How should the "quality" of seismic well ties and predictions of rock properties using machine learning be assessed?

John Castagna[1], Vi Kronberg[2], and Gabriel Gil[2]

https://doi.org/10.1190/tle-2025-1058

Abstract

The correlation coefficient has inadequacies as a metric for comparison of seismic data to presumed truth in the form of synthetic seismograms or well logs. It is not a direct measure of the information content of a prediction, nor the shared information between the prediction and perfect data. All else being equal, the information content (useful or not) of seismic data decreases monotonically with decreasing bandwidth, but the correlation coefficient may vary differently. Quantities from information theory, such as cross entropy, variation of information, and channel capacity, are measures that can aid in the assessment of the information content of a prediction and are amenable for use as loss functions in machine learning and other applications. These ideas were demonstrated on real and synthetic data.

Introduction

Two aspects in the assessment of the quality of ties between seismic and well data are (1) accuracy and (2) information content. It has become standard practice to use the correlation coefficient as an indication of the accuracy of seismic well ties (see, e.g., Lines and Newrick, 1985). However, the correlation coefficient measures only collinearity, not information content. It is thus an incomplete measure of quality, as correlation can be maximized by reducing the amount of information considered. All else being equal, it is easier to match a simple signal than a complex one. Sometimes, it may be worthwhile to accept poorer correlation to capture more information, provided that the additional content is useful. The purpose of this paper is to show that measures of information content can be used in conjunction with the correlation coefficient or related measures to better indicate quality in comparisons between seismic data and well logs or synthetics.

We should be surprised by seismic data or quantities derived from that data (e.g., inverted impedances or rock property estimates from machine learning). From the standpoint of information theory, *if data were perfectly predictable with no surprise, it would convey no information*. The predictability of a noiseless signal is strongly related to the complexity and thus the bandwidth of that signal. If the seismic trace had no bandwidth (i.e., a single-frequency sinusoid), forward prediction would be perfect, with no surprise and little value. For such monochromatic data, the correlation coefficient (r) between perfect synthetic and real noisy data, filtered to the same frequency and shifted to the best alignment, is unity ($r = 1$). Alternatively, if the seismic wavelet were a spike (infinite bandwidth), the best correlation between real data and perfect synthetic data would be lower (due to greater information complexity, but also as a practical matter, a consequence of the unachievable requirement, at infinite bandwidth, for there to be perfect alignment between seismic and well-log spike series).

It has long been recognized that the correlation coefficient can be improved upon as a measure of accuracy, for example, in evaluating time-lapse repeatability (Kragh and Christie, 2002; Cantillo, 2012) and amplitude variation with offset analysis (Coléou et al., 2013). Thus, putting aside that synthetic data contain errors for a variety of reasons, one can firmly conclude that (1) the correlation coefficient alone is inadequate to measure the quality of a synthetic tie resulting from any seismic process and (2) the information content aspect of quality needs more attention. For simplicity, in this paper, we deal only with synthetic and real data that have been previously corrected to have the same scale factor, so improvements in accuracy measures accounting for differences in root mean square (RMS) amplitudes need not be considered here.

Explorationists know that the best way to judge a well-log tie is using one's eyes and experience. However, if we are looking for quantitative measures, e.g., (1) to evaluate the match to synthetic seismic data, (2) to assess seismic rock properties predictions, or (3) as loss functions in optimization schemes, we need additional metrics that are more fit for purpose than the correlation coefficient. Furthermore, when seismic data are categorized or clustered, we need means of quantitatively measuring similarity between distributions of predicted and presumed perfect categories, such as facies logs (see Pendrel and Schouten, 2019).

In this paper, we present some basic concepts from information science, compute a variety of measures from that discipline, and compare them to the correlation coefficient for instructive synthetic and real data examples. A variety of measures from information theory, first developed for telecommunications (Shannon, 1948; Shannon and Weaver, 1949), include entropy, cross entropy, divergence, mutual information, variation of information, and channel capacity, as well as traditional statistical measures, such as the F-test and p-value.

Manuscript received 12 December 2025; revision received 17 February 2026; accepted 7 April 2026; published ahead of production 23 April 2026.
[1]University of Houston, Department of Earth and Atmospheric Sciences, Houston, Texas, USA. E-mail: jpcastag@central.uh.edu.
[2]Lumina Geophysical, Houston, Texas, USA. E-mail: vi.kronberg@luminageo.com; gabriel.gil@luminageo.com.

In this paper, we focus on entropy, cross entropy, variation of information, and channel capacity, with additional discussion provided in the Appendices. Though these measures are rarely used to date in evaluating seismic ties to wells, there is some geophysical literature applying information theory in a variety of applications (see, e.g., Sax, 1966; Petty, 2018; Pendrel and Schouten, 2019; Pasten et al., 2022). Giannakis and Mendel (1987) use entropy to evaluate deconvolution methods, White (1988) minimizes entropy to correct seismic phase, and Mukerji et al. (2001) use information theory measures to select attributes that most reduce uncertainty in predicting reservoir properties. A general introduction to information theory can be found in books such as Kinchin (1957) and Sivia and Skilling (2006).

As seismic inversion and rock property prediction using machine learning or deterministic methods are nonunique, a simple question illustrates the importance of information theory: For the case of two different predictions that have the same correlation and mean-squared error (MSE) relative to presumed truth, which is better? We can add a level of complexity to this question and, noting that filtering back to a single-frequency maximizes correlation, ask which bandwidth of the predictions is best? In this regard, we argue that information measures are useful, and we recommend that cross entropy, variation of information, and channel capacity can help determine the quality of synthetic ties and predicted well logs. The first step is to measure the information content of noisy seismic data (or imperfect predictions from that data) and compare it to that of presumed perfect synthetic data or well logs. The presumption of perfect synthetic or well-log data is an important issue that extends beyond the scope of this paper. We make the observation that information theory can be used to readily assess other, perhaps more meaningful, measures of ground truth (pay/wet, productive/tight, etc.).

Shannon entropy

In his landmark 1948 paper "A Mathematical Theory of Communication," C. E. Shannon recognized that the more "surprising" (less predictable) a data series is, the more information it conveys. The information content is obtained by considering a series of values (as a function of time or depth in our case) as random variables and computing a probability density function (PDF) estimate from the sampled population of values. We outline the key ideas here and provide the full mathematical framework in Appendix A.

Let $g(x)$ be a data trace, where x denotes time or depth. The trace takes on a range of amplitude values that we discretize into n bins centered at v_i, for $i = 1, \ldots, n$. The probability that $g(x)$ takes a value within the bin centered on v_i is estimated using kernel-density estimation, in which a Gaussian kernel is placed at each data point and the contributions are summed to produce a smooth, continuous density estimate (see Appendix A for details). The resulting density is evaluated at the bin centers v_i to yield discrete probabilities $G(v_i)$ (strictly a discrete probability mass function, though we retain the common term PDF throughout as it is an estimate of the population PDF). The information content of the discretized signal, measured in bits, is given by the Shannon entropy (H):

$$H(g) = -\sum_{i=1}^{n} G(v_i) \log_2 G(v_i), \tag{1}$$

where n is the number of possible v_i. As any $G(v_i)$ approaches unity, the entropy approaches zero. Thus, if $g(x)$ is constant (all amplitude values fall in a single bin), $H(g)$ is zero. On the other hand, if all $G(v_i)$ have equal probability, $H(g)$ takes the maximum value $\log_2 n$. To place the entropy on a scale from 0 to 1, it can be normalized by this maximum value.

Unlike the correlation coefficient, which is not affected by scalar multipliers, entropy and derived quantities change with the scalar multiplier unless the data are first standardized to unit variance. Notably, v_i can be categories rather than quantitative values, allowing entropy and related measures to be applied to categorical predictions, such as facies or lithology (see, e.g., Pendrel and Schouten, 2019). In the following, we consider measures involving PDFs of presumed perfect signals (synthetics and well logs) and seismic data and predictions (such as estimated impedances or rock properties).

Cross entropy

Cross entropy is a measure of the information cost when a predicted distribution replaces the true one. The lower the cross entropy, the more accurately the second distribution approximates the true one. Here, we presume the data from wells, $t(x)$, provide the true PDF $T(v_i)$, whereas the data from seismic $m(x)$ provide the distribution of the predicted (modeled) values $M(v_i)$. The cross entropy $H(t, m)$, in bits, is then

$$H(t, m) = -\sum_{i=1}^{n} T(v_i) \log_2 M(v_i), \tag{2}$$

where the PDFs are calculated on the same support (v_i and n). This equation is not symmetrical, as $H(t, m)$ does not generally equal $H(m, t)$, so it can be useful to compare the two directions.

Given two predictions of $t(x)$, the one with lower cross entropy is the better approximation of its distribution. Two predictions with the same cross entropy relative to $t(x)$ can have different MSE values, so cross entropy should be used in conjunction with other measures. When used as a loss function, cross entropy has desirable mathematical characteristics, such as good convergence and strong penalization of large confident mistakes; see, e.g., Terven et al. (2025).

