DOI: 10.61782/fa.2025.0663

FORUM ACUSTICUM EURONOISE 2025

OPTIMIZATION OF SPARSE SAMPLING AND INTERPOLATION FOR SEISMIC DATA COMPRESSION USING A GENETIC ALGORITHM

Nicolas Giovanardi[1*] **Lucas Lobato**[1] **Josafat Ribeiro**[2]
Stephan Paul[1]
[1] Laboratory of Vibration and Acoustics, Florianópolis, Brazil
[2] Software and Hardware Integration Laboratory, Florianópolis, Brazil

ABSTRACT

The acquisition of seismic data in the oil and gas industry generates large volumes of information, which presents substantial challenges for both storage and transmission. Efficient data compression is imperative to address these challenges while preserving data integrity. This study revisits the block-sorting and decimation method for near-lossless seismic data compression, which combines data sorting and non-uniform sparse sampling for data reduction, followed by linear interpolation and reordering for data reconstruction. Traditionally, interpolation points are selected using a stretching transform based on the extrema of Chebyshev polynomials. In contrast, this study proposes a data-driven optimization procedure that leverages a genetic algorithm to refine the selection of interpolation points. This approach exhibits high adaptability to the varying characteristics of different seismic signals. The performance of the proposed approach is evaluated using active seismic data from a permanent reservoir monitoring survey. Results demonstrate a significant reduction in data reconstruction error, albeit with increased computational cost due to the iterative optimization process. Overall, the findings highlight the potential of this approach for near-lossless seismic data compression in applications demanding high data integrity in the oil and gas industry.

Keywords: *data compression, sparse sampling, genetic algorithm*

**Corresponding author: nicolasgiovanardi@gmail.com.*

1. INTRODUCTION

Oil and gas exploration is based on geophysical surveying techniques to identify reservoirs, and among the different techniques employed, seismic reflection surveys are the most widely used [1]. In marine seismic reflection surveys, high-pressure air discharges generate elastic waves that propagate through the water column and into the marine subsurface, reflecting at different interfaces. These reflected waves are captured by sensors, namely hydrophones and geophones, which transmit data for further processing and interpretation. As the industry transitions from cable-based to wireless transmission, the massive volume of data generated during these surveys presents substantial challenges for both storage and transmission.

To address these challenges, numerous seismic data compression methods have been proposed in recent years. These methods can be categorized into lossless and lossy compression techniques. In lossless compression methods, data can be fully reconstructed with no loss of information. Conversely, in lossy compression methods, some information is discarded, and thus data cannot be fully reconstructed. Despite this limitation, lossy compression provides significantly higher data reduction, achieving compression ratios of up to 1000:1 through dimensionality reduction techniques [2]. Although this makes lossy compression highly valuable in many scenarios, lossless compression is still used when data integrity is a priority.

In this context, Junior and Paul [3] have proposed a near-lossless method for compression of seismic data. In the compression stage, the method involves sorting data in a monotonically increasing sequence and applying non-uniform sparse sampling. In the decompression stage, data is reconstructed through interpolation of the transmitted samples and reordering of the signal amplitudes, resulting in a reconstruction error that represents a small percentage of the signal amplitude values. Their findings, encompassing both synthetic active seismic data and real passive seismic data, demonstrate the method's high efficiency for seismic data compression, owing to its low computational cost and high data reconstruction accuracy.

This study extends the near-lossless seismic data compression method proposed by Junior and Paul [3] by integrating a genetic algorithm to optimize the selection of interpolation points within the sparse sampling procedure during the compression stage. Although the genetic algorithm optimization introduces additional computational cost, it significantly enhances data reconstruction accuracy, an essential feature in scenarios demanding high data integrity. Moreover, the optimization process is entirely data-driven, ensuring its adaptability to a wide range of seismic signals, as the optimal set of interpolation points varies across traces depending on the signal's amplitudes.

