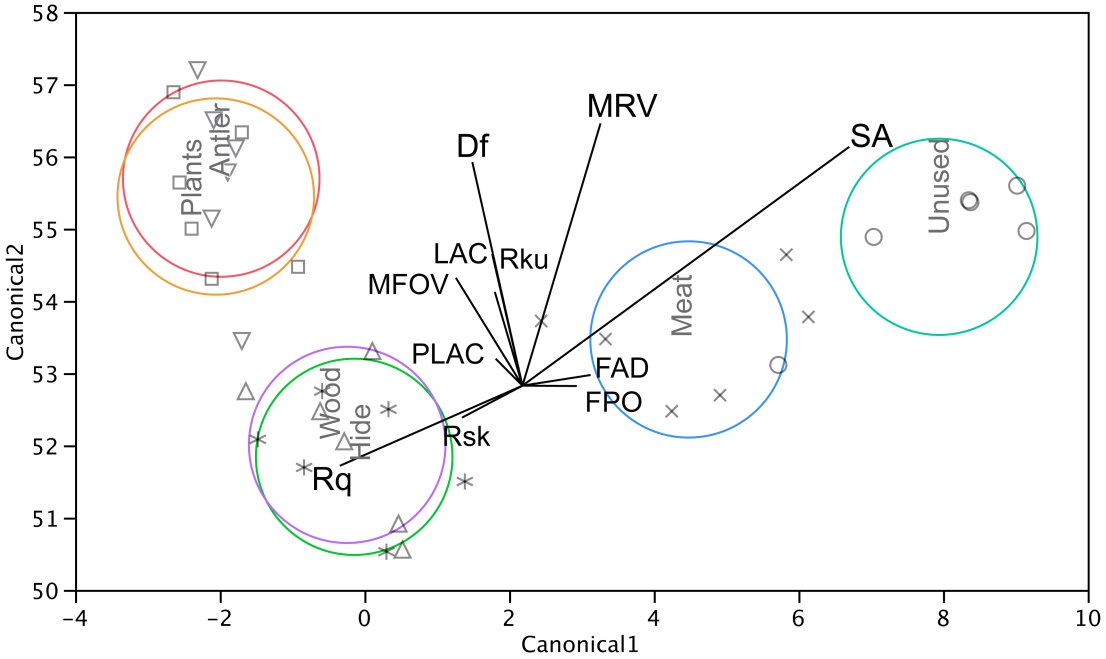

Figure 3.

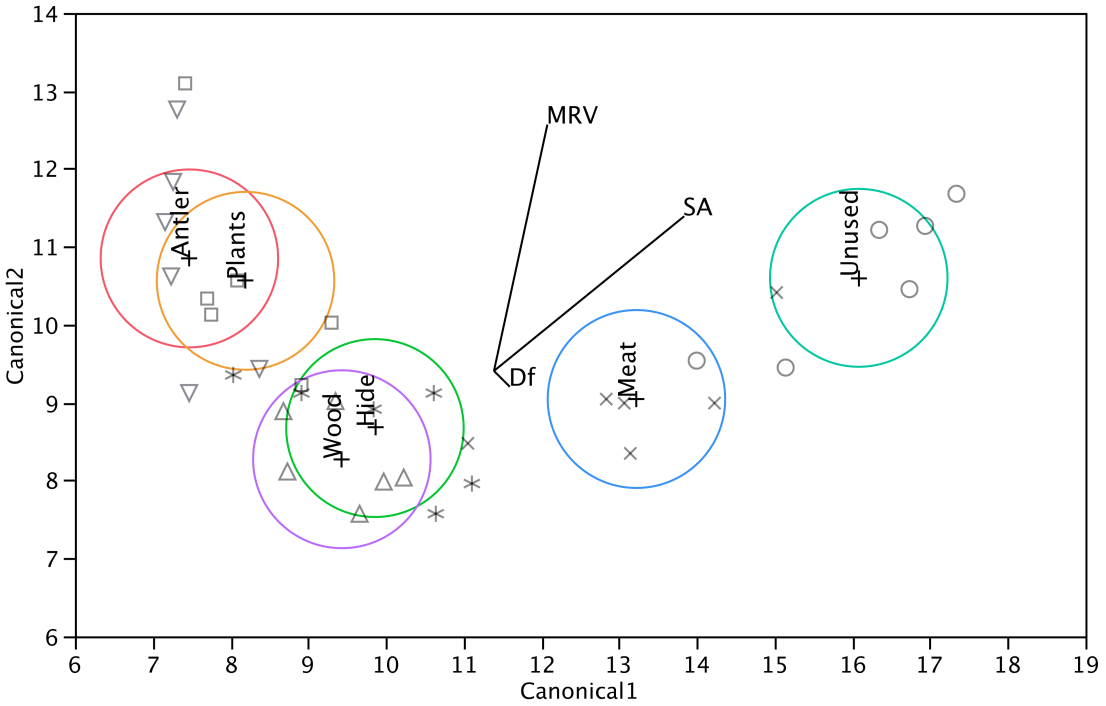

Figure 4.