The minimum possible cross entropy is zero (when $t(x)$ and $m(x)$ have the same constant value, thereby yielding the same single-valued PDF for T and M). More typically, T and M will have some spread, and $H(t, m)$ is greater than zero. Although the theoretical maximum cross entropy is infinity (when, for a particular v_i, T is finite and M is zero), as long as this condition is avoided (see Appendix A), it is also convenient to normalize the cross entropy by $\log_2 n$.

Variation of information

The variation of information $V(t,m)$ is a measure of the amount of information change between the prediction and perfect truth. It is given by

$$V(t,m) = H(t) + H(m) - 2I(t,m), \tag{3}$$

where $H(t)$ and $H(m)$ are the entropies and $I(t,m)$ is the mutual information (see Appendix B). The lower the variation of information, the closer the information content of the prediction $m(x)$ is to that of the truth $t(x)$. The variation of information represents the distance between clusters and, unlike cross entropy, qualifies as a metric as $V(t,m) = V(m,t)$. It must be less than the sum of entropies, so it is convenient to normalize it by $2\log_2 n$. As entropy increases with noise, so does the variation of information.

When using standardized PDFs, as we do in this paper, the variation of information is insensitive to multiplicative scalars, as is the correlation coefficient. It is not a measure of accuracy but rather quantifies the difference in information content, which can be due in part to noise.

Channel capacity

Channel capacity is the maximum rate at which information can be communicated by a noisy channel of information of a given bandwidth. The preceding measures treat the truth $t(x)$ and the prediction $m(x)$ as static distributions and compare their information content. Channel capacity reframes the problem in terms of information transfer: $t(x)$ is the message we aim to recover, and $m(x)$ — obtained via, e.g., inversion or machine learning estimation — is the received signal from a channel of communication. Everything that causes $m(x)$ to differ from $t(x)$ (limited bandwidth, noise, modeling errors, etc.) collectively affects the channel capacity, and these factors may be time variant. One can always compute information measures over sliding windows, but the fewer the number of samples, the less likely that the sampled data will be a good representation of the true population. To use the largest number of samples, and for illustrative purposes, we have therefore opted to compute the information measures over the entire trace in this paper. Channel capacity then provides an upper bound on the rate of information transfer about $t(x)$ that can be preserved in $m(x)$ over the length of the trace given the channel's bandwidth and noise characteristics — although this maximum could vary locally due to variable signal-to-noise (S/N) levels.

More formally, channel capacity (C) is the maximum rate at which information can be transmitted through such a channel, with a given bandwidth (B) and S/N ratio. It is given by

$$C = B\log_2\left(1 + \frac{\mathrm{S}}{\mathrm{N}}\right). \tag{4}$$

When B is given in Hz, the units of channel capacity are bits/second. This quantity captures the idea that, for a given S/N ratio, the broader the bandwidth, the more information is transmitted.

The S/N ratio can be expressed in terms of the correlation coefficient (r) between the noisy prediction $m(x)$ and the perfect truth $t(x)$ so that

$$C = B\log_2\left(1 + \frac{r^2}{1 - r^2}\right). \tag{5}$$

Channel capacity is zero if the S/N ratio is zero and goes to infinity as the S/N ratio does. It increases monotonically with bandwidth if the S/N ratio is unchanged, as broader band signals can transmit more information.

Numerical example

We take a presumed perfect band-limited true impedance well log $t(x)$ and hypothesize a band-limited imperfect prediction of impedance $m(x)$ by adding white noise to $t(x)$ (emulating a seismic inversion output). We then ask, *up to what frequency does the prediction provide useful information about the true band-limited impedance?* For these computations, $v_i \in [-10, 10]$ for all computations to ensure a consistent resolution and fair comparisons between PDFs. This matches the full observed range of the unfiltered (broadband) signals.

Figure 1 shows the spectra for $t(x)$ and $m(x)$. The prediction $m(x)$ suffers from poor S/N at high frequencies. Comparing only amplitude spectra, the frequency bands where no useful information

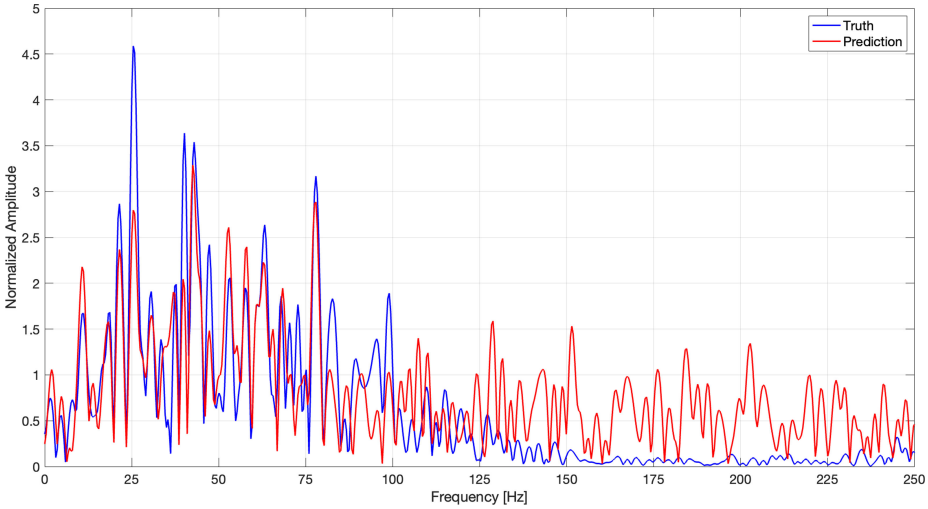

Figure 1. Amplitude spectra for a synthetic band-limited and detrended impedance log (blue line) and emulated seismic band-limited inverted impedance obtained by adding white noise (red line). The spectra are normalized by their respective RMS amplitudes. Noise dominates above 100 Hz.

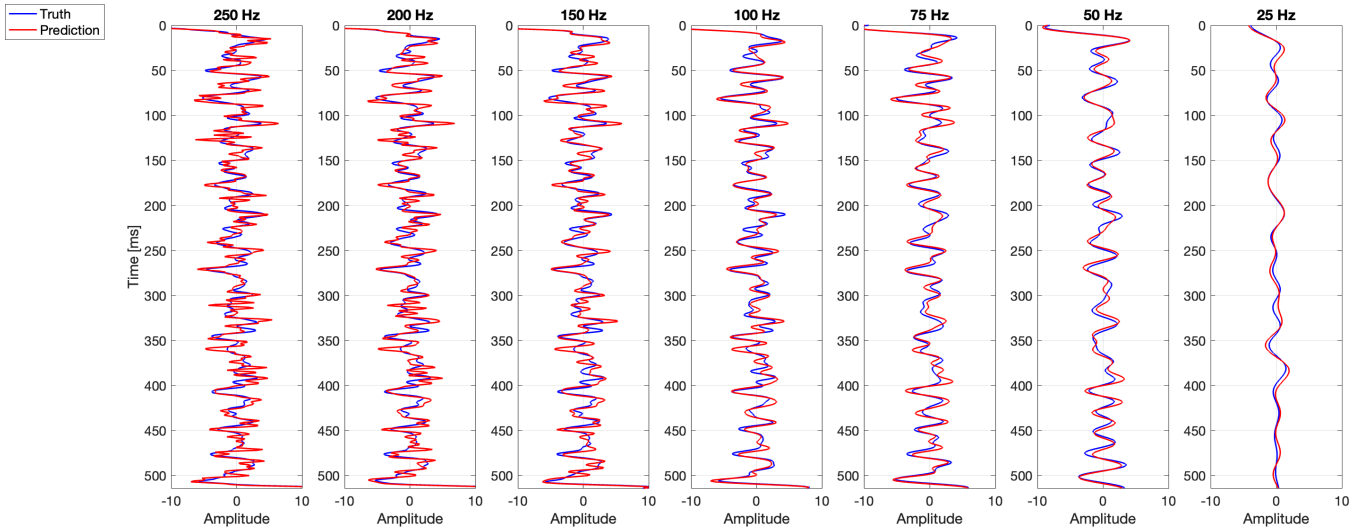

Figure 2. Lowpass filtered perfect (blue) and emulated imperfectly predicted (red) impedance logs for different bandwidths obtained by varying the high cut frequency. The bandpass high cut is indicated above each panel.

