Colapinto Giulia, Pedol Enrico, Quadrelli Emanuele

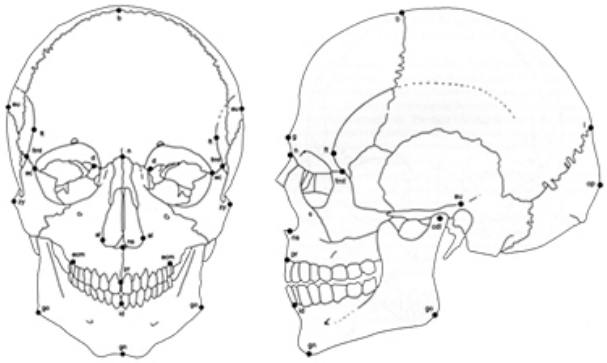
January 10, 2026

# CRANIAL CLUSTERS

## INTRODUCTION

The research activity conducted focused on identifying and characterizing demographic subgroups within a large morphometric database. The primary objective was to examine how biological and geographical variations influence the structure of craniometric data.

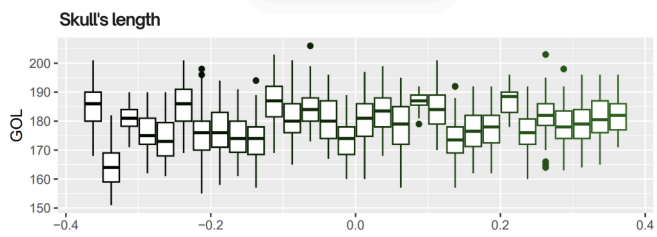
The dataset under analysis features high dimensionality, with 82 specific craniometric measurements recorded for each statistical unit, along with labels for the sex and population of origin of each individual. Additionally, the evaluation set (Test-set) includes supplementary variables (such as the recovery site and descriptive notes) which, while providing archaeological or anthropological context, were excluded from the predictive models as they were considered non-influential for the purposes of morphometric classification.



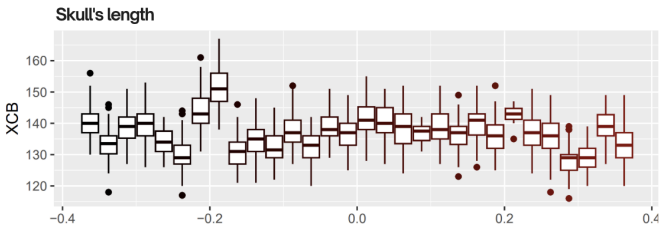
## DESCRIPTIVE ANALYSIS

Preliminary data exploration confirmed that the categorical variables **Sex** and **Population** represent the primary factors of variability within the dataset.

### Influence of Population

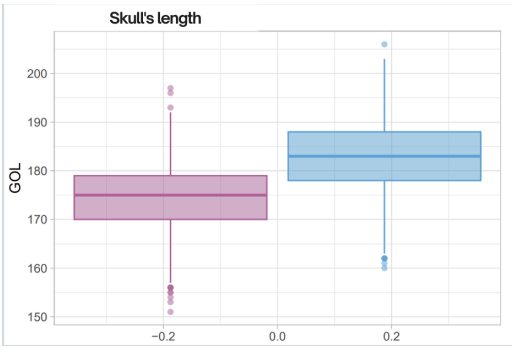



Graphical analyses indicate that geographical origin significantly impacts fundamental parameters such as maximum skull length (GOL) and maximum skull width (XCB).




However, given the complexity of discriminating between 28 distinct populations with a sample of approximately 2,500 observations, the analysis focused primarily on sexual differentiation to ensure greater statistical robustness.

## Sexual Dimorphism




In all major dimensions recorded, a clear dimorphism emerges: male skulls, on average, exhibit larger dimensions than female skulls.

This dimensional discrepancy, clearly visualized through box plots of key variables, provides the foundation for the development of subsequent classification and clustering models.

## DATA CLEANSING

First, variables with a large number of 0 observations (such as BSA, SLA, etc.) were eliminated. This is likely because synostosis often occurs in the sutures of elderly skulls: the sutures fuse completely and the bone becomes smooth, making measurement impossible.

