Zooplankton Classification in Ontario Lakes Using Environmental and Geometric Features

STA2453

Daihao Wu

Department of Statistical Sciences
University of Toronto

March 2025

Abstract

Zooplankton play a fundamental role in aquatic ecosystems, serving as indicators of water quality and contributing to the food web. This study explores the classification of zooplankton species using a combination of geometric image features and environmental factors. Our dataset, collected from Lake Huron and Lake Simcoe, consists of thousands of observations mapped to species labels, physical attributes, and environmental metadata. Through exploratory data analysis (EDA), we identified key features influencing species differentiation and discussed the implications for classification modeling. Given the extreme class imbalance and feature correlations observed, we propose methods to improve prediction accuracy, including feature selection, class rebalancing, and machine learning-based classification. The findings of this study aim to support biodiversity monitoring and ecosystem management efforts.

Contents

1 Introduction **3**

2 Data **3**

2.1 Data Preprocessing 3

2.2 Zooplankton Data Overview 4

3 Methodology **6**

3.1 Multiclass Logistic Regression (Baseline) 6

3.2 Random Forest 6

3.3 XGBoost 7

4 Results **7**

4.1 Evaluation Metrics 7

4.2 Final Result 8

5 Discussion **9**

A Data Overview **11**

A.1 List of All Zooplankton Classes 11

A.2 Selected Variables from Environmental/Geometric Dataset 11

A.3 Count Specific for Each Zooplankton Class 13

1 Introduction

Zooplankton are microscopic aquatic organisms that serve as primary food sources for fish and other aquatic life. Their distribution and population dynamics are influenced by environmental conditions such as temperature, water depth, and nutrient levels (The Land Between, 2019). Accurately classifying different zooplankton species is crucial for ecological studies, as certain species act as indicators of water quality and ecosystem health. However, manual classification is labor-intensive and prone to errors, necessitating the development of automated classification techniques.

This study targets the seven specific zooplankton species: *Calanoid_1, Bosmina_1, Herpacticoida, Chironomid, Chydoridae, and Daphnia* out of 32 classes (Refer to Appendix A.1 for list of classes). By leveraging a dataset containing both environmental descriptors and geometric properties extracted from image data, we aim to build predictive models capable of distinguishing between these species. This research explores which features contribute most to classification accuracy and how data preprocessing and modeling strategies can be optimized to enhance performance. We explored the Logistic Regression model as our baseline and tree-based models to capture more complex relationships between features.

2 Data

The dataset used in this study consists of zooplankton samples collected from various aquatic sites in Ontario, specifically from Lake Huron and Lake Simcoe. The dataset contains three primary components: (1) a Master Table File, which records metadata about each sampling event, including location, environmental conditions, and links to corresponding image data; (2) CSV files containing geometric properties of zooplankton particles, extracted from mosaic images using automated image analysis techniques; and (3) Mosaic TIF images, which include raw image data of zooplankton specimens.

Each zooplankton observation is associated with both environmental factors (such as water temperature, depth, precipitation, and distance from shore) and geometric features (including area, aspect ratio, circularity, and perimeter). This multi-source dataset allows for the integration of both contextual and morphological characteristics to improve classification performance. Given the dataset's complexity and volume, careful preprocessing is required to ensure consistency and enhance model performance.

2.1 Data Preprocessing

For the analysis and modelling in our research, the dataset undergoes several preprocessing steps to ensure data quality and usability.

The first step involved identifying and extracting the relevant zooplankton observations from two separate folders, HURONOvlerap_csv and SIMC.Overlap_csv, which contained numerous CSV files. Using the master table file, we identified the corresponding 474 CSV files that were valid for analysis. Any CSV files that were not listed in the master table were excluded from further processing to maintain dataset integrity.