The proposed approach is evaluated using active seismic data provided by Petrobras S.A., demonstrating its applicability to real-world scenarios in the oil and gas industry. With fixing the genetic algorithm parameters, the optimization process is initially applied to a single trace for demonstration purposes and then extended to multiple traces from the same station. Results show that the optimized sparse sampling procedure outperforms the traditional method, yielding improved data reconstruction accuracy. This highlights the optimization procedure as a valuable extension to the original method, particularly in scenarios in which additional computational resources can be allocated to ensure data integrity during compression.

2. SPARSE SAMPLING AND INTERPOLATION

The block-sorting and decimation operation for seismic data compression, proposed by Junior and Paul [3], is a lightweight method that reduces the data size of each trace individually while operating entirely in the time domain. Instead of transmitting the original time-series signal composed of floating-point values, it applies non-uniform sparse sampling and interpolation to encode primarily integer values, achieving a compression ratio around 1.7 by reducing the number of bits required for transmission. As it maintains a relatively low data reconstruction error, the method is classified as near-lossless. This section details the compression and decompression stages of the method.

In the compression stage, the N signal samples are reordered into a monotonically increasing sequence, transforming the signal into a well-behaved curve. This step is essential, as the goal is to interpolate the curve using a limited number of points. For accurate data reconstruction, the original sample indices must be transmitted alongside the j selected interpolation points. These points x are distributed within the interval $[-1, 1]$ using non-uniform sampling based on the extrema of the Chebyshev polynomial, as defined by Eqn. (1) [4]. Since the original signal may span a different interval, the selected points have to be adjusted accordingly to fit the signal's actual interval.

$$x_j = \cos\left(\frac{j\pi}{N}\right), \quad j = 0, \dots, N-1. \tag{1}$$

In the decompression stage, the j transmitted interpolation points are linearly interpolated to recover all N signal samples, while the original sample indices are used to restore the correct ordering. As a result, the reconstructed signal consists of j original samples and $N - j$ interpolated samples, which introduce a reconstruction error. Although it would be possible to transmit only the amplitudes of the interpolation points, along with the number of points, leveraging the known Chebyshev polynomial extrema to infer their indices, this approach was not considered probably to avoid additional computational cost. Therefore, both the indices and amplitudes of the interpolation points are transmitted during the compression stage.

With this procedure, $N + j$ integer values (indices of the original signal and indices of the interpolation points) are transmitted along with j floating-point values (amplitudes of the interpolation points), instead of transmitting the full signal composed of N floating-point values. Assuming $j \ll N$, data reduction is achieved, as integer values require fewer bits for transmission compared to floating-point values. It is important to note that, logically, a smaller number of interpolation points j results in an increased data reduction but also leads to a higher reconstruction error, as the interpolation points would be more spaced apart. A schematic overview of both the compression and decompression stages are illustrated in Fig. 1.

Since specific intervals of the signal may exhibit higher reconstruction errors due to nonlinearities in the sorted curve, a suitable mapping function, defined by Eqn. (2), can be employed during compression to concentrate interpolation points in targeted intervals of the sorted trace [5]. In particular, approximately half of the points, $r_j = r(x_j)$, are placed within the interval $0 \leq r \leq r_c$, with more points around $r = r_c/2$. This mapping function enables a user-controlled data reconstruction error by al-

