A near-lossless alternative for OBN data compression through block-sorting and decimation.

Francisco Carlos Lajús Junior, Stephan Paul
Universidade Federal de Santa Catarina - UFSC

This paper was prepared for presentation at the 18th International Congress of the Brazilian Geophysical Society, held in Rio de Janeiro, Brazil, 8 November to 11 November, 2023.

Abstract

When using OBN for seismic surveys, data recovery or transmission has to deal with large data volume and inefficient transmission channels, thus sustaining the development of compression schemes to reduce the storage amount and/or the information to be transmitted. In this work, we propose a near-lossless alternative for seismic data in OBN-type acquisition geometries, that does not require transmitted residues for signal recovery, as is normally carried out with classic predictive coding alternatives. Our proposal involves sorting a selected data chunk by their amplitude values and segmenting the signal into three partitions: negative, zeroed, and positive amplitudes. A non-uniform subset of samples from the negative and positive partitions are then selected to approximate the overall amplitude shape of the well-behaved monotonically increasing sorted signal, resulting in a significantly reduced number of data floating-points to be compressed. The sample selection process is determined by their corresponding positions with a stretching transform of Chebyshev-Lobatto points, which also allows for user-controlled clustering at a desired range of amplitudes or a varying recovery precision with fixed compression rate. Decompression is performed through interpolation, which introduces a type of relative error (or adaptive precision) along the signal, standing as a small percentage of the local amplitude values at places with reduced number of selected samples. The proposed method is found to achieve a compression ratio around 1.7, with a computationally low-cost compression/decompression alternative, displaying an R-squared (interpolated) value very close to unity.

Introduction

Predictive coding is a technique used in various fields to reduce the amount of data that needs to be stored or transmitted while preserving the most important information of the signal (Makhoul, 1975; Fout and Ma, 2012). Particularly in the context of seismic data compression, the application of predictive coding was a first-choice for the development of *lossless*-type compression algorithms to reduce the amount of data required to accurately represent a seismic signal (Stearns et al., 1993; McCoy et al., 1994; Stearns, 1995; Nijim et al., 1996; Mandyam et al., 1996), also considering ocean bottom sensors (Bordley, 1983). The underlying principle of such predictive coding approach is to use some previously known information of the signal to predict other sampled values. Specifically, classical predictive coding algorithms adopted in seismic data compression considered a causal-type framework, i. e., that previous samples of the seismic waveform can be used to predict the next samples. At a given discrete time t_n the difference between the predicted value and the actual value is then encoded, rather than the signal's value at t_n itself. By encoding the difference, along with some model parameters, the overall amount of data that needs to be stored and/or transmitted can be reduced. Usually, these residues are directed to a second stage, related to entropic coding (Stearns et al., 1993; Savazzi and Spagnolini, 2011; Payani et al., 2018), but this issue will not be discussed here, as we will restrict our attention strictly to the first stage, only related to prediction.

Typical least complex predictors include simply using the difference of previous sample values (Ahern et al., 2012), or a linear prediction with fixed coefficients determined *a priori* (McCoy et al., 1994; Nijim et al., 1996). Increased performance is expected by selecting optimized prediction coefficients, obtained for separated sample blocks (Stearns et al., 1993; Nijim et al., 2000). In some limiting cases, however, with more rigorous requirements regarding the computation power (say in deployed devices adopted in real-time sensor networks), these optimal approaches may not be applicable, and adaptive alternatives are preferred (Kiely et al., 2010). In this case, the prediction algorithm can adjust itself based on the characteristics of the signal being compressed, featuring lower computational complexity than fixed coefficient approaches (Magotra et al., 1995; Mandyam et al., 1996; Kiely et al., 2010). More recently, machine learning techniques have been applied to seismic data compression to improve the accuracy and efficiency of predictive coding methods (Payani et al., 2018; Nuha et al., 2019; Helal et al., 2021). However these are mostly focused on lossy-type of compression schemes. A lossless approach was proposed by Payani et al. (2019), with a two-dimensional Recurrent Neural Networks (RNN), with several workarounds to reduce the increased computational complexity of their proposal. Unfortunately, the obtained results are only briefly presented and discussed.