After filtering the relevant files, the next step involved selecting key variables for analysis. The selected features included environmental factors such as water temperature (WaterT), average depth (AvgDepth), and distance from shore (Distshore), as well as geometric properties of zooplankton such as aspect ratio, circularity, and perimeter. These features were extracted from each CSV file and standardized to ensure uniform column structures across all datasets. (Detailed descriptions of the selected key variables are listed in Appendix A.2)

Once the key variables were selected, the extraction of geometric and environmental data was next. This was done by matching the Filename column in the processed CSV files with the csvfile column in the master table. The master table contained additional environmental metadata, including sampling location, date, and environmental conditions. A successful merge ensured that each observation retained both its geometric properties and the corresponding ecological factors recorded during data collection.

To address the severe class imbalance across different species of zooplankton, we applied a class balancing step (Mancuso, 2022). Specifically, we set an upper limit of 100,000 observations per class to prevent overrepresentation by dominant species. Furthermore, for classes other than the 7 target species with fewer than 2,500 observations, we grouped them into a single category labeled `other` to preserve analytical significance while limiting sparsity. The class labeled `too small`, which generally consisted of poorly resolved or noise-prone observations, was entirely removed from the dataset to improve signal clarity and overall model performance.

2.2 Zooplankton Data Overview

To understand the structure and potential patterns in the dataset, exploratory data analysis (EDA) is conducted, focusing on both class distribution and feature relationships.

There are in total $400,955$ observations in our final merged dataset from the 474 CSV files. However, the majority of the Zooplankton species fall into class Floc_1, LargeZ1, Calanoid_1 and Cyclopoid_1. (Refer to Appendix A.3 for specific counts for each class)

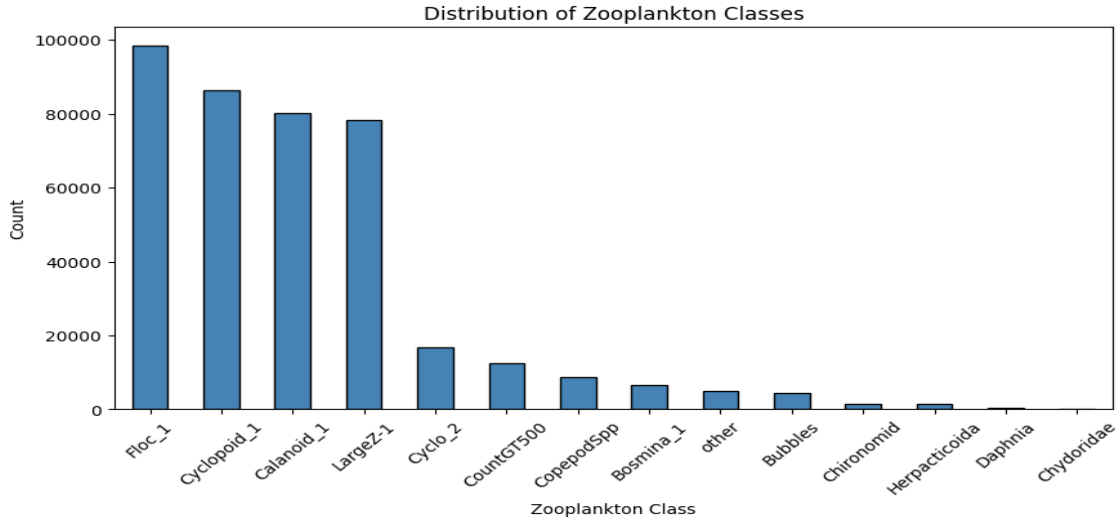


Figure 1: Distribution of Zooplankton Classes

Figure 1 shows the distribution of zooplankton species in the dataset. The extreme imbalance in class representation is evident, with Calanoid_1 comprising the majority of samples, while species like Chydoridae and Daphnia have significantly fewer observations. This imbalance suggests the use of oversampling techniques or class-weighted learning methods to prevent the model from being biased toward dominant species.

