

# **Model-Based Clustering of World Craniometric Variation**

by Dienekes Pontikos

## **Abstract**

Model-based clustering is applied to 2,504 crania of 28 populations of recent *Homo sapiens* using 57 cranial metric variates. This technique uses no a priori knowledge about the population affiliation of each skull. Model-based clustering varies the number and form of the clusters and selects a “good” model, showing a balance of data fit and parsimony, using the Bayes Information Criterion. Fourteen separate clusters were identified in the best run, each of which corresponds strongly to either one of the original populations, or to a racial group. It is shown that cranial variation can be used to infer ethno-racial affiliation.

Recently, Rosenberg et al. (1) applied model-based clustering to 377 autosomal microsatellite loci in 1,056 individuals from 52 world populations. Using this approach, and without any prior information about the population affiliation of each individual, they were able to infer several clusters corresponding to the major continental regions, thus proving the geographical stratification of human genetic variation. In this article I apply model-based clustering to 2,504 crania from 28 populations based on 57 metric traits to investigate whether the geographic structure of human biodiversity can also be inferred by the methods of physical anthropology.

Traditional works on human variation depended on ad-hoc racial typology (2, 3). The availability of abundant computing power and the development of multivariate statistics has made it possible to search for patterns in human variation, especially the skeletal variation of ancient populations in a systematic manner. Howells' work (4, 5, 6) has been particularly important in this regard, demonstrating the power of the new quantitative approach. I revisit Howell's dataset, applying a powerful statistical technique which was not previously available, and has not -to my knowledge- been applied to this material before.

## **MATERIALS & METHODS**

The 2,504 skulls used in this paper are part of Howells' world craniometric database and belong to 28 original populations. Several measurements are available for each skull; I have used the 57 traits that were also used in (5). The 1,348 male and 1,156 female skulls were pooled to increase the sample size. The source for the data was the original dataset made available by Howells (7) which was obtained from an online source (8).

To remove sexual dimorphism sex-specific  $Z$ -scores were obtained for each skull by subtracting the sex-specific average for each measurement and dividing by the sex-specific standard deviation. This made male and female measurements directly comparable by eliminating size and variability differences between the sexes.

Model-based clustering (9) is a powerful statistical technique which fits a mixture model with  $k$  multivariate normal components, varying  $k$  and the structure of the covariance matrices of the components. Six different covariance structures are each evaluated over a choice of  $k$  ranging from 1 to 16. Thus, in total 96 different models are evaluated. To balance parsimony with data fit the Bayes Information Criterion (BIC), a measure of clustering quality is reported for each of these models. Informally, as  $k$  increases so does data fit, but parsimony is reduced. Using the BIC, the goodness of the subdivision of the 2,504 skulls into  $k$  groups can be assessed.

For each model, each skull is assigned a  $k$ -long vector of membership probabilities for the  $k$  different clusters. No prior information about its population affiliation is used in this analysis. A posteriori the correspondence of clusters with population groups is assessed by calculating the average probability of membership in each cluster for all the skulls of each of the 28 original populations.

All experiments were run in the R statistical computing system (10) using the adaptation of the mclust software (11) for this platform. Population codes are given in Table 1.

