An Analysis of the Integer Wavelet Transform as a Decorrelation Strategy for Lossless Seismic Data Compression

Julio Nobre Lopes*[1], Renata Nascimento[1], [1]Tecgraf Institute/PUC-Rio

Abstract

Recent decades have seen an exponential increase in the amount of data to be collected, stored, transmitted and processed. In particular with regard to seismic data, data compression is becoming an indispensable tool to meet this challenge. In many cases, lossless compression of seismic data is indicated or even necessary. Many ideas and solutions are being proposed for data compression in all areas. In this context, the present study deals with lossless seismic data compression, analyzing the use of integer wavelet transforms (IWT) as a decorrelation strategy in order to obtain better compression results. The study compares the variation in entropy and compression ratio (CR) between the original samples and wavelet coefficients for a set of Cohen–Daubechies–Feauveau family's wavelets. The effectiveness of the IWT at the decorrelation stage is analyzed for seismic volumes in combination with traditional and modern encoding algorithms.

Introduction

The increase in the efficiency of acquisition and the quality of data representation has facilitated the high availability of massively large data, which is becoming more frequent and essential in the most diverse contexts of seismic data processing and analysis. Additionally, in a scenario where communication and processing are already distributed, transferring information between processing entities must be carried out efficiently, ensuring the integrity of the data of interest.

Although several compression techniques are available in the literature, lossless techniques are strongly desired since for various seismic applications, losses may be unacceptable.

In general, a lossless compression scheme has two main steps: decorrelation (or transformation) and encoding. Among decorrelation strategies, the wavelet transform is commonly used (Stigant et al., 1995, Khène et al., 2000) since mapping a signal in the time domain to a time-frequency representation can often lead to lower entropy and effectively achieve better compression ratios. Although efficient, this approach cannot be widely used once it is usually lossy. To avoid losses, an integer wavelet transform (IWT) turns out to be a suitable approach, as described in Giurcaneanu et al. (1999). The IWT of integer-valued signals decomposes the signal into integer coefficients and the inverse integer wavelet transform can then be applied to these coefficients, recovering the original data without loss.

Concerning the coding step, a wide variety of algorithms are available, ranging from traditional methods such as Huffman and Arithmetic coding, to more sophisticated strategies which are employed in modern compression software (Gupta et al., 2017).

The IWT with wavelet CDF (2, 2) has already shown effectiveness as a decorrelation strategy for seismic data, through multiple compression algorithms (Lopes et al., 2021). The present work investigates several wavelets from the Cohen–Daubechies–Feauveau family, evaluating how entropy and CR change as the wavelet's numbers of vanishing moments increase.

Theory

The Wavelet Transform

Mallat (1989) introduces the concept of multiresolution analysis showing that the difference of information between approximations of a signal at levels of resolution 2^{j+1} and 2^j, for $j \in Z$, can be extracted by decomposing the signal on a wavelet orthonormal basis. Computation of the wavelet representation may then be accomplished by convolutions with quadrature mirror filters. The wavelet transform can therefore be performed via successive applications of a 2-channel filter bank, as illustrated in Figure 1, consisting of a high-pass filter (i.e., wavelets) and a low-pass filter (i.e., scaling functions): the detail coefficients (resulting from application of the high-pass) of every iteration step are kept apart, and the iteration starts again with the remaining approximation coefficients (from application of the low-pass filter).

Scaling functions and wavelets are related through equations:

$$\phi(t) = \sqrt{2} \sum_k h(k)\phi(2t - k) \text{ and}$$

$$\psi(t) = \sqrt{2} \sum_k g(k)\phi(2t - k),$$

where $h(k)$ and $g(k)$ are the quadrature mirror filters (low-pass and high-pass, respectively).

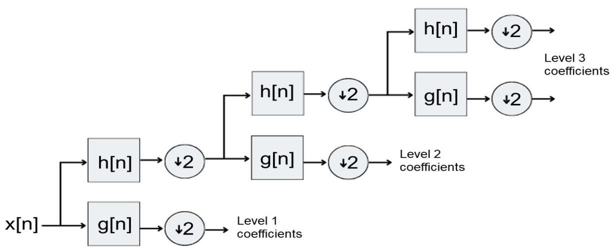

Figure 1: Three level filter bank for DWT decomposition.