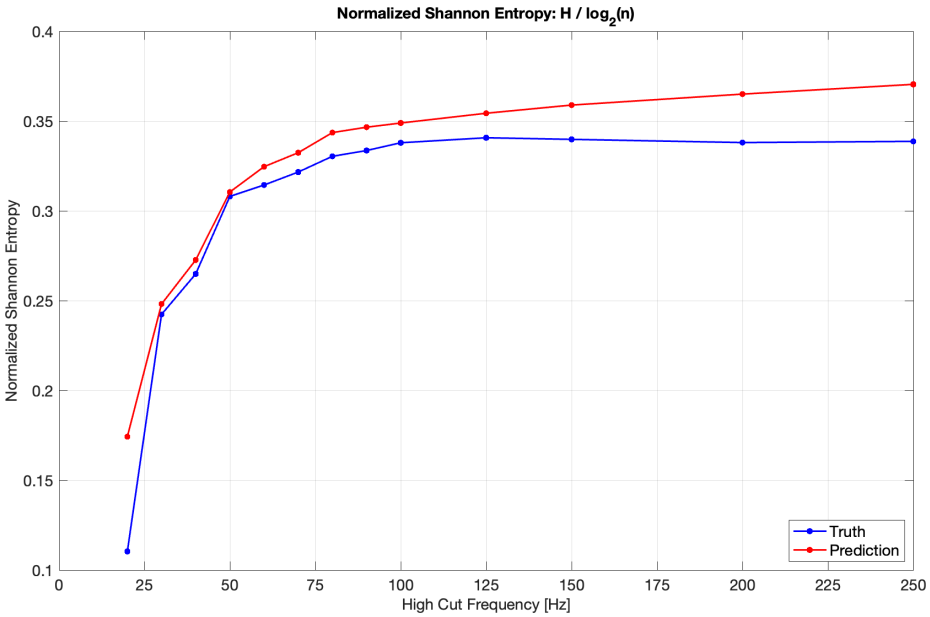

Figure 3. Normalized Shannon entropies for perfect band-limited impedance with zero noise (blue) and emulated imperfect prediction with added white noise (red) versus high pass frequency for filtered logs shown in **Figure 2**. Entropy increases with bandwidth but does not distinguish the signal from noise. The true Shannon entropy levels off at high frequency where there is little signal, whereas the entropy of the prediction increases due to noise at those frequencies.

exists are ambiguous. Even when amplitude spectra are consistent, there may be phase differences. Phase errors usually increase entropy, particularly if reflectivity is sparse (White, 1988).

A series of filter panels (Figure 2) are applied to the true and predicted band-limited impedance logs by varying the high cut frequency, and again, the useful high cut is ambiguous. The bandpass filtering smooths the impedance log to the same resolution as the seismic prediction. The PDFs associated with each filter panel are shown in Figure A 1.

The Shannon entropies (Figure 3) increase with bandwidth as there is more information (which includes noise). The imperfectly predicted band-limited impedance log has higher entropy than the perfect log because of the added noise. As frequency increases, as evident in Figure 1, the S/N ratio decreases, and the difference between perfect and imperfect entropies increases.

For this example, although there is energy in the spectrum of $t(x)$ beyond 125 Hz (Figure 1), the Shannon entropy (Figure 3) flattens out, suggesting that the high frequencies are not adding significant information content. Below 125 Hz, the entropy and cross entropy (Figure 4) increase with frequency, indicating increasing information content as frequencies are added. Along with the increasing information content, the variation of information also increases with frequency, indicating that the underlying PDFs are becoming flatter and also more dissimilar due to reduced accuracy, as evidenced by the reduction in the correlation coefficient and S/(S + N). Above 125 Hz, cross entropy plateaus, whereas r and S/(S + N) decrease (Figure 4), indicating that minimal additional useful information in the prediction exists at higher frequencies. Here, we can see that the correlation coefficient and S/(S + N) track each other, both indicating the reduced accuracy of high frequencies. The variation of information continues to

increase mildly above 125 Hz due to noise in the prediction. If S/N were constant over all frequencies, the channel capacity would increase monotonically with bandwidth. Here, monotonic increase occurs only below 80 Hz where it reaches its maximum value. Beyond 125 Hz, the channel capacity drastically decreases due to deteriorating S/(S + N), which overwhelms the bandwidth increase.

We conclude from this numerical example, in conjunction with the correlation coefficient and the S/(S + N) ratio, that the information theory measures are readily interpretable in terms of the amount of information in the prediction while facilitating quantitative determination of the prediction's useful bandwidth. The 125-Hz upper limit is consistent with what can be qualitatively observed from the true and predicted spectra (Figure 1), whereas quantitative determination of the useful bandwidth limit is not readily determined from spectra alone.

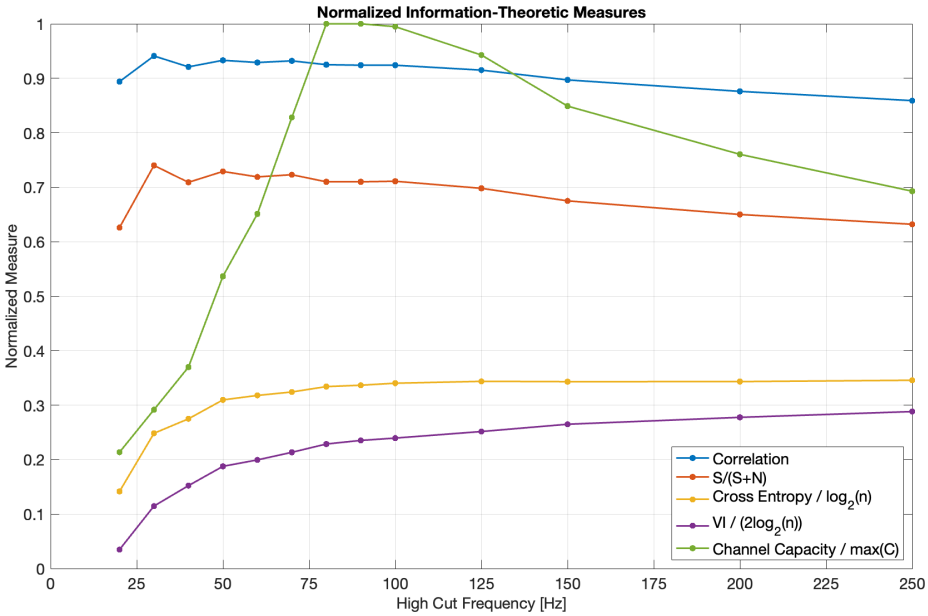

Figure 4. Correlation coefficient r (blue), $S/(S + N)$ (red), cross entropy normalized by $\log_2 n$ (yellow), variation of information normalized by $2 \log_2 n$ (purple), and channel capacity normalized by its maximum value (green), between perfect broad-band impedance and the emulated noisy prediction versus high cut frequency for lowpass filtered logs (Figure 2).

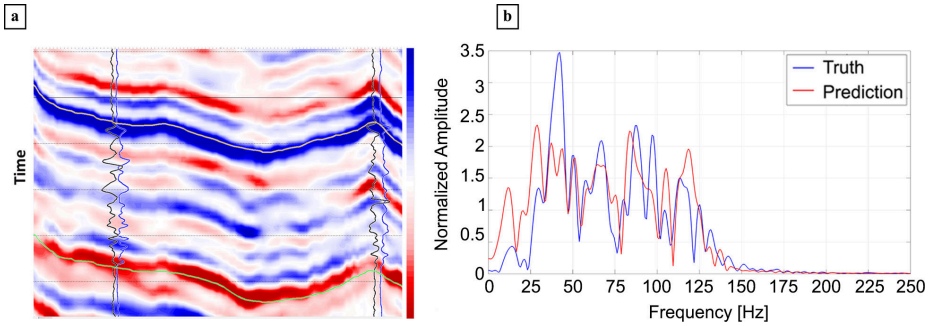

Figure 5. (a) High-resolution seismic line in the DaQing Field with overlain gamma-ray (black) and impedance (blue) logs at two well locations. The timing lines are at 10 ms increments. Amplitudes are relative (blue = positive, red = negative). Picked horizons represent the top (yellow) and base (green) of the reservoir interval. (b) Amplitude spectra for synthetic trace (blue) and seismic trace (red) at the leftmost well normalized by RMS amplitude. In this case, the spectra do not match well at low frequencies.

Synthetic tie examples

Let us now consider synthetic well ties to field seismic data. Figure 5 shows an arbitrary seismic line and real and synthetic data spectra from a high-resolution 3D cube in the DaQing Field in China. The real and synthetic traces do not match well at low frequencies as evidenced by the spectral mismatch (Figure 5b) and the low-frequency synthetic tie (Figure 6). If the reflectivity is sparse, low-frequency phase error not evident in the amplitude spectrum would also increase the cross entropy and variation of information, as it will increase $H(m)$.

The quality measures (Figure 7) indicate that below 50 Hz, there is a large discrepancy between the synthetic and real data. This is best seen by comparing the cross entropy $H(t, m)$ with the reverse cross entropy $H(m, t)$. The disparity is a consequence of the model's overestimation of low frequencies below 50 Hz, which, for $H(m, t)$, magnifies the expected surprise of the true distribution by incorrectly high probabilities in the model distribution.