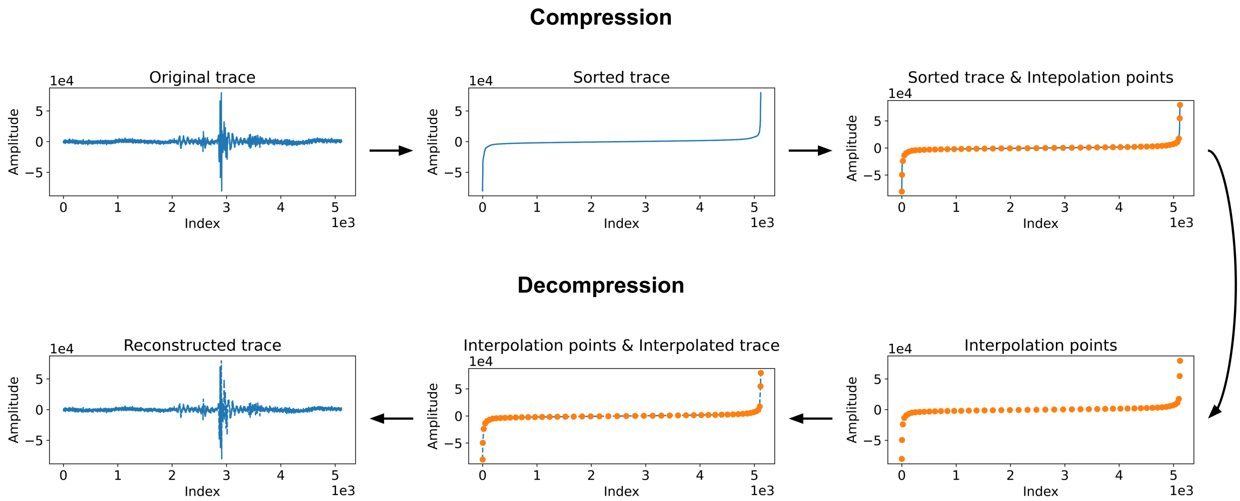

Figure 1. Compression and decompression stages.

lowing the spatial distribution of interpolation points to be adjusted, while keeping the total number of points fixed.

$$r(x) = r_c \frac{1-x}{1-x^2+2r_c/r_{\max}}, \quad (2)$$

3. OPTIMIZATION PROCEDURE

Interpolation points distributed according to the Chebyshev polynomial extrema follow a general procedure, providing a robust signal-independent pattern that yields accurate data reconstruction. To further enhance data reconstruction accuracy, the placement of these points can be refined individually for each trace, as they depend on signal amplitudes and thus vary across different traces. Although the mapping function allows for adjustments, this study aims to determine the optimal or near-optimal interpolation points for each trace independently. As this constitutes a nonconvex optimization problem, a genetic algorithm is used [6], providing a data-driven approach.

The genetic algorithm, introduced in 1975 [7], is a metaheuristic inspired by the process of natural evolution. It iteratively refines a population of candidate solutions based on a fitness function. In each iteration, known as a generation, the fittest solutions are selected from the current population, and the genetic operators of selection, crossover, and mutation are applied to create the next generation. Through this evolutionary process, the algorithm continuously improves solutions, yielding better fitness values. The process ends when a predefined number of iterations is reached or a specific fitness threshold is met.

In this study, the population of candidate solutions consists of a set of interpolation points. In each iteration, these candidates are used to interpolate the sorted trace and are then evaluated based on the mean absolute error (MAE), an error metric associated with the L^1 norm, calculated between the original samples and the interpolated ones. To form the next generation, a selected proportion of the best-performing interpolation points, determined by their fitness, are retained and used to generate new solutions through genetic operations. The L^1 norm is chosen over the L^2 norm due to its robustness against outliers [8].

Genetic algorithm optimization may not always converge to a local or global minimum, as there are no theorems guaranteeing this, and the process can be computationally demanding. Therefore, the interpolation points determined by the Chebyshev polynomial extrema can be used for creating the initial population of candidate solutions, as they already provide good data reconstruction results. This is done by introducing slight deviations, that is, small shifts in the points determined by the Chebyshev polynomial extrema, so that the algorithm starts with a near-optimal set of interpolation points and iteratively refine them according to the characteristics of each trace.

Therefore, this refinement process, based on the initial positioning of interpolation points proposed by Junior and Paul [3], requires fewer generations to enhance the results than it normally would starting with a randomly generated initial population. Moreover, the optimization procedure is highly configurable, allowing the user to adjust parameters such as population size, number of generations, and the initial population itself to achieve variable outcomes. This flexibility provides a balance between data reconstruction accuracy and computational cost, enabling the approach to be tailored to specific application requirements and the available computational resources.