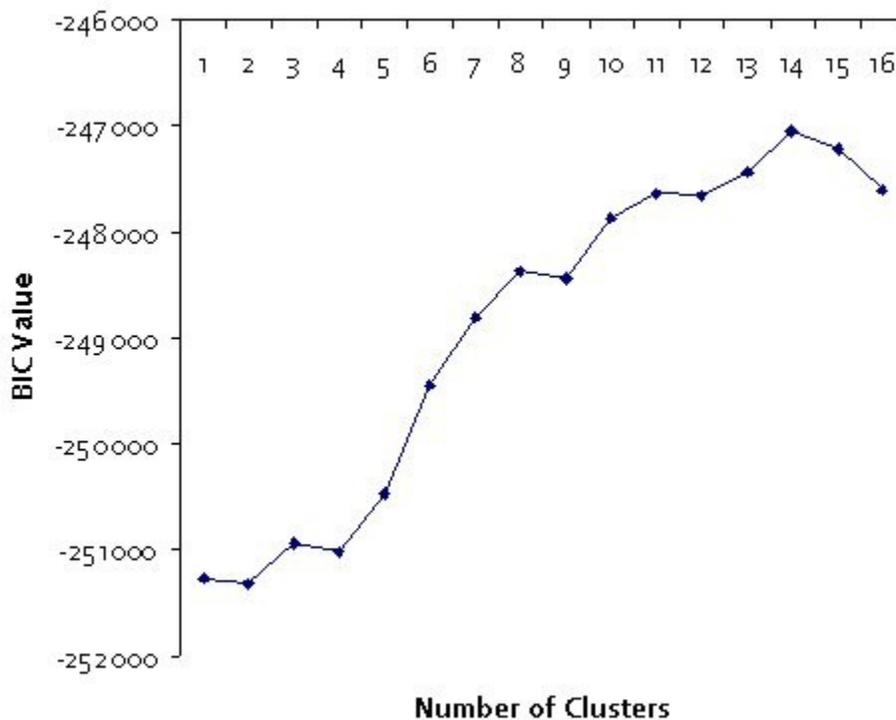
1. Norse: (Medieval), Europe, Oslo
2. Zalavar: Central Europe, Hungary
3. Berg: Central Europe, Carinthia, Austria
4. Teita: East Africa, Kenya
5. Dogon: West Africa, Mali
6. Zulu: South Africa
7. Lake Alexandrina Tribes: South Australia
8. Tasmanian: Tasmania
9. Tolai: Melanesia, New Britain
10. Mokapu: Oahu, Hawaii, Polynesia
11. Easter Island: Polynesia
12. Moriori: Chatham Islands, Polynesia
13. Arikara: (Early) North America
14. Santa Cruz Island: California, N. America
15. Yauyos: Peru, South America
16. Hokkaido: North Japan
17. North Kyushu: South Japan

18. Hainan: Haikou City, China
19. Atayal: Taiwan Aborigines
20. Phillipine: Phillipine Islands
21. Guam: Latte Period
22. Egypt: Gizeh, 26th-30th Dynasties
23. San: South Africa
24. Andaman Islands: Andaman Islands
25. Ainu: S. and SE. Hokkaido, Japan
26. Buriat: Siberia
27. Eskimo: Inugsuk, Greenland
28. Anyang: Shang Dynasty, China

**Table 1.** Numerical codes for 28 populations.

## RESULTS

For all values of  $k$  the model with ellipsoidal components with equal covariance matrices (“EEE”) maximized the BIC. In Figure 1 the BIC value is plotted as a function of  $k$ . With higher values of  $k > 16$ , the BIC steadily decreases, indicating that these more detailed models are not very well supported by the available data.



**Figure 1.** Bayes Information Criterion with varying  $k$ .

The model with k=3 is the first to show an improvement over the assignment of all skulls in a single component (k=1). Greater successive improvements are shown for k=5, 6, 7, 8 and subsequently the BIC is maximized for k=14.

A posteriori inspection of the membership coefficients of the 28 populations in each of the k cluster indicates that the correspondence between the two is astounding, with most skulls of a population belonging to the same cluster, and each cluster including populations of known “racial” affinity. In Table 2 these coefficients (in %) are shown for the optimal value of k=14.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1.	95	1	1						1	1	1			
2.	90	2	3				2	2	1	1				
3.	89	3							3	1		1	1	1
4.	4		1				2	1		91		1	1	
5.			2							96			2	
6.	2		3				4	1		87		3		
7.	1					2	96							
8.	2			3		75	17			2				
9.			2				95			2	1			
10.			4	92							4			
11.			1	98										
12.		1	1	5					1		93			
13.	13	54	13	1					11		5		2	
14.	3	3	1						94					
15.	2	92	2						3				1	
16.	1	1	94						1	1			1	
17.	2	1	96							1				
18.		1	96		1					1				
19.	1		88					1		5		1	3	
20.	6	2	81	2						6			2	
21.	1		93	4							1			
22.	94		1					1		4			1	
23.	1		1							8		89		
24.	3		4							1		3	89	
25.	3		15	2				77		2				
26.	2		2											96
27.					99									
28.	1	2	97											

