Wavelet decomposition for passive data compression and processing

Hafiz Issah & Eileen R. Martin

Center for Wave Phenomena and Dept. of Geophysics, Colorado School of Mines, Golden CO 80401
email aissah@mines.edu

ABSTRACT

As seismic data acquisition techniques improve, the problems associated with having large amounts of data become more apparent. We aim to reduce the cost and footprint of data storage, as well as improving the efficiency of passive seismic data processing. As a contribution to the solving the storage problem, we consider wavelet decomposition and thresholding as one technique for lossy compression of passive seismic data. By calculating the distribution of wavelet coefficients for passive data, we can compare strategies to determine the compressibility and optimum threshold for adaptive on-the-fly lossy compression. We compare compressibility and error bounds when using one-dimensional versus two-dimensional wavelet decomposition. Because sparse cross-correlation in the wavelet domain is more scalable than in the frequency domain, we are particularly interested in understanding the implications of compression on these. We consider error analysis to quantify how errors in data compression propagate through cross-correlation analysis (e.g. for template matching or ambient noise interferometry).

Key words: DAS,Lossy Compression, Wavelets, Seismic Data

1 INTRODUCTION

The advent of Distributed Acoustic Sensing (DAS) made it easier to gather high density seismic data, resulting in a vast increase in the volume of data recorded (Lindsey and Martin, 2021). While there is more data to work with, there is also the problem of managing the storage and computational costs that comes with having large amounts of data. Lossy compression provides a way of compressing the data and thereby reducing storage cost with little effect to the data integrity.

One such lossy compression technique is wavelet compression, in which one represents data using a wavelet decomposition, and then sets all small coefficients to zero. Wavelet bases are formed by the translation and scaling of wavelets which are little snippets of functions which are confined in time and frequency. Wavelet decomposition has been used in multiple areas of seismic processing including changepoint detection and associated error estimation (Simon et al., 2020); estimation of phase arrival-time (Tibuleac et al., 2003);study of anisotropy (Bear et al., 1999); de-noising (Donoho, 1995); feature extraction for earthquake detection algorithm (Yoon et al., 2015).

Due to the localized support of wavelets, wavelet decomposition produces few high amplitude coefficients corresponding to irregularities, strong signals and points of singularities within the data (Mallat, 2008). This means that there is a lot of sparsity in wavelet decomposed data if the data has few of these points. This property often makes wavelet decomposition a good way to do lossy compression. One of our goals is to explore this potential for different types of seismic data and to understand the efficiency and accuracy tradeoffs in this type of compression that can be achieved in different cases. Wavelet decomposition has been used in the compression of seismic data recorded for an active source (Villasenor et al., 1996). Here, we look at its suitability for compressing data from passive sensing. Another area of concern is the computational costs associated with decomposition and reconstruction between storage and processing. Our prior work yielded some strategies to circumvent these costs by doing more processing in the wavelet domain(Kump, 2021). Our goal ultimately is to provide more ways of carrying out more of the seismic processing steps in the wavelet domain.

In this report, we provide background on: wavelet decomposition in section 2, and cross-correlation and wavelet domain cross-correlation in section 3. We describe our new contributions in section 4, detail experiments with real data in section 5, and finally talk about our results, implications and future directions in section 6.

2 BACKGROUND ON WAVELET COMPRESSION OF SEISMIC DATA

Wavelet decomposition is a tool that can be used for multi-resolution analyses of signals. It allows the expression of signals as the sum of wavelets which are localized in time and frequency (Daubechies, 1992). A mother wavelet is a function with a zero average and from this, a wavelet basis can be constructed (Mallat, 2008). The wavelet basis is constructed by dilating and translating the mother wavelet in time. With this we represent a signal at different scales referred to as levels. The lower levels correspond to small scale or high resolution and can also be thought of as high frequency and the scale increases in higher levels. At the highest level, the rest of the signal is approximated by the father wavelet. Given mother wavelet, ψ and father wavelet, ϕ which are translated by n and scaled by j to form $\psi_{j,n}$ and $\phi_{j,n}$ respectively, the J level wavelet decomposition of a signal, f of length $N = 2^{-L}$ can be written as,

$$f = \sum_{j=L+1}^{J} \sum_{n=0}^{2^{-J}-1} d_j[n]\psi_{j,n} + \sum_{n=0}^{2^{-J}-1} a_j[n]\phi_{J,n} \tag{1}$$

Where the inner products, $d_j[n] = \langle f, \psi_{j,n} \rangle$ and $a_j[n] = \langle f, \phi_{J,n} \rangle$ are referred to as the wavelet coefficients and the wavelets are defined as

$$\psi_{j,n}(x) = \frac{1}{\sqrt{2^j}} \psi\left(\frac{x-n}{2^j}\right) \qquad \text{and} \qquad \phi_{j,n}(x) = \frac{1}{\sqrt{2^j}} \phi\left(\frac{x-n}{2^j}\right)$$

