Neural Network Classification of Skeletal Remains

Suzanne Bell
Department of Chemistry, Eastern Washington University
226 Science Building
Cheney, WA, USA
Phone 509-359-7932 - Fax 509-359-6973 - E-mail: Suzanne.Bell@mail.ewu.edu

Richard Jantz
Department of Anthropology
University of Tennessee, Knoxville, USA

Abstract: Neural networks are data analysis tools well suited to archaeological applications. With both supervised and unsupervised architectures available, networks are flexible and complement existing multivariate analytical techniques. In the present application, networks were used to classify skeletal remains based on common osteological measurements. Both a contemporary (mid-1800s to 1970) and an older database (~1600-1800) were investigated to learn if networks could classify based on several different criteria including race, sex, time frame, phase, and date of birth. Use of genetic algorithms and variable contribution measures allowed an evaluation of relative importance and predictive power of variables retained in the models. Results were compared with traditional multivariate techniques and while performance was not, in general, dramatically different, interesting and potentially significant differences in variable contribution measures were noted. This emphasizes the value of networks as complimentary tools for statistical analysis of archaeological data.

Key words: osteological analysis, neural networks, genetic algorithms, cranial measurements.

Introduction and Background

Multivariate statistical techniques have been used to classify skeletal remains based on age at death, sex, and race. In the present example, results from such statistical analyses were compared to and contrasted with results obtained using back propagation neural networks coupled to genetic algorithms. Two different data sets, one an archaeological and the other a forensic, were used to evaluate the feasibility of classifying remains based on skeletal cranial measurements. Although specific applications and models varied, each study had four overriding goals:

1) To determine if a neural network model could successfully predict the attribute in question.

2) To determine which variables were the most and least important for classification and prediction.

3) To evaluate trends and relationships among and between these critical variables

4) To compare the results with those obtained using traditional multivariate analysis

Although neural networks have becoming an increasingly common tool in business, engineering, and sciences, few archaeological applications have been reported (for example, Barcelo 1993 and 1997, Bell 1998, Reeler, 1999). While a detailed description of network theory, design, and applications is beyond the scope of this paper, general references and texts are available in most university libraries. A good overview is presented by Wasserman (Wasserman 1989).

Several characteristics lend networks to complex problems as often encountered in physical anthropology and archaeology. Networks are capable of uncovering intricate, non-linear relationships among variables that might be overlooked or missed by statistical techniques such as principle component analysis. Networks do not require that the underlying variable distributions be normal. Finally, a variety of network types are now available in commercial PC-based software, providing the investigator with a number of options in both supervised and unsupervised learning techniques.

Datasets and Methods

Two databases were examined, both containing cranial measurements exclusively. A contemporary database, spanning the years AD 1850-1970, was derived from anatomical and forensic investigations. This database was comprised of measurements from 885 individuals classified by sex and race (whites versus blacks). Most of the individuals with 19th century birth dates were obtained from the Terry and Todd anatomical collections, while those with 20th century birth dates were obtained from the Forensic Anthropology Data Bank (Ousley and Jantz 1997) Fifteen cranial measurements were reported for each. The second collection was archaeological, representing 559 individuals from various sites occupied by the Arikara Indians located in the Middle Missouri region of South Dakota (Key 1983 and McKeown 2000). Years represented ranged from AD 1600-1817 and data was collected from 25 excavations. For each individual, 66 cranial measurements were reported.

The archaeological data was classified by phase, time frame,

and bank of the river on which the site was located. For time frame, the identifier "Arikara" refers to the historic and protohistoric sites, while "EC" refers to the prehistoric. A summary is presented in Table 1. Although the data in the forensic database was evenly distributed, this archaeological database was not. When broken down by time frame, 392 of the skeletons were Arikara while 158 were EC. Similar imbalance was seen based on riverbank, with 403 skeletons representing the left bank and 147 representing the right bank. Further complications were introduced by the uneven distribution of site dates over time. For example, with right and left bank, only data from the right bank was dispersed enough over time to allow for an investigation of how cranial measures evolved. When considering phase, imbalances were even more significant. As a result, models created to predict phase were limited to distinguishing the Le Beau 3 phase from all others. While such imbalances in data distribution do not preclude statistical or neural network evaluation, it is important to keep it in mind when interpreting the models.

