An Integrated Imaging and Lab Analysis Dataset for Invasive Lake Mussels in Lake Ontario and Lake Erie

Dominique Brunet

November 26, 2019

1 Data Acquisition

1.1 Field Sampling

Divers go to specific sites (denoted by PSN code) to image and then harvest mussel samples. The diver first estimate the overall percentage coverage of mussels over the site extent. Three square quadrats (metal frames with yellow and black stripes with inside area of 0.15 squared meters) are laid down on the lake bed at three representative locations in term of mussel coverage for the site. The diver then estimates visually the percentage of live mussels (full) and the percentage of empty mussels inside each quadrat. If cladophora is present, images/videos are generally acquired before and after manual harvesting of cladophora.

Mussels are harvested by scrapping them from the lake bed surface and aspiring them in a tube. In rough water conditions, some mussels might break or be carried by the current. Divers make a visual estimate of sampling efficiency (percentage of mussels in the quadrats that are collected for analysis). Images/videos are acquired before and in some cases during the harvesting of mussels.

1.2 Lab Analysis

Mussels (and cladophora) are freeze-dried and brought back to the lab for analysis. The mussels are then manually counted, sorted by size with a sieve and weighted (biomass).

1.3 Imaging Data

Underwater video data was acquired between 2012-2018 by a hand-held camera for most of the dives. From these videos, trimmed videos were manually extracted for each portion over a quadrat. Still images were either acquired by direct image or extracted from videos. Some images were acquired by a camera system attached to a metal frame on top of quadrats (2016-2018). When multiple images were acquired for the same quadrat, the last image is deemed to be of highest quality. However, in some instances all the images available are taken before harvesting of cladophora, which might grow on top of mussels.

2 Data Preparation

Python scripts to generate the data described above from the data provided by WQMSD (Megan McCusker).

2.1 Imaging Data Naming Convention

Original video data file names were standardized to

`f"{PSN}_{YYYY}-{MM}-{DD}.{EXT}"`

and saved on an external hard-drive in the folder

`f"MusselsCurated/GLNI_2012-2018/GLNI_diver_videos/{PSN}/"`

In total, 493 videos were acquired for a total of 165GB of data. Format (EXT) of the videos were either .mp4, .mpg, .avi or .wmv.

Names of raw images acquired from a camera mounted on a frame were standardized to

`f"GLNI_{PSN}-{Quadrat#}_{YYYY}-{MM}-{DD}_image-{Image#}.nef"`

and saved on an external hard-drive in the folder

`f"MusselsCurated/GLNI_2012-2018/GLNI_quadrat_raw_images/"`

The .nef file format stands for the raw Nikon image format. In total, there are 921 raw image files taking 36.2GB of disk space.

Videos were trimmed manually by quadrat using the Windows Photo App and saved as

`f"GLNI_{PSN}-{Quadrat#}_{YYYY}-{MM}-{DD}_video-{File#}.mp4"`

in the folder

`f"MusselsCurated/GLNI_2012-2018/GLNI_quadrats_stills_images_videos/"`

and subfolder

`f"{PSN}/{YYYY}/{Mmm}.{DD}/Videos/Quad{Quadrat#}/"`

where:

PSN is the diving location code.

YYYY is the four digits year.

MM is the two digits month (with leading zero).

DD is the two digits day (with leading zero).

Mmm is the first three letters of the month (with first letter capitalized).

Quadrat# is the quadrat number (1, 2 or 3).

File# is the video.image/still number when multiple sequences of the same quadrat were taken.