4. RESULTS

For the development and evaluation of the approach proposed in this study, Petrobras S.A. provided an active seismic datasets acquired during a permanent reservoir monitoring survey conducted at the Jubarte field between 2013 and 2014. These datasets were recorded as common receiver gather stations and stored in SEG-Y files containing traces of sound pressure and three-dimensional acceleration. Each trace has a duration of 10.24 seconds and was acquired at a sampling rate of 500 Hz, resulting in 5120 samples per trace. For the purposes of this study, station 258 was randomly selected to assess the data reconstruction accuracy of the proposed optimization procedure.

The genetic algorithm was configured with an initial population size of 100, a total of 1000 generations, and a mutation rate of 0.1. For demonstration purposes, the first sound pressure trace from station 258 was selected to illustrate the optimization outcomes, using 50 interpolation points. Fig. 2 compares the interpolation points obtained through Chebyshev polynomial extrema with those obtained through genetic algorithm optimization, plotted alongside the sorted trace. Fig. 3 presents the evolution of fitness values (measured by the MAE between original and interpolated curves) over generations. Finally, Fig. 4 illustrates the original and reconstructed traces, along with the absolute error, indicating near-zero information loss.

To further evaluate the approach on a broader scale while using the same genetic algorithm parameters, the first 441 sound pressure traces from station 258 were selected for analysis, as they constitute a single receiver gather. Fig. 5 compares the MAE of the Chebyshev polynomial extrema-based sparse sampling with that of the optimized sparse sampling using a genetic algorithm, showing that the optimization significantly enhances data reconstruction across all analyzed traces. Fig. 6 presents the original traces and their respective reconstructions using the optimized interpolation points, confirming the high reconstruction accuracy achieved by the proposed approach.

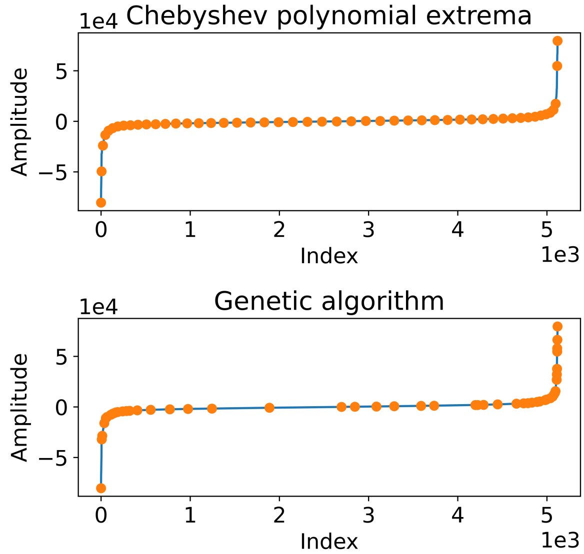

Figure 2. Sparse sampling procedure comparison.

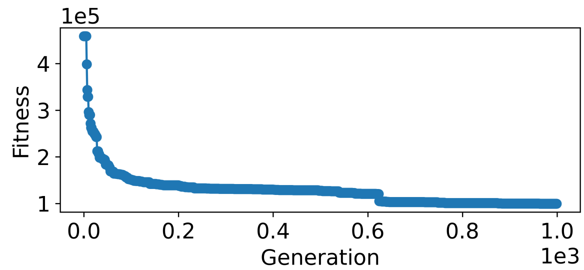

Figure 3. Evolution of fitness.

The results demonstrate that incorporating optimization through a genetic algorithm not only significantly improves data reconstruction accuracy but also reduces the variability of reconstruction error across traces. It is also important to note that the proposed approach does not impact the compression ratio of the original near-lossless seismic data compression method, which remains approximately 1.7. Moreover, as illustrated in Fig. 2, the genetic algorithm allocates interpolation points in regions of nonlinear behavior, exactly where linear interpolation is more likely to introduce errors, thereby highlighting the adaptability of the approach to varying trace characteristics.