The data were then centered, not globally using all observations, but grouped by population to which they belong, so that the potentially unique characteristics of each population were not lost.

For the various analyses, i.e. supervised and unsupervised classification, different combinations were subsequently selected, to explore multiple methodologies and verify whether variable selection was essential, or whether different combinations could obtain virtually identical results.

## SUPERVISED CLASSIFICATION (DISCRIMINANT ANALYSIS)

The analysis was conducted separately on the training and test sets, after data selection and centering. The classification model chosen was: **V-Fold Cross Validation**.

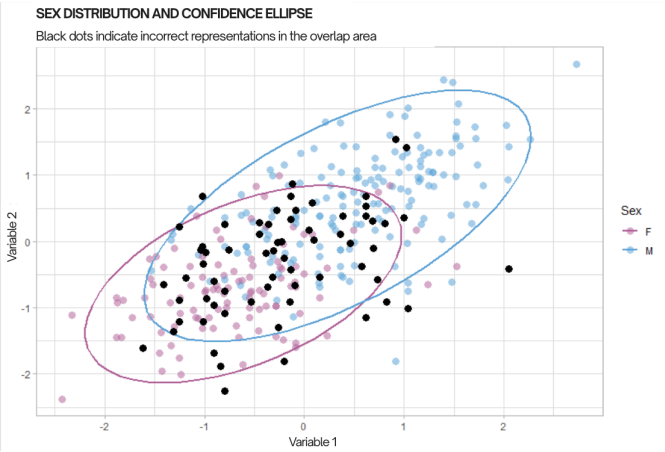
## V-FOLD CROSS VALIDATION CON 18 VARIABILI

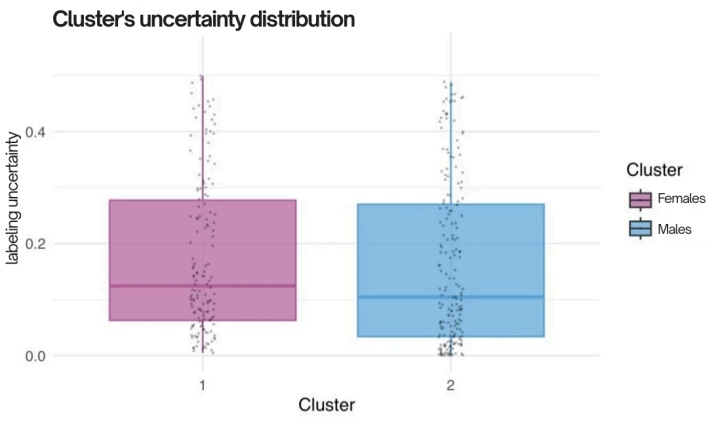
The first step taken was the selection of the variables: through the analysis of the loadings only on the first principal component. The following were identified: **18 variables** that contribute most to the variance of the dataset (including NAR, GOL, JUB, ZYB, and BNL). This selection

allowed us to reduce the dimensionality without losing large amounts of information and maintaining the effectiveness of the model.

The choice to test different segmentations (**V = 5, 8, 10, 12, 15**) is used to map the stability of sexual dimorphism in the sample. By dividing the training dataset into V blocks, the script uses V-1 blocks in rotation for training and the remaining block for validation. During the entire model estimation process, the accuracy value is between **80% and 83%** in the test set, with CV values around 0.125. The maximum average model accuracy was found with a division into 10 blocks. Only models with free proportions were evaluated within the script given their best fit to the data. The script highlights a competition between three main Gaussian mixture models:

- Gaussian_pk_Lk_C (**WATER**) e Gaussian_pk_L_D_Ak_D (**EVE**): Models better at maximizing test set accuracy;
- Gaussian_pk_L_C (**EEE**): Best model in minimizing the training test CV.




## V-FOLD CROSS VALIDATION CON PCA SCORES

In this phase a procedure similar to the previous analysis was used but using **scores of the first two principal components**. Despite the drastic reduction in the two principal components (PC1 and PC2), the classification accuracy in the test set remains stable at around 82%.

To verify the robustness of the model and to avoid that the results are influenced by a particular subdivision of the data, the V-Fold Cross Validation process was performed with a variable number of blocks: **10, 20, 60, 100**.