Neural network analysis was conducted using two software packages: Neurogenetic Optimizer (NGO, Biocomp Systems, Redmond, WA USA) and Statistica Neural Networks (Statsoft, Tulsa, OK USA). Both of these packages utilize genetic algorithms for network optimization and variable selection. Multivariate statistical analysis was conducted using Statistica. The databases were converted to Excel and comma-delimited text file format for network analysis.

Neural networks are layered structures of interconnected mathematical processing elements linked together as shown in figure 1. Note that the first layer consists of one processing element per input variable and that these processing elements do not perform any mathematical operations. Rather, they pass the value of the variable on to each element in the hidden layer. Nodes in the hidden and output layers perform the mathematical operations illustrated in figure 2. Each of these nodes has associated with it a set of weights (equal to the number of inputs) that are initially set to small random values. The processing element performs two operations. First, the input values are multiplied by their corresponding weight and the value summed. Second, this sum is passed onto a transfer or "squashing" function the compresses the value between 0 and 1 or −1 and +1, depending on the transfer function selected. This is where the analogy to the nervous system can be seen; the processing element, also called a neuron, will either "fire" or not fire (be activated or not) based on the sum of the inputs. A common feature of most transfer functions is the abrupt transition between a relatively large value (analogous to a nerve cell activating) and a relatively small value (analogous to an inactive neuron). If the node is in the hidden layer, the output from the transfer function is passed to the next layer.

The size of the hidden layer is determined by optimization experiments but in general, the smaller this layer is, the better the model will generalize. Although any number of nodes can be used in the output layer, in the present case, all networks models utilized one. Outputs in these models could be continuous (such as year of birth for the forensic database) or discrete (such as which bank of the river the individual was recovered from). In such cases, a binary encoding was used. For example, 1 would encode for a skeleton recovered from a site on the left bank while 0 would encode for a skeleton recovered on the right bank. Numerical results from the output node were scaled accordingly between 0 and 1.

Training of a neural network involves iterative adjustment of the weights until all training data presented to the network results in a correct response within the error threshold specified by the user. Individually, inputs (here a set of cranial measurements for one individual) are presented to the network and calculations are performed. The result is compared to the desired value and the resulting error is back propagated through the weights and used to make corrections. In practice, error is usually accumulated over one complete pass through the training data (commonly called an epoch) before weights are adjusted. This process is repeated through hundreds or thousands of epochs until training is complete. Progress is monitored not by the training data itself, but by a representative cross section removed prior to training. The training process continues for a set number of epochs (for example 100), and then the test data is run through the network and an error calculated. Training is resumed for another 100 epochs and the test is repeated. Training is halted if one of three conditions is met: 1) test performance meets the user-defined threshold; 2) test performance is no longer improving; or 3) the maximum number of epochs has been reached.

Note that the training and test data are both involved in the training process and as such do not provide an independent evaluation of network performance. For this, a separate set of validation data is needed, which consists of a representative cross-section removed from the original dataset before training. The performance of a network can be reliably measured using this validation data, and this is the metric reported throughout this paper. Figure 3 summarizes how data was divided, and how these subdivisions were used to generate the five replicates for each model constructed. Each replicate run represents a reshuffling of training, test, and validation data and these replicates were used to gauge the reproducibility and uncertainty (at the 95% confidence interval) of the models.

