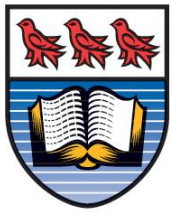

University of Victoria

Department of Physics and Astronomy

Soundscape Metrics Data Product

Spencer Plovie

spencerwplovie@uvic.ca

In partial fulfillment of the requirements of

the Physics and Astronomy Co-op Program

Fall 2022

Work Term #3 of 4

Junior Scientific Programmer

Confidentiality notice: This report is not confidential and may be shared with other co-op students.

This report will be handled by the UVic Co-op staff and will be read by one assigned report marker who may be a co-op staff member within the Physics and Astronomy Science Co-operative Educative Program, or a UVic faculty member or teaching assistant. The report will be retained and available to the student or, subject to the student's right to appeal a grade, held for one year after which it will be deleted.

I approve the release of this report to the University of Victoria for evaluation purposes only.

Supervisor's Signature: _____ Position: Senior Scientific Programmer Date: January 12, 2023

Supervisor's Name (print): Ben Biffard Email: bbiffard@oceannetworks.ca

Company: Ocean Networks Canada

Contents

I. Introduction

Ocean Networks Canada (ONC) is a data pipeline designed to deliver cabled, mobile, and community-based ocean data to the general public. Established as a not-for-profit in 2007, the data is freely available through ONC's data management system (DMAS), Oceans 3.0[1]. This archived and continuous data acquisition is available in a wide variety of formats. The scientific programmer is primarily responsible for creating, maintaining, and improving the data products created from ONC's data framework and works with the Data Products team under the larger Software Development team, which is partly responsible for maintaining the DMAS framework.

Part of the work performed in this position was done using MATLAB, a programming language developed by MathWorks that is widely used by scientists and engineers[2]. When a user goes to the Data Search page on the ONC website, they can select a device they want data from and choose the data product(s) they are interested in, then add this to their cart and 'checkout', where the data is externally retrieved and then processed through the DMAS framework and returned to the user as a compressed file, with a search ID associated with the parameters the user specified. The main function in this process that all MATLAB-based searches use, *oncmatlabsearch*, calls a wide variety of functions designed to process data from different devices – fluorometers, seismometers, spectrometers, and much more.

One such device type is an underwater microphone used to detect and record ocean sounds, known as a hydrophone. ONC has a wide array of hydrophones with over 10 different types and growing. These devices can pick up on various abiotic, biological, and anthropogenic components that make up the ocean soundscape[3]. Searching for events through audio data can take a while, which is where spectrogram images come in handy to accumulate and analyze large amounts of data. Spectrograms are a form of time series analysis where the different frequencies in audio are broken up into a visual spectrum. This allows us to easily analyze high or low frequency sounds that the human ear cannot detect. Another useful visualization tool are Spectral Probability Density (SPD) plots and datasets[4]. The data used for this project included weekly-accumulated SPD matlab files derived from hydrophone data in several locations within the Strait of Georgia region, and was processed in Jupyter notebooks, an open-source Python interface project developed from the iPython Project in 2014[5]. The creation and usage of these weekly SPD files will be discussed in more detail in the *Summary* section of this report.

During this term, part of the work completed included working on this project, where the goal was to design and implement a soundscape metrics (SSM) data product along several frequency bands as an ongoing time series plot, similarly to the State of the Ocean (SOO) plots[6] that ONC provides, where these plots would be automatically updated every week.

II. Summary

In the early stages of development, the sample dataset used for this project were previously-downloaded weekly SPD files from box hydrophone arrays in the Strait of Georgia East, Cascadia Basin, and Barkley regions. The latter two regions are part of the Pacific, and the former is found in the Salish Sea[7]. These metrics, derived by Svein Vagle and his research group, include a 1-minute linear mean Sound Pressure Level (SPL_{lin}), hourly/daily/monthly Power Spectral Densities (PSD), as well as 1st, 5th, 50th, 95th, and 99th percentiles, to describe the ocean soundscape. These metrics are used for all frequency bands – decade, Southern Resident Killer Whales (SRKW), 1/3-octave European vessel (anthropogenic), and a few additional bands for various abiotic soundscape components (Burnham et al., 2021). The frequency bands chosen for this data product included the decadal frequencies as described in Table 1 in the appendix section of this report.

