Geo-Archeological Analysis of Middens on the Island of Oronsay: What multivariate Statistics and Machine Learning Reveal

Jonathon Sevy

Department of Geological Sciences

Brigham Young University, Provo

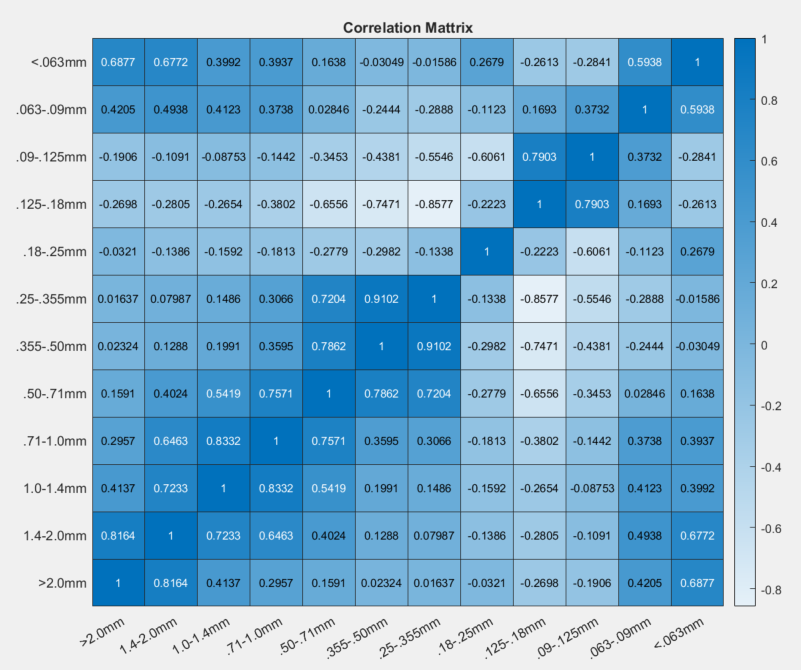
# Abstract and Introduction

Although Archology has previously discussed the cultural importance of Midden burial sites it is still unknown if middens were buried in beaches or in sand dunes behind the beaches. To solve this multivariate statistics and machine learning were employed. Modern beach and dune sands were collected and sieved into size fractions. These size fractions were consolidated using principal component analysis and machine learning in MATLAB and analyzed. Using Geologic principles to examine the results, it is concluded that the middens were most likely berried in beaches.

# Methods and Results

Samples were fractionated by grainsize into 12 categories and weighed. These were then z-scored. Z-scoring allows variables that have very diferent units to be compared easily to each other, without the order of magnitude throwing off the statistical processes. A correlation matrix was made. The correlation matrix helps show variables that are correlated, both positively and negatively, to each other. The correlation matrix as shown in figure 1 give a good idea of the variables that can be condensed in a principal component analysis to make a fewer axis.

Figure 1. Correlation matrix or grain size fractions. This shows correlation coefficient between the 12 grain size fractions. This give a good idea that there are some good correlation variables that Principal Component Analysis will likely work well.



Principal component analysis was then preformed on the z-scored data using the MATLAB build in function ‘pca’. Principal components are combinations of the orrigional variables combined to maximize variance in the minimal number of axes. This allows us to look at fewer variables without losing variance, or information. The first three principal components of our analysis accounted for 39.8%, 26.4%, and 16.2% of the variance, respectively. In other words, 66.2% of variance can be accounted for by the first 2 principal components, and 82.4% of the variance can be accounted for by the first 3 principal components. By this method, Principal component analysis allows us to graph and view a majority of information (variance) in 2 and three dimensions, which is a more instinctive way to view and understand data.

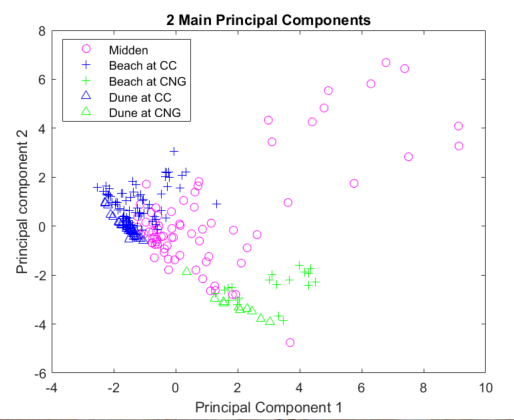


Figure 2. Graph of the first two principal components. These components account for 66.2% of variance. Although quite a bit of variance is missing, it gives a good idea of how dune, beach, and midden sands are behaving. Note should be taken that the dunes plot along a very close line, with a negative slope, and beach and midden sand starts at that line and then trails up with increasing principal component 1 and 2.