Until recently, optimization of the many parameters and settings associated with a given neural network model, as well as the selection for variables to include, has been problematical. Settings such as number of hidden layers, number of nodes in the hidden layer, transfer functions, learning rates, etc. all must be determined, yet little more than experience and rules-of-thumb existed for guidance. In the last ten years, an increasing number of neural network packages have incorporated genetic algorithms, which have greatly increased the utility of the technique. To implement a genetic algorithm optimization, the first step is to create a chromosome (figure 4), which is a binary encoding of all the parameters necessary to construct a neural network. This can include type of network, number of hidden layers, number of hidden nodes per layer, etc., but also includes instructions as to which variables are to be included. Genetic algorithm optimization (figure 5) begins with the generation of a pool of such chromosomes, typically 30-50. For each generation, a network is built according to the instructions encoded on each chromosome and the network is trained long enough to gauge its fitness relative to other networks. Fitness takes

into account accuracy and size; smaller networks tend to generalize better and are favored. At the end of the training and comparison step, the fittest chromosomes (the ones generating the best performing networks) are allowed to pass to the next generation. The candidate pool is brought back to original size by allowing the fittest chromosomes to mate and mutate, and by adding additional, randomly generated chromosomes into the pool. The cycle repeats until one of three conditions is met: 1) no further improvement in fitness is occurring; 2) a user-defined level of performance is achieved; or 3) the maximum number of allowed generations has elapsed. At the end of this process, the algorithm will have converged on an optimal solution; however only replicate analyses (as described previously) will verify that the solution is reproducible and reliable. All network models discussed in this paper were obtained using genetic algorithm optimization as described here.

Because the outputs of all the networks were binary, the issue of cutoff or threshold values must be addressed and is best illustrated through an example. With the archaeological database, one set of models was constructed to predict bank of origin based on cranial measurements. Thus, each set of measurements was associated with either the right bank (coded as 1) or the left bank (coded as 0) of the Missouri river. Accordingly, the value of the output node varied continuously from 0 to 1. A decision must be reached as to what error threshold will be accepted as a correct response. Commonly, values of 0.51 and above are considered to be 1, while values of 0.49 or less are considered to be 0, and the cutoff value in this case is 0.50. Clearly, this is a generous threshold and may or may not be reasonable based on the problem at hand. Rather than set an absolute threshold to start with, results were evaluated for three thresholds: 0.90, 0.75, and 0.60. In all cases, the metric used to monitor network performance was the percentage of validation cases that were correctly classified by the model based on the threshold value selected.

Once working models were constructed and optimized, the next step was to determine which cranial measurements were the most important, how much each contributed, how they were related to each other, and how they were changing over time. Terms such as variable contribution, sensitivity, weight, and importance are often cited in the literature, but there are no universally accepted definitions. Within this study, the methods of calculating variable contribution were used as defined below:

1) **Sensitivity**: Determined using the trained network and validation data. The sensitivity of a variable is calculated by dithering its value +/- 10% of the original and averaging the change in network output across all validation cases. This measurement, while not truly multivariate, helps to identify variables than can have a large impact on network output.

2) **Number of occurrences**: Each network model was built using five replicates as discussed above. Accordingly, the genetic algorithm was repeated five times as well. A variable that is important to achieving optimal network performance should be included in the model more often than variables that are not important. Accordingly, the number of occurrences of a variable (0-5) is one measure of its importance. Moreover, this metric is multivariate in that the genetic algorithm will converge on an optimal variable combination.

It is important to note that a variable can be important and not sensitive and vice versa. Consider a variable that is included by the genetic algorithm in most of the models, but small changes in that variable do not result in large changes in network output. However, removal of this hypothetical variable results in dramatic degradation of network performance. Such a variable is an example of an important variable that is not sensitive.

Both of these criteria (sensitivity and number of occurrences) were combined in a quantity called the variable score calculated as follows:

Score = (0.33(normalized sensitivity) + 0.67(normalized number of occurrences)) * 100

The weighting factors were assigned somewhat arbitrarily but reflect the greater weight attributed to the multivariate measure. Since networks excel at finding complex, nonlinear, multivariate relationships among variables, the metric used to gauge variable importance should be slanted in favor of the multivariate measurement.