The weekly SPD files were used in the March 2020-2022 date range and are calculated based on the Power Spectral Density (PSD), a frequency-based visualization of power per Hertz. For each frequency bin, a histogram of the sound decibel level is created and normalized by the noise power bandwidth of the window function used. These histograms are then combined to form a matrix across all frequencies (Barton et al., 2013). The result is a three-dimensional plot, typically shown as a heatmap, of the PSD plotted against the logarithmic frequency (Figure 1). The heatmap's distribution is dependent on the relative occurrence of the data and is a function of the horizontal and vertical bin sizes[4]. This data is then broken into weekly chunks and can be output as a matlab file, with the percentiles and linear mean SPL calculated as per frequency bin. This is the data that was used for the SSM project.

This is where processing in Python began. For a given file (week), the percentiles were organized into decadal bands using their frequency bin centre values to determine the bounds. For each percentile, the mean value and standard deviation across a particular band were calculated, with the latter used for error bars. Note that the mean and standard deviation were calculated outside of log space, then converted back to decibels. From here the data was ready to be plotted and included a subplot for each percentile with the four decadal bands chosen. A subplot was also included for SPL_{lin}, following a similar processing pattern as the percentile data.

After the plots were created as shown in Figure 2, the code was made functional across SPD data from any hydrophone. The Oceans 3.0 API[8] was used to request, run, and download the weekly SPD files. This later switched to using the Python ONC client library[9] to order the data product, which integrates the code used in the Oceans API into a more compact, user-friendly format. This allowed for the notebook to be run on a command line interface, with the date, location, and other parameters specified, so that the SPD files could be downloaded, processed to produce the SSM data product, and saved to a specified location in one manual command.

III. Discussion

Currently, the SSM project is under construction and there are several points of improvement before it can be considered as an addition to ONC's data products. There are a few more steps needed to create an ongoing time series available to the public – the Python code would need to be running on a schedule and would need to automatically append new data points on a weekly basis without regenerating the years of data that has already been plotted. Most of the SOO and other Data Preview plots that accomplish this are automated with the use of a scheduled task in the Scheduler Console, but rely on MATLAB code, not Python, to do so. The importance of the Scheduler Console is that the task parameters can be updated, the schedule set, the job enabled, and then left to run. One possible alternative included using the Oceans 3.0 Sandbox[10] to upload and run the python script as a scheduled job, save the output plots in a data product FTP directory, then add them to Data Preview. However, the Sandbox can be difficult to maintain, and the directories are intended for temporary use, with generated files purged two weeks after they're run. This is not ideal for a cumulative time series data product. Another possible alternative would be to run the scripts through MATLAB with the *pyrun* function – this is still a viable option but requires an update to a more recent version of MATLAB to use this function. However, a reliance on other programming languages to run Python notebooks might not be the best solution, as further support for Python-based data products in the future might result in more demand for a platform that supports automated Python code.

Another aspect of this project that needs to be improved upon is in the SSM code itself – it is important to ensure that the metrics laid out by Svein Vagle and his group are being replicated faithfully and efficiently with the final data product created. For instance, in Figure 2 we might expect to see that the error spread would be largest for the bands with the most values given that this is calculated based on the standard deviation, and while the largest frequency band is consistent with this expectation, the smaller bands do not follow this pattern. There could be aspects of the dataset contributing to this effect that have not yet been considered, but it is still reasonable to consider it a point of suspicion at this stage in the process. Other specific improvements to be made are likely to be found after a discussion with the Data Acquisition and Analytics team regarding this project. It is also worth noting that the weekly SPD mat files used to create the SSM do not need to be generated on-the-fly from the PSDs, as there are quite a few of these files archived – this would save a lot of time when the SPD data files are being requested and downloaded, which takes up a majority of the runtime for the notebook.