There is a somewhat potential trend to extend lossless approaches to a *near-lossless* compression option (i.e., lossy compression with user-defined absolute and/or relative error limits in the reconstructed signal), allowing substantially smaller compressed file sizes when a small amount of distortion can be tolerated (Campobello et al., 2021; Hernández-Cabronero et al., 2021; Lindstrom, 2014). Therefore, the goal of predictive coding is to achieve efficient data compression while preserving a considerable amount of the most important information in the signal. In the case of OBN data, this refers to the information that allows detection of important features at specific parts of a seismic signal, which are related to changes in subsurface geology.

In this work, we propose a near-lossless alternative for OBN seismic data compression, which considers data block-sorting and non-uniform selection of points for later signal decompression, that does not requires the typical transmission of floating point residues for proper signal recovery, but instead mostly integer values (indexes) that refers to their locations along the signal. The paper is organized as follows. First we describe the mathematical basis of the suggested approach. We then present the obtained results for synthetic (active source) and a realistic (passive source) OBN-type data. Lastly, we present the main conclusions.

Near-lossless compression through block-sorting and decimation

Here we describe the approach based on the data sorting and non-uniform selection of points (that for now on we term as decimation) for the compression phase, which is later recovered through interpolation and correct rearrangement of the signal amplitudes. In the following, we develop some quantitative notion for the minimal required information needed to recover the signal, compared with its original size.

Our approach is as follows: given N samples of the waveform seismic data, we can sort this samples by their increasing amplitudes, forming $N^- + N^0 + N^+ = N$ samples, which stands for N^- ordered values up to zero amplitude, N^0 zeros, and N^+ ordered samples with positive amplitudes. Additionally, the same amount of I indexes should be known, thus increasing the total N number of samples by $I^- + I^0 + I^+$ values necessary for the original signal recovery. It worth noting, however, that while there are $N^- + N^0 + N^+$ values related to a monotonically increasing amplitude (which are hoped to be considerable easier to predict than the original waveform) in one hand, on the other there are $I^- + I^0 + I^+$ indexes corresponding to a somewhat random sequence of integers, that are worst to compress but use less bits to be represented. The potential of such measure, therefore, turns out to depend on how good one is able to approximate two well-behaved segments with a significantly smaller number of floating points, while the sorted (integers) indexes must be accounted. Thus, to be truly applicable for data compression, the following inequality of sorted data volume must hold:

$$\overbrace{|N^- + N^0 + N^+ + I^- + I^0 + I^+|}^{sorted} < \overbrace{|N|}^{original}, \tag{1}$$

which can be further simplified. By knowing the total number of samples $\#N$ (a single floating point value, in this case) and the I^- and I^+ indexes, there is no need to store the I^0 indexes. Also, we avoid storage of/retention of N^0 zero amplitudes. So, to recover the signal

$$\overbrace{|\#N + N^- + N^+ + I^- + I^+|}^{sorted} < \overbrace{|N|}^{original}. \tag{2}$$

Here we foresee a possibility to explore this inequality, through non-uniform decimation at each monotonically increasing N^- and N^+ segment, significantly reducing the number of higher precision samples that should be considered for further decompression of the original signal.

This is pursued herein by a stretching map of ε_j Chebyshev-Lobatto points (Trefethen, 2019; Boyd, 2001),

$$\varepsilon_j = \cos\left(\frac{j\pi}{N_{A,B}-1}\right), \qquad j = 1, 2, ..., N_{A,B}-1; \tag{3}$$

transforming the amplitude segment (N^- or N^+), onto the Chebyshev interval $-1 \leq \varepsilon \leq 1$, where they are discretized in less collocation points (N_A or N_B, respectively). For the problem at hand, a suitable mapping function $\varepsilon(r)$ is considered to concentrate most collocation points within a user-defined amplitude region (Lesshafft and Huerre, 2007), with the following two-parameter transformation

$$r(\varepsilon) = r_c \frac{1-\varepsilon}{1-\varepsilon^2 + 2r_c/r_{max}}, \tag{4}$$

where approximately half of the points $r_j = r(\varepsilon_j)$ are placed in the interval $0 \leq r \leq r_c$, concentrated around $r = r_c/2$. Values of r_c, from $0.025 \rightarrow 0.1$, and $r_{max} = 1$ have been considered in the calculated results presented subsequently. This mapping provides an interesting compression potential, with a user controlled error related to the interpolation recovery. In other words, one can fix the number of points to be used to approximate the monotonically increasing positive and negative segments, but distribute these differently, thus having distinct errors on the decompression phase. The interpolation error may be seen as some sort of relative error (or adaptive precision), which displays a percentage of the local values found in the original signal, being also bounded by their limiting points in each segment. A way to overcome this issue, if a more strict criteria for the error is required, is to increase the number of samples around the regions of maximum absolute interpolation error. However, one must properly value this decision to carry on a fixed small error throughout the entire signal. *Is it necessary to ensure, say a 10^{-4} precision, in regions where amplitudes are varying at 10^5?*