Results and Discussion

Results for the archaeological database are summarized in figures 6-8. Not surprisingly, under the most stringent threshold (0.90), none of the models predicted their target characteristic as well as chance. However, at a threshold of 0.75, the models were able to predict time frame slightly better than chance (looking at the average % correctly classified). At the 0.60 threshold, all three characteristics were predicted better than chance (even considering uncertainties) and the difference in performance between networks predicting time frame and bank were statistically significantly. Across all models and all thresholds, there were no statistically significant differences when males and females were considered separately. Given the nature of the predictions and the inevitable uncertainties and inconsistencies that arise from different investigators collecting data, it was assumed that the 0.60 threshold was the most reasonable one from which to draw conclusions. Thus, results suggest that the river served was an effective barrier between sites and peoples since cranial measurements were sufficient to distinguish them. Certainly, other factors play a role and investigation is continuing.

Once models were optimized, the next step was to evaluate variable contribution and importance. Using the variable score defined above, the ten most important variables for each of the three models (predict time frame, predict bank, and predict phase) were determined across five replicate networks. These variables were then examined to evaluate changes over time and to study the relationships between them. The ten most important variables for each model are summarized in Tables 2,3, and 4. For predicting time frame, the most important cranial measurement is Nlh, followed by a group of similar scores for Ocs, Aub, Ocr, Sos, Asb, Mls, and Xcb. The trend listed in the

table is how the measurement appears to change over time and is discussed in detail below. Note that Ocs was also among the top ten for predicting phase. For predicting time frame, Occ was the most important variable with the next closest (Avr) nearly ten points lower. Finally, for predicting phase, Mdb was even more dominant than Occ was for predicting time frame. However, since the imbalanced dataset limited what could be done with phase, this finding has limited application.

With important variables identified, the next step was to study how these variables are changing and how they are related to each other. Since the dates associated with the skeletons were discrete and not continuous, it was necessary to group individuals into time spans as shown in Table 5. This grouping was used for plotting purposes only. Examples plots are presented in figures 9-11, with examples of increasing, decreasing, and indeterminate patterns. Several difficulties were encountered in this phase of the work. First, given the imbalance in the data set (as per Table 1), it was difficult to obtain a representative number of individuals in many of the time spans for a given subset. Thus, while the networks were able to create good predictive models, and important variables were clearly identified, the exact nature of the relationships among these cranial measures cannot be deciphered with sufficient confidence to draw firm conclusions.

Analysis of the Forensic Database

Neural network models were also applied to a contemporary database comprised of an anatomical database from the nineteenth century combined with a forensic database from the twentieth. While the detailed results of this investigation are pending, a comparison of results with those obtained above is useful. The primary goal of this study was to determine if a network model could predict race, sex, and the year of birth based on cranial measurements. Previous work with multivariate statistics and regression models resulted in a mean square error of ~ 25 years, with a 95% confidence interval of +/- ~ 50 years.

Neural network models were constructed using the general techniques previously described with network performance and variable importance measured as before. Optimized models were able to predict race from cranial measurements ~86-92% of the time while models to predict sex were correct ~81-88% of the time. Models to predict year of birth with an uncertainty of ~ 20 years, better than obtained with regression. Interestingly, the genetic algorithm retained most variables in the models used to predict race and sex while fewer variables were retained for models used to predict year of birth. This finding suggests that information regarding sex and race is widely dispersed across the cranial measurements while information on year of birth is concentrated in fewer variables.

For predicting sex, Zyb was most important followed by Nph. For predicting race, the order was Bnl, Bpl, Nph, Nlh, and Nlb. For year of birth models and across all race and sex divisions, Bbh, Zyb, Aub, and Pac were the most important. With the interesting exception of Aub, these variable rankings coincide with results from multivariate statistics, in which Aub was not flagged as important. In the network models, Aub appeared to be sex-linked and more useful for predicting year of birth for females as opposed to males. In males, results suggested that Aub was a sensitive variable, but not an important one, a distinction that was discussed above.

