TrawlExpert: Tool for Watershed Biological Research

Lab section: L01
Christopher W. Schankula
Student ID: 400026650
schankuc@mcmaster.ca

January 29, 2018

Abstract

A powerful tool to enable researchers to analyze and filter large datasets from fish trawl surveys in order to perform environmental research on fish and / or invertebrate populations. TrawlExpert gives researchers the ability to intelligently filter and query datasets based on biological classification such as family, genus or species, or based on location or timeframe. Outputs include statistical data on relative population abundance as a function of time and geographical distribution. Additionally, TrawlExpert gives an advanced tool for finding local subpopulations within a larger query. A dataset of thousands of Great Lakes trawl surveys from 1958-2016 will be used as a demonstration of TrawlExpert's capability to help researchers narrow down large datasets and glean data which pertains to their research. TrawlExpert will be designed to be used easily and effectively as a first step in a groundbreaking climate and ecological research pipeline.

1 Objective

Statistical and visual tool for the analysis of water ecosystems, based on scientific water trawl data. Provides researchers with tools to analyze large datasets to find patterns in fish populations, including the plotting of historical population data on a map, analyzing population trends over time and finding subpopulations of a certain species, genus, family, etc.

2 Motivation

The diminishing of fish populations in the Great Lakes became a problem in the latter half of the 20th century, with the total prey fish biomass declining in Lakes Superior, Michigan, Huron and Ontario between 1978 and 2015 (Kinnunen, 2017). Annual bottom trawl surveys involve using specialized equipment to sweep an area and are used to determine the relative temporal variation in stock size, mortality and birth rates of different fish species (Walsh, 1997). These surveys are performed annually and often have hundreds of thousands of records, making manual analysis infeasible. The ongoing protection and development of the Great Lakes water basins is considered an important topic for scientists in both Canada and the United States, as evidenced by grants such as the *Michigan Sea Grant* (of Michigan; Michigan State University, 2018).

TrawlExpert will give researchers tools to filter through these large amounts of data by allowing them to search through data based on class, order, genus, family or species. This will help support scientific researchers and fishing companies as they study fish populations in order to launch initiatives to preserve fish populations and perform their business in an environmentally friendly way going forward. As more data is collected on an annual basis, TrawlExpert can easily be injected with the new data and will adjust and scale accordingly, combining the new data with the old data for continued analysis.

TrawlExpert will also analyze the trawl data to find connected subpopulations within the data, giving researchers tools to analyze the portions of the water body that contain different populations and even track these specific subpopulations over time.

The focus of the project will be to develop these unique data searching and querying tools as a first step in a complete trawl survey analysis. For a complete analysis, tools like stratified statistical analysis would be required by the researcher (Walsh, 1997). For purposes of maintaining a scope for this project, the implementation of advanced trawl survey scientific and statistical analysis tools will be relegated to future developments.

3 Prior Work

From preliminary research, past software has been developed to analyze trawl survey data. One such piece of software is MIXFISH which is part of the LFSA package written in BASIC in 1987 (Sparre, 1987). MIXFISH / LFSA has been used by several researchers, (e.g. Levi et al. (1993), Chakraborty (1996)), while performing trawl survey analyses. Most of these were confined to running on smaller samples of data (e.g. 1000 samples for a given species); however, with modern hardware and using a modern programming language, TrawlExpert will allow processing of much larger datasets.

Other studies, e.g Swartzman et al. (1992), make use of software programs to perform statistical analysis but do not use software specialized to analyzing trawl survey data. Although the main focus of this semester's project will be the searching and selection of data, TrawlExpert's goal will be to give researchers an all-in-one platform in order to perform their analysis, in each step of the analysis pipeline.

4 Input / Output and Proposed Solutions

4.1 Datasets

The test dataset that will be used for purposes of this project is the *USGS Great Lakes Science Center Research Vessel Catch Information System Trawl* published by the United States Geological Survey. Compiled on yearly operations taking place from early spring to late fall from 1958 until 2016, the dataset contains over 283,000 trawl occurrences in the five Great Lakes, including the latitude and longitude co-ordinates and biological classification such as family, genus and species. A complementary dataset contains 45,000 entries detailing the operations and parameters such as equipment.

4.2 Outputs

This section describes in detail the types of data analysis and queries that TrawlExpert will be able to perform. Section 4.3 describes a potential use case for a researcher studying a body of water.

4.2.1 Output 1: Historical Distribution

Researchers will be able to query for historical population data of a certain genus, family, etc, outputted in text format or generated as a figure.

4.2.2 Output 2: Geographical Distribution

For a given species over a given timeframe, the program will output the locations of all records matching that query in terms of the recorded location at which the fish were found.

4.2.3 Output 3: Geographical Subgroupings

The third main type of output will be to classify a given search into subgroupings of highly clustered populations (see section 5.3 and figure 1).