Biorthogonal Wavelets

A requirement for wavelets to be orthogonal can place strong limitations on the system (Burrus et al., 1998), such as preventing linear phase analysis and synthesis filter banks. Biorthogonal wavelet systems using a nonorthogonal basis and dual basis are more flexible and free from this constraint.

For a 2-channel biorthogonal filter bank, the relationship between scaling functions and wavelets is given by:

$$\phi(t) = \sqrt{2}\sum_k h(k)\phi(2t - k)$$

$$\psi(t) = \sqrt{2}\sum_k g(k)\phi(2t - k)$$

$$\tilde{\phi}(t) = \sqrt{2}\sum_k \tilde{h}(k)\, \tilde{\phi}(2t - k)$$

$$\tilde{\psi}(t) = \sqrt{2}\sum_k \tilde{g}(k)\, \tilde{\phi}(2t - k),$$

where and $\tilde{\phi}(t)$ and $\tilde{\psi}(t)$ are the dual scaling functions and wavelets, respectively.

Lifting Scheme

Introduced by Sweldens (1998), lifting is a general framework for constructing second generation wavelets (Daubechies and Sweldens, 1998). It also provides a natural framework for constructing reversible integer wavelet transforms for lossless compression of integer samples.

The previous sections described the wavelet transform of a signal as a multiresolution representation of that signal with wavelets as basis functions. At each level of resolution, the (low-pass part of the) signal is split into a high-pass and a low-pass part. Lifting provides an efficient implementation of these filtering operations (Uytterhoeven et al., 1999).

Let λ_{j+1} be a dataset at level of resolution $j + 1$. This set is to be transformed into two other sets at level j: the low-resolution part λ_j and the high-resolution part γ_j. First, λ_{j+1} is split into subsets of even and odd samples. Next, the two subsets are recombined in subsequent lifting steps: prediction and update.

The three stages of lifting can be summarized as follows:

Splitting (lazy wavelet transform): Split dataset λ_{j+1} into distinct datasets λ_j and γ_j.

Prediction (dual lifting): Predict values of γ_j from values of λ_j and store prediction error: $\gamma_j = \gamma_j - P(\lambda_j)$.

Update (primal lifting): Update values of λ_j with values from γ_j: $\lambda_j = \lambda_j + U(\gamma_j)$.

Mapping Integers to Integers

The steps above can be easily modified to create a transform that maps integers to integers and is still reversible (Uytterhoeven et al., 1997). This is done by adding rounding operations (indicated by curly braces), at the expense of introducing non-linearity in the transform. The changes in the prediction and update steps of the forward transform are, respectively:

$$\gamma_j = \gamma_j - \{P(\lambda_j)\} \text{ and } \lambda_j = \lambda_j + \{U(\gamma_j)\}.$$

Naturally, the inverse update and prediction steps become:

$$\lambda_j = \lambda_j - \{U(\gamma_j)\} \text{ and } \gamma_j = \gamma_j + \{P(\lambda_j)\},$$

which assures the reversibility of the transform.

Cohen-Daubechies-Feauveau Wavelets

The Cohen–Daubechies–Feauveau (CDF) wavelets are a widely used family of biorthogonal wavelet bases. The nomenclature within the family (e.g., CDF 2,2) refers to the numbers of vanishing moments.

A wavelet basis function ψ has N vanishing moments if:

$$\int_{-\infty}^{\infty} t^p \psi(t)dt = 0,$$

for $0 \le p < N$. This property can also be translated into a certain kind of ”blindness” of the basis function with respect to polynomials up to degree $N - 1$ (Jansen and Oonincx, 2005).

For biorthogonal wavelet systems with basis functions $\psi_{j,k}$ and $\tilde{\psi}_{jk}$, the series expansion of a signal *s* is given by:

$$s(t) = \sum_j \sum_k \langle s, \tilde{\psi}_{jk} \rangle \psi_{jk}$$

So, signals that are well approximated by polynomials will produce small coefficients if the dual basis functions have a sufficiently high number of vanishing moments.