Between 50 and 80 Hz, the correlation coefficient oddly increases with frequency along with the other information measures, suggesting problems in the data at the low end of the spectrum, which are ameliorated somewhat by a stronger and more accurate signal being added above 50 Hz. Above 80 Hz, correlation decreases again along with a more gradual increase in the information measures. This indicates that above 80 Hz, there is additional information but with lower accuracy. Above 125 Hz, the channel capacity degrades, suggesting that transmission of information is less reliable at higher frequency. All the information measures flatten out above 150 Hz, above which there is little energy. We conclude that the useful information content of the prediction is somewhere between 50 and 150 Hz with frequencies above 125 Hz being suspect due to decreasing channel capacity. This is consistent with visual inspection of the amplitude spectra (Figure 5b), which track best over that range.

Rock property prediction by machine learning

We address the problem of porosity prediction for a Permian Basin 3D seismic volume with five wells available for training and

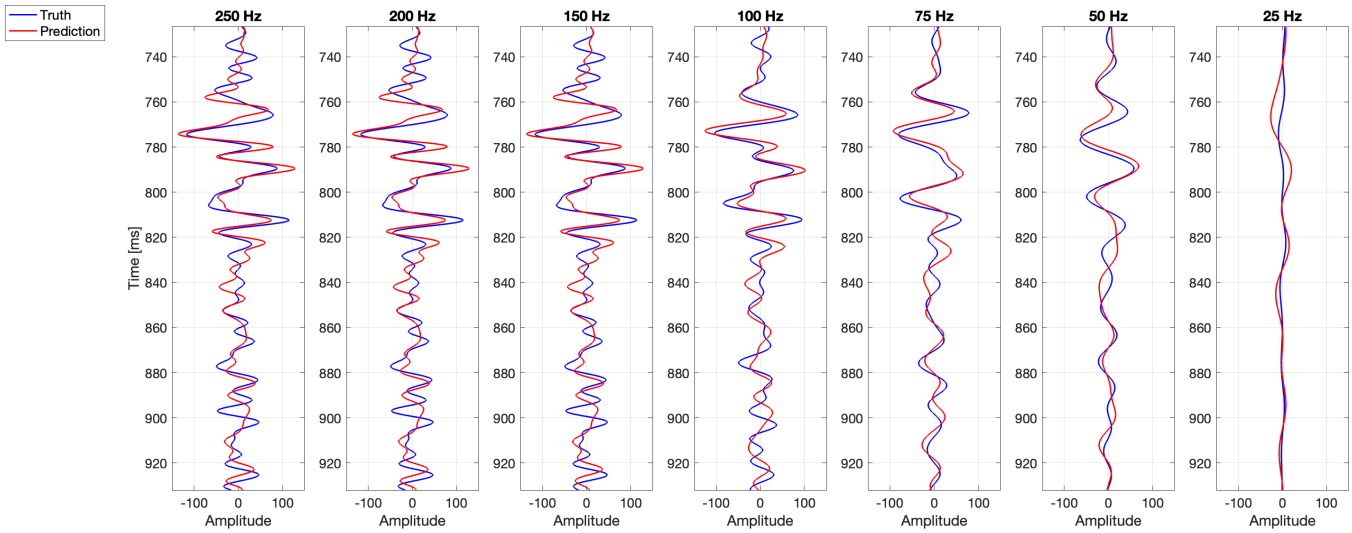

Figure 6. Lowpass filtered synthetic (blue) and seismic (red) traces in the DaQing Field for different bandwidths obtained by varying the high cut frequency. The bandpass high cut is indicated above each panel. The match is poor at low frequency and changes little above 150 Hz.

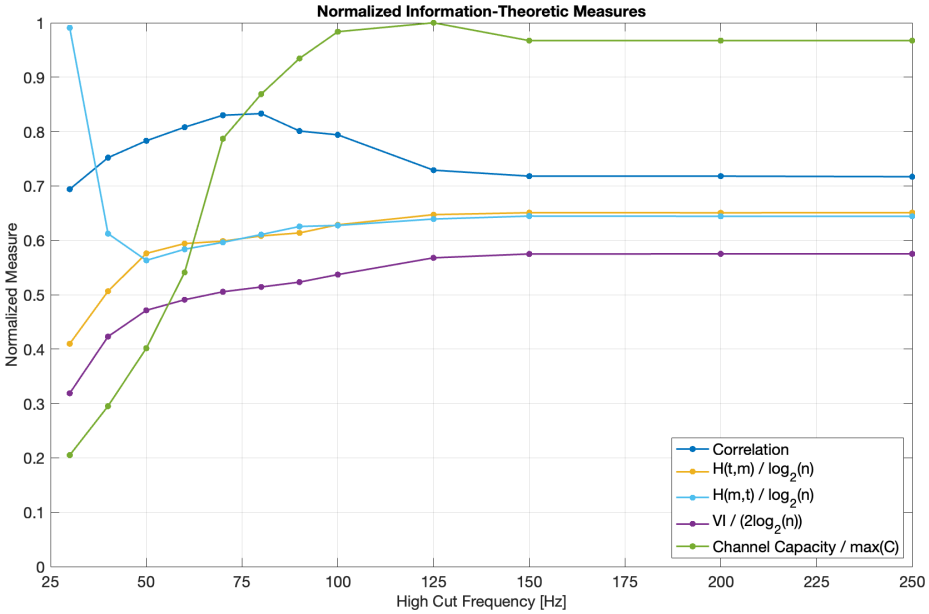

Figure 7. Correlation coefficient r (blue), cross entropy $H(t, m)$ (yellow), reverse cross entropy $H(m, t)$ (light blue) both normalized by $\log_2 n$, variation of information normalized by $2 \log_2 n$ (purple), and channel capacity normalized by its maximum value (green), between seismic and synthetic data versus high cut frequency for lowpass filtered data (Figure 6).

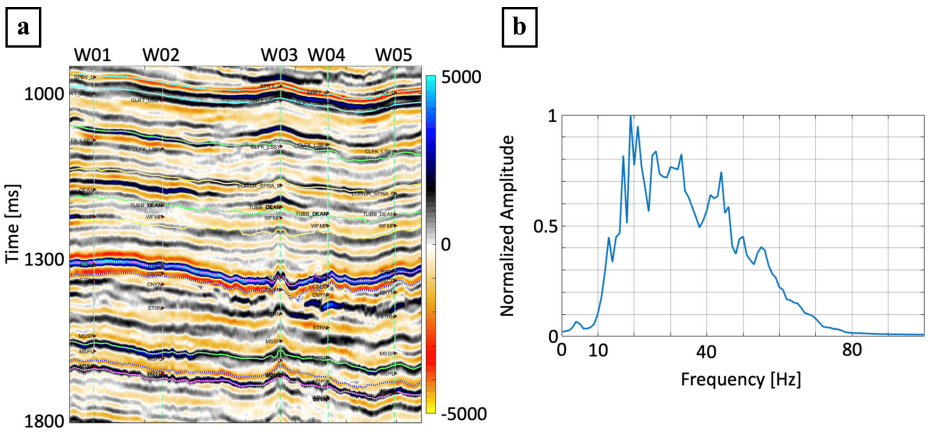

Figure 8. (a) Arbitrary line from a Permian Basin 3D seismic survey through five wells. (b) Seismic data spectrum.

validation. An arbitrary line and data spectrum are shown in Figure 8.

A variety of seismic attributes output from spectral decomposition and combined via principal component analysis are used to predict porosity by sparse multiple nonlinear regression using minimum support criteria (Kronberg et al., 2025). Here, nonlinear regression is used to identify useful attributes, but a variety of other machine learning algorithms can be subsequently applied to possibly make better predictions (for a general introduction to the approach, see the pioneering paper by Hampson et al., 2001). Out-of-sample validation curves and training curves for various models used are displayed in Figure 9. Each model tested adds an additional principal component from multiscale spectrally decomposed data or a traditional seismic attribute (*n*th order derivative, envelope, Hilbert transform of trace, etc.) and the lags (operator length) of those selected by minimum support. As the total number of lags used differs for each model, the *F*-test and *p*-value (discussed in Appendix B) are applied to account for the variable number of degrees of freedom in the model predictions. The training and validation curves are not monotonic as the number of

lags selected by minimum support for each attribute added varies across models.

The number of attributes to use in the porosity prediction is selected by finding the best combination of attributes based on the following weighted criteria: maximize the coefficient of determination (r^2), channel capacity (C), and F-statistic, and minimize the standard error, p-value, cross entropy, and variation of information. It is good practice to limit the number of attributes and lags used as too many degrees of freedom can lead to overtraining and poor predictions. For this reason, the number of inversion parameters used (sum of the number of lags for all attributes used) is displayed as quality control. Model 7 requires more parameters than model 5 or 6, which is a concern that can be mitigated somewhat by the high F-test and low p-value for model 7. These quantities account for the degrees of freedom used to make the prediction. On the other hand, model 5 has lower variation of information, particularly during training, and uses fewer parameters; it may therefore be more robust. In the end, the prediction from any model needs to be displayed in section and map view, and the interpreter needs to decide which model best meets their needs and is geologically sound.

In this case, we deemed model 7 the best compromise between these criteria. The model is composed of a weighted sum of principal components 1, 2, 5, 10, 18, and 20, as well as the first derivative of the original seismic data.