Graphing principal component 1 versus principal component 2 (see figure 2) shows the dune at both locations graph more or less in a line with a negative slope. Beach sands start at the line where the dune sands graph and tail up in the positive direction. Midden data, similar to the dunes plots positively sloping from the same area but with larger values. Adding a third dimension to the graph, as seen with figure 3, turns the line where all datasets are to a plane, with all the dune data staying nearly on the plane, and the beach data extending away from the plane, decreasing in principal component 3 as it moves away. Midden data also extends out from the plane, but in two distinct ways. The first tails out at a very similar angle to the beach samples but elevated along the 3rd principal component. The Second tails away from the plane, but instead of being sloped downward along the 3rd principal component, it stays nearly constant.

Lastly, Machine learning was used through MATLAB’s classification learner application to build a predictive model. Three models were chosen, ensemble bagged trees, Weighted KNN, and Trilayered Neural Network. All of these models scored above 80% in the training runs and are differing in method of how they split the data. These can all be easily run through the included “Model\_PCA” script, to for the given result of each. These all tend toward guessing dune sands.

# Conclusion

From looking at the result of the 2 dimensional and 3 dimensional principal components, id appears that although all the sand has the same base size, the processes that sort the sand can be inferred. The linear/planar nature of the dune sands is interpreted to reflect the narrow range of sorting that takes place in dunes, with very little variation. The beach sand however has both high and low energy acting on it. The range of energy which the beach experiences allows different grain size fractionations. The tailing of beach data in the higher principal component 1 and 2 direction is interpreted as this multimodal energy which beaches experience.

The models included here all predict dunes, if there is only one option for middens. This would seem to be a classification mis representation because the tailing away on higher 1st and 2nd principal components would be controlled by energy, like the dune’s tailing. This is likely a miss classification and is perhaps caused because although many of the dunes follow the same trend as beaches, since the middens are elevated (along PC3) there is no overlap, so there was no comparable data to middens in the training data. In this case the process that the principal components describe is interpreted as being more telling than the predicted ability of these models. Better models would be ones that can distinguish on trend on, or away from the plane that forms in the data. Until better models can be formed and trained, these results are discarded and middens being made in beaches is still excepted as the most probable solution.

Lastly, the presence of two distinct midden trends indicates that perhaps not all middens were the same. This diference could be one of location, such as middens being in the bay versus an unprotected part of the island. Or the diference could be two different forms of creating or use of the midden. Although this study does not go into enough detail to resolve which of these it could be, future research should focus on determining the cause of the bi-trend nature of the midden data in the 3rd principal component dimension.

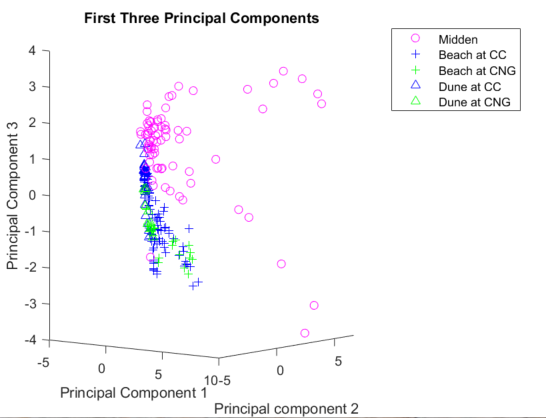


Figure 3. Graph of the first three principal components. These components account for 82.4% of variance. This graph gives a good idea of how dune, beach, and midden sands are behaving. Note should be taken that the dunes plot along a very thin plane (nearly vertical at this angle, extending in and out of the board). Additionally, beach sands are tailing in a negative PC3 trend away from the dune plane. midden sand starts at that line and trail away in two bodies, one similar to beach sands, but elevated, the other with almost constant PC3 values.