**Table 2.** Membership coefficients for 28 populations in k=14 clusters.

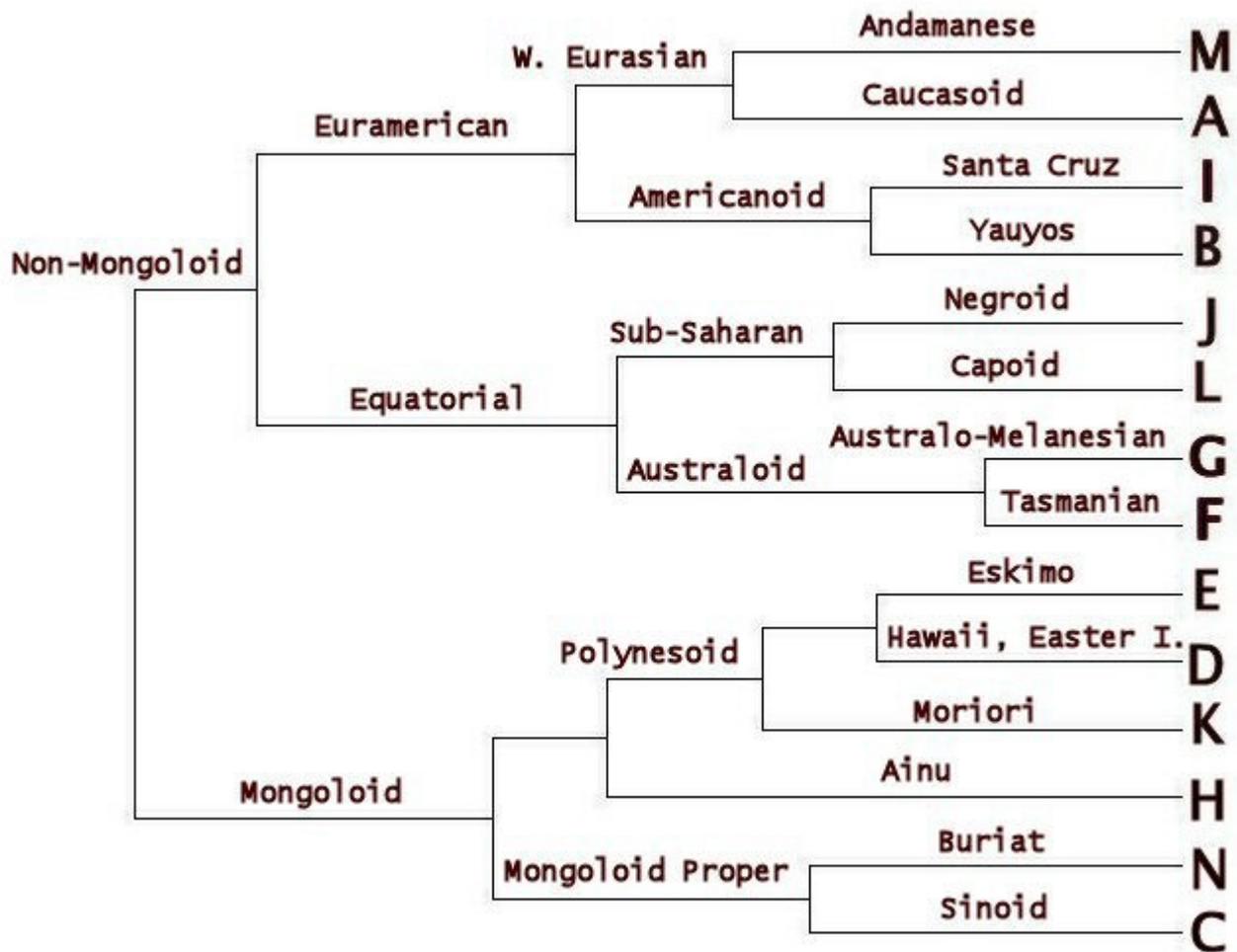
Some of the inferred clusters, e.g., cluster E, correspond closely to an individual population, in this case with the Eskimos of Greenland. It is striking that such population-specific clusters were identified even though the algorithm makes no use of population affiliations, indicating that some populations are craniometrically distinctive, with all their members sharing a high degree of within-group similarity.