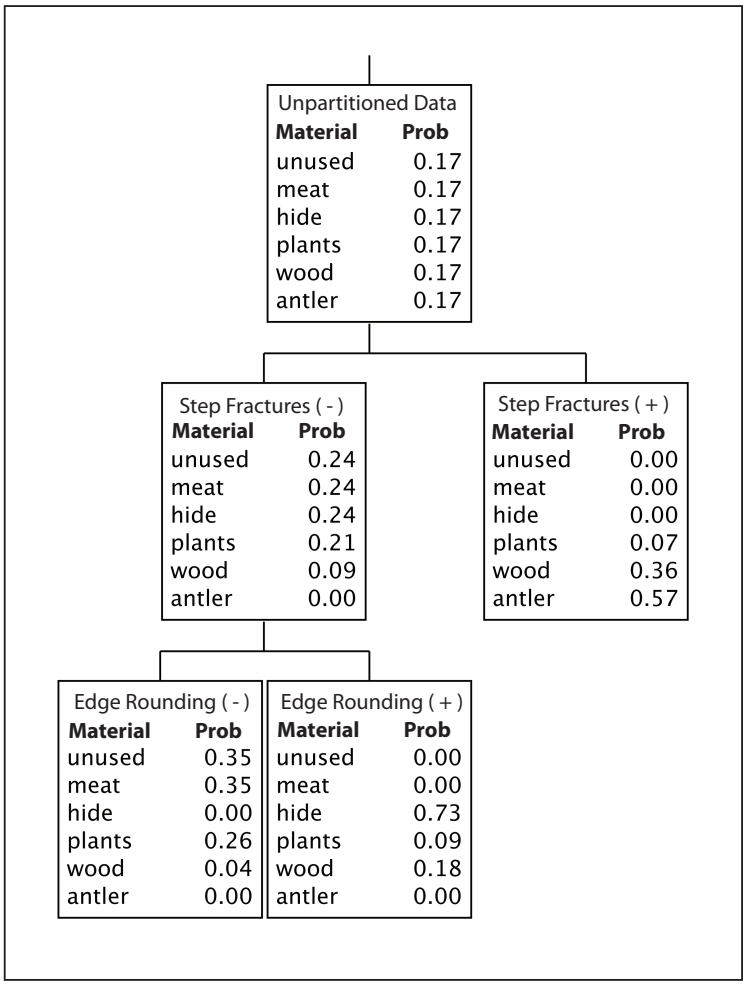
Unpartitioned Data
Material Prob
unused 0.17
meat 0.17
hide 0.17
plants 0.17
wood 0.17
antler 0.17
Step Fractures (-)
Material Prob
unused 0.24
meat 0.24
hide 0.24
plants 0.21
wood 0.09
antler 0.00
Step Fractures (+)
Material Prob
unused 0.00
meat 0.00
hide 0.00
plants 0.07
wood 0.36
antler 0.57
Edge Rounding (-)
Material Prob
unused 0.35
meat 0.35
hide 0.00
plants 0.26
wood 0.04
antler 0.00
Edge Rounding (+)
Material Prob
unused 0.00
meat 0.00
hide 0.73
plants 0.09
wood 0.18
antler 0.00

Figure 5.

Artifact	Actual	Analyst	Edge D. Only	Polish Only	Combined	Probability
3	unused	**unused**[b]	meat	**unused**	**unused**	0.66
7	unused	**unused**[b]	meat	**unused**	**unused**	0.66
104	hide	**hide**[b]	**hide**	wood	**hide**	0.49
118	wood	hide[b]	unused	hide	hide	0.28
119	meat	**meat**[b]	unused	unused	unused	0.67
121	hide	antler[b]	**hide**	plants	plants	0.40
123	antler	**antler**[b]	**antler**	plants	**antler**	0.41
124	plants	**plants**[b]	hide	antler	hide	0.41
126	antler	wood[b]	**antler**	**antler**	**antler**	0.69
127	hide	meat[a]	**hide**	wood	**hide**	0.41
128	unused	**unused**[a]	meat	**unused**	**unused**	0.66
129	antler	**antler**[a]	**wood**	**antler**	**antler**	0.61
130	meat	wood[a]	unused	**meat**	**meat**	0.65
131	antler	**antler**[a]	**antler**	**antler**	**antler**	0.65
132	wood	**wood**[a]	antler	plants	antler	0.46
133	wood	plants[a]	hide	hide	hide	0.62
134	unused	**unused**[a]	meat	meat	meat	0.67
135	hide	**hide**[a]	**hide**	wood	**hide**	0.50
136	wood	hide[a]	hide	**wood**	hide	0.54
137	wood	**wood**[b]	antler	hide	**wood**	0.35
Correct		65%	40%	40%	60%	

Table 1.