Data Compression

Data compression involves encoding information using fewer bits than the original representation. Two main categories are commonly used when describing data compression in general: Lossy and lossless.

Lossy vs Lossless

Lossy techniques allow for some information loss. Data that has been compressed that way will likely not be recovered exactly. In many applications, exact reconstruction is not necessary. When transmitting speech for example, the exact value of each audio sample is not required as some loss in quality may be imperceptible to the human ear. In general, higher compression ratios (CRs) can be obtained from lossy compression.

Lossless compression techniques involve no loss of information. If data has been losslessly compressed, the original data can be recovered exactly.

Compression algorithms often consist of two conceptually separated stages: data decorrelation (or transformation) and encoding. A lossy algorithm may also include a stage for quantization, as illustrated in Figure 2.

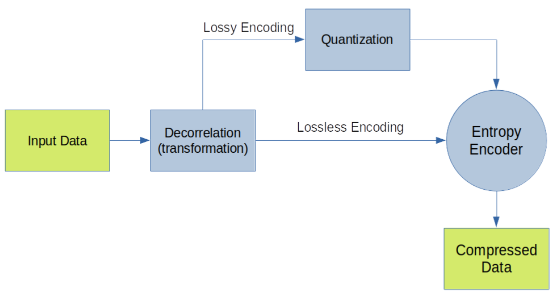

Figure 2: A general schema for data compression.

The decorrelation stage exploits the fact that neighboring samples tend to have similar values. Decorrelated samples normally exhibit less variability, i.e., lower entropy rates. The encoding stage then compacts decorrelated samples in a reversible way. The output compressed data size depends directly on the entropy of those samples, hence the importance of effective decorrelation (Hernández-Cabronero et al. 2020).

Compression Algorithms

Numerous techniques have been developed for data compression over the last decades. A general description of traditional coding algorithms, such as Huffman coding, Arithmetic coding and dictionary based methods, can be found in Uthayakumar et al. (2018). Most modern algorithms employ strategies that consist of a combination of classical techniques and often outperform them in both compression ratio and speed (Gupta et al., op. cit.).

Examples of such algorithms are Zstandard, TurboPFor and ZLib.

Zstandard uses two types of entropy encoding: Finite State Entropy (FSE) and Huffman coding (Zstandard, 2016).

Turbopfor is a family of integer compression codecs based on the Patched Frame of Reference (PFOR) strategy (TurboPFor, 2023).

ZLib is based on the DEFLATE method, which combines Huffman coding and LZ77 compression (ZLib, 2022).

Results

In this study, we investigate the entropy and compression ratios for the original samples and wavelet coefficients of four different datasets, using wavelets of the Cohen–Daubechies–Feauveau family.

Five compression strategies were used to calculate the CR: ZStandard (ZSTD), TurboPFor (TPFOR), ZLib (ZLIB), Huffman coding (HUFF) and Arithmetic coding (ARITH).

The seismic datasets used were: Nornefield (Norway), Waka-3D (New Zealand), Tui-3D (New Zealand) and Penobscot (Canada). These datasets are open geophysical data, free and available to the public.

Six wavelets of the CDF family were evaluated. These wavelets have an equal number of vanishing moments for the synthesis and analysis, as biorthogonal filters of more similar length are obtained when the number of vanishing moments is the same (Mallat, 2009) and, for data compression, filters of similar lengths are generally preferred (Van Fleet, 2019).

For each dataset and wavelet, coefficients were calculated and a representation of the dataset associated with that particular wavelet was produced. Each representation was then encoded through different techniques. Figure 3 illustrates this process for a single dataset.

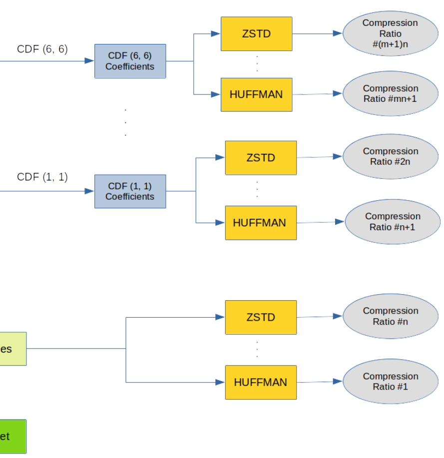

Figure 3: Process for calculating CR from dataset's samples and wavelet coefficients.