Two-dimensional wavelet decomposition can also be done on two-dimensional data-sets with wavelets that are derived from products of one-dimensional mother wavelets. When recording continuous DAS data, we typically work with files that each hold a 2D matrix representing between 0.5 and 10 minutes of data across all channels. On this data we could do either a one-dimensional or two-dimensional wavelet decomposition. Then we can choose to keep only a small percentage of the coefficients and still reconstruct an acceptable version of our initial data. When we keep only the largest coefficients, and set smaller coefficients to zero, this process is known as thresholding and this allows as to do lossy compression using wavelet decomposition. The localization of frequency and time of the wavelet basis also makes it possible to keep the time and frequency resolution in the original data.

While these properties have been exploited using wavelet decomposition to compress active source siesmic data(Villasenor et al., 1996), we instead explore the suitability for continuously recorded passive seismic data. The factors considered in evaluating the suitability include the relationship between the distribution of coefficients, sparsity and compressibility for different forms of passive seismic data. At least one prior study has used the distribution of time domain data values to classify various types of signal and noise in continuously recorded DAS data (Dumont et al., 2020); we study the distribution of coefficients in the wavelet domain as an indication of compressibility and to determine the optimum level of thresholding. Previous works have studied threshold estimators for noise with Gaussian distribution based on the standard deviation and number of coefficients, and the associated expected error (Donoho and Johnstone, 1994). Building on this, we look at which probability distribution would best fit passive seismic data and how we can use it as a basis for threshold and error estimation. We explore one-dimensional and two-dimensional wavelet compression.

3 WAVELET CROSS-CORRELATION

Cross-correlation has several application in seismic imaging including receiver function analysis ; template matching; It also has application in seismic interferometry for exploration seismology, near surface geophysics, ultrasonics and underwater acoustics (Snieder et al., 2009). It has also been used in the migration of free-surface multiples in common depth point (CDP) data and imaging of source distributions using passive seismic data (Schuster et al., 2004). Thus we are motivated to ensure that we can perform cross-correlations efficiently on compressed data.

Wavelet domain cross-correlation presents a way to do cross-correlation on one-dimensional wavelet transformed data without the need to reconstruct the data first (Kump, 2021). If we have two time series signals d_r and d_s with wavelet coefficients stored in vectors $x^{(r)}$ and $x^{(s)}$ respectfully, the cross-correlation between these two signals at time lag τ can be expressed as a vector-matrix

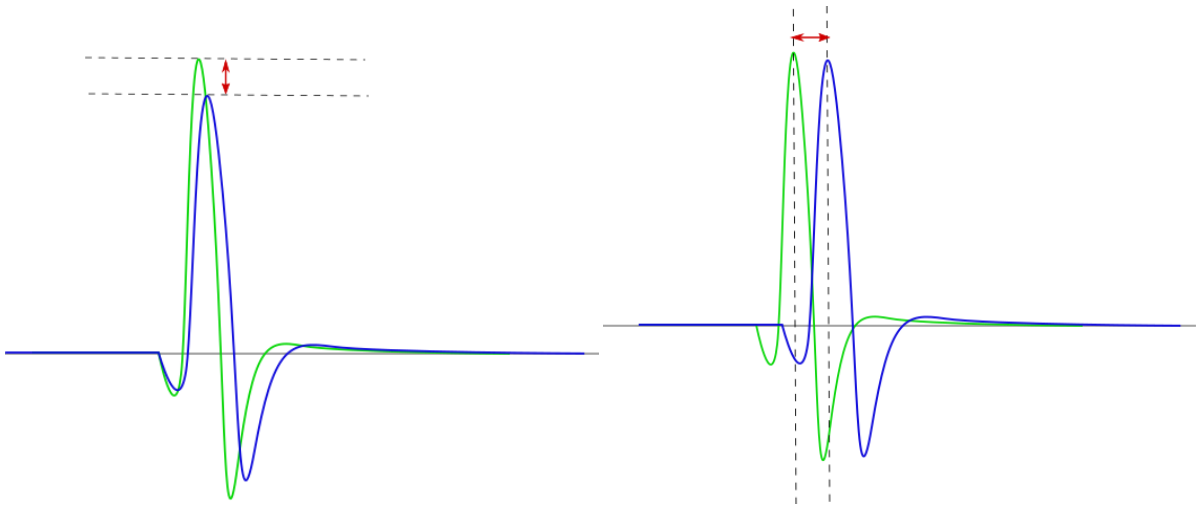

Figure 1. The two kind of errors that we can have when computing cross-correlation: error in amplitude (left) and error in peak location (right)

product:

$$(d_r \star d_s)(\tau) = x^{(r)T} W^{(\tau)} x^{(s)} = \sum_{j=1}^{J+1} \sum_{k=1}^{J+1} x_j^{(r)T} W_{j,k}^{(\tau)} x_k^{(s)}, \quad x_r, x_s \in \mathbb{R}^N \tag{2}$$

where W is a matrix containing the cross-correlations between the wavelet basis functions That is,

$$\mathbf{W} = \begin{bmatrix} \phi_1 \star \phi_1 & \phi_1 \star \phi_2 & \phi_1 \star \phi_3 & \cdots & \phi_1 \star \phi_N \\ \phi_2 \star \phi_1 & \phi_2 \star \phi_2 & \phi_2 \star \phi_3 & & \phi_2 \star \phi_N \\ \phi_3 \star \phi_1 & \phi_3 \star \phi_3 & \phi_3 \star \phi_3 & & \phi_3 \star \phi_N \\ \vdots & & \ddots & \ddots & \\ \phi_N \star \phi_1 & \phi_N \star \phi_3 & \phi_N \star \phi_3 & \cdots & \phi_N \star \phi_N \end{bmatrix} \tag{3}$$

And $\phi_n,\ n \in 1, 2, 3...N$ represents the wavelet basis functions used to decompose the signals. The method takes advantage of properties in this representation including sparsity and redundancies to reduce the number of computations that need to be done and increase efficiency. W can also be precomputed and used for any data transformed using the same wavelet basis. There are two types of errors that may be incurred in calculating the cross-correlation function. First, is the error in the amplitude of the cross-correlation function (left image, figure 1) for which the initial work for wavelet domain cross-correlation provides an error bound (equation 5). To derive this error bound, we first let $\hat{x}^{(r)}$ and $\hat{x}^{(s)}$ represent the thresholded vectors of wavelet coefficients corresponding to signals d_r and d_s, $\hat{d}_r$ and $\hat{d}_s$ represent the time-domain approximate data reconstructed from $\hat{x}^{(r)}$ and $\hat{x}^{(s)}$, and all other variables are as defined previously. By splitting up the zero-ed and nonzero coefficients of the compressed representation, we can prove that:

$$(d_r \star d_s)(\tau) - (\hat{d}_r \star \hat{d}_s)(\tau) = x^{(r)T} W^{(\tau)} x^{(s)} - \hat{x}^{(r)T} W^{(\tau)} \hat{x}^{(s)} \tag{4}$$

$$|(d_r \star d_s)(\tau) - (\hat{d}_r \star \hat{d}_s)(\tau)| \leq \|W^{(\tau)}\|_2 \frac{N\sqrt{c-1}}{c} (\sqrt{c-1} T^r T^s + T^r B^s + B^r T^r) \tag{5}$$

where N is the length of $x^{(r)}$ and $x^{(s)}$, c is the compression factor (where $\frac{1}{c}$ represents the proportion of coefficients that were not thresholded), T^r and T^s are the maximum magnitude of coefficients corresponding to $x^{(r)}$ and $x^{(s)}$ respectively. B^r and B^s are also the maximum magnitude of coefficients removed after compression corresponding to $x^{(r)}$ and $x^{(s)}$.

While this is one type of error analysis, for some applications of cross-correlation we are more concerned with the lag at which various peaks are detected in the cross-correlation function. Therefore, there is the need to consider errors that may result in a shift of the time lag of peaks illustrated in the right image in figure 1. We make some new contributions towards finding an upper bound for this time-lag error in the following section.

4 NOVEL THEORY DEVELOPED

To investigate the error in locations of peaks that may be introduced in wavelet domain cross-correlation, we look at the magnitude of the difference in error we can achieve at different lags. That is, for two different time lags,τ_1 and τ_2, each would have some error bound which can be defined by equation 4. $\Delta((d_r \star d_s)(\tau_1))$ and $\Delta((d_r \star d_s)(\tau_2))$ are used to here to represent the errors (that is, $\Delta((d_r \star d_s)(\tau_1)) = (d_r \star d_s)(\tau_*) - (\hat{d_r} \star \hat{d_s}(\tau_*))$ for τ_1 and τ_2 respectively. Then the difference in errors would be

$$\Delta((d_r \star d_s)(\tau_1)) - \Delta(d_r \star d_s)(\tau_2) = (x^{(r)T} W^{(\tau_1)} x^{(s)} - \hat{x}^{(r)T} W^{(\tau_1)} \hat{x}^{(s)}) - (x^{(r)T} W^{(\tau_2)} x^{(s)} - \hat{x}^{(r)T} W^{(\tau_2)} \hat{x}^{(s)})$$

Under the assumption that the points closest to the peak will be the ones mostly likely to have enough error to become the new peaks, we consider small changes in τ and take a derivative of the errors with respect to τ.

