Annex 3. PCA Dimensionality and Clustering Diagnostics

A3.1 Determination of Principal Components

Method. Scree plot of eigenvalues to choose dimensionality.
Result. PC1+PC2 explain ∼56% of variance (example: 36.9% + 19.1%). The elbow (scree) plot of within-cluster sum of squares (WCSS) showed an inflection at k = 3, selected as optimal.
Interpretation. Additional components contributed marginal variance (<10%), confirming that PC1–PC2 sufficiently represent the nutrient space for visualization and clustering. **T**hree groups offered the best balance between compactness and interpretability, aligning with biological expectations of lean, oily, and outlier species.

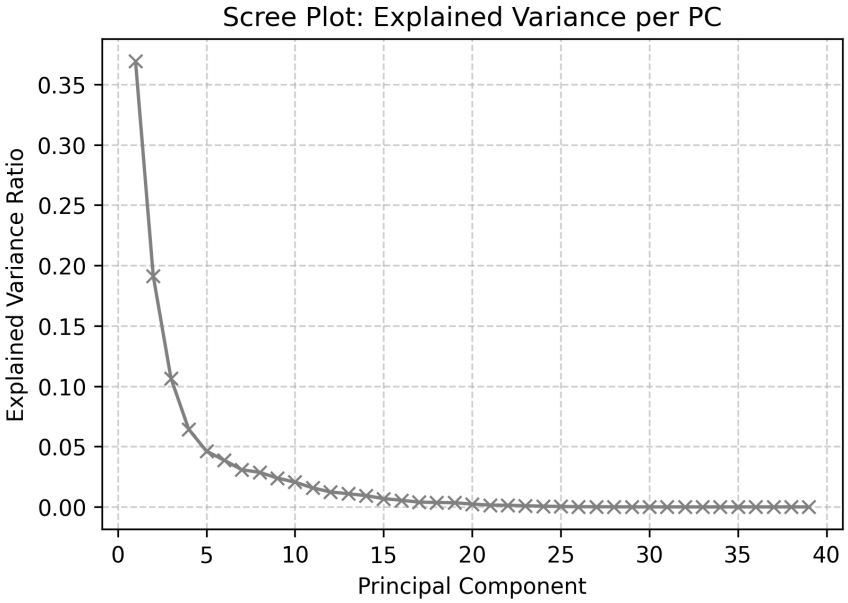
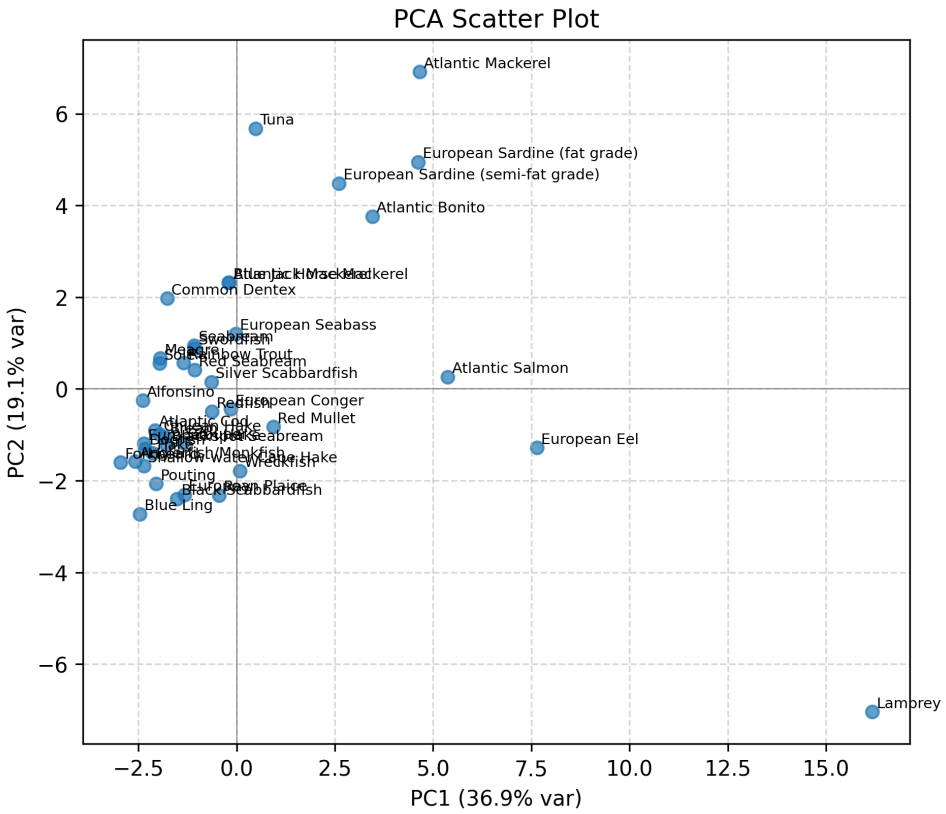