Artifact	P[Antler]	P[Hide]	P[Meat]	P[Plants]	P[Unused]	P[Wood]	Pred Material	Actual
3	0.00	0.00	0.18	0.14	**0.66**	0.02	**unused**	unused
7	0.00	0.00	0.18	0.14	**0.66**	0.02	**unused**	unused
104	0.00	**0.49**	0.01	0.05	0.00	0.45	**hide**	hide
118	0.00	**0.28**	0.18	0.14	0.18	0.22	hide	wood
119	0.00	0.00	0.17	0.14	**0.67**	0.02	unused	meat
121	0.07	0.37	0.00	**0.40**	0.00	0.15	plants	hide
123	**0.41**	0.04	0.00	0.22	0.00	0.33	**antler**	antler
124	0.34	**0.41**	0.00	0.11	0.00	0.14	hide	plants
126	**0.69**	0.00	0.00	0.11	0.00	0.20	**antler**	antler
127	0.08	**0.41**	0.00	0.21	0.00	0.29	**hide**	hide
128	0.00	0.00	0.18	0.14	**0.66**	0.02	**unused**	unused
129	**0.61**	0.00	0.00	0.20	0.00	0.19	**antler**	antler
130	0.00	0.00	**0.65**	0.14	0.18	0.02	**meat**	meat
131	**0.65**	0.00	0.00	0.16	0.00	0.19	**antler**	antler
132	**0.46**	0.03	0.00	0.25	0.00	0.27	antler	wood
133	0.02	**0.62**	0.00	0.15	0.00	0.22	hide	wood
134	0.00	0.00	**0.67**	0.14	0.17	0.02	meat	unused
135	0.00	**0.50**	0.08	0.05	0.00	0.37	**hide**	hide
136	0.00	**0.54**	0.00	0.06	0.00	0.39	hide	wood
137	0.32	0.25	0.00	0.09	0.00	**0.35**	**wood**	wood

Table 2.

Artifact	P[Antler]	P[Hide]	P[Meat]	P[Plants]	P[Unused]	P[Wood]	Pred Material	Analyst
963-5	0.29	0.30	0.07	0.04	0.00	**0.31**	**hard material**	hard material
964-8	0.29	0.16	0.00	0.05	0.00	**0.50**	**wood**	hard material
964-11	0.00	0.41	0.00	0.05	0.00	**0.54**	wood	soft plants
971-10	0.29	0.27	0.00	0.04	0.00	**0.40**	**hard material**	wood
974-4	0.29	0.20	0.00	0.04	0.00	**0.47**	**hard material**	hard material
980-4	0.02	**0.50**	0.00	0.09	0.00	0.39	**hide**	soft material
981-3	0.29	0.24	0.00	0.05	0.00	**0.43**	**hard material**	wood
984-2	0.00	0.37	**0.49**	0.05	0.00	0.09	**soft material**	meat
987-5	**0.29**	0.15	0.26	0.04	0.00	0.27	**hard material**	hard material
988-10	0.00	0.38	**0.48**	0.05	0.00	0.09	**soft material**	soft material
996-3	0.29	0.00	**0.47**	0.04	0.03	0.18	**soft material**	soft material
1549-6	0.01	**0.49**	0.00	0.06	0.00	0.44	**soft material**	dry hide
1554-1	0.06	**0.46**	0.00	0.13	0.00	0.36	soft material	hard material
1573-4	0.29	0.21	0.00	0.04	0.00	**0.46**	hard material	meat
1598-3	0.29	0.00	0.00	0.04	**0.50**	0.18	unused	hard material
1616-1a	0.00	**0.52**	0.00	0.05	0.00	0.43	**hide***	indet, cutting
1616-1b	0.29	0.00	**0.33**	0.04	0.17	0.18	soft material	hard material
980-4int	0.00	0.06	**0.59**	0.13	0.17	0.05	meat	unused
964-8int	0.00	0.00	0.17	0.13	**0.67**	0.02	**unused**	unused
974-4int	0.00	0.00	**0.52**	0.13	0.32	0.02	meat	unused

Table 3.