In any case, here we are left with a reduced number of floating-points samples (and integer indexes) that are necessary to represent the negative $|N_A + I_A| \ll |N^-|$, and positive $|N_B + I_B| \ll |N^+|$ segments, resulting in

$$\overbrace{|\#N + N_A + N_B + I_A + I_B + I^- + I^+|}^{sorted+decimated} \ll \overbrace{|N|}^{original}. \tag{5}$$

For the signal recovery, first the N_A (N_B) samples, with their associated I_A (I_B) indexes, are interpolated along the I^- (I^+) indexes, which are then relocated into their original positions, in a vector of length N. So, here we expect to obtain compression when the total of $N_A + N_B$ floating points, along with $I_A + I_B + I^- + I^+$ integers for associated indexes, plus the knowledge of 1 floating point number ($\#N$), is found to be less then N floating points. It also worth noting that sorting and interpolation, normally present computational complexity of $\mathscr{O}(N)$ each, thus being computationally fast. Here, compression is associated with $\mathscr{O}(N)$ for sorting, while decompression takes $\mathscr{O}(N)$ for interpolation.

Results

Figure 1 presents two possible OBN scenarios for the evaluation of this proposed method. In the first case, pressure data of a synthetic simulation (active survey), from a single OBN receiver is gathered into a single trace, which is then sorted by their amplitude values, where positive and negative segments are approximated with a reduced selection of points. Reconstruction error is then shown, along a single stacked data trace. For the realistic OBN case, we consider the data from a passive PRM (Permanent Reservoir Monitoring) system deployed in 2012 at the Jubarte oil field (Goertz et al., 2015; Thedy et al., 2015; Bulcão et al., 2019), with a fully fiber-optic system deployed at 1300 meters water depth, with 712 four component (4C) receivers. Here we exemplify the results obtained for two components (25 traces of 15000 samples, for pressure and one displacement-related component).

Table 1 and 2 shows some of the obtained results with our proposal, after a final stage compression with Burrows-Wheeler transform (bzip2), respectively for synthetic and realistic OBN scenarios. Thus, for the synthetic case, original data X directly compressed with state-of-the-art WinRAR technology produces 2327 KB file size. For the sorted file X^*, we found a great improvement on file size reduction (768 KB). However the correct sample locations must be also accounted for the signal recovery, adding 1459 KB. Lastly, with the presented proposal (BS+D), $|N_A + I_A + N_B + I_B + \#N|$ returns a 5 KB file, $|I^-|$ a 630 KB file , and $|I^+|$ a 750 KB file; which in total results in a file with 60% of the original data size. The Compression Ratios (CR), and their inverse, are presented for each case.

Table 1: Compression ratio of synthetic OBN data.

	size (KB)	CR	1/CR
X	2327	1.000	1.000
X^* + indexes	2391	0.973	1.023
BS + D	1393	1.671	0.598

** sorted*

In the realistic OBN-type situation, having a passive source, original data of pressure (X_1) and a displacement-related (X_2) component, directly compressed with WinRAR technology produce 1386 KB and 1449 KB, respectively. Again, the proposed method is found to approximately reduce each file size to something close to 60% of the original.

Table 2: Compression ratio of realistic OBN data.

	size (KB)	CR	1/CR
X_1	1386	1.000	1.000
X_1^* + index	1386	1.000	1.000
BS + D (X_1)	825	1.680	0.595
X_2	1449	1.000	1.000
X_2^* + index	1472	0.984	1.015
BS + D (X_2)	873	1.659	0.602

** sorted*

A more detailed picture of the relative error behavior is shown in Figure 2. A closer look on the synthetic data case, for instance, exemplifies how the signal recovery is able to properly display even small amplitude events, such as those appearing right after the direct wave, although some level of error is tolerated at higher amplitude events. The same behavior is verified in the sample values for 2 components in Jubarte passive data, although in this mostly noisy case such values are more dispersed. Maximum relative amplitude errors of these databases are found to be, respectively, -0.0132, 0.0212 and 0.0377, occurring at high amplitude regions.