Summary and Conclusions

The above studies illustrate the value of neural networks in archaeological and osteological applications, particularly when combined with traditional statistical methods. Networks excel at identifying complex, multivariate, and nonlinear relationships among variables and often identify important variables that other techniques such as regression analysis or principle component analysis overlook. However, because of the complex nature of the variable interactions, it is sometimes difficult to completely characterize these relationships. In the present case, much of this difficulty arises not from the networks, but from the nature of the data itself. Nonetheless, neural networks, particularly using up-to-date PC software and genetic algorithms, represent a valuable addition to the mathematical toolbox available to archaeologists and physical anthropologists.

Acknowledgements

Dr Bell's work was funded by the US National Science Foundation, POWRE program award.

Dr. Jantz' work on the forensic anthropology data bank was supported by a grant from the National Institute of Justice.

References

Barcelo, JA. 1993. Back-propagation algorithms to compute similarity relationships among archaeological artifacts, in *Computer Applications and Quantitative Methods in Archaeology, 1993*, BAR International Series, Oxford.

Barcelo, JA and Faura, JM. 1997. Time series and neural networks in archaeological seriation. An example on early pottery from the near east, in *Compuer Applications and Quantitative Methods in Archaeology 1997*, BAR International Series, Oxford.

Bell, S. 1998. *Artificial Neural Networks as a Tool for Archaeological Data Analysis*, Archaeometry 40(1), 139-151.

Key, P.J. 1983. *Craniometric Relationships Among Plains Indians: Culture-Historical and Evolutionary Implications, Report of Investigations, No. 34*, University of Tennessee, Knoxville.

McKeown, A.H. 2000. *Investigating Variation Among Arikara Crania Using Geometric Morphology, Doctoral dissertation*, University of Tennessee, Knoxville.

Ousley, SD and Jantz, RL. 1997. The Forensic Data Bank: Documenting skeletal trends in the United States. In: *Forensic Osteology*. 2nd ed. (Ed: Reichs, KJ) Charles C. Thomas, Spring-

field, IL, 441-458.

Reeler, C. 1999. Neural networks and fuzzy logic analysis in archaeology, in *Computer Applications and Quantitative Methods in Archaeology*, BAR International Series (supplementary), 750 (1998), 150.

Wasserman, PD. 1989. *Neural Computing: Theory and Practice*, New York: Van Nostrand Reinhold.

Tables

Table 1. Distribution of archaeological data

Time frame	Phase	River Bank	Period	n
Arikara	Arikara	Right	1817	52
Arikara	Bad River 1	Right	1705-1707	10
Arikara	Bad River 2	Right	1766-1768	39
Extended Coalescent (EC)	La Roche	Both	1612-1663	158
Arikara	Le Beau 1	Left	1687-1726	21
Arikara	Le Beau 2	Both	1700-1766	50
Arikara	Le Beau 3	Both	1688-1728	220

Table 2.Ten most important variables for predicting bank

Cranial measurement	Score	Trend[1]
Nlh	87	Increasing
Ocs[2]	80	Indeterminate
Aub	78	Indeterminate
Ocr	77	Decreasing
Sos	76	Indeterminate
Asb	76	Decreasing
Mls	75	Increasing
Gol	71	Decreasing
Xcb	70	Increasing
Fmb	62	Indeterminate

1: Trend over time; see text for discussion
2: Important variable in more than one model (predict bank/phase/time frame)

Table 3.Ten most important variables for predicting time frame

Cranial measurement	Score	Trend[1]
Occ	100	Indeterminate
Avr	91	Indeterminate
Frs	88	Increasing
Sts	78	Decreasing
Mdb	78	Decreasing
Frf	77	Indeterminate
Nlb	77	Indeterminate
Jub	76	Increasing
Frc	71	Indeterminate
Iml	71	Indeterminate

1: Trend over time; see text for discussion

Table 4. Ten most important variables for predicting phase.