The results appear constant regardless of the number of partitions set: the prediction accuracy in the test set does not undergo significant fluctuations, stabilizing around **80-83%.** While for the CV of the training set the average value is **0,145**. The predominant models are:

- Gaussian_pk_L_D_Ak_D (**EVE**): It appears more frequently, especially in the 10- and 20-block partitions, in line with the results of the previous model;
- Gaussian_pk_L_C (**EEE**): Significantly present in 60-block partitions;

- Gaussian_pk_Lk_Dk_A_Dk (EEV): Although less common as a “best CV criterion model”, this model often emerges when seeking maximum accuracy in 60-block and 100-block partitions.

In conclusion, in both methods used the results are very similar:

- The chosen model is *Gaussian_pk_L_D_Ak_D (EVE)* or alternatively *Gaussian_pk_Lk_C (VEE)*;
- *Accuracy* is around the *82%*;
- The *CV* is between *0.12 and 0.15*.

The resulting model can be considered sufficiently accurate in predicting the test set data.

**UNSUPERVISED CLUSTERING (K-MEANS)**

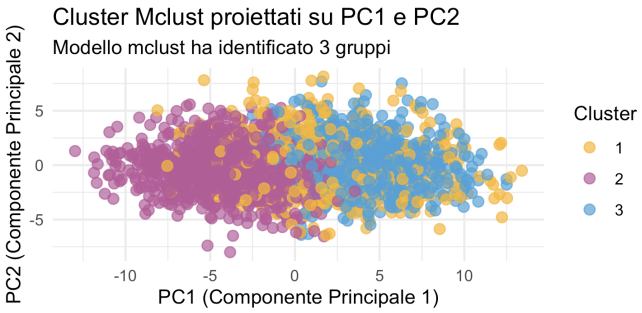
K-means clustering performs poorly, with an Adjusted Rand Index (ARI) of 0.32 and an error of 21%. This is due to the significant overlap between the two groups, which confirms how geographic origin or individual genetic variability can generate similarities between subjects of different sexes, complicating the separation between the two groups.

**MCLUST WITH PCA SELECTION**

Once again, for the choice of variables on which to apply the model of *clustering,* principal components were used. It was decided to select the ten variables with the highest weights in the first principal component (those with *loadings* greater than 0.15).

Once selected, the variables were inserted into a tibble used in the function *mclust* without imposing any constraints. The model identified in the absence of restrictions classifies the observations into three different groups, for which, however, no clear discriminating characteristic can be found. Looking at the statistical units, it is assumed that the model correctly identifies skulls belonging to the female and male genders; however, it appears uncertain regarding the extreme observations of the two sexes, which are not classified into either of the two main groups.

Having taken note of this, it was decided to impose on *mclust* the two-group limit. The result is that the best model is “VEE” (*ellipsoidal, equal shape and orientation*), which has a low entropy; through a short Monte Carlo simulation it was found that*error rate*between 13.5% and 14%. Although the results are remarkable, the procedure used was not considered satisfactory, since without imposing constraints the optimal model found three unidentifiable groups.

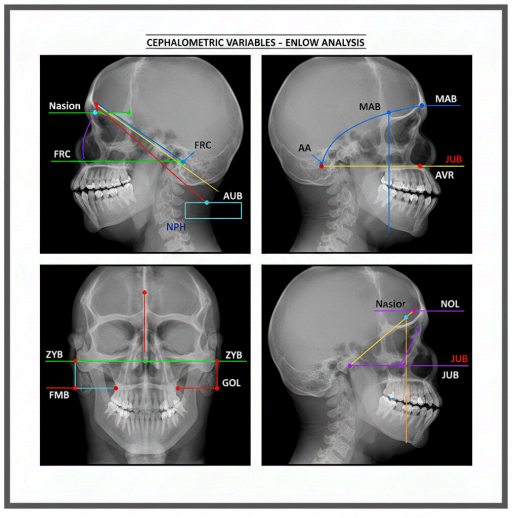



## MCLUST WITH STOCHASTIC FEATURE SELECTION

As a solution to the problem encountered, It was decided to apply a random variable selection method, called *Stochastic feature subset selection.* Once again, the variables with the *loadings are higher* than the first principal component. It was decided to initially select twenty variables and exclude those highly correlated with those already present.