Conclusion

In this work we have presented preliminary results of an alternative approach for near-lossless compression in OBN-type seismic data. The proposal potential is closely related to the conversion of floating-point samples to integer indexes, as the sorted amplitudes can be significantly reduced by considering less number of data points. This selection was here explored through a two-parameter stretching map for Chebyshev-Lobatto grid points. Future directions are hoped to provide a more clear connection between the stretching function and the recovery errors, potentially leading to a better choice of the stretching parameters for an optimized compression, with improved compression ratios.

Acknowledgements

The authors are grateful to Alexandre Augusto Cardoso da Silva, who provided the OBN synthetic database and the Jubarte-PRM passive data considered in this work. This research was supported by Petrobras through “OBNZip - Compressor Inteligente de Dados Sísmicos para OBN” project with Universidade Federal of Santa Catarina (UFSC), and ANP through the R&D levy regulation, process: 2022/00099-6.

References

Ahern, T.; Casey, R.; Barnes, D.; Benson, R., and Knight, T., 2012, Seed reference manual–standard for the exchange of earthquake data: International Federation of Digital Seismograph Networks Incorporated Research Institutions for Seismology United States Geological Survey, version, **2**.

Bordley, T., 1983, Linear predictive coding of marine seismic data: IEEE transactions on acoustics, speech, and signal processing, **31**, 828–835.

Boyd, J. P., 2001, Chebyshev and fourier spectral methods: Courier Corporation.

Bulcão, A.; Alves, G.; Dias, B.; Soares Filho, D., and Cardoso da Silva, A., 2019, Methodologies for passive seismic event location: based on wave propagation: Presented at the 16th International Congress of the Brazilian Geophysical Society.

Campobello, G.; Quercia, A.; Gugliandolo, G.; Segreto, A.; Tatti, E.; Ghilardi, M. F.; Crupi, G.; Quartarone, A., and Donato, N., 2021, An efficient near-lossless compression algorithm for multichannel eeg signals: 2021 IEEE International Symposium on Medical Measurements and Applications (MeMeA), 1–6.

Fout, N. and Ma, K.-L., 2012, An adaptive prediction-based approach to lossless compression of floating-point volume data: IEEE Transactions on Visualization and Computer Graphics, **18**, 2295–2304.

