Exploratory Data Analysis for Zooplankton

Daihao Wu

February 10, 2024

Introduction:

Zooplankton are microscopic aquatic organisms that play a crucial role in freshwater ecosystems, serving as primary food source in the food web and acting as indicators of water quality and environmental health. These organisms are highly sensitive to changes in water conditions, making them valuable for monitoring ecological conditions. Different species of zooplankton exhibit unique characteristics, which can be influenced by factors such as water temperature, nutrient availability, and habitat conditions.

Lakes and other freshwater bodies, including **Lake Huron** and **Lake Simcoe**, support diverse zooplankton populations. However, variations in environmental conditions can affect species distribution and abundance over time. The classification of zooplankton species based on their physical and environmental characteristics is essential for understanding aquatic ecosystem dynamics. Despite their importance, accurately distinguishing between different zooplankton species remains a challenge due to overlapping features and excessive human labor.

Our project aims to classify six specific species of zooplankton—*Calanoid_1, Bosmina_1, Herpacticoida, Chironomid, Chydoridae, and Daphnia*— using a combination of image data and environmental descriptors. By analyzing the dataset, we seek to identify key factors influencing zooplankton classification, assess potential data imbalances, and examine how environmental variables contribute to species differentiation. The findings from this study could provide insights into aquatic biodiversity monitoring and ecosystem management strategies.

Research Question:

How can zooplankton species be accurately classified based on their image features and environmental/geometric properties, and what key factors influence their distribution across different freshwater sites?

Data:

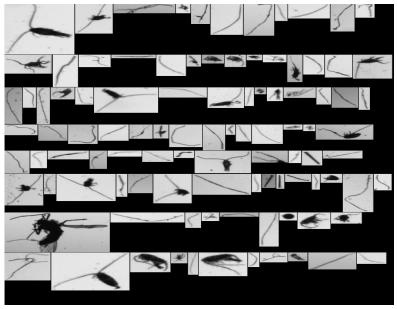


Image 1: Sample Mosaic tif Image of Zooplankton

tifffile	csvfile	Loc	DOY	WaterT	LAT0	LON1	avgdepth t
04072021_Huron_6_	20210407_FISHI_00	FISHI	97	6.1	44.983	-81.3877	5.25
04072021_Huron_1C	20210407_FISHI_01	FISHI	97	6.4	44.97633	-81.3758	2.95
04072021_Huron_1C	20210407_FISHI_01	FISHI	97	6.4	44.97633	-81.3758	2.95
04072021_Huron_1C	20210407_FISHI_01	FISHI	97	6.4	44.97633	-81.3758	2.95
04072021_Huron_6_	20210407_FISHI_00	FISHI	97	6.1	44.983	-81.3877	5.25
04072021_Huron_6_	20210407_FISHI_00	FISHI	97	6.1	44.983	-81.3877	5.25
04072021_Huron_6_	20210407_FISHI_00	FISHI	97	6.1	44.983	-81.3877	5.25
04142021_Huron_0C	20210414_NOTTA_(	NOTTA	104	5	44.552	-80.2812	5.65
04142021_Huron_0C	20210414_NOTTA_(	NOTTA	104	5	44.552	-80.2812	5.65
04142021_Huron_0C	20210414_NOTTA_(	NOTTA	104	5	44.552	-80.2812	5.65

Table 1: Master File with environment factors and mapping of tif image, CSV files

Class.Parti	Class	Area..ABD	Area..Fille	Aspect.Rat	Calibration	Calibration	Camera
1	Floc_1	56491.12	56491.12	0.1141	9.47	1	1
2	Floc_1	102380.1	103182.4	0.3057	9.47	1	1
3	Floc_1	38690.74	38690.74	0.2089	9.47	1	1
4	Floc_1	139744.8	139744.8	0.1841	9.47	1	1
5	Floc_1	97413.38	97413.38	0.3896	9.47	1	1
1	TooSmall	34496.92	34496.92	0.7204	9.47	1	1
2	TooSmall	59514.53	59514.53	0.8574	9.47	1	1
6	Floc_1	45735.07	45735.07	0.2239	9.47	1	1
3	TooSmall	39594.41	39669.79	0.4115	9.47	1	1
4	TooSmall	106073.3	106073.3	0.5574	9.47	1	1

Table 2: CSV file with tif Particle Property

Our dataset mainly consists of three main data types as follows: *'MasterTable_AI_FlowCAM.xlsx'*, *'HURONOvlerap_csv'* and *'SIMC.Overlap_csv'*, and 'Mosaic TIF' Images.