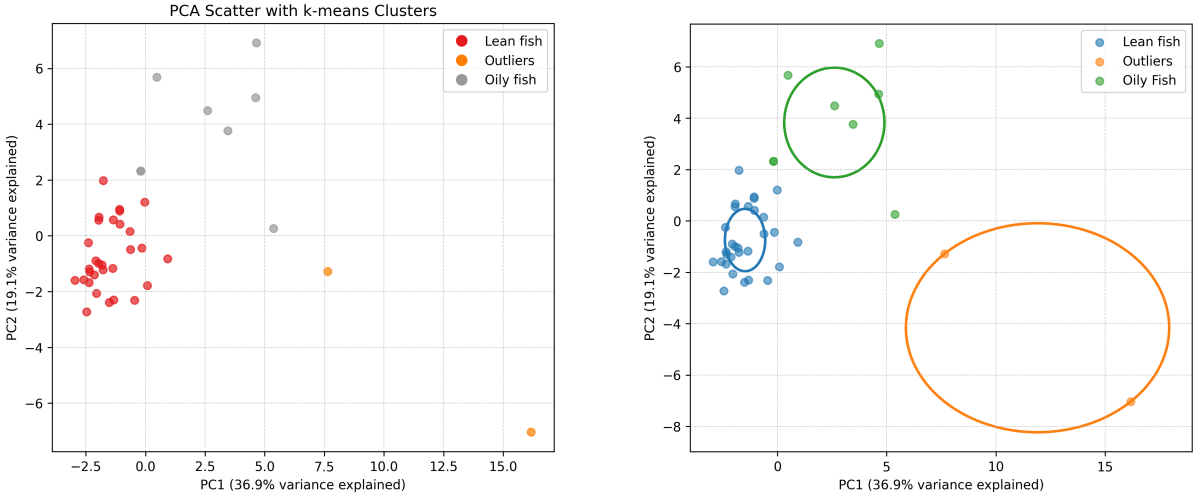
Figure 1: Explained variance ratio by principal component (scree plot).

A3.2 Visual Assessment of Species Distribution

Plots. (i) PC1 vs PC2 scatter; (ii) same with k-means colors ($k = 3$); (iii) clusters with 1 SD ellipses.



(a) PC1–PC2 scores (all species).



(b) Scores colored by k=3 clusters.

(c) Cluster ellipses (1 SD) on PC1–PC2.

Figure 2: PCA score-space diagnostics.

A3.3 Nutrient Loadings

Method. Loadings extracted from PCA on standardized nutrients.
Criterion. Report top contributors by absolute loading (e.g., $|\ell| \geq 0.40$).
PC1 ("fatness-energy axis"). Total lipids, SFA, MUFA, PUFA, energy (kcal), moisture (negative).
PC2 ("micronutrient axis"). Vitamin B12, niacin equivalents, vitamin B6, phosphorus.

Table 1: Top nutrient loadings for PC1 and PC2.

PC1 ("fatness–energy axis")			PC2 ("micronutrient axis")		
Rank	Nutrient	Loading	Rank	Nutrient	Loading
1	Energy [kcal]	0.9554	1	Niancin Eq	0.8563
2	Energy [kJ]	0.9551	2	Niacin	0.8316
3	Lipids	0.9529	3	Vitamin B6	0.7847
4	Water	0.9509	4	Vitamin B12	0.7201
5	SFA	0.8960	5	Phosphorus	0.6478
6	MUFA	0.8955	6	trans Fatty Acids	0.6421
7	PUFA	0.8774	7	Tryptophan	0.6325
8	Riboflavin	0.8667	8	Salt	0.5964

A3.5 Notes

For the raw data used int his analysis please see Annex 2. The PCA was performed on standardized nutrient data per 100 g edible portion to identify the principal axes of compositional variation. K-means clustering was then applied to group species into nutritionally coherent categories without relying on arbitrary cut-offs, ensuring consistency with established nutritional terminology. The analytical matrix comprised energy, macronutrients, lipid fractions (SFA, MUFA, PUFA, trans fatty acids, cholesterol), selected vitamins (A, D, E, B-group, C, folate, tryptophan), and minerals (Na, K, Ca, P, Mg, Fe, Zn). EPA and DHA were excluded, as their values originated from external databases (Norwegian, Dutch, U.S.), potentially introducing methodological heterogeneity.

Software: All analyses and visualizations were conducted in JupyterLab (Python 3.11). Data handling and computation were performed using pandas and numpy. Principal component analysis (PCA) and k-means clustering were implemented with scikit-learn (modules: decomposition.PCA, preprocessing.StandardScaler, cluster.KMeans). Figures were produced using matplotlib.pyplot, with graphical elements generated through matplotlib.patches.Ellipse and matplotlib.lines.