The resulting porosity prediction (Figure 10) ties the filtered porosity logs well (as quantified by the validation criteria for model 7 in Figure 9) and shows geologically plausible spatial variation. The filtering of the logs to the prediction bandwidth is a nontrivial process designed to gracefully handle edge effects at the top and the bottom of the logged sections.

A compelling visual example of porosity prediction validation previously published by Mora et al. (2020) in the DaQing Field is shown in Figures 11 and 12. The channel capacity is not reported by Mora et al. (2020) but is computed from the provided r^2 versus high frequency cutoff. It is obvious to the eye that the porosity prediction from bandwidth-extended data is geologically superior to that from the original seismic data. Relative to bandpass filtered well-log porosity in 20 wells, for the original seismic data, r = 0.69 and C = 52. Meanwhile, for bandwidth-extended seismic data, r = 0.66 and C = 113. Although r is slightly lower for the bandwidth-extended prediction, the channel capacity is more than double, indicating that much more information is transmitted despite the small S/N ratio reduction. Mora et al. (2020) also report improved MSE and F-test at high frequency for the bandwidth extended data.

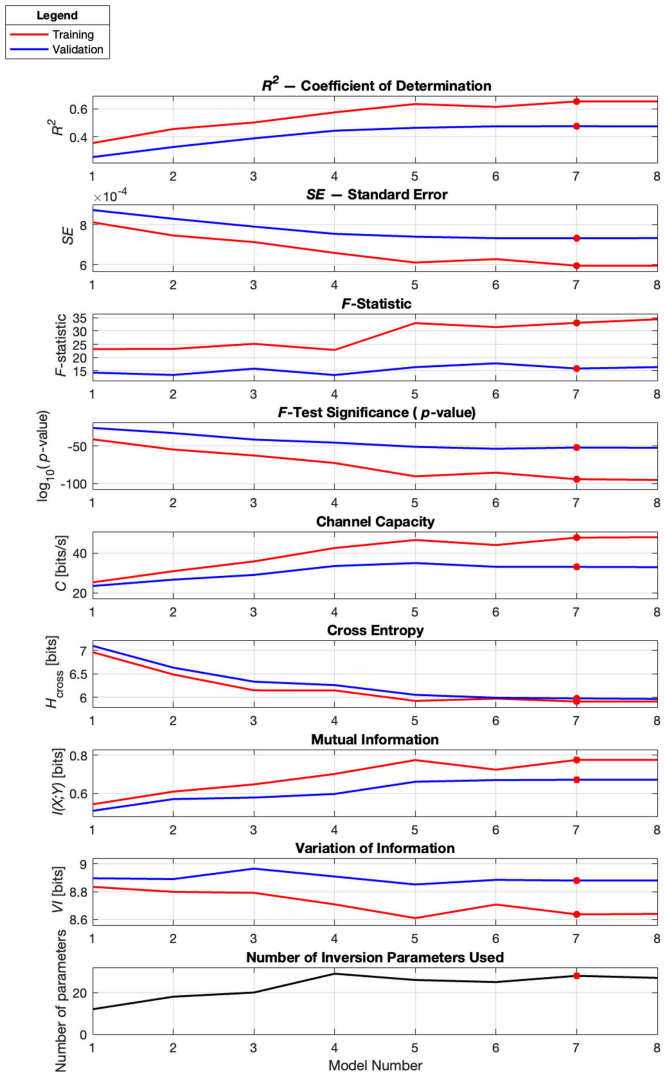

Figure 9. Training (red) and validation (blue) curves of various measures versus the number of attributes used (model number) for porosity prediction from the Permian Basin 3D seismic data set. The number of inversion parameters used is the sum of the number of lags for each attribute used plus an additive constant. Red filled circles represent the selected best number of attributes to use in predicting porosity.

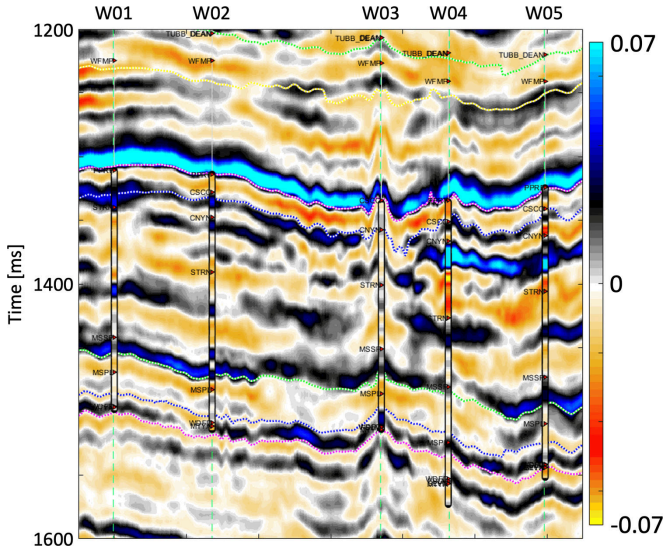

Figure 10. Seismic band-limited relative total porosity prediction (in decimal porosity units) using validation criteria shown in Figure 9 compared to band-limited well logs (colored insets) from the Permian Basin data set. The light blue event at about 1300 ms (above the training window) is a soft shale with low effective porosity but higher total porosity.

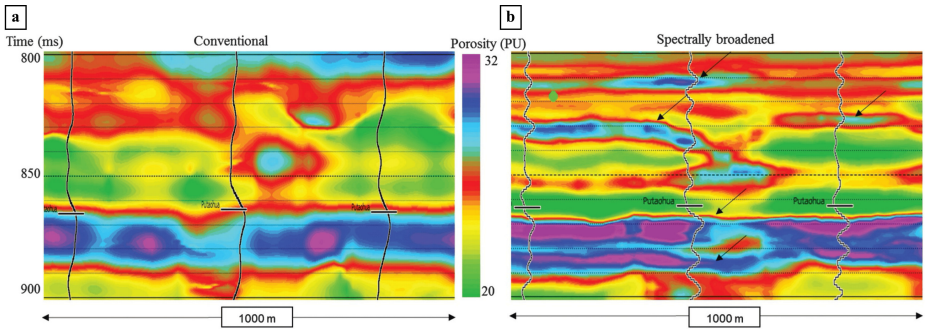

Figure 11. Seismic porosity prediction in the DaQing Oil Field (from **Mora et al., 2020**; reproduced with permission) using machine learning for (a) original seismic data and (b) bandwidth-extended seismic data. Well-log porosities (black curves) are bandpass filtered to match the seismic bandwidth. Black arrows point out features that tie the porosity log but are unresolved on the porosity prediction from the original seismic.

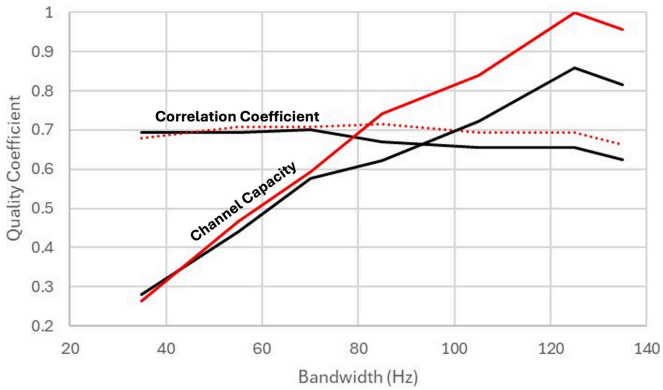

Figure 12. Correlation coefficient and channel capacity (normalized by the maximum value) versus bandwidth (of seismic prediction and well logs) for machine learning porosity prediction from seismic in the DaQing Field by **Mora et al. (2020)** for spectrally broadened data (red dashed and solid lines) and original seismic data (black lines).

Discussion

Historically, as evidenced by common commercial seismic interpretation packages, exploration geophysicists have somewhat cavalierly used the correlation coefficient almost exclusively as a quantitative measure of the quality of well ties despite the limitations of that metric. In machine learning, there is a need to quantitatively measure goodness of predictions to (1) assess well ties (particularly as a function of bandwidth) needed for training and (2) provide quantitative loss functions used by prediction algorithms and methodologies for quantitative and categorical predictions.

Of course, implicit in all the measures discussed is the idea of "perfect" truth, requiring perfect measurements, which never really exist. Well logs are seriously prone to error, and time–depth (or image-to-well log depth) functions are never perfectly correct. Time or depth shifts particularly degrade higher frequency comparisons, particularly as corrections are usually determined to maximize low frequency ties. However, the information theory measures do not require exact time or depth ties as they are based only on distributions of values, which do not change with minor shifts in alignment. So, despite their ability to measure information content in certain applications, they can be of limited utility in aligning series with each other.

Conclusions

Even with perfect adjustment to well-log depths, the correlation coefficient alone is not sufficient to determine the quality of fit between a prediction and presumed perfect truth because it does not provide a measure of the information content of the prediction or that of the ground truth. Information theory provides additional measures of information content and comparisons between predicted and presumed perfectly known distributions. Thus, by quantifying information content, information theory measures offer an additional dimension for assessing prediction quality. However, information theory does not directly indicate the value of the information, only the amount of information, which includes noise.