$$\frac{d(\Delta((d_r \star d_s)(\tau_1)))}{d\tau} = \frac{d(x^{(r)T} W^{(\tau)} x^{(s)})}{d\tau} - \frac{d(\hat{x}^{(r)T} W^{(\tau)} \hat{x}^{(s)})}{d\tau} \tag{6}$$

The wavelet coefficients have no dependence on the time lag, τ. Hence this derivative only depends on $W^{(\tau)}$.That is $\frac{d(W^{(\tau)})}{d\tau}$. The components of this matrix are of the form,

$$\frac{d(\phi_n \star \phi_m)}{d\tau} = \phi_n \star \frac{d(\phi_m)}{d\tau} \tag{7}$$

The derivative of the wavelet bases can be computed for any wavelet assuming it is adequately smooth. Hence, $\frac{d(W^{(\tau)})}{d\tau}$ can be known and calculated. With this, we can come up with an error bound similar to equation 5.

$$|\frac{d(\Delta((d_r \star d_s)(\tau_1)))}{d\tau}| \leq \|\frac{d(W^{(\tau)})}{d\tau}\|_2 \frac{N\sqrt{c-1}}{c}(\sqrt{c-1}T^r T^s + T^r B^s + B^r T^r) \tag{8}$$

This bound gives the maximum difference in error that can be attained from one lag to the next. The general idea is that if this error is bigger than the slope between a peak and its immediate values, then there is the possibility of the peak being shifted in lag. The details of the implementation of this error bound are still being developed.

Further expanding on the wavelet domain cross correlation, we consider how to implement it efficiently when we have multiple channels and we are trying to cross correlate each channel with all the other channels. In this case, the first consideration is whether to decomposed each channel separately or to do a two dimensional wavelet decomposition in space and time. Towards this, we do some comparison between one and two dimensional wavelet decomposition taking into consideration the error between the compressed and original data, the changes in the power spectrum, compressibility and event detection.

Sparsity of the wavelet coefficients plays a big part in the compressibility of the data. The more coefficients we have equal to or close to zero, the more the number of coefficients we can get rid of with little cost to the integrity of the data. We investigate the sparsity patterns in different data and try to draw a comparison between the optimum threshold we can achieve and the how it may relate to the distribution of the coefficients.

5 COMPUTATIONAL EXPERIMENTS

We use openly available DAS data from the Penn State Fiber-Optic foR Environment SEnsEing (FORESEE) project (Zhu et al., 2021). The data was recorded at a sampling rate of 500 samples per second with 2137 channels at 2m spacing. The amount channels were reduced to 2120 by removing potentially problematic channels. Figure 2 shows examples of the data we worked with; data recording mainly uncorrelated background noise (left) and data containing an event and some anthropogenic noise(right).

To do a comparison between one dimensional and two dimensional wavelet compression, we did applied these to the same data, did some thresholding and reconstructed the data. On the reconstructed data we did the following comparisons.

- Comparison of the errors
- Comparison of the amount of coefficients
- Power spectrum of the data
- Event detection using Short-Time-Average/Long-Time-Average (STALTA)

The part of the data we used for for the examples in the figures was recorded between 03:33:35 – 03:34:35 UTC on April, 15, 2019; this data records a thunder strike that happens close by (Hone and Zhu, 2021).

To investigate the changes in compressibility of data, we compare the amount of coefficients equal to or close to zero. To quantify this we use the standard deviation of the distribution of the wavelet coefficients. The higher the proportion of coefficients above the standard deviation, the less the proportion close to zero and hence the less compressible the data is. We do this comparison

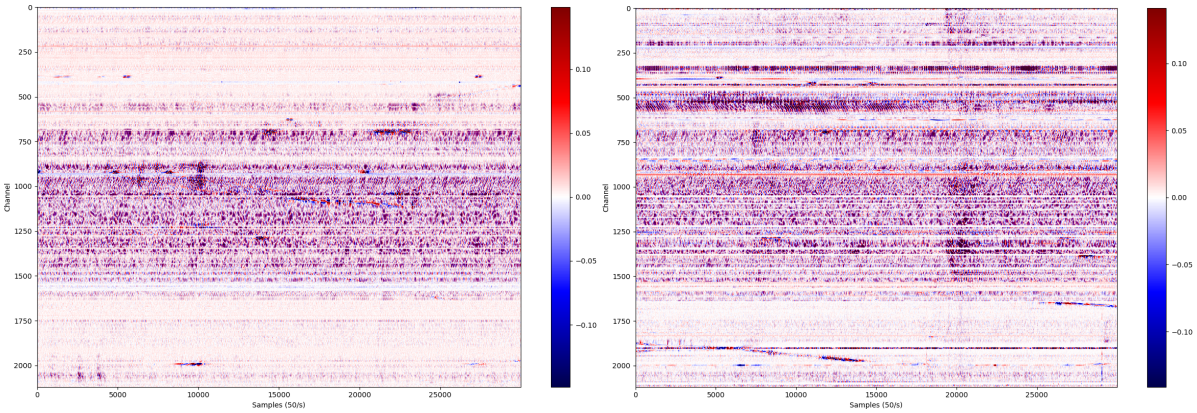

Figure 2. The two types of data we worked with: Mainly uncorrelated background noise (left) and data containing an event and some anthropogenic noise(right).