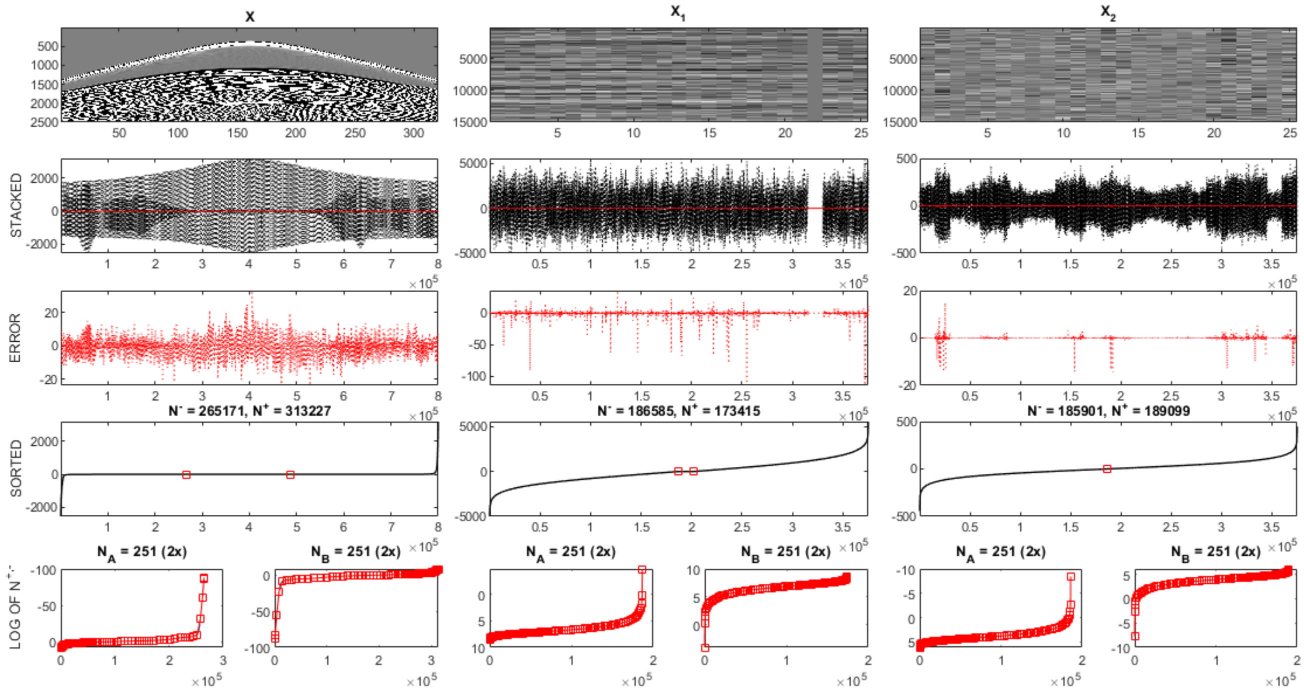

Figure 1: First row: (X) pressure results in a Common Receiver Gather (CRG) from the synthetic data, (X_1) sample of obtained pressure from Jubarte PRM pilot, and (X_2) a sample of the displacement-related component. Second row: the stacked version of each data displayed above, also showing the recovery error (red line). Third row: recovery error for each of the samples in the stacked data. Forth row: signal sorted by its amplitude values, with red markers delimiting the zero amplitude segment. The numbers of elements in N^- and N^+ segments are indicated. Fifth row: logarithmic scale for the amplitudes of each N^- and N^+ segments, with N_A and N_B markers indicating the selected collocation points considered for later signal recovery. Here, the number of points was fixed at 251 points, to all segments.

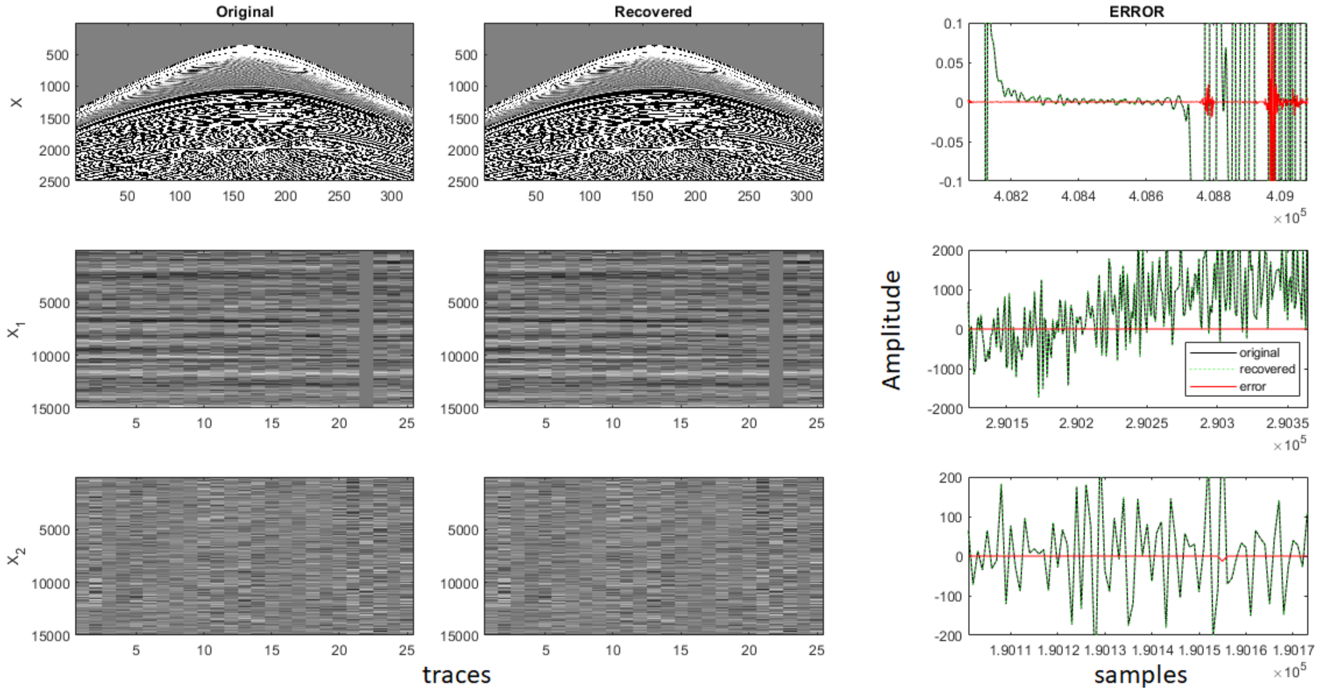

Figure 2: Original and recovered signals, with a detailed view of the relative error behavior of three randomly choosen parts of the signal. Original signals are displayed with solid black lines, recovered signal as dotted green, and error as solid red.