The change in entropy between original samples and wavelet coefficients for an increasing number of vanishing moments is shown in Figure 4.

Entropy was calculated as:

$$H(X) =-\sum p(x)log_2p(x)$$

where $p(x)$ is the probability of each specific symbol in the dataset (Cover and Thomas, 1991).

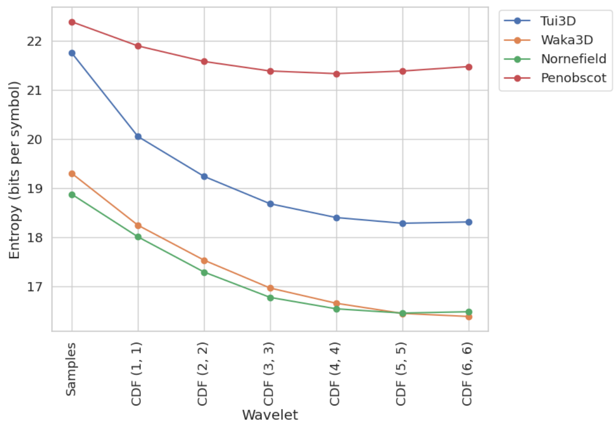

Figure 4: Entropy by wavelet with increasing numbers of vanishing moments.

Applying the IWT has led to a lower entropy representation of the data in all cases. The four datasets show a similar pattern of reduction in entropy following the progressive increase in the number of vanishing moments. The curves tend to a horizontal line or may even invert the tendency, as is clearly the case of Penobscot (with turning point at CDF (4, 4)). Nornefield and Tui-3D reached minimum entropy with CDF (5, 5). Waka continued to improve (although marginally) within the interval of vanishing moments investigated.

Figures 5-8 show compression ratios for samples and wavelet coefficients, through different compression algorithms.

The compression ratios were calculated as CR% = 100 * (Size after compression / Size before compression), following Salomon and Motta (2010).

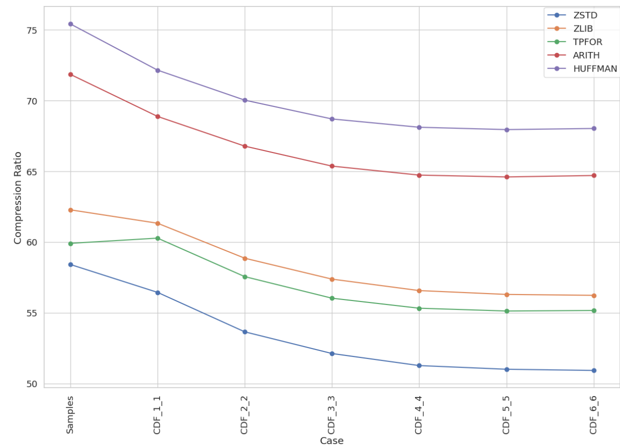

Figure 5: Nornefield - CR by wavelet and encoder.

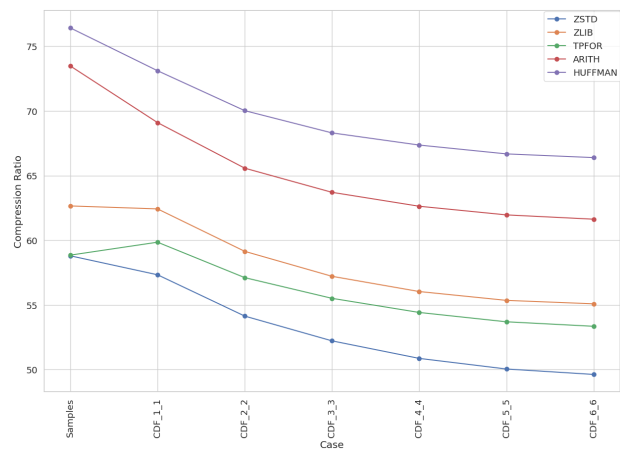

Figure 6: Waka-3D - CR by wavelet and encoder.

