Project Proposal Machine Learning Class

Project lead [Parker Sprinkle, dpsprink@uw.edu,github.com/spri902]:
Project lead [Verónica G Elgueta, velgueta@uw.edu, github.com/velgueta]:
Project GitHub:https://github.com/UW-ESS-DS/ML-Geo-Project-Veronica-Parker
Date: 11/9/2022

1. Scientific Motivation

Veronica's section on using f-k plots and data dimensionality reduction for the Whidbey Island DAS channels in the sea floor.

The database utilized in this project is from a fiber optic cable called Whidbey island. This cable has around 2000 channels, but for this project, we will use around 200 channels that are located under the ocean. The data has been recorded since May 2022 until now. The purpose of working with this data set is to characterize the interaction between the seafloor and Scholte waves. The Scholte waves are the surface waves that interact between a fluid and an elastic solid medium. We are interested in doing this characterization as the first step of our research project, which involved data DAS from Agustine's volcano which is located in Alaska in the middle of the ocean. The final goal of that project is to be able to measure the seismicity associated with Agustine, which requires that we first identify what part of the signal in the data is coming from Scholte wave interaction. Consequently, to do this characterization with the data from Whidbey cable, we need to build the f-k plot. The f-k plots are the Fourier transform in two dimensions of the data. As a result of f-k plots, we will obtain knowledge about the dominant frequencies and the velocities of the waves captured with the fiber optic. To study the f-k plots across time, we need to reduce the dimensionality of our data set. We will save all f-k plots along the time in one D matrix. Using the PCA algorithm we should be able to reduce the dimensionality of D.
Finally, after using PCA it should be easier to analyze and interpret what is happening on the seafloor. Once we have a characterization of Scholte waves with data DAS, this work can then be applied more broadly in research topics in the ocean, being able to distinguish more than one phenomenon in the seafloor.

Parker’s section on the DAS data in the wells at the BC cribs desiccation experiment.

Distributed acoustic sensing (DAS) data were recorded for approximately 5 months between May and October of 2022 during a vadose zone soil desiccation experiment located at the Hanford Site north of Richland, WA. The purpose of the desiccation experiment is to remove soil-moisture in the shallow subsurface (~ 15 m) located above a site contaminated with low-level nuclear waste and thereby preventing plume migration downward into the groundwater.
Traditional approaches to subsurface monitoring require expensive in-situ monitoring techniques (e.g. drilling monitoring wells and installing dense arrays of a variety of environmental and geophysical sensors). Thus, it would be beneficial to monitor the desiccation process “remotely” from the surface or with a minimal number of monitoring wells. Previously, electrical resistivity tomography (ERT) has been shown to provide accurate estimates of soil moisture during a similar desiccation experiment. However, as the soil volume becomes increasingly dry, changes in the electrical conductivity signal taper off and no longer provide soil moisture measurements at the desired resolution. Laboratory experiments on soil columns taken from the Hanford Site show that, conversely, seismic wave velocities are highly sensitive to changes in moisture content in the regime of extremely dry conditions. This result provided the motivation for this field scale experiment of monitoring shallow subsurface velocity changes with ambient noise recorded by a distributed fiber optic sensing array.

References:

Compressional wave velocity and effective stress in unsaturated soil: Potential application for monitoring moisture conditions in vadose zone sediments
DC Linneman, CE Strickland, AR Mangel - Vadose Zone Journal, 2021

Stork, Anna L., et al. "Application of machine learning to microseismic event detection in distributed acoustic sensing data." *Geophysics* 85.5 (2020): KS149-KS160
https://doi.org/10.1190/geo2019-0774.1

Williams, Ethan F., et al. "Distributed sensing of microseisms and teleseisms with submarine dark fibers." *Nature communications* 10.1 (2019): 1-11.
https://doi.org/10.1190/geo2019-0774.1

2. Scientific Workflow

Verónica

Data processing

- Make the matrix D of the f-k plots along a period of time.
- Calculate Singular value decomposition.
- Applied PCA to interpret the data

Parker

Data processing

- Locate the data channels from the tap tests
- Pre-process data (detrend, taper, filter, whiten)
- FFT the data for crosscorrelation in the frequency domain
- Cross-correlate the data from the DAS channels in the wells
- Stack the resulting cross-correlations (hours, minutes, days?)

GitHub Project and structure

The project is described in the readme file and any necessary code made available at this GitHub repository. There are *Data_p* and *Data_v* directories with a small amount of sample data. The codes for processing and creating figures are located in the *src_p* and *src_v* directories. An *environment* file is also included to inform users of any necessary dependencies. We use an short and simple permissive MIT license.

Data access

The data sets are saved in a local and private server. The data is to heavy to save that in the cloud. All code to reproduce processing and figures will live on the GitHub repository

3. Data

Verónica

The raw data were extracted from a Sintel Onyx distributed acoustic sensing (DAS) interrogator and saved as HDF5 files on a local network attached storage device the the data is saved in the cloud (Sermeq). The data is recorded each 1 min for around 6 months.

Parker

The raw data were extracted from a Sintel Onyx distributed acoustic sensing (DAS) interrogator and saved as HDF5 files on a local network attached storage device. Continuous data were recorded in 1 minute files and are approximately 28 MB in size. Over 100,000 1-minute files were recorded over the 5 month experiment for a total ~ 2 TB of ambient noise data.