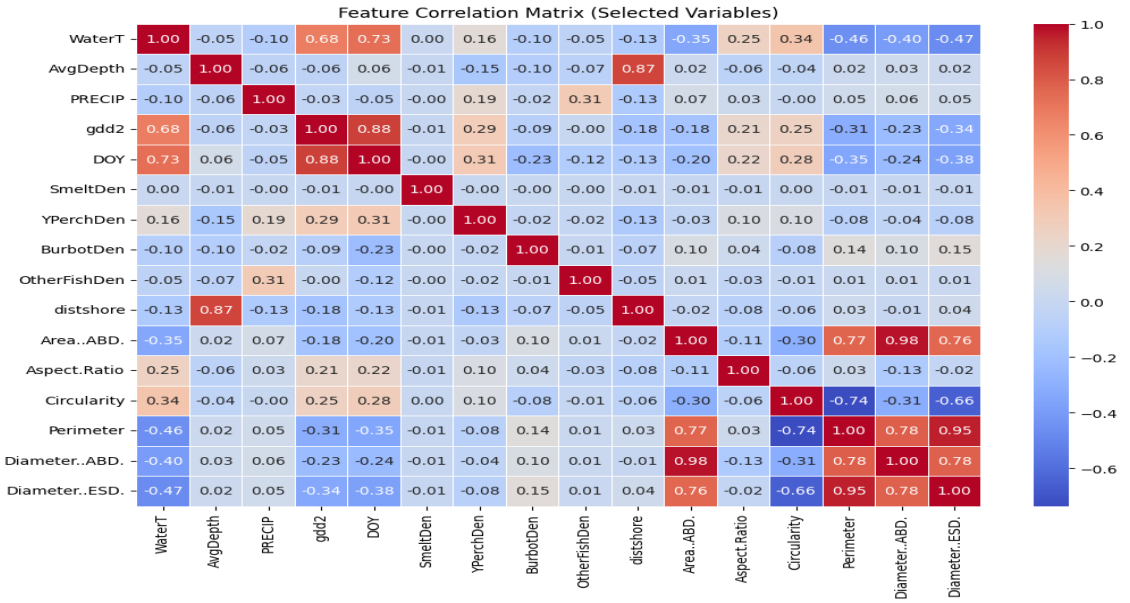
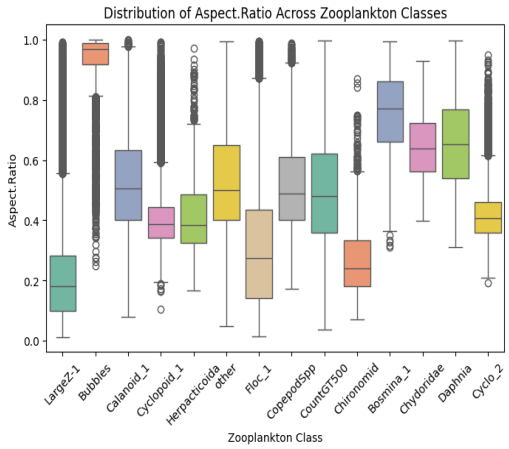


Figure 2: Correlation Heatmap

A correlation heatmap (Figure 2) highlights strong relationships between key variables. Water temperature and DOY exhibit a strong positive correlation (0.73), suggesting that seasonal effects influence temperature variation. Depth and distance from shore also show a high correlation (0.87), indicating that deeper water samples tend to be further offshore.



(a) Boxplot of Aspect Ratio Property Across Classes

(b) Boxplot of Circularity Property Across Classes

Figure 3: Boxplots of Geometric Information of the TIF Images

Boxplots of zooplankton geometric properties across different classes (Figures 3) reveal impor-

tant shape and size differences among species. Aspect Ratio and Circularity exhibit significant variation across classes, suggesting that shape-based features are highly discriminative. For instance, Bosmina_1 and Chydoridae have high aspect ratios (0.8-1.0), indicating elongated forms, while Chironomid has much lower aspect ratios (0.2-0.5), reflecting a rounder morphology. Similarly, Circularity differs significantly across species, further supporting the inclusion of these features in classification models.

Overall, This highlights key differences in both environmental conditions and morphological characteristics across species, guiding feature selection and model development. The insights gained inform subsequent modeling steps, ensuring that the classification framework is built on meaningful and well-preprocessed data.