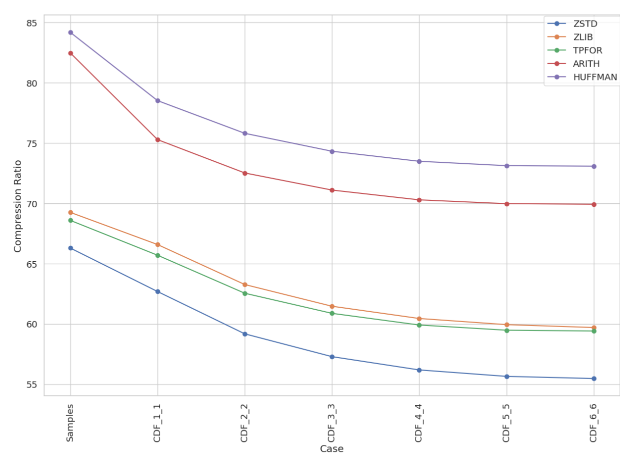

Figure 7: Tui-3D - CR by wavelet and encoder.

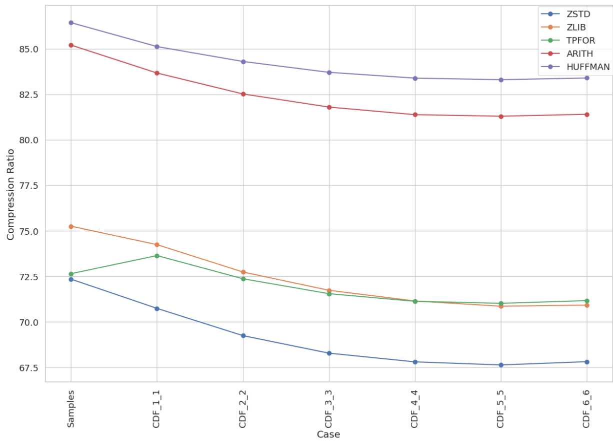

Figure 8: Penobscot - CR by wavelet and encoder.

Compression algorithms produced better CR for lower entropy representations of the data in general. Results show a common pattern of improvement in CR similar to the evolution of entropy previously discussed. Despite the similar shape of the curves for different coders, they are distributed in different levels of compression, due to the difference in efficiency between encoding techniques.

Some characteristics are common to the four datasets: the two traditional techniques HUFF and ARITH are always less efficient with HUFF as the least efficient followed by ARITH. ZSTD is the most efficient, producing the best CR in all cases.

Table 1 shows the CR obtained with compressor ZSTD for each dataset's original samples and CDF (5, 5) coefficients, and the difference in percentage points between them.

Dataset	CR - Samples	CR - CDF (5, 5)	Difference in percentage points
Nornefield	58.41	51.00	7.41
Tui-3D	66.29	55.64	10.65
Waka-3D	58.80	50.04	8.76
Penobscot	72.34	67.63	4.71

Table 1: Comparison between CR of original samples and CDF (5, 5) coefficients with ZSTD

Although the CR of all datasets improved from application of the IWT with wavelet CDF (5, 5), the gain varies significantly between datasets. Penobscot benefited the least, improving only 4.71 percentage points, while Tui-3D benefited the most, with a difference of 10.65 percentage points.

Conclusions

The results of this study show that the Integer Wavelet Transform with CDF wavelets have produced lower entropy representations of the seismic datasets and better CRs in comparison to the original data. Choosing wavelets with an incrementing number of vanishing moments caused a progressive reduction in entropy and compression was improved.

This reduction, however, was not linear, with its decrease becoming progressively less significant as the number of vanishing moments grew (in some cases, eventually reversing direction). The reason for this trend reversal is possibly due to the trade-off between important properties of CDF wavelets, which are the number of vanishing moments and support size (Pereyra and Ward, 2012).

Finally, these results confirm that the Integer Wavelet Transform is a suitable decorrelation strategy for lossless compression of seismic data.

Acknowledgments

The authors would like to thank Petrobras for the financial support.

References

BURRUS, C.S.; GOPINATH, R.A.; GUO, H. 1998. Introduction to Wavelets and Wavelet Transform - A Primer, Recherche 67, Jan. 1998.