IV. Future Considerations

There are other aspects of improvement for this project to consider in the future. Along with the multi-band plot as described above, another separate data product for a low-frequency single bin plot needs to be created. This would feature the 50th percentile SPD data for the 50 Hz frequency bin, with multiple hydrophones in an array on one plot. Both the multi-band and single bin data products should be made available in Data Preview, as previously mentioned, but would require a new tab in Data Preview to show data over all available time for a device / set of devices (in the case of the single bin plots).

General considerations for work in the following co-op term include the continuation of my main duties as junior scientific programmer, picking up other projects that may arise, as well as working on the SSM data product project based on the points of improvement highlighted above. This includes creating a SOO generator-style job for Python notebooks to have the code run on a schedule and update weekly, ensure that the metrics used to produce these data products are being reliably reproduced, relying on archived weekly SPD files to reduce notebook runtime, creating the low-frequency single bin data product, and making these new data products available in Data Preview with a new option to view over all available time.

V. References

T. R. Barton, P. M. Thompson, E. Pirotta. *Spectral probability density as a tool for ambient noise analysis*. The Journal of the Acoustical Society of America 133, EL262 (2013); https://doi.org/10.1121/1.4794934

R.E. Burnham, S. Vagle, C. O'Neill. *Spatiotemporal patterns in the natural and anthropogenic additions to the soundscape in parts of the Salish Sea, British Columbia, 2018-2020.* Marine Pollution Bulletin 170 112647 (2021).

Websites

[1] Ocean Networks Canada, *About ONC*, accessed 18 August 2022. *Online.* Available: https://www.oceannetworks.ca/about-onc/

[2] Wikipedia, *MATLAB*, accessed 18 August 2022. *Online*. Available: https://en.wikipedia.org/wiki/MATLAB

[3] Ocean Networks Canada Wiki, *Hydrophone Spectral Data*, accessed 3 January 2023. *Online*. Available: https://wiki.oceannetworks.ca/display/DP/45

[4] Ocean Networks Canada Wiki, *Hydrophone Spectral Probability Density*, accessed 3 January 2023. *Online*. Available: https://wiki.oceannetworks.ca/display/DP/51

[5] Jupyter, *About Us*, accessed 3 January 2022. Available: https://jupyter.org/about

[6] Ocean Networks Canada Wiki, *State of the Ocean and Environment Data Products*, accessed 3 January 2023. Available: https://wiki.oceannetworks.ca/display/DP/State+of+Ocean+and+Environment+Data+Products

[7] Ocean Networks Canada Wiki, *ONC Hydrophone Location Codes & Data Types*, accessed 4 January 2023. Available: https://wiki.oceannetworks.ca/pages/viewpage.action?pageId=72548584

[8] Ocean Networks Canada Wiki, *Oceans 3.0 API Home*, accessed 4 January 2023. Available: https://wiki.oceannetworks.ca/display/O2A/Oceans+3.0+API+Home

[9] Ocean Networks Canada Wiki, *Data Product Download Methods*, accessed 4 January 2023. Available: https://wiki.oceannetworks.ca/display/O2A/Data+product+download+methods

[10] Oceans Networks Canada Wiki, *The Oceans 3.0 Sandbox*, accessed 4 January 2023. Available: https://wiki.oceannetworks.ca/display/O2A/The+Oceans+3.0+Sandbox

VI. Appendix

Frequency Band	Frequency Ranges in the Band (Hz)
Southern Resident Killer Whales	• 10-100,000 • 500-15,000 • 15,000-100,000
Decade bands	• 10-100 • 100-1000 • 1000-10,000 • 10,000-100,000
1/3-octave European bands, centered	• 63 • 125
Various additional bands	• 49,500-50,500 • 7,500-8,500 • 19,500-20,500

Table 1. The frequency bands used in part to create the soundscape metrics data product.