3 Methodology

To classify the zooplankton species based on both environmental and geometric features, we employed three supervised machine learning models: multiclass logistic regression, random forest, and XGBoost. These models were selected to represent a spectrum of algorithmic complexity and learning strategies, allowing for both interpretability and performance benchmarking. Logistic regression served as a simple, interpretable baseline; random forest introduced ensemble-based decision-tree modeling with inherent feature selection and robustness; and XGBoost provided a state-of-the-art gradient boosting framework that is well-suited for structured tabular data.

3.1 Multiclass Logistic Regression (Baseline)

We implemented a multinomial logistic regression model as our baseline due to its simplicity and interpretability. It assumes a linear relationship between input features and the log-odds of each class, which allows us to understand the directional influence of each predictor (scikit-learn developers, 2024b). Given the high dimensionality of our feature space, particularly due to the inclusion of geometric descriptors, we applied Principal Component Analysis (PCA) to reduce dimensionality before fitting the model. PCA helped mitigate multicollinearity and reduced the risk of overfitting by projecting features into a lower-dimensional space while preserving most of the variance.

3.2 Random Forest

Random forest was selected as a non-linear, ensemble-based classifier that is robust to noise and outliers, and does not require feature scaling (scikit-learn developers, 2024a). It works by constructing a multitude of decision trees and aggregating their outputs through majority voting. To properly handle the categorical variables `Loc` and `SITE`, we applied one-hot encoding to convert these features into a set of binary indicator variables. This encoding ensures that the random forest model does not impose any ordinal relationship between the categories, which could happen if integer encoding were used instead. Preserving the categorical nature of these variables without introducing artificial ordering helps the model to learn class-specific patterns more effectively.

We tested the random forest model under two settings: with and without the application of SMOTE (Synthetic Minority Oversampling Technique). SMOTE was applied after the train-test split to address class imbalance by generating synthetic examples of minority classes. Without

SMOTE, the model tended to underperform in underrepresented classes. By including the SMOTE-resampled variant, we aimed to evaluate the trade-off between model complexity, training time, and classification performance on rare classes.

3.3 XGBoost

XGBoost was employed as a high-performance gradient boosting algorithm that improves upon traditional decision-tree ensembles by optimizing both accuracy and computational efficiency (XGBoost developers, 2023). It is particularly effective for structured, tabular data with mixed feature types. Like the random forest model, XGBoost was trained and evaluated in two versions: with and without SMOTE applied post train-test split. The SMOTE-enhanced model demonstrated improved recall and F1-scores for minority classes while maintaining high overall accuracy. We also utilized label encoding for categorical features (such as site and location) to ensure compatibility with the resampling and modeling pipeline.

4 Results

4.1 Evaluation Metrics

To assess the performance of the zooplankton classification models, a combination of evaluation metrics was used to account for both overall accuracy and the challenges posed by class imbalance. Given that certain species are significantly underrepresented in the dataset, traditional accuracy alone would be insufficient in determining the true effectiveness of the model. Therefore, weighted classification metrics were prioritized to ensure fair evaluation across all species (Brownlee, 2020).

The accuracy metric provides a simple measure of overall model correctness by computing the proportion of correctly classified instances out of the total number of predictions. However, in an imbalanced dataset, accuracy alone can be misleading, as it may be artificially high if the model simply predicts the most frequent species more often. To provide a more nuanced evaluation, we also compute macro-averaged and weighted F1-scores.

The macro F1-score is the unweighted mean of the F1-scores for each species and is computed as:

$$\text{Macro-F1} = \frac{1}{K} \sum_{i=1}^{K} F1_i$$

where K represents the number of classes (i.e., 14 in this case). This metric treats all species equally, regardless of their frequency in the dataset, making it particularly useful for measuring performance across both dominant and minority species.

Conversely, the weighted F1-score accounts for class imbalance by weighting each species' F1-score based on its occurrence in the dataset:

$$\text{Weighted-F1} = \sum_{i=1}^{K} w_i \times F1_i$$

where w_i is the proportion of samples belonging to species i. This ensures that common species contribute more to the overall score while still incorporating performance on rare species. A high weighted F1-score indicates that the model maintains strong predictive capability across all species while still prioritizing accuracy for the most frequently occurring classes.