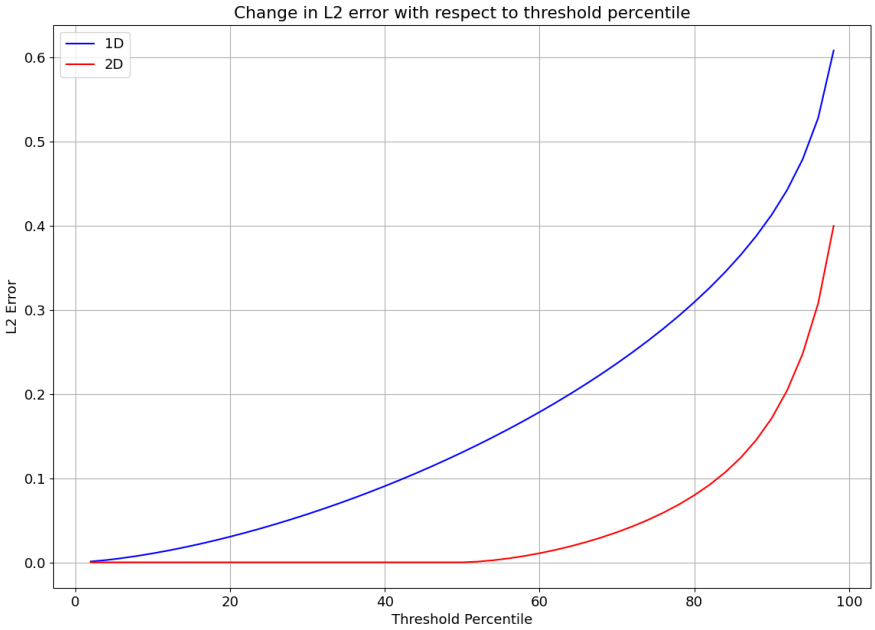

Figure 3. Comparison of the errors encountered for different levels of thresholding for one dimensional and two dimensional wavelet decomposition of the same data. This error is averaged across the multiple channels of the data

for a day's worth of data to see how this changes over the course of a day (August, 1, 2019). The results of these experiments and their implications are discussed in the next section.

6 CONCLUSIONS AND DISCUSSION

In this report, we investigate errors that arise in cross-correlation analysis of wavelet-compressed passive seismic data. First, we investigate this theoretically, and report on progress towards finding an error bound for the shift in peaks that may occur when we used thresholded coefficients for wavelet domain cross-correlation (equation 8). Additionally, we study the wavelet coefficient distribution of real passive DAS data to investigate this error bound for specific wavelets. In our future work, we will identify some wavelet properties that may reduce these errors identified, look at fitting the data with a generalised normal distribution using more than one parameter and develop an algorithm for doing cross-correlation in the wavelet domain on two dimensional wavelet decomposed data.

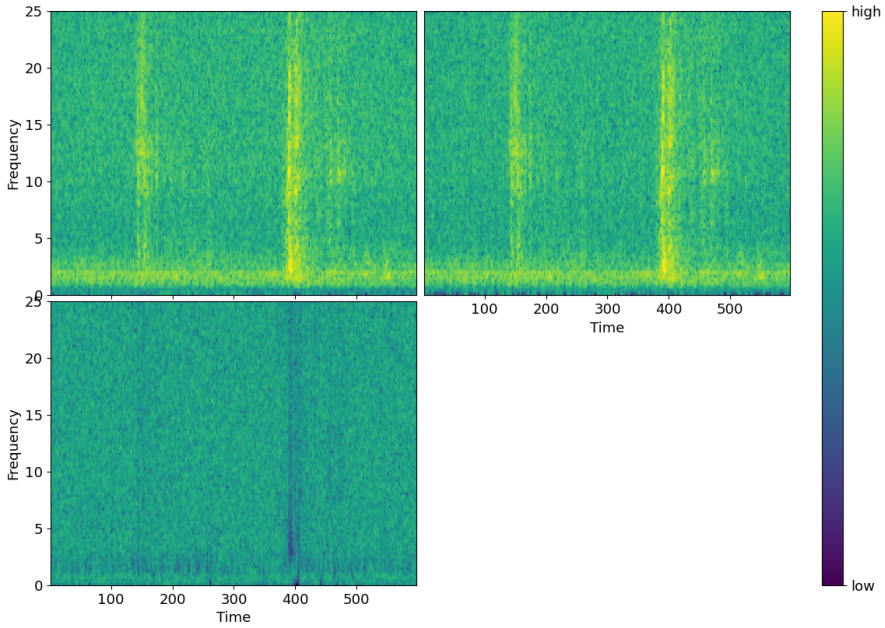

Figure 4. Comparison of the power spectrum of the original data (upper left), the wavelet compressed data after reconstruction (upper right) and the residual (lower left). This power spectrum is of a single channel in the data. Other channels show the same consistency