We have shown numerical and real data examples comparing a variety of measures used to assess well ties or predictions of rock properties and found cross entropy, variation of information, and channel capacity to be useful. Variation of information is a true metric that measures the distance between predicted and perfectly known distributions, whereas the channel capacity is a measure of the rate at which information is conveyed by a noisy channel of information. Further discussion of these and additional measures from information theory and statistics is provided in the Appendices. TLE

Data and materials availability

Data associated with this research are confidential and cannot be released.

Appendix A – PDF estimation, entropy, and practical considerations

This appendix covers more of the mathematical details around how to estimate PDFs from seismic traces or well logs. As seismic amplitudes and well-log values are digitized samples of continuous random variables, discretization into numerical categories is accomplished by binning over ranges of values as discussed below.

A continuous signal as a random variable. Let us start with the idea of representing a signal as a random variable. Let x be time or depth, restricted to some domain $[a, b]$. Next, let $g(x)$ be a signal, such as a seismic data trace. Then, following Kinchin (1957) let g be a complete set of events, such that one and only one of them must occur at each trial, i.e., for each value in time or depth.

More formally, let $g:[a,b] \to R$ be a continuous, piecewise-monotone function representing a recorded signal (e.g., a seismic trace, where the independent variable x denotes time or depth). Define a random variable $X^* \sim \mathcal{U}(a,b)$, i.e., X^* is drawn uniformly on the interval $[a,b]$ from the uniform distribution $\mathcal{U}$, and let

$$V = g(X^*). \tag{A-1}$$

Because X^* is drawn uniformly, the cumulative distribution function (CDF) of V is

$$F_V(v) = P(V \le v) = \frac{\lambda(\{x \in [a,b]: g(x) \le v\})}{b-a}, \tag{A-2}$$

where $\lambda(\cdot)$ denotes the Lebesgue measure. The CDF satisfies $F_V(v) = 0$ for $v < v_{\min}$ and $F_V(v) = 1$ for $v > v_{\max}$, with $v_{\min} := \min_{x \in [a,b]} g(x)$ and $v_{\max} := \max_{x \in [a,b]} g(x)$.

Under the piecewise-monotonicity assumption, F_V is absolutely continuous and therefore possesses a PDF

$$f_V(v) = F_V'(v), \tag{A-3}$$

defined almost everywhere on $[v_{\min}, v_{\max}]$ and satisfying $\int_{v_{\min}}^{v_{\max}} f_V(v)\,dv = 1$.

Caveat: Continuity of g alone guarantees that F_V is continous but not that a classical density exists. The additional assumption of piecewise monotonicity ensures that the preimage $\{x: g(x) \le v\}$ is a finite union of intervals for each v, from which the absolute continuity of F_V follows. For practical signals sampled at finite resolution, this condition is not restrictive.

Discretization and Shannon entropy. Recorded signals are digitized: the range $[v_{\min}, v_{\max}]$ is partitioned into n bins of equal width

$$\Delta v = \frac{v_{\max} - v_{\min}}{n}, \tag{A-4}$$

centered at values $v_i = v_{\min} + \left(i - \frac{1}{2}\right)\Delta v$ for $i = 1, \ldots, n$. The probability that V falls in the ith bin is

$$p_i = P\left(V \in \left(v_i - \frac{\Delta v}{2}, v_i + \frac{\Delta v}{2}\right]\right) = \int_{v_i - \Delta v/2}^{v_i + \Delta v/2} f_V(v)\,dv \approx f_V(v_i)\,\Delta v. \tag{A-5}$$

The approximation becomes exact in the limit $\Delta v \to 0$ and is accurate whenever f_V is approximately constant within each bin. By construction, p_i are nonnegative and satisfy $\sum_{i=1}^{n} p_i = 1$, so they define a probability mass function (PMF).

The information content of the discretized signal, measured in bits, is given by the Shannon entropy of this PMF:

$$H_{\Delta v}(V) = -\sum_{i=1}^{n} p_i \log_2 p_i, \tag{A-6}$$

with the convention $0 \log_2 0 = 0$. The entropy is bounded by $0 \le H_{\Delta v}(V) \le \log_2 n$, attaining its maximum when all bins are equally likely (i.e., V is uniformly distributed) and its maximum when the signal occupies a single bin.

Kernel density estimation. In practice, the density f_V is unknown and must be estimated from the observed sample values $\{v^{(1)}, \ldots, v^{(N)}\}$, where $v^{(j)} = g(x_j)$ and x_j are uniformly spaced sample positions in $[a,b]$. Note that the superscript notation distinguishes the N observed data points from the n bin centers v_i introduced in the discretization and Shannon entropy section; in general, $N \ne n$. A histogram provides a simple estimator of f_V, but its discontinuities and sensitivity to bin placement make it unsuitable when a smooth density estimate is required, for instance, when computing entropy or comparing distributions across traces.

Kernel density estimation (KDE) replaces the histogram bins with smooth kernel functions centered on each observation. The KDE of f_V is

$$\hat{f}_V(v) = \frac{1}{Nh} \sum_{j=1}^{N} K\left(\frac{v - v^{(j)}}{h}\right) \tag{A-7}$$

where K is a kernel function satisfying $K(u) \ge 0$ and $\int_{-\infty}^{\infty} K(u)\,du = 1$, and $h > 0$ is the kernel width (smoothing parameter). We shall be using the familiar Gaussian kernel

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right), \tag{A-8}$$

which yields a density estimate that is infinitely differentiable and strictly positive on $\mathbb{R}$.

Evaluation support. The KDE must be evaluated over a fixed interval $[v_{\text{lo}}, v_{\text{hi}}]$. When comparing entropy or divergence measures across multiple signals (e.g., a predicted and an observed seismic trace), it is essential that all densities are evaluated on a common support. If each signal were evaluated on its own range, the resulting bin centers and widths would differ, and the information-theoretic quantities would not be directly comparable. In this manuscript, the support $[v_{\text{lo}}, v_{\text{hi}}]$ is therefore fixed for each data set (example) and applied uniformly to all signals within that data set.

The bounds were chosen to encompass the full range of the unfiltered (broadband) signals and chosen to be symmetric around the mean of the signal. When bandpass-filtered versions of the signals are subsequently analyzed, the same support is retained even though the filtering may reduce the amplitude range considerably. This ensures that the quantities remain comparable across filter bands: a narrowband signal that concentrates its probability mass to a small portion of the evaluation range will correctly yield a lower entropy than

the corresponding broadband signal, reflecting the genuine reduction in information content.

Kernel width selection. The kernel width h controls the bias-variance trade-off of the estimator: a small h produces a spiky estimate that tracks individual samples (low bias, high variance), whereas a large h oversmooths and obscures genuine structure (high bias, low variance).

A classical default is Silverman's rule of thumb,

$$h_S = 0.9\min\left(\widehat{\sigma}, \frac{\widehat{Q}}{1.34}\right) N^{-\frac{1}{5}}, \tag{A-9}$$

where $\widehat{\sigma}$ is the sample standard deviation and $\widehat{Q}$ is the sample interquartile range. This rule of thumb is optimal (in the sense of minimizing the mean integrated square error) when the underlying density is Gaussian and serves as a reasonable starting point for other unimodal densities. See Silverman (1986) for a comprehensive treatment of KDE.

In this manuscript, we instead set the kernel width as a fixed fraction of the sample standard deviation,

$$h = c\widehat{\sigma}, \tag{A-10}$$

where c is a hardcoded constant (we use $c = 0.2$). This formulation decouples the kernel width from the sample size N, which is appropriate when comparing signals of equal length, and provides direct control over the degree of smoothing relative to the spread of each individual signal. The choice $c = 0.2$ was found for seismic data to give a good balance between the resolving multimodal structure on the amplitude distributions and suppressing sampling noise.

For distributions that are multimodal or strongly skewed (as seismic impedance distributions can be), data-driven methods, such as least-squares cross-validation or the improved Sheather–Jones plug-in estimator, offer alternatives to Silverman's rule and the fixed-fraction approach (Rudemo, 1982; Sheather and Jones, 1991; Botev et al., 2010).

Properties. The KDE $\widehat{f}_V$ inherits the smoothness of the kernel: with a Gaussian kernel, it is a valid probability density ($\widehat{f}_V(v) \geq 0$ and $\int_{-\infty}^{\infty} \widehat{f}_V(v)\,dv = 1$) that can be evaluated at arbitrary points v. As $N \to \infty$ with $h \to 0$ and $Nh \to \infty$, the estimator is consistent, i.e., $\widehat{f}_V(v) \to f_V(v)$ in probability at every continuity point of f_V.