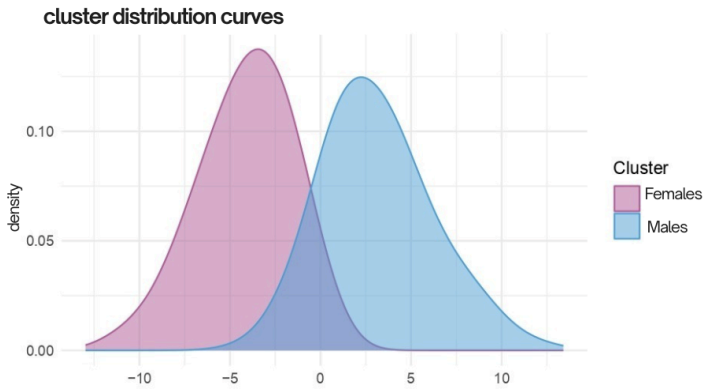
Once the starting set was defined, a simulation was built that randomly extracted ten of the initial variables and, with that sample, executed the function six times *mclust,* storing in a list the name of the variables used, the best model identified and the average of the relative *error rate* on six simulations.

The result indicated the best model as "*VVE*", obtained using the variables "*MAB", "FRC", "AUB", "GOL", "NOL", "AVR", "JUB", "ZYB", "NPH"* and *"FMB"*. This time the model found recognizes the two groups without the need to impose further constraints. Carrying out a further verification using the Monte Carlo method, it emerges that the *error rate* is between 12.9% and 13.4%.

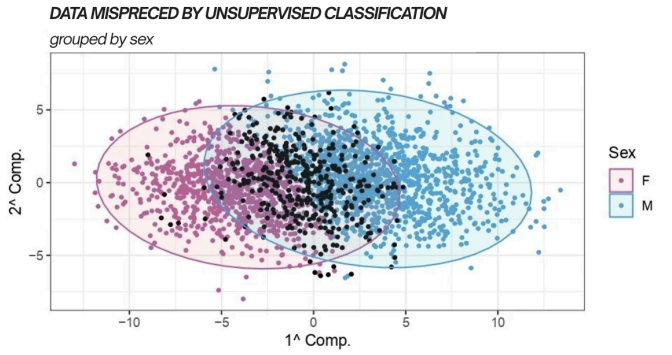


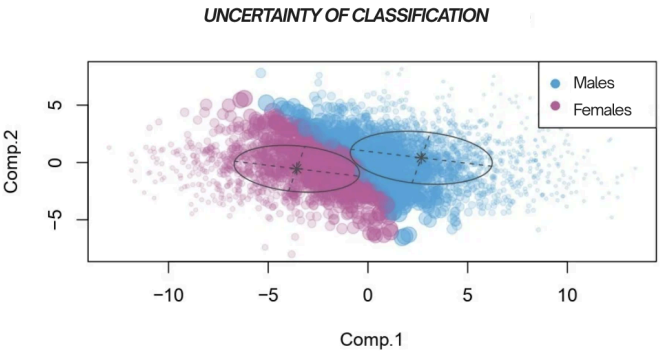
Further confirmation comes from the accuracy calculation of around 87% and from the Goodness-of-fit indices *$R^2$ Determinant and $R^2$ Trace,* which are 0.31 and 0.57 respectively. The Kullback-Leibler divergence was calculated, yielding a value of 5.84, which is very satisfactory.

Finally, the model was applied to predict new observations from the dataset *Test.*After matching the variables with those used in the model, the predict function was used. The results are satisfactory, as, out of 303 observations, the model correctly predicted 252 with an accuracy of 83%.




Ultimately, the resulting model interprets well the biological differences between male and female skulls, with physiological errors due to the natural individual morphological variability present in individuals, and demonstrates excellent predictive capacity based on new observations.




Works cited

Howells, W.W. (1973) Cranial variation in man: A study by multivariate analysis of patterns difference among recent human populations.

Cambridge, MA: Peabody Museum of Archeology and Ethnology, Harvard University