4.2 Final Result

To evaluate the performance of the three classification models, we compared their prediction accuracy, macro-averaged F1 scores, and weighted F1 scores using the held-out testing dataset. The results are summarized in Figure 4.

The logistic regression model, used as a baseline, achieved an accuracy of 81.7%, a macro F1

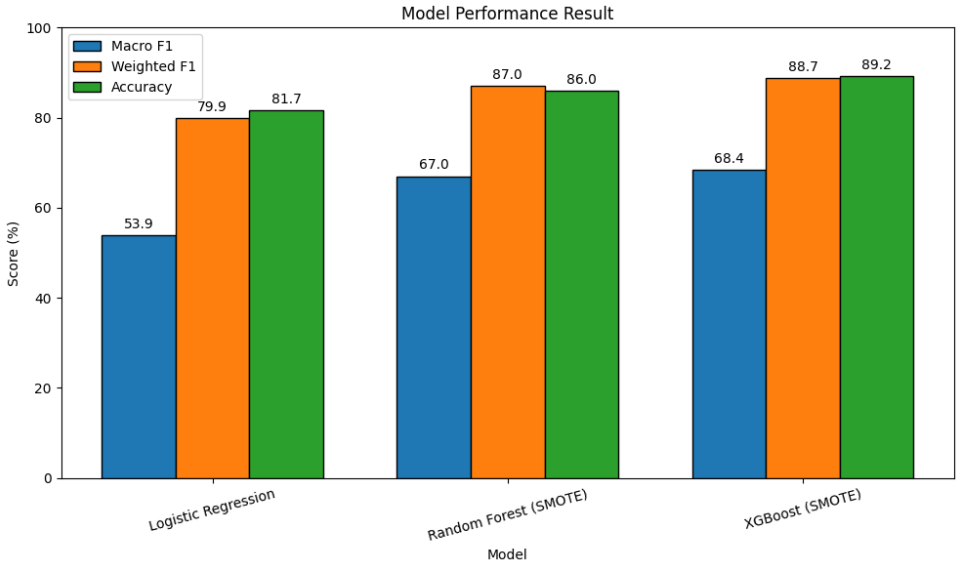


Figure 4: Comparison of model performance metrics: accuracy, macro F1, and weighted F1 for each classification method.

score of 53.9%, and a weighted F1 score of 79.9%. While its overall accuracy is reasonably high, the lower macro F1 score indicates poor performance on the minority classes. This result reflects the model's limited capacity to capture the nonlinear relationships and class imbalance inherent in the data.

The random forest model, trained with SMOTE to address class imbalance, significantly outperformed the logistic regression baseline. It achieved an accuracy of 86.0%, a macro F1 score of 67.0%, and a weighted F1 score of 87.0%. The substantial improvement in macro F1 suggests enhanced sensitivity to underrepresented classes while maintaining strong overall predictive performance.

The XGBoost model, also trained with SMOTE, delivered the best performance among all models. It achieved an accuracy of 89.2%, a macro F1 score of 68.4%, and a weighted F1 score of 88.7%. The high macro F1 and weighted F1 scores demonstrate the model's robustness and its ability to generalize across both frequent and rare zooplankton classes.

Overall, these results highlight the importance of using ensemble methods and class balancing strategies for complex, imbalanced ecological datasets. XGBoost with SMOTE proves to be the most effective method for zooplankton classification in our study, striking a strong balance between accuracy and fairness across classes.

5 Discussion

While our proposed approach demonstrates strong overall performance in classifying zooplankton species based on environmental and geometric features, several limitations remain that warrant further consideration.

One notable limitation lies in the comparatively lower macro-averaged F1 scores across all models, particularly in the logistic regression baseline. Although ensemble models like Random Forest and XGBoost achieved high accuracy and weighted F1 scores, the macro F1 scores indicate that the models still struggle to perform equally well across all classes—especially the rare ones. Since macro F1 gives equal weight to each class regardless of frequency, the lower values suggest that rare zooplankton species are often misclassified. This poses a challenge for ecological monitoring, where less abundant species may be of particular interest due to their sensitivity to environmental changes.