In other cases, e.g., cluster A, correspond to a “racial” group of traditional anthropology, in this case with the

Caucasoid race, encompassing Ancient Egyptians from Gizeh as well as medieval European populations from Oslo, Berg and Zalavar.

Inspection of clusters for both lower and higher values of k are similar in their coherence, indicating that human craniometric variation is hierarchical and that it can be mapped both to macro-groups (“races”) as well as finer-grained subdivisions.

To gain some further insight into the morphometric similarity of the k=14 inferred clusters I calculated the Pearson distances between the cluster means, defined as 1 minus the Pearson correlation coefficient between two vectors. An Unweighted Pair Group Method with Arithmetic Mean (UPGMA) dendrogram of the distance matrix, is shown in Figure 2.



**Figure 2.** Clustering of Pearson Distances for k=14 clusters.

Cluster branches were assigned descriptive labels. For example, cluster A which encompasses Caucasoids is linked with cluster M which is specific to the Andamanese, corresponding to a “West Eurasian” cluster despite

apparent phenotypical differences between Caucasoids and the Andamanese. The main distinction in this dendrogram is between mainly Mongoloid and non-Mongoloid populations, broadly similar to results obtained by (1). It should be noted however that this dendrogram is a summary of phenetic distances and must not be interpreted phylogenetically.

## **DISCUSSION**

According to Livingstone (12) “there are no races, there are only clines.” Howells (6) complemented this emphatic pronouncement by declaring that “there are no races, there are only populations.” He arrived at this conclusion by clustering individual skulls, using a Euclidean-distance iterative algorithm, and failing to see the emergence of clear racial clusters.

The inability to infer meaningful clusters can be due either to (i) the non-existence of such clusters, or (ii) to the inability of the clustering algorithm to discover them. The results of this article show that racial clusters can be recognized using a powerful model-based technique. So, even though both the “cline” and “population” concepts capture the global distribution of craniometric variation, the “race” concept can still be justified as representing a group of phenetically similar populations.

## **REFERENCES**

1. Rosenberg N.A., Pritchard J.K., Weber J.L., Cann H.M., Kidd K.K., Zhivotovsky L.A., Feldman M.W. (2002) *Science* 298, 2381–2385.
2. Deniker, J. (1926) *Les races et les peuples de la terre*, 2 ed. (Masson et Cie, Paris).
3. Eickstedt, E. von (1934) *Rassenkunde und Rassengeschichte der Menschheit*, (Ferdinand Enke, Stuttgart).
4. Howells, W.W. (1973) *Cranial Variation in Man: A Study by Multivariate Analysis of Patterns of Difference Among Recent Human Populations*, Papers of the Peabody Museum of Archaeology and Ethnology (Harvard Univ. Press, Cambridge, MA), Vol. 67.
5. Howells, W.W. (1989) *Skull Shapes and the Map: Craniometric Analyses in the Dispersion of Modern Homo*, Papers of the Peabody Museum of Archaeology and Ethnology (Harvard Univ. Press, Cambridge, MA), Vol. 79.
6. Howells, W.W. (1995) *Who’s Who in Skulls: Ethnic Identification of Crania from Measurements*, Papers of the Peabody Museum of Archaeology and Ethnology (Harvard Univ. Press, Cambridge, MA), Vol. 82.
7. Howells, W.W. (1996) *Am. J. Phys. Anthropol.* 101, 441–442.
8. L.W. Konigsberg’s Homepage, URL: <http://konig.la.utk.edu/howells.htm>.
9. Fraley C., Raftery A.E. (2002) *J Am Stat Assoc.* 97, 611–631.

10. Model-Based Clustering Software (MCLUST / EMCLUST), URL: <http://www.stat.washington.edu/fraley/mclust/>
11. R Development Core Team (2003). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL: <http://www.R-project.org>.
12. Livingstone, F.B. (1962) Curr. Anthropol. 3, 279-281.

**The citation of this paper is:**

Pontikos, D. Model-Based Clustering of World Craniometric Variation. September 2004.

<http://dienekes.angeltowns.net/articles/anthropologica/clustering.html>

<http://dienekes.angeltowns.net/articles/anthropologica/clustering.pdf>