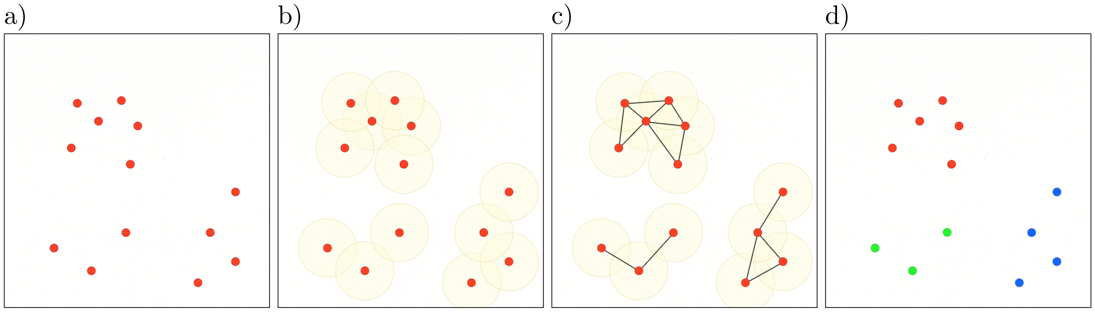

Figure 1: A visualization of the proposed algorithm for output 3. In a), there are some species sightings that the user wishes to group. In b), a radius of similarity is chosen by the user. In c), overlapping regions determine connected points, creating a graph structure with groups of connected points. In d), the points are classified into groups by finding connected components of the graph structure.

4.3 Family Subgroupings: A Use Case Example

The dataset presented in section 4.1 will be used as the main input in order to generate these outputs. The location and temporal data given in the dataset will be used to help generate outputs 1 and 2. In order to illustrate this, the following example use case will be presented and analyzed:

A researcher is studying the decline of all species in the family *Cyprinidae*. She has a large dataset of trawl data and wishes to obtain information about the related subpopulations in the data. Therefore, she uses the TrawlExpert to generate an output of type 3 which will give her the output she needs for her research. By recursing down a pre-built tree structure built from the data (described in more detail in section 5.3), the program will determine which genera, species and potentially subspecies are included in this family of organisms. It will locate all entries for this expanded search. These entries each contain the latitude and longitude at which the sightings were made. The program will then cluster geographically close results and return a list of these subgroupings, either in text format or visualized as a figure.

The researcher can now continue her scientific analysis of the data having easily and intelligently narrowed down to relevant data.

5 Algorithmic Challenges and Opportunities

5.1 Searching Algorithms

A modified form of binary search will be used for quickly locating the first of a given key in the large dataset and will be a crucial building block of all three main types of output. This will allow all entries of a given query to be found.

5.2 Sorting Algorithms

Sorting will be crucial for both ordering historical data in chronological order as well as being the basis for binary search to work, since it requires data to be sorted. The mergesort algorithm will be advantageous due to its fast and predictable runtime ($N \lg N$).

5.3 Graph Algorithms

Graph algorithms will be used for the advanced searching features of the program. Firstly, the biological classification of each organism forms a tree from which species in the same genus, for example, can be located.

Secondly, a graph algorithm will be used to find connected components for generating output of type 3 described in section 4.2.3. Entries form nodes which can be connected together based on their distance to surrounding points and a breadth-first search algorithm will be used to determine connected components (Broder et al., 2000). This is visualized in figure 1.

6 Project Plan

The following milestones will help inform our progress towards completing the goals. The team should be divided up into two subteams for maximum efficiency:

Milestone	Subteam A	Subteam B
Milestone 1 ("Bedrock") (End of Week 1)	Finished parsing module for .csv data to create Java objects that can be used for analysis; start data cleansing	General binary search module underway
Milestone 2 ("Quartz") (End of Week 2)	Cleansed the data to remove or correct entries not containing all of the columns; start generating biological classification tree module	Finished and tested binary search; start mergesort
Milestone 3 ("Granite") (End of Week 4)	Finished and tested classification tree; start data visualization or formatted text output tools	Finished and tested mergesort; start writing query module for output type 1 and 2, using mergesort and binary search to get results from data
Milestone 4 ("Sandstone") (End of Week 6)	Continue data visualization or text output tools; start helping subteam B if needed	Continue query module: finished output 1 & 2; start output type 3
Milestone 5 ("Diamond") (End of Week 8)	Finished data visualization or text output tools; start using them to display outputs; work on keynote presentation	Finished output type 3; work on keynote presentation

While this schedule provides a good reference and a way to monitor progress, it may have to be modified throughout. For example, if a milestone is reached before its given date, the next milestone should start development early. Approximately 1-2 weeks has been purposely left as padding at the end in case of unforeseen circumstances.

References

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph structure in the web. *Computer networks*, 33(1-6):309–320.

Chakraborty, S. (1996). Stock assessment of sin croaker johnieops sina (cuvier) from bombay waters. *Journal of the Indian Fisheries Association*, 26:9–15.

Kinnunen, R. (2017). Great lakes prey fish populations declining.

Levi, D., Andreoli, M., and Giusto, G. (1993). An analysis based on trawl-survey data of the state of the 'italian'stock of mullus barbatus in the sicilian channel, including management advice. *Fisheries research*, 17(3-4):333–341.

Michigan State University, U. (2018). Michigan sea grant.

Sparre, P. (1987). *Computer programs for fish stock assessment: Length-based fish stock assessment for Apple II computers*, volume 2. Food & Agriculture Org.

Swartzman, G., Huang, C., and Kaluzny, S. (1992). Spatial analysis of bering sea groundfish survey data using generalized additive models. *Canadian Journal of Fisheries and Aquatic Sciences*, 49(7):1366–1378.

Walsh, S. J. (1997). Efficiency of bottom sampling trawls in deriving survey abundance indices. *Oceanographic Literature Review*, 7(44):748.