In addition, short videos of the metadata recorded by divers on a white board was saved in the subfolder

`f"{PSN}/{YYYY}/{Mmm}.{DD}/Videos/Meta/"`

under the name

`f"GLNI_{PSN}-Meta_{YYYY}-{MM}-{DD}_video-1.mp4"`

From these videos, a limited number of still images were extracted using the Windows Photo App and saved as

`f"GLNI_{PSN}-{Quadrat#}_{YYYY}-{MM}-{DD}_still-{File#}.jpg"`

in the subfolder

`f"{PSN}/{YYYY}/{Mmm}.{DD}/Stills/Quad{Quadrat#}/"`

Also, raw images were postprocessed with default parameters and compressed to JPEG images with quality factor 90 using a Python script. These compressed images were saved as

`f"GLNI_{PSN}-{Quadrat#}_{YYYY}-{MM}-{DD}_image-{File#}.jpg"`

in the subfolder

`f"{PSN}/{YYYY}/{Mmm}.{DD}/Images/Quad{Quadrat#}/"`

For this project, the difference between "stills" and "images" is the source: stills are extracted from video files while images are taken directly from a camera. In total, there are 2089 trimmed video, compressed images or video still files taking a total of 14.6GB of space.

2.2 Table Data

The source Excel spreadsheet obtained from Megan McCusker of the Water Quality Monitoring and Surveillance Division (WQMDS) is stored in the external hard-drive as

`"GLNI_Miussel_Analysis_Data_2012-2018.xlsx"`

in the folder

`f"MusselsCurated/Tables/"`

This spreadsheet was reformatted into three tables: "Sites.csv", "Dives.csv" and "Analysis.csv". The "Sites.csv" table contains the following columns:

PSN Unique code for site identification

Lake Name of lake

Name Name of site

Latitude decimal latidunal (North-South) position (GPS coordinates)

Longitude decimal longitidunal (East-West) position (GPS coordinates)

The Sites table is extracted from the "site names" sheet in the source Excel spreadsheet. The following entries were added under "ERIE SITES" to represent locations for "Nanticoke Shoal" (July 19, 2016 dive):

PSN	*Latitude*	*Longitude*	*Name*
502	42.7728	−79.9694	*PeacockPoint*
504	42.7881	−79.9838	*PeacockPoint*

Note: the "Name" field in the "Sites.csv" table should be changed to "Site Name" to avoid confusion with "Name" field from images, videos or stills.

The "Dives.csv" table contains information that stays constant for a given dive at a specific site and day. It includes the following columns:

Dive Index Unique index for the dive

CSN Code for dive number (unique for each dive index)

Cruise # Code for the cruise (not-unique)

PSN Code for site identification (match PSN from "Sites.csv")

Date Date of the dive (YYYY-MM-DD)

Depth (m) Depth of water at the site

Overall Coverage Percentage mussel coverage at the site as visually estimated by diver

Silt (%) 1/512 mm to 1/64 mm in diameter.

Clay (%) Less than 1/512 mm in diameter.

Sand (%) 1/16 to 2 mm in diameter.

Gravel (%) 2 to 64 mm in diameter.

Cobble (%) 64 to 256 mm in diameter.

Rock (%) Not clear, maybe something of size between cobble and boulders.

Bedrock (%) Underlying hard rock structure.

Boulders (%) Greater than 256 mm in diameter.

Shale (%) Underlying brittle rock structure.

Underlying Substrate Type Type of rock found under the sediments.

Underlying Substrate Depth (cm) Depth under lake bed.

Note that each substrate type was independently visually estimated by the diver, so the total percentage does not always add up to 100%.

The "Analysis.csv" table contains information pertaining to a particular quadrat, which is a square metal frame of area of $0.15m^2$. It contains the following columns:

Analysis Index Unique index for the analysis.

Dive Index Index of the dive associated with the analysis.

Quadrat Quadrat number (unique for each dive index).

LiveCoverage Percentage of area with live (full) mussels within the quadrat as visually estimated by diver.

EmptyCoverage Percentage of area with empty/broken mussels within the quadrat.

Biomass Total mass (g) of mussels tissue within the quadrat as measured in the lab.

Count Total number of live mussels within the quadrat as counted in the lab.

Xmm Percentage of mussels not passing the X mm sieve (estimated mussel size of more than X mm in diameter).

Sieve sizes are 16 mm, 14 mm, 12.5 mm, 10 mm, 8 mm, 6.3 mm, 4 mm and 2 mm. Note that in some cases, the sum of live and empty coverage exceeds 100%. This inconsistency might either due to errors in estimation of the coverage or by considering live mussels underneath empty shells.