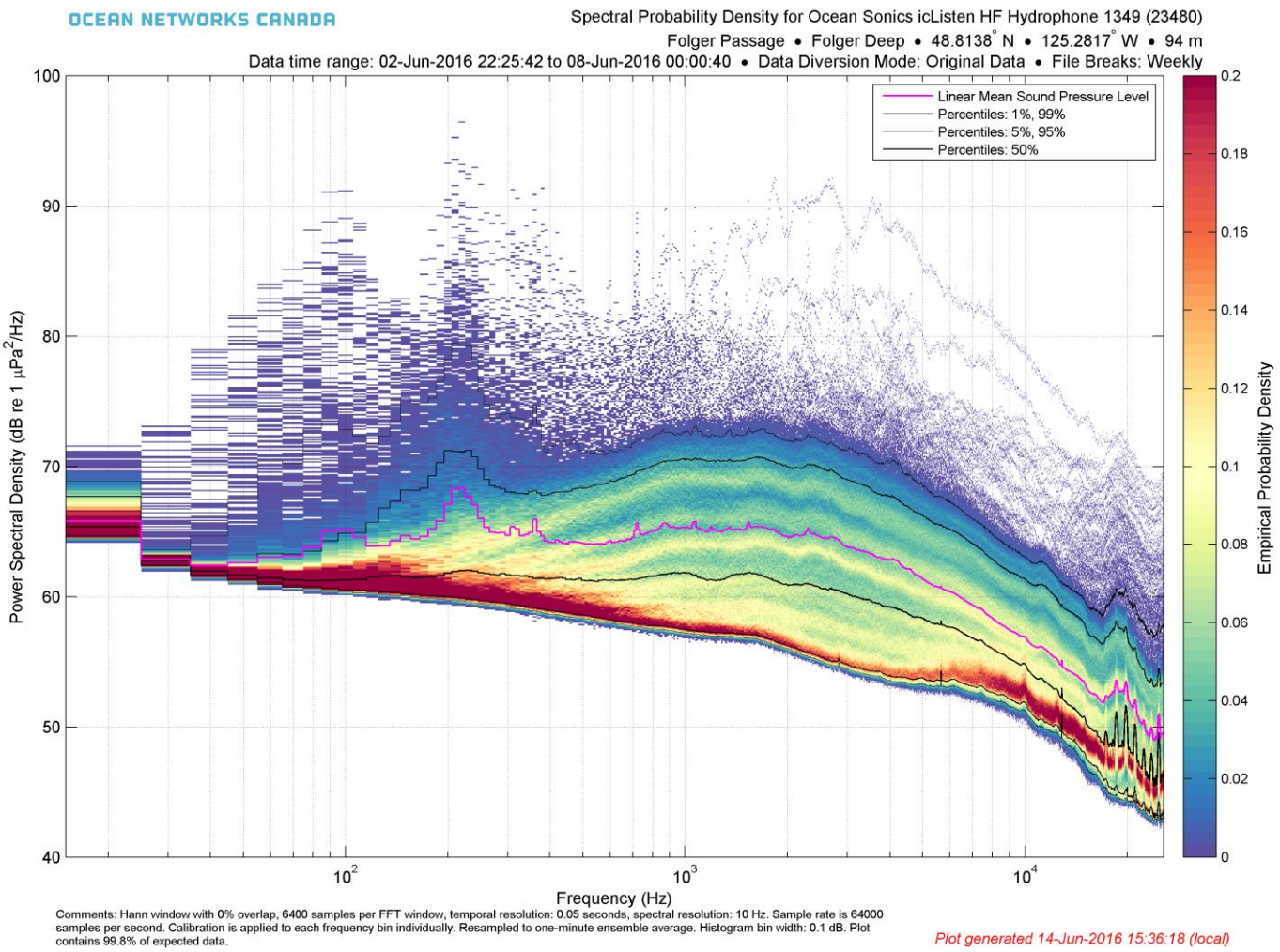

Figure 1. Example of a Spectral Probability Density (SPD) plot, generated from ONC's hydrophone spectral data[4].