Goertz, A.; Richardson, J.; Faragher, J.; Remington, C.; Morton, P.; Barros, P., and Theodoro, C., 2015, Microseismicity recorded by a fiberoptic ocean-bottom prm installation offshore brazil: Third EAGE Workshop on Permanent Reservoir Monitoring 2015, 1–5.

Helal, E. B.; Saad, O. M.; Hafez, A. G.; Chen, Y., and Dousoky, G. M., 2021, Seismic data compression using deep learning: IEEE Access, **9**, 58161–58169.

Hernández-Cabronero, M.; Kiely, A. B.; Klimesh, M.; Blanes, I.; Ligo, J.; Magli, E., and Serra-Sagrista, J., 2021, The ccsds 123.0-b-2 ”low-complexity lossless and near-lossless multispectral and hyperspectral image compression” standard: A comprehensive review: IEEE Geoscience and Remote Sensing Magazine, **9**, 102–119.

Kiely, A.; Xu, M.; Song, W.-Z.; Huang, R., and Shirazi, B., 2010, Adaptive linear filtering compression on realtime sensor networks: The Computer Journal, **53**, 1606–1620.

Lesshafft, L. and Huerre, P., 2007, Linear impulse response in hot round jets: Physics of Fluids, **19**, 024102.

Lindstrom, P., 2014, Fixed-rate compressed floating-point arrays: IEEE transactions on visualization and computer graphics, **20**, 2674–2683.

Magotra, N.; McCoy, W.; Livingston, F., and Stearns, S., 1995, Lossless data compression using adaptive filters: 1995 International Conference on Acoustics, Speech, and Signal Processing, 1217–1220.

Makhoul, J., 1975, Linear prediction: A tutorial review: Proceedings of the IEEE, **63**, 561–580.

Mandyam, G.; Magotra, N., and McCoy, W., 1996, Lossless seismic data compression using adaptive linear prediction: IGARSS'96. 1996 International Geoscience and Remote Sensing Symposium, 1029–1031.

McCoy, J.; Magotra, N., and Stearns, S., 1994, Lossless predictive coding: Proceedings of 1994 37th Midwest Symposium on Circuits and Systems, 927–930.

Nijim, Y. W.; Stearns, S. D., and Mikhael, W. B., 1996, Lossless compression of seismic signals using differentiation: IEEE transactions on geoscience and remote sensing, **34**, 52–56.

———, 2000, Quantitative performance evaluation of the lossless compression approach using pole-zero modeling: IEEE transactions on geoscience and remote sensing, **38**, 39–43.

Nuha, H. H.; Balghonaim, A.; Liu, B.; Mohandes, M., and Fekri, F., 2019, Seismic data compression using deep neural network predictors: Presented at the SEG International Exposition and Annual Meeting.

Payani, A.; Abdi, A.; Tian, X.; Fekri, F., and Mohandes, M., 2018, Advances in seismic data compression via learning from data: Compression for seismic data acquisition: IEEE Signal Processing Magazine, **35**, 51–61.

Payani, A.; Fekri, F.; Alregib, G.; Mohandes, M., and Deriche, M., 2019, Compression of seismic signals via recurrent neural networks: Lossy and lossless algorithms, *in* SEG Technical Program Expanded Abstracts 2019, 4082–4086, Society of Exploration Geophysicists.

Savazzi, S. and Spagnolini, U., 2011, Compression and coding for cable-free land acquisition systems: Geophysics, **76**, Q29–Q39.

Stearns, S. D., 1995, Arithmetic coding in lossless waveform compression: IEEE Transactions on Signal Processing, **43**, 1874–1879.

Stearns, S. D.; Tan, L.-Z., and Magotra, N., 1993, Lossless compression of waveform data for efficient storage and transmission: IEEE Transactions on Geoscience and Remote Sensing, **31**, 645–654.

Thedy, E.; Dariva, P.; Ramos Filho, W.; Maciel Jr, P.; Silva, F., and Zorzanelli, I., 2015, First results on reservoir monitoring in jubarte prm-offshore brazil: Third EAGE Workshop on Permanent Reservoir Monitoring 2015, 1–5.

Trefethen, L. N., 2019, Approximation theory and approximation practice, extended edition: SIAM.