Cranial measurement	Score
Mdb[1]	100
Prr	75
Ocs	72
Wmh	71
Nar	62
Nph	62
Wcb	61
Osr	61
Lar	61
Pac	60

1: Important variable in more than one model (predict bank/phase/time frame)
Note that trend is not included since there were not enough representative measurements from different groups

Table 5: Grouping of time spans

Period	Midpoint used for graphing
1612-1688	1650
1700-1708	1704
1725-1728	1726
1766-1777	1766
1817	1817

Figures

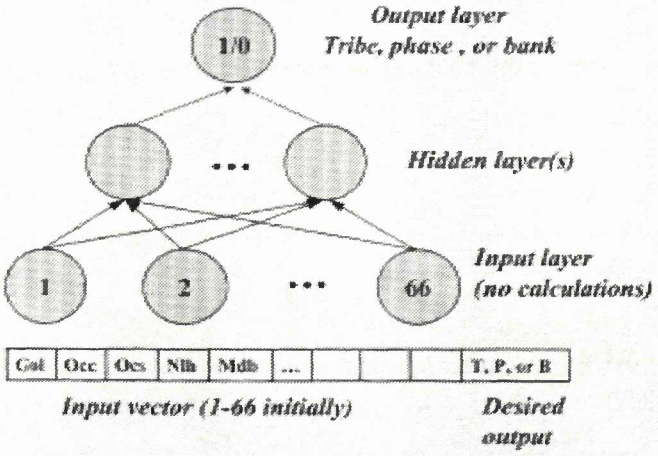

Figure 1. Back propagation neural network.

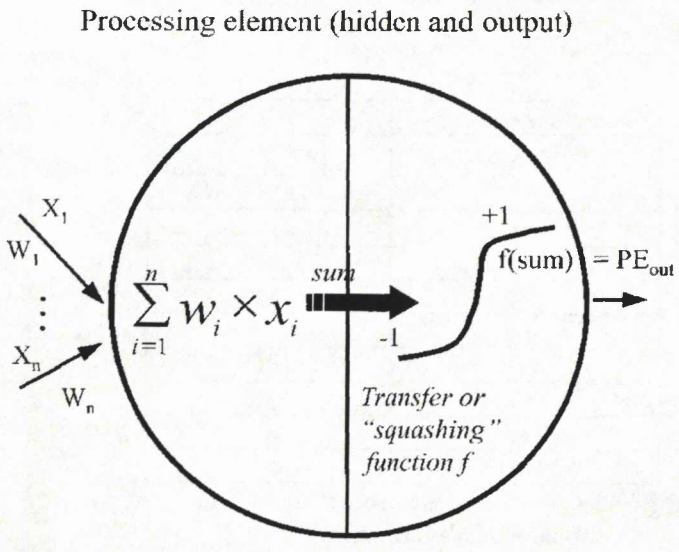

Figure 2. Processing elements, hidden and output layers.

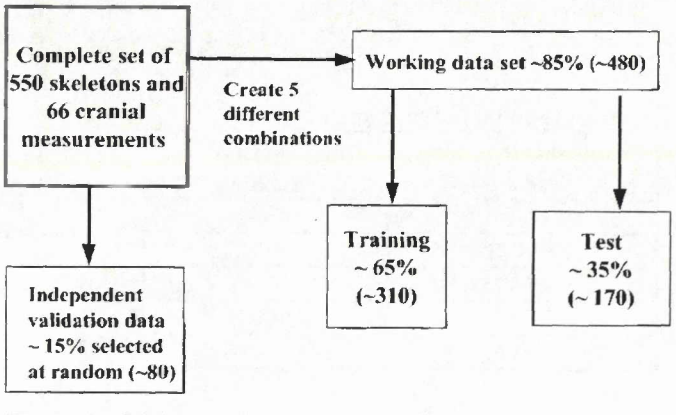

Figure 3. Division of data for network training and validation.