Additionally, certain limitations exist in our data preprocessing pipeline. Although SMOTE was applied to improve class balance in the training data, it is constrained by the small number of samples in some underrepresented classes, which limits its effectiveness in generating realistic synthetic examples. Moreover, noise-prone classes such as *TooSmall* were removed to enhance data quality, but this step also introduces a trade-off by potentially discarding biologically meaningful information. The heavy reliance on engineered features extracted from CSV files also limits the model's ability to capture nuanced morphological details that may be present in the original image data.

Looking ahead, a promising direction for future work involves incorporating raw image data (e.g., TIFF mosaic files) into the classification pipeline. By combining image-based features using convolutional neural networks (CNNs) with the structured environmental and geometric features currently used, a hybrid model could be developed to better capture both spatial and contextual information. This would allow for a more comprehensive and potentially more accurate classification of zooplankton species. Furthermore, integrating temporal patterns and seasonal trends using time-aware models could enhance the ecological interpretability of the predictions.

In summary, while this study demonstrates the feasibility of using machine learning to classify zooplankton from structured data, future advancements in multimodal modeling and data enrichment will be key to improving classification fairness and ecological relevance.

References

Brownlee, J. (2020). *A tour of evaluation metrics for imbalanced classification* [Accessed: 2024-04-17]. https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/

Mancuso, W. (2022). *Class imbalance strategies: A visual guide with code* [Accessed: 2024-04-17]. https://medium.com/data-science/class-imbalance-strategies-a-visual-guide-with-code-8bc8fae71e1a

scikit-learn developers. (2024a). Sklearn.ensemble.randomforestclassifier [Accessed: 2024-04-17].

scikit-learn developers. (2024b). Sklearn.linear_model.logisticregression [Accessed: 2024-04-17].

The Land Between. (2019). *Zooplankton – small but mighty* [Accessed: 2024-04-17]. https://www.thelandbetween.ca/2019/08/zooplankton-small-but-mighty/

XGBoost developers. (2023). *Xgboost documentation* [Accessed: 2024-04-17]. XGBoost. https://xgboost.readthedocs.io/en/release_3.0.0/

Appendix

A Data Overview

A.1 List of All Zooplankton Classes

Class	Count	Class	Count
Bubbles	5,440	Calanoid_1	248,390
CopepodSpp	9,721	CountGT500	14,691
Cyclopoid_1	197,893	Floc_1	300,000
Herpacticoida	1,739	LargeZ-1	81,325
Nauplii	2,209	TooSmall	300,000
Unknown	1,265	Bosmina_1	7,225
Chironomid	1,540	Nematode	30
Eggs	103	Naididae	36
InsectLarvae	164	Chydoridae	107
CladoceraSpp	883	Sididae	477
Rotifer	8	Daphnia	560
Holopedidae	84	Insecta	57
Polyphemidae	3	Leptodoridae	1
Floc_2	101	Trombidiforme	1
Holopididae	12	Cercopagididae	5
Cyclo_2	19,881	Holopediidae	5

Table 1: Zooplankton Class Distribution (Raw Counts)

A.2 Selected Variables from Environmental/Geometric Dataset

The selected variables from Environmental data are as follows:

- **LAT0, LAT1 (Latitude Bounds)** – Represent the minimum and maximum latitude coordinates of the sampling trawl, used to approximate the north-south spatial extent of data collection.
- **LON0, LON1 (Longitude Bounds)** – Represent the minimum and maximum longitude coordinates of the sampling trawl, indicating the east-west extent of the sampling transect.
- **XANGLE (Net Angle)** – Describes the angle of the net or camera relative to the horizontal plane during sampling, which can influence the quality and orientation of captured images or sample precision.
- **PRECIP (Precipitation)** – Measures the amount of rainfall (mm) on the sampling date, which may impact water quality, nutrient levels, and zooplankton populations.
- **XWAVEHT (Wave Height)** – Indicates the average wave height (m) at the sampling site, potentially affecting the stability of the water column and zooplankton vertical distribution.
- **WIND (Wind Speed and Direction)** – Captures the wind direction (degrees) and speed (knots) at the time of sampling, both of which can influence surface mixing, circulation patterns, and zooplankton dispersion.