Master Table File:

'MasterTable_AI_FlowCAM.xlsx'

The master table file recorded the general information regarding data collection factors such as location, date, wind speed, surface water temperature for a specific Mosaic observation. This file also maps the corresponding Mosaic tif image to each sub csv file that contains factors and information about a specific Mosaic tif image particle properties such as area, aspect ratio etc.

CSV File with Particle Property:

'HURONOvlerap_csv' & *'SIMC.Overlap_csv'*

The above two folders contain numerous csv files that the master table file maps to. Huron folder represents that the Mosaic data from this folder is collected in Lake Huron. On the other hand, SIMC folder represents that the data is collected in Lake Simcoe. However, these csv files contain similar information such as the labels/classes, image cropping properties and extra geometric information about each Zooplankton in the Mosaic image.

Data Preprocessing:

To ensure data consistency and proper alignment of environmental and geometric features, several preprocessing steps were applied to the dataset. 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 CSV file names 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), precipitation (Precip), and distance from shore (Distshore), as well as geometric properties of zooplankton such as area, aspect ratio, circularity, and perimeter. These features were extracted from each individual CSV file and standardized to ensure uniform column structures across all datasets.

Once the key variables were selected, the next step was merging the extracted geometric and environmental data. 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 environmental factors recorded during data collection.

Data Analysis:

This report analyzes key features in the dataset to assess their potential impact on zooplankton classification. The dataset consists of environmental factors such as temperature, depth, and precipitation, along with geometric features describing the zooplankton species, including area, aspect ratio, and circularity. The goal is to identify patterns and insights that can guide the development of a predictive model.