The "FullVideos.csv" table links video names with "Dive Index". The "Video Name" column, which corresponds to the basename of each video file, serves as a unique index. It is associated with the "Dive Index" using the PSN number and the date. Finally, "Video Path" corresponds to the full file path of the original (names not curated) video file. It should eventually be changed to the file path of the standardized video files. Eventually, more metadata could be added in the table, such as file size, video duration, frame rate and frame resolution.

The "QuadratVideos.csv" (and "QuadratStills.csv") tables links trimmed video files (extracted stills) per quadrat to an "Analysis Index". It contains the folllowing columns:

Source Video The basename of the full video from which the trimmed video (stilll) was extracted. It corresponds to "Video Name" in the "Fullvideos.csv" table.

Name The basename of the trimmed (per quadrat) video file.

Quadrat Video Path The full path to the trimmed video (in the external hard-drive, not in the Google Drive). For "QuadratVideos.csv" only.

Quadrat Still Path The full path to the still (in the external hard-drive, not in the Google Drive). For "QuadratStills.csv" only.

Analysis Index The index of the anlaysis associated with the trimmed video. It is found from the PSN, date and quadrat number.

The "ImageTable.csv" table links raw images to the "Analysis Index". It contains the folllowing columns:

Analysis Index

Raw Image Path The raw image path in the external hard-drive.

Name The basename of the raw image path.

Timestamp The timestamp when the image was acquired, as recorded in image metadata.

Note that compressed images (.jpg) are stored in the Google Drive. The "ImageTable.csv" is assembled in three steps. First, a table is made from image names from 2017-2018. A second table for images from 2016 (without quadrat information) is populated, with quadrat information manually inputed. Finally, the two tables are merged into "ImageTable.csv".

The Google Colab Notebook "ImagingDataInventory.ipynb" compares the imaging data listed in "ImageTable.csv", "QuadratVideos.csv" and "QuadratStills.csv" with the imaging data stored in the Google Drive. The list of file names for which any difference is found is returned for each of the three tables. As all imaging data is updated on the Google Drive, the difference should be an empty set.

2.3 Merging Tables

The Google Colab Notebook "TableDataPreparation.ipynb" generates two merged tables: a large un-processed "MergedTable.csv" and a smaller pre-processed "SimplifiedImagingAnalysis.csv".

The "MergedTable.csv" table is assembled in two steps. A first table is obtained by merging "Analysis.csv" and "Dives.csv" on "Dive Index" and then merging the resulting table with "Sites.csv" on "PSN". This first table contains all the data analysis information. A second table is generated by appending "ImageTable.csv", "QuadratVideos.csv" and "QuadratStlls.csv" (which share the "Name" field). This second table links all imaging information to the "Analysis Index". The two tables are then merged on "Analysis Index" to generate "MergedTable.csv". Note that "Name" has to be changed to "Site Name" in "Sites.csv" to differentiate from the "Name" column in the other tables. The

merged table should be used in a read-only mode as it has multiple redundant information spread out through many rows.

For the "SimplifiedImagingAnalysis.csv" table, a single imaging data per analysis is selected. Pre-processing steps include:

- Replacing null values to 0 for "Shale (%)" in "Dives.csv"
- Replacing null values of "Live Coverage" in "Analysis.csv" to the corresponding "Overall Coverage" value from "Dives.csv"
- Replacing percentage values to count for mussel size by replacing null values to "0" in "Xmm" column and by multiplying "Xmm" column with "Count" column in "Analysis.csv".

The last imaging data available is selected for each "Analysis Index" with order of preference: i) raw images, ii) quadrat videos, iii) stills. Only a subset of columns useful for image analysis are kept.

3 Table Data Exploration

The Google Colab Notebook "TableDataExploration.ipynb" generates statistics and summaries for different columns of table data. In particular, it displays the diving sites on a map, count the number of dives and boat cruises per day, per month and per year, and displays histograms for mussels count, biomass, size distribution, depth, and substrate type percentage. Finallly, it provides a count for the number of analysis and the number of analysis with associated imaging data and the number of unique analysis associated with imaging data.

The Google Colab Notebook "TableDataAnalysis.ipynb" performs a correlation analysis between different columns of table data.