- **CLOUD_PC (Cloud Cover)** – Reports the percentage of cloud cover during sampling, potentially related to light availability and photosynthetic activity, indirectly affecting zooplankton behavior.
- **WaterT (Water Temperature)** – Represents the recorded water temperature (°C) at the time of sample collection, influencing zooplankton distribution and metabolic activity.
- **AvgDepth (Average Depth)** – The average depth (m) at which the sample was collected, affecting species composition as different zooplankton thrive at varying depths.
- **PRECIP (Precipitation)** – Measures the amount of rainfall (mm) on the sampling date, which may impact water quality, nutrient levels, and zooplankton populations.
- **gdd2 (Growing Degree Days)** – A cumulative temperature-based metric representing thermal accumulation over time, which can indicate seasonal growth conditions affecting zooplankton life cycles.
- **DOY (Day of Year)** – The numeric representation of the sampling date within a calendar year, used to track seasonal changes in zooplankton abundance and diversity.
- **LOC** - The location of Zooplankton Collection site (Location : FISHI, LSIMC, NOTTA, BLIND, FATFI)
- **distshore (Distance from Shore)** – The measured distance (m) from the sampling site to the nearest shoreline, which may correlate with habitat type and species distribution.

The selected variables from Geometric data are as follows:

- **Area..ABD. (Zooplankton Area - ABD)** – The measured surface area (in pixels) of individual zooplankton particles in the image, providing an indicator of species size.
- **Aspect.Ratio (Aspect Ratio)** – The ratio of the length to width of a zooplankton particle, distinguishing elongated species from more compact ones.
- **Circularity** – A shape descriptor that quantifies how close a zooplankton particle is to a perfect circle, helping differentiate between rounded and irregularly shaped species.
- **Perimeter** – The boundary length of a zooplankton particle, useful for analyzing species morphology and distinguishing between species with complex or simple shapes.
- **Compactness** – Indicates the compactness of the zooplankton particle by comparing area to perimeter, helping to distinguish denser-bodied species.
- **Convexity** – Represents the ratio between the convex hull perimeter and the actual perimeter, reflecting the smoothness or irregularity of the organism's shape.
- **Elongation** – Measures the aspect ratio of the zooplankton's bounding box, used to characterize long or stretched organisms versus rounder ones.
- **Intensity** – Refers to the average pixel intensity of the zooplankton image, often related to its internal structure or optical density.

- **Sigma.Intensity** – The standard deviation of pixel intensities, indicating texture or internal variation in the organism's body.
- **Roughness** – Quantifies the surface irregularity of the particle contour, useful in distinguishing between smooth and spiny species.
- **Transparency** – Measures the optical transparency of the organism, which may correlate with species identity or developmental stage.
- **Diameter..ABD. (Diameter - ABD)** – The estimated diameter of the zooplankton particle based on area measurements, often used as a proxy for organism size.
- **Diameter..ESD. (Equivalent Spherical Diameter - ESD)** – A standardized measure of particle size that approximates the diameter if the zooplankton were a perfect sphere, providing a comparable size metric across species.

Other dataset property related variables such as Image.File and particle.ID columns are omitted in the above list for simplicity.

A.3 Count Specific for Each Zooplankton Class

Class	Count
Floc_1	98,476
Cyclopoid_1	86,409
Calanoid_1	80,125
LargeZ-1	78,309
Cyclo_2	16,673
CountGT500	12,490
CopepodSpp	8,814
Bosmina_1	6,623
other	5,025
Bubbles	4,522
Chironomid	1,522
Herpacticoida	1,468
Daphnia	404
Chydoridae	95

Table 2: Zooplankton Class Counts After Preprocessing