Table 1: Summary Statistics of Features

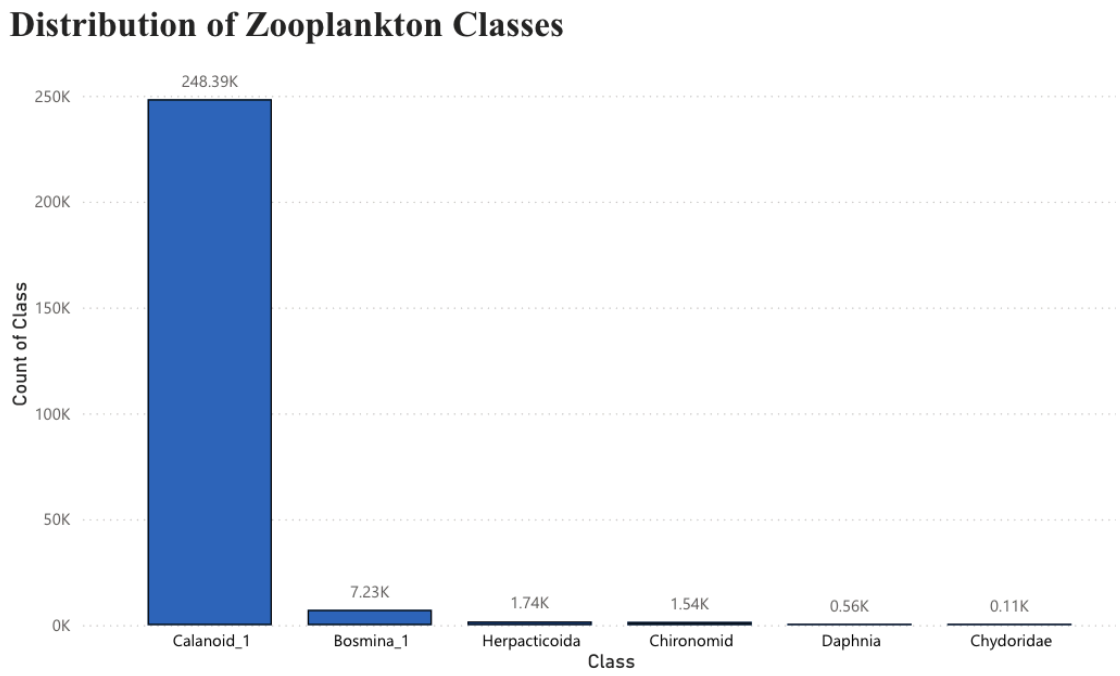
Feature	Mean	Std Dev	Min	25%	50% (Median)	75%	Max
WaterT (°C)	11.54	5.09	3.5	7.4	10.6	15.6	22.6
AvgDepth (m)	12.97	8.74	1.25	5.1	9.4	20.35	38.3
Precip (mm)	2.49	10.55	0.0	0.0	0.0	0.0	61.0
GDD2 (Growing Degree Days)	175.69	126.49	11.72	59.23	150.51	279.71	475.24
DOY (Day of Year)	138.12	12.93	97.0	128.0	136.0	149.0	168.0
SmeltDen (Density)	0.000003	0.00031	0.0	0.0	0.0	0.0	0.036
YPerchDen (Density)	0.026	0.108	0.0	0.0	0.0	0.0	1.232
BurbotDen (Density)	0.00005	0.00055	0.0	0.0	0.0	0.0	0.006
OtherFishDen	0.00009	0.00129	0.0	0.0	0.0	0.0	0.020
Distshore (m)	1848.48	2231.43	32.35	337.08	683.11	2623.88	7473.93
Area (ABD)	190,909	92,333	31,453	125,931	185,244	238,529	4,312,578
Aspect Ratio	0.516	0.168	0.072	0.383	0.490	0.628	0.999

Table 1: Summary Statistics of Features							
Feature	**Mean**	**Std Dev**	**Min**	**25%**	**50% (Median)**	**75%**	**Max**
Circularity	0.257	0.133	0.018	0.159	0.229	0.324	0.882
Perimeter	3,634	1,434	1,259	2,491	3,368	4,531	34,625
Diameter (ABD)	479.84	113.27	200.1	400.4	485.7	551.1	2,343
Diameter (ESD)	786.92	274.56	333.3	576.9	739.9	935.9	6,815

From table 1, while fish density remains generally low, occasional spikes, particularly in Yellow Perch, may influence plankton dynamics. Zooplankton exhibit significant variation in size, shape, and body structure, providing valuable features for classification. However, the dataset primarily covers spring and early summer, potentially limiting insights into seasonal dynamics.

Class Distribution

Figure 1:

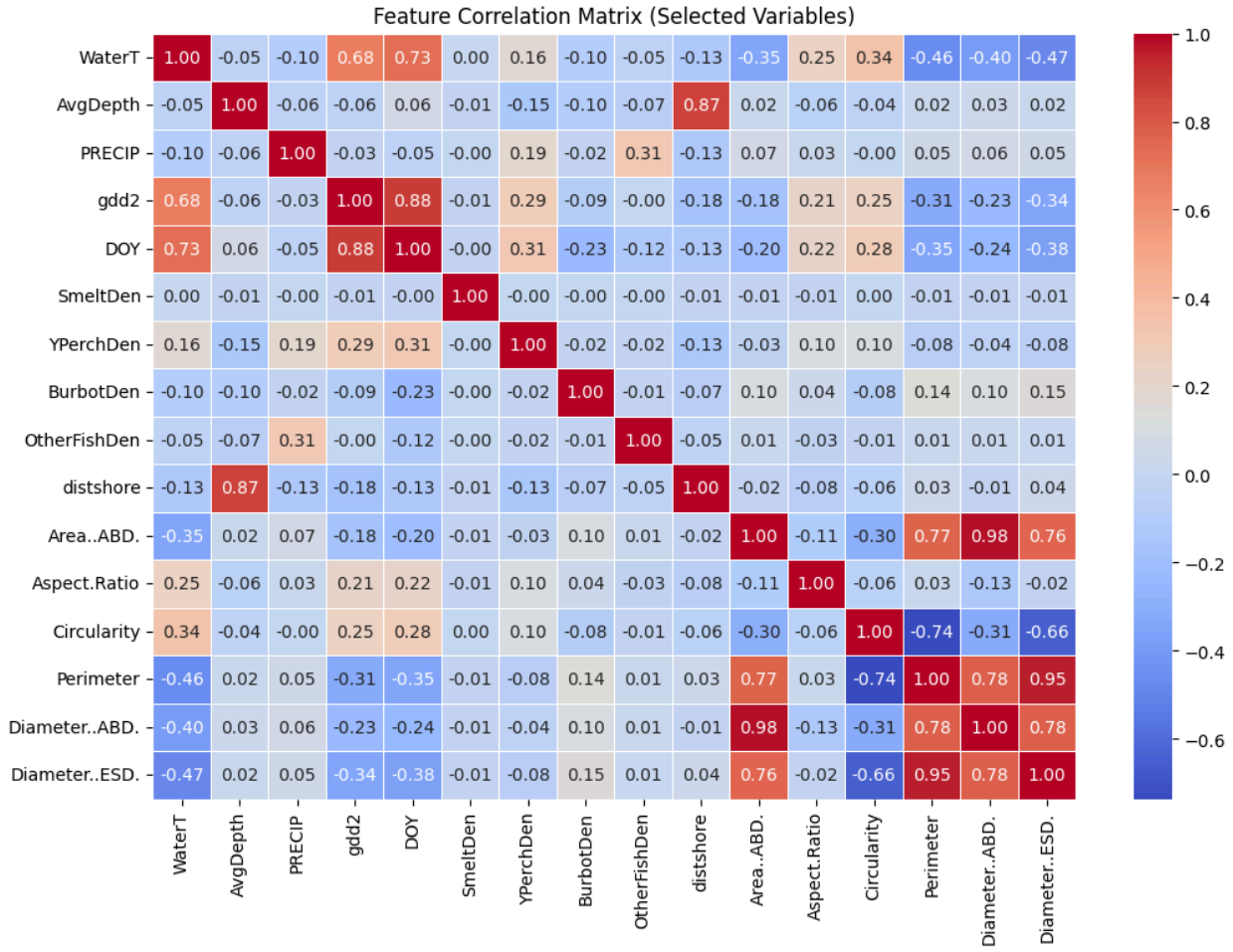


The distribution of zooplankton classes is highly imbalanced, as shown in the class distribution chart. The majority of samples belong to the Calanoid_1 class, which has approximately 248K observations, while the remaining species, such as Chydoridae, have significantly fewer occurrences, with less than 1K observations. This extreme imbalance suggests that classification models may struggle to accurately identify minority classes. To address this challenge, various strategies such as oversampling the minority

classes, applying class weighting in model training, or implementing specialized loss functions like focal loss can be explored.