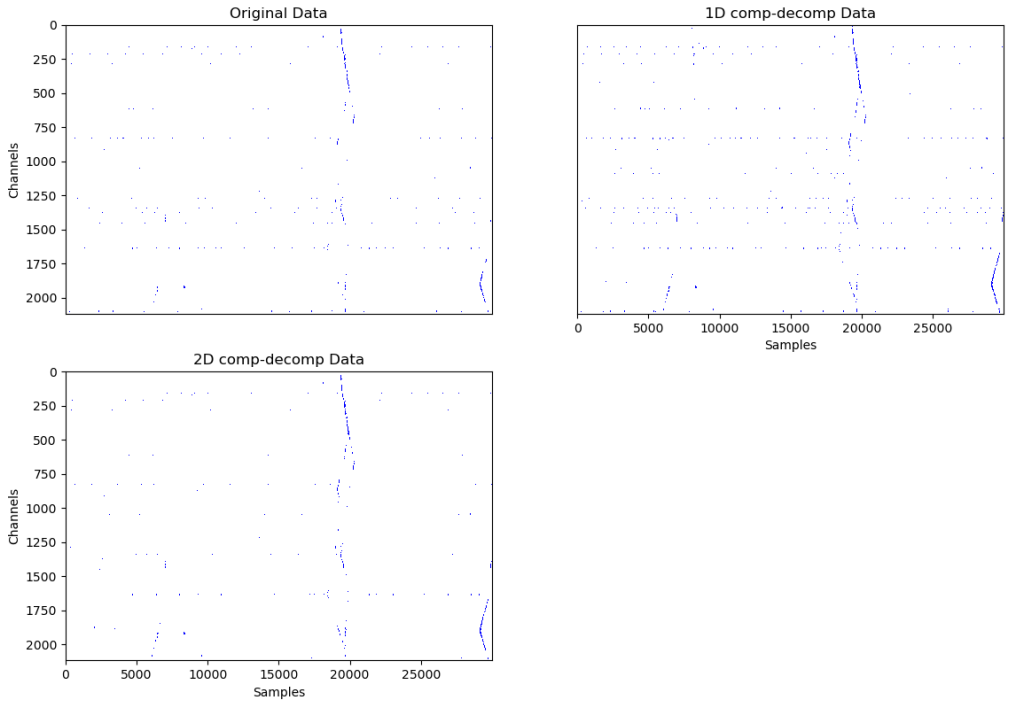

Figure 5. Event detection comparison using STA/LTA for original data (upper left), 1-D compressed data (upper right) and 2-D compressed data (lower left)

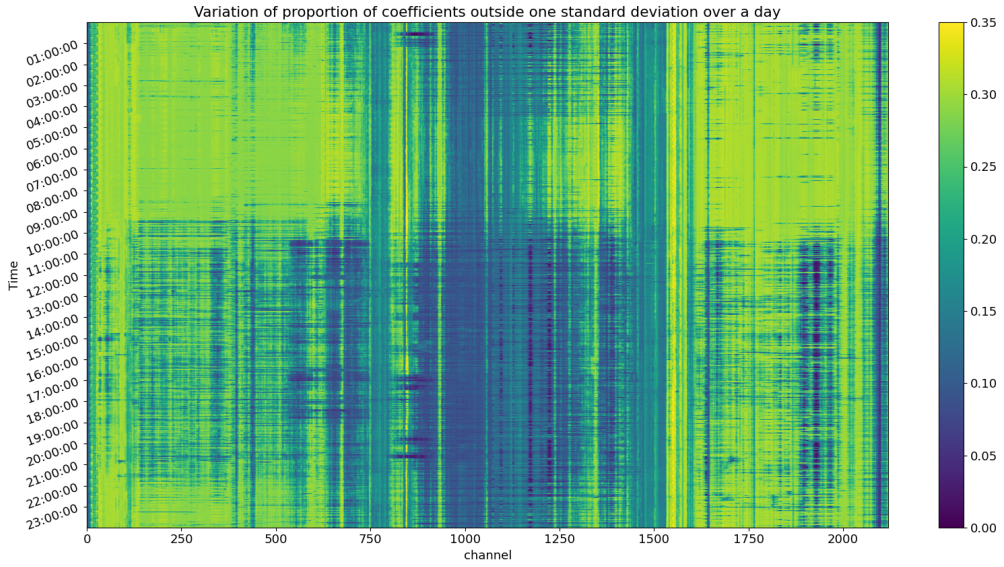

Figure 6. Proportion of wavelet coefficients outside one standard deviation for data collected from 2120 channels over the course of a day (08/01/2019). The higher this amount the less compressible it is. There are some channels that show overall low quality data but in there other channels a pattern can be seen. The data recorded at the hours around midnight and early morning where there was little antropogenic activity have the least compressibility indicating less redundancy and potentially a more "white" distribution of sources