COVER, T.M.; THOMAS, J.A. 1991, Elements of Information Theory. John Wiley & Sons, Inc, 1991.

DAUBECHIES, I.; SWELDENS, W. 1998, Factoring wavelet transforms into lifting steps. The Journal of Fourier Analysis and Applications. Volume 4, Issue 3, 1998.

GIURCANEANU, C.; TABUS, I.; ASTOLA, J. 1999. Integer wavelet transform based lossless audio compression, Proc. of the IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing, Jun. 1999, 378-382.

GUPTA, A.; BANSAL, A.; KHANDUJA, V. 2017. Modern Lossless Compression Techniques: Review, Comparison and Analysis. In: 2nd IEEE International Conference on Electrical, Computer and Communication Technologies, vol. 3, Feb. 2017,.

HERNÁNDEZ-CABRONERO, M.; PORTELL, J.; BLANES, I.; SERRA-SAGRISTÀ, J. 2020. High-Performance Lossless Compression of Hyperspectral Remote Sensing Scenes Based on Spectral Decorrelation. Remote Sensing; 12(18):2955, 2020.

JANSEN, M.; OONINCX, P. 2005. Second Generation Wavelets and Applications, Springer-Verlag London Limited, 2005.

KHÈNE, M.F.; ABDUL-JAUWAD, S.H. 2000. Efficient seismic compression using the lifting scheme. In: Annual

International Meeting, Soc. of Expl. Geophysicists, Exp. abstracts, 2052-2054, Aug. 2000.

LOPES, J. N.; FREDERICK, P.; IVISON, P.; NASCIMENTO, R. 2021. A comparative study of the integer wavelet transform applied to seismic data lossless compression. In: 17th International Congress of the Brazilian Geophysical Society and Exposition, Nov 2021.

MALLAT, S. 1989. A theory for multiresolution signal decomposition: the wavelet representation. In IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 11, no. 7, 674-693, July 1989.

MALLAT, S. 2009. A Wavelet Tour of Signal Processing - The Sparse Way. Third Edition. Elsevier, 2009.

PEREYRA, M.C.; WARD, L.A. 2004, Harmonic analysis - From Fourier to wavelets. Student Mathematical Library IAS/Park City Mathematical Subseries, vol 63, 2004.

SALOMON, D.; MOTTA, G. (2010), Handbook of data compression. Springer-Verlag London Limited 2010.

STIGANT, J. P.; ERGAS, R. A.; DONOHO, P. L.; MINCHELLA, A. S.; GALIBERT, P. Y. 1995. Field trial of seismic compression for real time transmission. In: Annual International Meeting, Soc. of Expl. Geophysicists, Exp. abstracts, 960-962, Oct. 1995.

SWELDENS, W. 1998. The lifting scheme: a construction of second generation wavelets, SIAM J. Math. Anal. Vol. 29, N. 2, pp. 511-546.

TURBOPFOR (2023), TurboPFor-Integer-Compression. Available in: https://github.com/powturbo/TurboPFor-Integer-Compression

UTHAYAKUMAR, J.; VENGATTARAMAN, T.; DHAVACHELVAN, P. 2018. A Survey on Data Compression Techniques: From the Perspective of Data Quality, Coding Schemes, Data Type and Applications. Journal of King Saud University - Computer and Information Sciences, vol 33, 2018.

UYTTERHOEVEN, G.; ROOSE, D.; BULTHEEL, A. 1997. Wavelet transforms using the lifting scheme. ITA-Wavelets Report WP, 1, 1997.

UYTTERHOEVEN, G.; ROOSE, D.; BULTHEEL, A. 1999. A Wavelet Toolbox for Large Scale Image Processing. In: Zinterhof, P., Vajteršic, M., Uhl, A. (eds) Parallel Computation. Lecture Notes in Computer Science, vol 1557, 1999.

VAN FLEET, P.J. 2019. Discrete Wavelet Transformations - An Elementary Approach with Applications. Second Edition. John Wiley & Sons, Inc. 2019.

ZLIB (2022), Zlib Data Compression Library. Available in: https://zlib.net/

ZSTANDARD, 2016. Zstandard - Fast real-time compression algorithm. Available in: https://facebook.github.io/zstd/