To obtain the discrete probabilities p_i, required for, e.g., the Shannon entropy, the KDE is evaluated on the bin centers v_i and multiplied by the bin width:

$$\widehat{p}_i = \widehat{f}_V(v_i)\,\Delta v\,. \tag{A-11}$$

Numerical treatment of zero probabilities. The Shannon entropy and related quantities, such as the cross-entropy and divergence, involve the logarithm of probability values. In exact arithmetic, the Gaussian KDE assigns strictly positive density everywhere on $\mathbb{R}$, so the estimated probabilites $\widehat{p}_i$ are never exactly zero. In finite-precision floating-point arithmetic, however, bins that lie far from any data point can underflow to zero, particularly when the evaluation support is deliberately wider than the effective range of filtered signals as described above.

A standard mitigation is additive smoothing: a small constant $\epsilon > 0$ (e.g., $\epsilon = 10^{-9}$) is added to every bin probability, after which the PMF is renormalized to sum to unity. This ensures that all logarithmic terms remain finite while introducing a negligible bias.

Numerical example

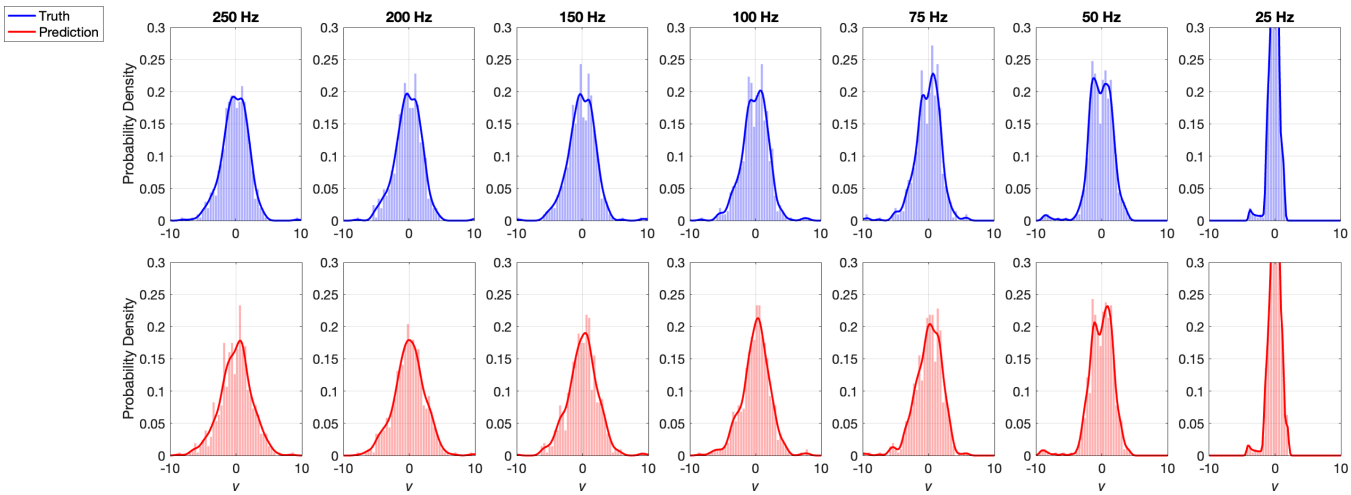

Figure A1. Histograms and KDE PDFs of lowpass filtered perfect (blue) and emulated imperfectly predicted (red) impedance logs for different bandwidths obtained by varying the high cut frequency. The logs are shown in **Figure 2**. The bandpass high cut is indicated above each panel.

Stationarity and sampling considerations. The information-theoretic measures used in this manuscript implicitly assume stationarity — that is, the statistical properties of the signal do not change over the analysis window. In practice, this assumption is an approximation: geologic properties vary laterally and vertically, so the underlying PDF is not truly stationary over large windows.

Furthermore, the probability density functions used in the calculation of entropy and related measures are those of sampled populations, which are not the true probability density functions. An implicit assumption is that the PDF of the sampled population approaches that of the full population as the number of samples increases. How well the sampled PDF used to compute Shannon entropy represents the actual continuous PDF of the infinite population is beyond the scope of this paper, though one can generalize that the larger the data window, usually the better.

It is of course possible to produce nonstationary information measures with moving vertical and lateral windows or perhaps guided by geologic horizons, at the expense of fewer samples and less confidence that the sampled PDF resembles the true PDF. Regularization of PDFs to better represent hypothetical full populations, perhaps guided by knowledge of depositional systems, is likely to be a fruitful area of inquiry, particularly as applied to optimization problems.

Properties of Shannon entropy. Shannon entropy is nonunique: different probability distributions can share the same entropy value. For example, the different distributions (0.25, 0.5, 0.25) and (0.5, 0.25, 0.25) yield an entropy of 1 bit. Also, series with the same PDF and entropy can have different values. The same entropy does not imply equality of the data. However, when used in a loss function, it can help steer predictions toward a correct probability distribution, which can be a useful additional constraint.

Multiplying $g(x)$ by a scalar stretches or squeezes $G(v_i)$ and thus changes the entropy unless v_i are standardized accordingly relative to σ. Adding a constant to $g(x)$ shifts $G(v_i)$, which does not change the entropy. Adding a linear trend to $g(x)$ increases the entropy, so it may be worthwhile to use detrended logs and predictions, particularly when trends are not well constrained by the seismic data.

As entropy does not distinguish signal from noise, adding white noise to $g(x)$ increases entropy. Thus, information content, right or wrong, increases by adding unpredictable noise. Convolving $g(x)$ with a stationary filter $f(x)$ changes the entropy. Spiking deconvolution, for example, increases entropy and reduces predictability. Unlike thermodynamic entropy, the Shannon entropy of a sum of random variables is not the sum of the entropies. Rather, the PDF of the sum is the convolution of the individual PDFs from which the entropy of the sum can be determined. For the special case of a known white-noise level having a specified Gaussian or uniform PDF, the PDF of signal plus noise is readily corrected by taking the ratio of the Fourier transforms of the PDFs and then applying the inverse transform, thereby yielding the noise-corrected PDF from which the entropy of the signal is determined.

Shannon entropy represents the irreducible complexity beyond which no data compression is valid. We can view $\log_2 G(v_i)$ as the "surprise" associated with a certain outcome. If $G(v_i)$ has 100% probability, $\log_2(1) = 0$, and there is no surprise associated with that outcome. Shannon entropy multiplies the surprise by $G(v_i)$ and sums over all v_i and is thus the expected value of the surprise. A great advantage of this measure is that v_i can just as well be categories as numerical values. So, entropy and related quantities allow, for example, predicted facies or lithology identifications to be compared to known geologic distributions (as illustrated by Pendrel and Schouten, 2019).

Appendix B – Additional information and statistical measures

A variety of statistical and information theory measures, in addition to those discussed elsewhere in this article, may provide additional insight into the information content of a prediction if perfect ground truth is known (or presumed to be so).

Additional information measures

Joint density estimation. Several of the measures below require the joint probability distribution of the truth $t(x)$ and the prediction $m(x)$. The joint density is estimated by a bivariate extension of the univariate KDE described in Appendix A. Given paired samples $\{(t^{(l)}, m^{(l)})\}_{l=1}^{N}$, the bivariate KDE is

$$\hat{f}_{T,M}(s,u) = \frac{1}{N}\sum_{l=1}^{N} \frac{1}{2\pi h_t h_m} \exp\left(-\frac{1}{2}\left[\left(\frac{s-t^{(l)}}{h_t}\right)^2 + \left(\frac{u-m^{(l)}}{h_m}\right)^2\right]\right) \quad \text{(B-1)}$$

where $h_t = c\hat{\sigma}_t$ and $h_m = c\hat{\sigma}_m$ are the bandwidths for each variable, computed using the same fixed-fraction rule as in the univariate case (kernel density estimation section in Appendix A). The joint density is evaluated on a 2D grid over the common support $[v_{\text{lo}}, v_{\text{hi}}]^2$ and discretized into bin probabilities $p_{ik} = \hat{f}_{T,M}(v_i, v_k)\,(\Delta v)^2$. The marginal distributions are recovered by summation: $p_i^T = \Sigma_k\, p_{ik}$ and $p_k^M = \Sigma_i\, p_{ik}$. The same additive smoothing described in Appendix A is applied to the joint PMF before computing any logarithmic quantities.

Mutual information. Mutual information, $I(t,m)$, is a measure of the shared information between two data series. For comparison of the PDF of noisy predictions, $M(v_i)$, to the PDF of perfect truth, $T(v_i)$, the mutual information is the uncertainty about $T(v_i)$ given $M(v_i)$ and is given by

$$I(t,m) = H(m) - H(m\,|\,t), \quad \text{(B-2)}$$

where $H(m)$ is the Shannon entropy of $m(x)$ and $H(m|\,t)$ is the conditional entropy of $m(x)$ given $t(x)$. The conditional entropy is readily calculated from the joint probability distribution as

$$H(m \mid t) = -\sum_i \sum_k p_{ik} \log_2 \frac{p_{ik}}{p_i^T}. \quad \text{(B-3)}$$

Equivalently, mutual information can be expressed directly in terms of the marginal and joint entropies:

$$I(t,m) = H(t) + H(m) - H(t,m), \tag{B-4}$$

where $H(t,m) = -\Sigma_i \Sigma_k\, p_{ik} \log_2 p_{ik}$ is the joint entropy. This identity can be a useful consistency check in the numerical implementation.