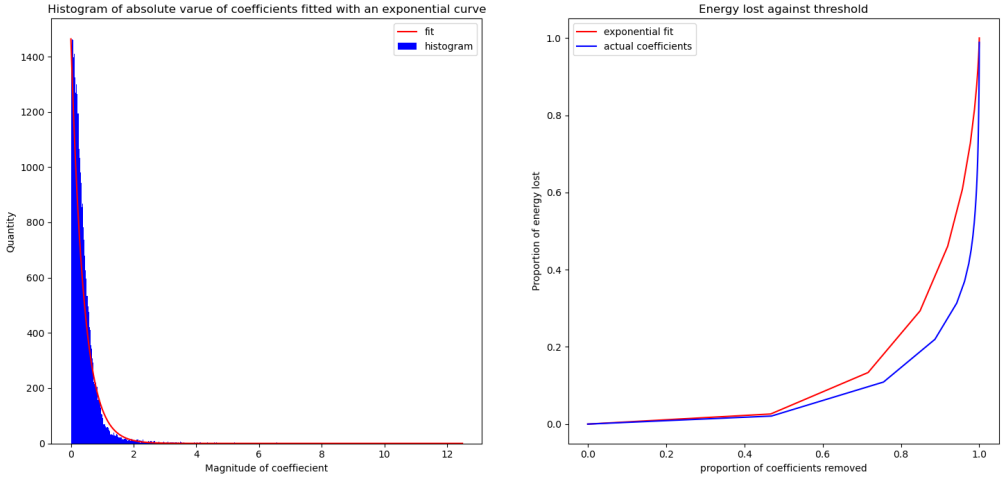

Figure 7. Data with recorded signal. On the left we have the distribution of absolute value of the coefficients from wavelet decomposition of data that had a visible signal recorded. Plotted on this distribution is an exponential curve fitted to the distribution. On the right is a plot of the loss of energy against the threshold percentile calculated from the coefficients (blue) compared with that calculated from the fitted exponential curve

We also looked at the distribution of coefficients for different wavelet transforms of passive seismic data. We see that, the data is more compressible in situations where there are some activities other than just uncorrelated background noise being recorded. This can be seen in figure 6 where data recorded during the night shows less compressibility than the data recorded during the data when there is more (often highly correlated) anthropogenic activities to be recorded.

To calculate a simple approximation of the distribution of coefficients, we also try to fit an exponential curve determined with just one parameter-the mean of the absolute of the coefficients- to the distribution of the magnitude of the coefficients. For the data with recorded activities, the fit underestimated the amount of coefficients of low magnitude but over estimated the amount of high magnitude coefficients as can be seen in the figure on the left in figure 7. In the image on the right in figure 7, we compute the error due to thresholding for the data and compare to that estimated from the fit. The fit always overestimates the error and can be seen as an upper bound for the actual error. In the image on the left in figure 8, we looked at how the exponential fit worked for data

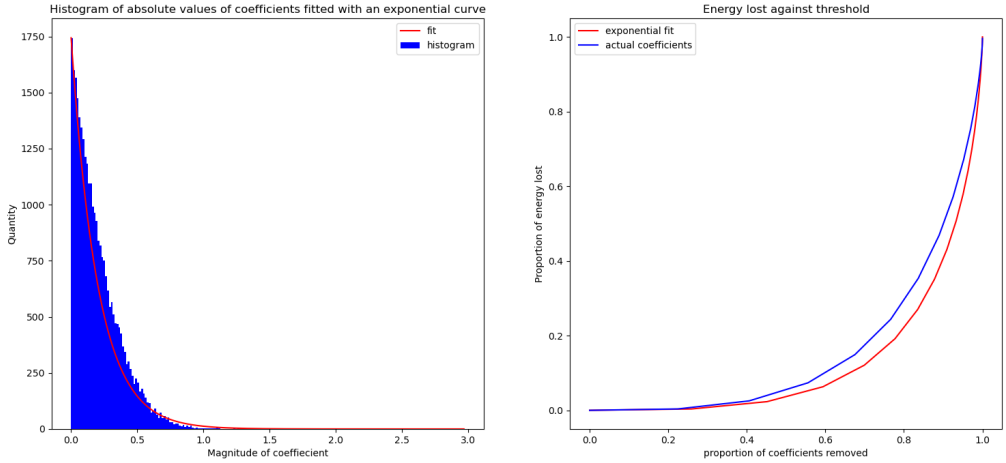

Figure 8. Data with recorded signal. On the left we have the distribution of absolute value of the coefficients from wavelet decomposition of data that had no visible signal recorded. Plotted on this distribution is an exponential curve fitted to the distribution. On the right is a plot of the loss of energy against the threshold percentile calculated from the coefficients (blue) compared with that calculated from the fitted exponential curve