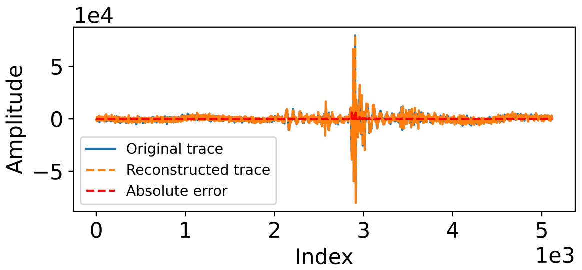

Figure 4. Single trace reconstruction.

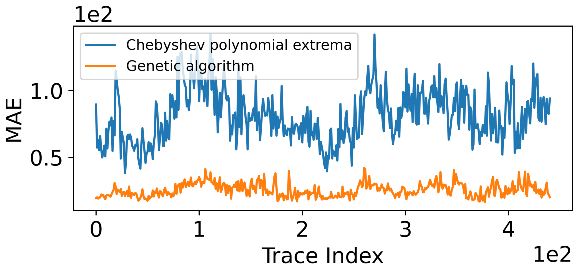

Figure 5. Reconstruction MAE across traces.

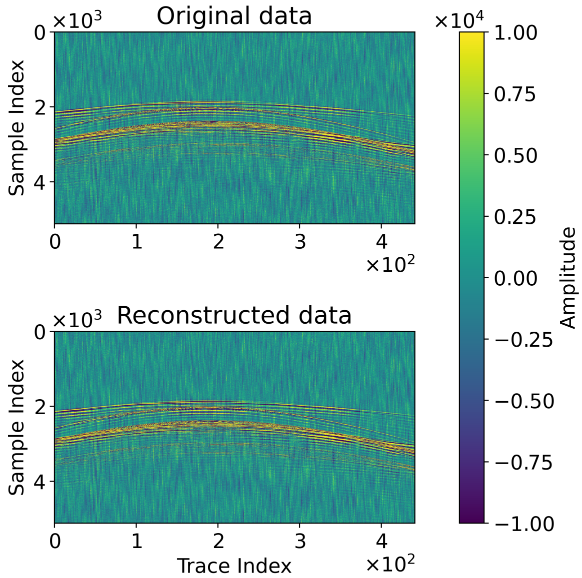

Figure 6. Multiple traces reconstruction.

5. CONCLUSION

This study proposed an extension to a near-lossless seismic data compression method by integrating a genetic algorithm to refine the selection of interpolation points within the sparse sampling procedure. This extension significantly enhanced data reconstruction accuracy compared to the original method, although it introduced additional computational demand. As a data-driven approach, it also demonstrated high adaptability to varying signal characteristics. Future research is expected to explore more efficient near-lossless data compression strategies that can further enhance both data reconstruction accuracy and compression ratio, without substantially increasing computational cost, thus broadening the method's applicability to real-time or resource-constrained environments.

6. REFERENCES

[1] W. A. Mousa, *Advanced Digital Signal Processing of Seismic Data*. Cambridge: Cambridge University Press, 2020.

[2] M. Wagner, M. M. de Lucena, J. L. Ribeiro, and A. A. Fröhlich, "Combining pca and dct for improved passive seismic data compression," *IEEE Access*, vol. 12, pp. 170648–170660, 2024.

[3] F. C. L. Junior and S. Paul, "A near-lossless alternative for obn data compression through block-sorting and decimation," in *International Congress of the Brazilian Geophysical Society*, (Rio de Janeiro), 2023.

[4] D. Gottlieb and S. A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*. Philadelphia: Society for Industrial and Applied Mathematics, 1977.

[5] L. Lesshafft and P. Huerre, "Linear impulse response in hot round jets," *Physics of Fluids*, vol. 19, no. 2, p. 024102, 2007.

[6] S. Sargsyan, S. L. Brunton, and J. N. Kutz, "Online interpolation point refinement for reduced order models using a genetic algorithm," *SIAM Journal on Scientific Computing*, vol. 40, no. 1, pp. B283–B304, 2018.

[7] J. H. Holland, *Adaptation in Natural and Artificial Systems*. Michigan: University of Michigan Press, 1975.

[8] S. L. Brunton and J. N. Kutz, *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge: Cambridge University Press, 2022.