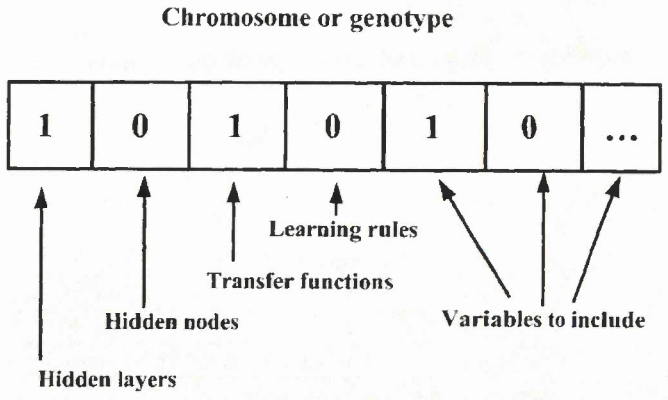

Figure 4. "Chromosome" used for genetic algorithms.

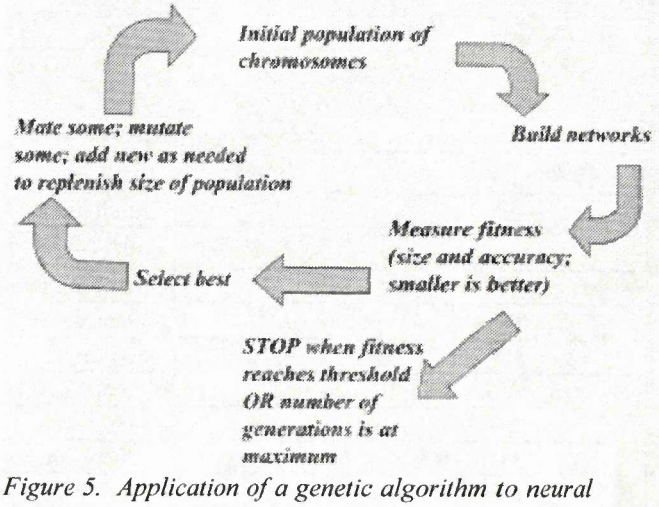

Figure 5. Application of a genetic algorithm to neural network optimization

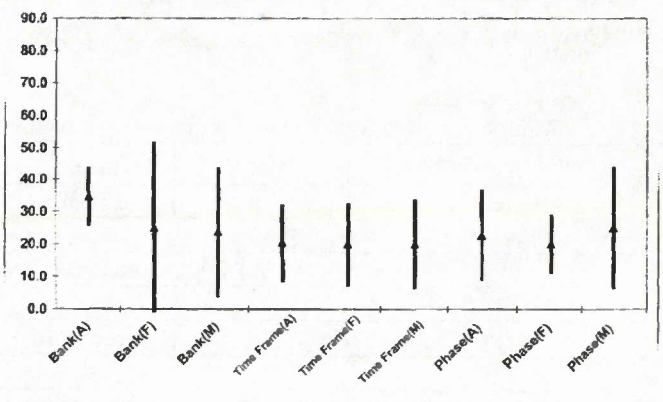

Figure 6. Summary of results, archaeological database, 0.90 threshold. From left to right, results are reported for models that predict bank (left vs. right); time frame (Arikara vs. EC); and Phase (Le Beau 3 vs. all others). Data is reported for all data (A), females only (f), and males only (m).

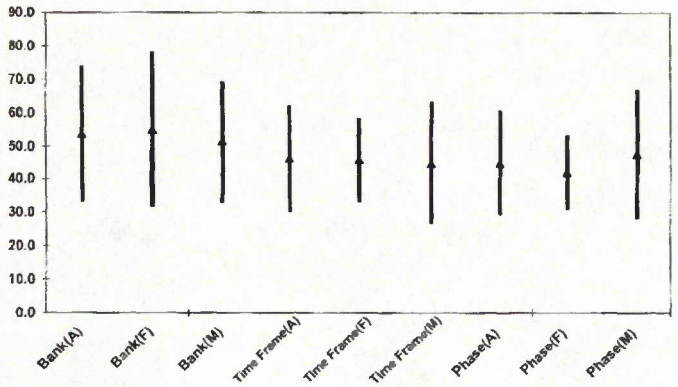

Figure 7. Summary of results, archaeological database, 0.75 threshold. Labels as in Figure 6.