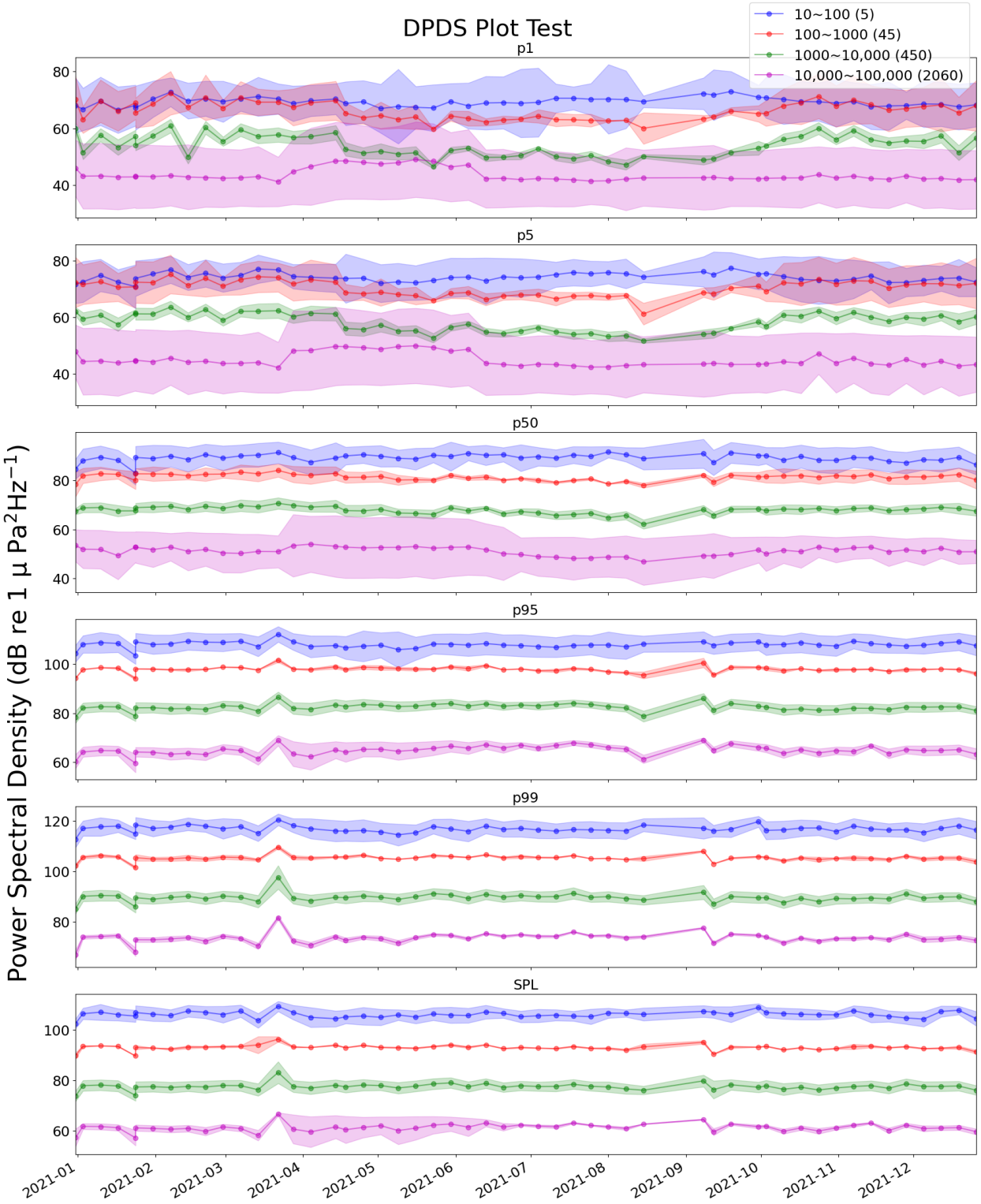

Figure 2. A test plot of the SSM data product using the Data Product Delivery Service (DPDS) code to request and download the SPD weekly files. Created using data from the Strait of Georgia East hydrophone A with a date range from January 2021 – 2022. Note the number of data points that went into each bin as the number in brackets next to each band in the legend.