that recorded mainly background noise. We see that the fit always underestimates the amount of coefficients. The error comparison shown in the image on the right in figure 8 also shows that estimated error from the fit always underestimates the actual error. Although in the case of the data with recorded activities, the fit can be used as an upper bound estimator for the errors, the fit in both cases could use some improvement.

We also compare how one and two dimensional wavelet decomposition perform when used to compress passive seismic data. We notice that the two dimensional case produces lower errors as shown in figure 3. In investigating the effects of wavelet compression on the frequency content in figure 4, we see that the frequency distribution in the original data (upper left) and the reconstructed compressed data (upper) is identical. The residual data (lower left) generally does not show any localized high amplitudes. This implies the compression does not remove selected frequencies which is a good thing if we want to retain the frequency content of our original data. Using basic STALTA, event detection in the data is not depleted even at compression as high as 96% for the two dimensional case and 90% for the one dimensional case (figure 5). However, there is more noise introduced in the one dimensional case.

7 ACKNOWLEDGEMENTS

We appreciate NSF Grant through NSF office of advanced cyberinfrastructure for funding this research. We also want to thank the Virginia Tech Advanced research computing for computing resources at the initial stages of this research. We thank the sponsor companies of the Center for Wave Phenomena.

REFERENCES

Bear, L. K., G. L. Pavlis, and G. H. R. Bokelmann, 1999, Multi-wavelet analysis of three-component seismic arrays: Application to measure effective anisotropy at Piñon Flats, California: Bulletin of the Seismological Society of America, **89**, 693–705.

Daubechies, I., 1992, Ten lectures on wavelets: Society for Industrial and Applied Mathematics.

Donoho, D., 1995, De-noising by soft-thresholding: IEEE Transactions on Information Theory, **41**, 613–627.

Donoho, D. L., and I. M. Johnstone, 1994, Ideal spatial adaptation by wavelet shrinkage: Biometrika, **81**, 425–455.

Dumont, V., V. R. Tribaldos, J. Ajo-Franklin, and K. Wu, 2020, Deep learning for surface wave identification in distributed acoustic sensing data: Presented at the 2020 IEEE International Conference on Big Data (Big Data), IEEE.

Hone, S., and T. Zhu, 2021, Seismic Observations of Four Thunderstorms Using an Underground Fiber-Optic Array: Seismological Research Letters, **92**, 2389–2398.

Kump, J. L., 2021, Efficient algorithms for data analytics in geophysical imaging: Master's thesis, Virginia Polytechnic Institute and State University.

Lindsey, N. J., and E. R. Martin, 2021, Fiber-optic seismology: Annual Review of Earth and Planetary Sciences, **49**, 309–336.

Mallat, S., 2008, A wavelet tour of signal processing: the sparse way: Academic press.

Schuster, G. T., J. Yu, J. Sheng, and J. Rickett, 2004, Interferometric/daylight seismic imaging: Geophysical Journal International, **157**, 838–852.

Simon, J. D., F. J. Simons, and G. Nolet, 2020, Multiscale Estimation of Event Arrival Times and Their Uncertainties in Hydroacoustic Records from Autonomous Oceanic Floats: Bulletin of the Seismological Society of America, **110**, 970–997.

Snieder, R., M. Miyazawa, E. C. Slob, I. Vasconcelos, and K. Wapenaar, 2009, A comparison of strategies for seismic interferometry: Surveys in Geophysics, **30**, 503–523.

Tibuleac, I. M., E. T. Herrin, J. M. Britton, R. Shumway, and A. C. Rosca, 2003, Automatic Determination of Secondary Seismic Phase Arrival Times Using Wavelet Transforms: Seismological Research Letters, **74**, 884–892.

Villasenor, J., R. Ergas, and P. Donoho, 1996, Seismic data compression using high-dimensional wavelet transforms: Proceedings of Data Compression Conference - DCC '96, 396–405.

Yoon, C. E., O. O'Reilly, K. J. Bergen, and G. C. Beroza, 2015, Earthquake detection through computationally efficient similarity search: Science Advances, **1**, e1501057.

Zhu, T., J. Shen, and E. R. Martin, 2021, Sensing earth and environment dynamics by telecommunication fiber-optic sensors: an urban experiment in pennsylvania, usa: Solid Earth, **12**, 219–235.