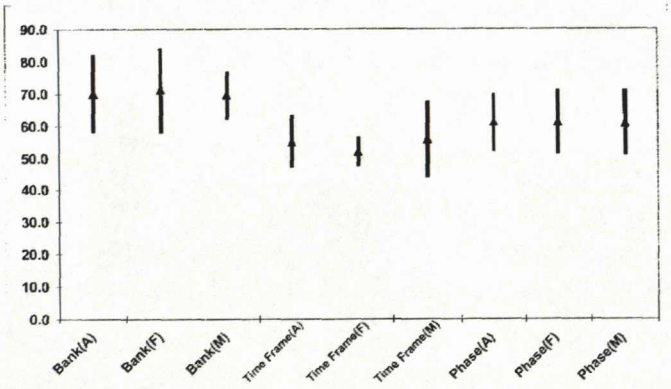

Figure 8. Summary of results, archaeological database, 0.60 threshold. Labels as in Figure 6.

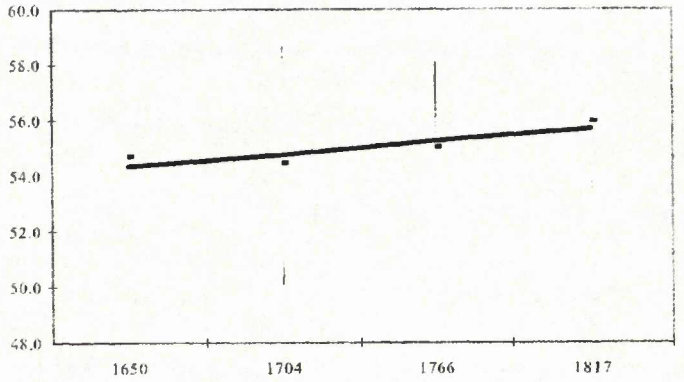

Figure 9. Example variable plot, Nlh (males only, right bank), predicting bank model. Note that the trend is generally increasing, although with the large uncertainties (+/-1 standard deviation), such observations are not conclusive. The x-axis is the midpoint date as presented in Table 5 and the y axis is the mean cranial measurement for that time period. The line is the best fit regression through the average.

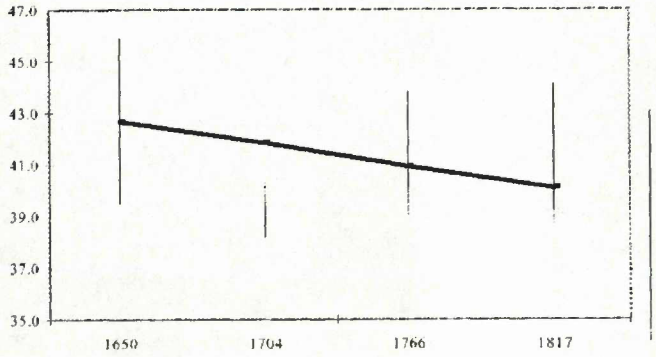

Figure 10. Example variable plot, Ocr (males only, right bank), predicting bank model. This is an example of a variable showing a general decrease. The same plot for the corresponding females (not shown) also shows a general decrease but it is not as uniform or smooth.

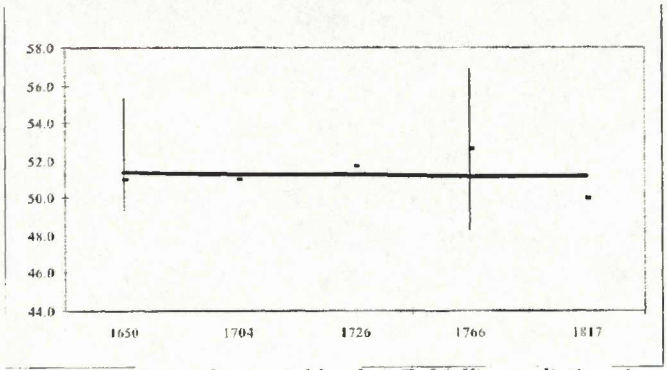

Figure 11. Examples variable plot, Frf (all), predicting time frame model. This is an example of a trend that was labeled as indeterminate.