The higher the mutual information, the more shared information, or information gain, is contained in the prediction. In evaluating well ties, we presume that the more shared information, the better. However, there is no implication about the value of that information. For example, if one is interested in predicting absolute rock properties over a long interval, low frequencies, if accurate, may have more value in that regard.

Divergence. Divergence, $D(t,m)$, also called relative entropy, is a measure of how different the predicted (M) and true (T) probability density functions are. It is given by

$$D(t,m) = -\sum_{i=1}^{n} T(v_i) \log_2 \frac{T(v_i)}{M(v_i)}. \tag{B-5}$$

When the predicted and true probability density functions are identical, the divergence is zero, whereas the maximum divergence is infinity that occurs when, at a given v_i, T has finite probability but M is zero. It differs from cross entropy in that it measures the relative information difference between distributions, whereas the cross entropy measures the total information needed to predict one distribution from the other. Divergence is not a true metric in that $D(t,m)$ does not generally equal $D(m,t)$.

Derivative of cross entropy. The derivative of cross entropy with respect to frequency, $\mathrm{d}H(t,m)/\mathrm{d}f$, where f is, for example, the bandpass high-cut frequency, identifies when increasing frequency content no longer adds useful information as $\mathrm{d}H(t,m)/\mathrm{d}f$ approaches zero.

Spectral entropy. A perfectly predictable series conveys minimal information. Thus, a series of constant values has zero entropy. However, a monochromatic sinusoid is also perfectly predictable yet has a bimodal probability density function (representing peaks and troughs) and, thus, has finite entropy. This complication can be avoided using spectral entropy, which is just the entropy of the positive frequency spectrum. Thus, an infinite sinusoid has a single spike in the frequency domain, and zero spectral entropy.

Perturbational entropy. Values are rank ordered to represent patterns that are treated as categories. For example, upward fining and upward coarsening categories would be distinguished.

Statistical measures. The correlation coefficient (r) and MSE are well-known criteria commonly used in training and are closely related to the S/N ratio. r^2 is the coefficient of determination, which is the proportion of the true variance explained by the prediction. However, neither r nor MSE addresses the statistical significance of the prediction, which could align with the true values by chance. The more seismic attributes used in making a prediction, the closer that prediction can be to training data, irrespective of whether any of those attributes convey additional information. If more attributes than datapoints are used, the training data can be matched perfectly, even if those attributes were purely random numbers. Despite the perfect match in training, such a combination of attributes has no predictive ability and will likely fail in validation, though it could accidentally pass the validation test. The F-test and p-value are used to determine whether the prediction is statistically significant. F is a ratio of explained to unexplained variance of the prediction (considering the number of attributes and degrees of freedom) and is related to the correlation coefficient by

$$F = \frac{r^2/K}{1 - r^2/(N - K - 1)}, \tag{B-6}$$

where K is the number of attributes used in the prediction and N is the number of datapoints to be matched in training. When $K = n - 1$, $F = 0$, no matter the correlation achieved. Usually, the higher F, the more significant the prediction is.

However, there is always the possibility that the observed correlation is achieved purely by random chance. The p-value (sometimes called *sig F*) is a measure of that probability. In many scientific applications a p-value less than 0.05 is considered good (corresponding to a 5% probability that the correlation is achieved by random chance). In seismic versus well-log comparisons, N is usually on the order of 10^3, so F can be very high and p very low. On the other hand, it is important to point out that if average reservoir porosity is the most important quantity to be predicted, there is only one datapoint per well, and the prediction of that average is unlikely to be statistically significant without a large number of wells. This means that seismic predictions from machine learning are better at showing changes in rock properties than in predicting average rock properties without other calibration. The F-test and p-value are relative measures of significance that degrade as the number of seismic attributes and lags of those attributes increases. In training and validation, F should be maximized and p minimized.

Corresponding author: vi.kronberg@luminageo.com

References

Botev, Z. I., J. F. Grotowski, and D. P. Kroese, 2010, Kernel density estimation via diffusion: The Annals of Statistics, **38**, no. 5, 2916–2957, https://doi.org/10.1214/10-AOS799.

Cantillo, J., 2012, Throwing a new light on time-lapse technology, metrics and 4D repeatability with SDR: The Leading Edge, **31**, no. 4, 405–413, https://doi.org/10.1190/tle31040405.1.

Coléou, T., J. P. Coulon, D. Carotti, P. Dépré, G. Robinson, and E. Hudgens, 2013, AVO QC during processing: 75th EAGE Conference and Exhibition, Extended Abstracts.

Giannakis, G. B., and J. M. Mendel, 1987, Entropy interpretation of maximum-likelihood deconvolution: Geophysics, **52**, 1621–1630, https://doi.org/10.1190/1.1442279.

Hampson, D. P., J. S. Schuelke, and J. A. Quirein, 2001, Use of multiattribute transforms to predict log properties from seismic data: Geophysics, **66**, 220–236, https://doi.org/10.1190/1.1444899.

Kinchin, A. I., 1957, Mathematical foundation of information theory. Translated by Silverman, R. A. and Friedman, M. D.: Dover Publications.

Kragh, E., and P. Christie, 2002, Seismic repeatability, normalized RMS, and predictability: The Leading Edge, **21**, no. 7, 640–647, https://doi.org/10.1190/1.1497316.

Kronberg, V., G. Gil, and J. P. Castagna, 2025, Incorporating the multiscale Fourier transform, principal components, and significance measures in seismic prediction of rock properties: Fifth International Meeting for Applied Geoscience and Energy, SEG/AAPG, Expanded Abstract, https://doi.org/10.1190/image2025-4307084.1.

Lines, L. R., and R. T. Newrick, 1985, Fundamentals of geophysical interpretation: Geophysical Monograph Series: SEG.

Mora, D., J. P. Castagna, R. Meza, S. Chen, and R. Jiang, 2020, Case study: Seismic resolution and reservoir characterization of thin sands using multiattribute analysis and bandwidth extension in the DaQing field, China: Interpretation, **8**, no. 1, T89–T102, https://doi.org/10.1190/INT-2019-0017.1.

Mukerji, T., P. Avseth, G. Mavko, I. Takahashi, and E. F. González, 2001, Statistical rock physics: Combining rock physics, information theory, and geostatistics to reduce uncertainty in seismic reservoir characterization: The Leading Edge, **20**, 313–319, https://doi.org/10.1190/1.1438938.

Pasten, D., G. Saravia, E. E. Vogel, and A. Posadas, 2022, Information theory and earthquakes: Depth propagation seismicity in northern Chile: Chaos, Solitons & Fractals, **165**, 112874, https://doi.org/10.1016/j.chaos.2022.112874.

Pendrel, J. V., and H. J. Schouten, 2019, Entropy QC for Bayesian facies estimations: 89th Annual International Meeting, SEG, Expanded Abstracts, https://doi.org/10.1190/segam2019-3204859.1.

Petty, G. W., 2018, On some shortcomings of Shannon entropy as a measure of information content in indirect measurements of continuous variables: Journal of Atmospheric and Oceanic Technology, **35**, no. 5, 1011–1021, https://doi.org/10.1175/JTECH-D-17-0056.1.

Rudemo, M., 1982, Empirical choice of histograms and kernel density estimators: Scandinavian Journal of Statistics, **9**, no. 2, 65–78.

Sax, R. L., 1966, Application of filter theory and information theory to the interpretation of gravity measurements: Geophysics, **31**, 570–575, https://doi.org/10.1190/1.1439794.

Shannon, C. E., 1948, A mathematical theory of communication: The Bell System Technical Journal, **27**, no. 3, 379–423, https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Shannon, C. E., and W. Weaver, 1949, The mathematical theory of communication: University of Illinois Press.

Sheather, S. J., and M. C. Jones, 1991, A reliable data-based bandwidth selection method for kernel density estimation: Journal of the Royal Statistical Society. Series B (Methodological), **53**, no. 3, 683–690, https://doi.org/10.1111/j.2517-6161.1991.tb01857.x.

Silverman, B. W., 1986, Density estimation for statistics and data analysis: Chapman & Hall.

Sivia, D. S., and J. Skilling, 2006, Data analysis: A Bayesian tutorial: Oxford University Press.

Terven, J., D. Cordova, J. Romero-Gonzalez, A. Ramirez-Pedraza, and E. A. Chavez-Urbiola, 2025, A comprehensive survey of loss functions and metrics in deep learning: Artificial Intelligence Review, **58**, no. 7, 195, https://doi.org/10.1007/s10462-025-11198-7.

White, R. E., 1988, Maximum kurtosis phase correction: Geophysical Journal, **95**, no. 2, 371–389, https://doi.org/10.1111/j.1365-246X.1988.tb00475.x.