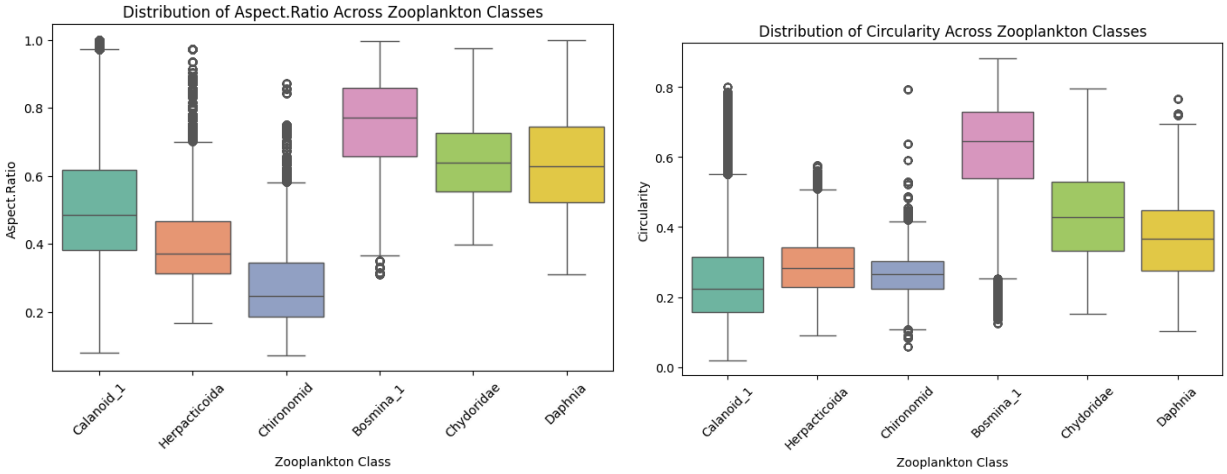
Environmental Factor Analysis

Figure 2:



The correlation heatmap provides insights into relationships between environmental variables. Water temperature (WaterT) and the day of the year (DOY) show a strong positive correlation (~0.73), confirming that temperature rises as the season progresses. Similarly, average depth (AvgDepth) and distance from shore (Distshore) exhibit a strong relationship (~0.87), indicating that deeper samples are generally collected further from shore. Precipitation appears sparse in the dataset, as evidenced in the pairplot analysis, suggesting that rainfall events are infrequent but might influence certain environmental conditions. The scatter plots further highlight clustering in temperature and depth, suggesting that species may be associated with distinct environmental conditions.

Geometric Feature Analysis



Examining the aspect ratio distributions across zooplankton classes, it is evident that Bosmina_1 and Chydoridae tend to have higher aspect ratios, indicating more elongated forms. In contrast, Chironomid has a much lower aspect ratio, suggesting a more rounded body structure. These distinctions indicate that aspect ratio could be a crucial feature for differentiating species. Similarly, circularity measurements reveal that Bosmina_1 tends to be the most circular species, while Chironomid has the lowest circularity. The inverse relationship between circularity and aspect ratio suggests that these features complement each other and should be retained for classification models.

Implication for Modelling:

The correlation heatmap indicates that some geometric features are highly redundant. Area, perimeter, and diameter (ABD & ESD) exhibit strong positive correlations (~0.77 - 0.95), suggesting that including all of them could introduce multicollinearity. To optimize model efficiency, one of these size-based features should be selected, while retaining shape-based features such as aspect ratio and circularity, as they provide distinct classification signals. Additionally, given the strong correlation between WaterT and DOY, only one of these variables should be retained in the model to avoid redundancy.

The extreme class imbalance observed in the dataset necessitates rebalancing strategies to ensure fair model performance. Applying oversampling techniques such as SMOTE, incorporating class-weighted loss functions, or using evaluation metrics like F1-score instead of accuracy could help mitigate classification bias toward dominant classes.

Considering the nature of the data, tree-based models such as Random Forest or XGBoost may perform well due to their ability to handle correlated features. If raw image data is incorporated, deep learning approaches like Convolutional Neural Networks (CNNs) could be tested. A hybrid approach that combines environmental and geometric features may further enhance classification accuracy.

Conclusion:

The exploratory data analysis provided critical insights into the structure and characteristics of the dataset, highlighting key factors that influence zooplankton classification. Environmental variables such as temperature, depth, and distance from shore show strong correlations and may play a role in species distribution. The geometric features of zooplankton, including aspect ratio, circularity, and area, reveal distinct patterns across species, making them valuable predictors for classification. However, the dataset exhibits extreme class imbalance, with Calanoid_1 dominating the observations, which poses challenges for model training. Additionally, the limited time frame of data collection suggests that seasonal variations in zooplankton distribution may not be fully captured. Moving forward, careful feature selection, class rebalancing techniques, and appropriate modeling approaches will be essential to building an effective classification model.