Research Article

Horizon Detection in Seismic Data: An Application of Linked Feature Detection from Multiple Time Series

Robert G. Aykroyd[1] and Fathi M. O. Hamed[2]

[1] *Department of Statistics, University of Leeds, Leeds LS2 9JT, UK*
[2] *Department of Statistics, University of Benghazi, P.O. Box 1308, Benghazi, Libya*

Correspondence should be addressed to Robert G. Aykroyd; r.g.aykroyd@leeds.ac.uk

Seismic studies are a key stage in the search for large scale underground features such as water reserves, gas pockets, or oil fields. Sound waves, generated on the earth's surface, travel through the ground before being partially reflected at interfaces between regions with high contrast in acoustic properties such as between liquid and solid. After returning to the surface, the reflected signals are recorded by acoustic sensors. Importantly, reflections from different depths return at different times, and hence the data contain depth information as well as position. A strong reflecting interface, called a horizon, indicates a stratigraphic boundary between two different regions, and it is the location of these horizons which is of key importance. This paper proposes a simple approach for the automatic identification of horizons, which avoids computationally complex and time consuming 3D reconstruction. The new approach combines nonparametric smoothing and classification techniques which are applied directly to the seismic data, with novel graphical representations of the intermediate steps introduced. For each sensor position, potential horizon locations are identified along the corresponding time-series traces. These candidate locations are then examined across all traces and when consistent patterns occur the points are linked together to form coherent horizons.

1. Introduction

There are many applications with multiple time-series data recorded at high frequency, such as environmental science, geophysics, financial trading, and internet marketing. Often the aim of the analysis is to identify events which occur across many data series but not necessarily at the same time. Rapid acquisition means that a full analysis may be impractical and so simple yet reliable methods are needed. An exemplary application is geophysical surveying where large datasets are obtained but only limited questions need answering, such as the following: is there an oil field? And if so, what is its location? In such cases, there is no need to perform a full 3D reconstruction of the study area and then interpret the reconstruction. Instead, it is possible to produce an answer directly from the data. A seismic study will form motivation for the proposed approach, but the method could be equally applied to other situations where the aim is to identify coherent features across multiple time series.

A seismic dataset consists of records of reflected signals measured by a number of seismic sensors placed at the nodes of a lattice on the earth surface. Figure 1 presents a diagram of the forming of reflected seismic signals. For details and a review of seismology, see, for example, Shearer [1], Sheriff and Geldart [2], Waters [3], and Hatton et al. [4]. The data can be treated as adjacent time series, called traces, indicating the arrival time of artificially generated sound waves reflected by underground features. An incident sound wave is reflected primarily by boundaries between layers with differing physical properties. Strong reflections are named horizons and indicate boundaries between distinct layers, or strata.

The main purpose of the analysis is to identify geometric features automatically, which represent strong reflections, and describe the horizons. The tracking of horizons is not easy in many cases, for example, when dealing with noisy measurements or across faults. A fault is a vertical crack in the rock which is recognized in seismic data by discontinuities in
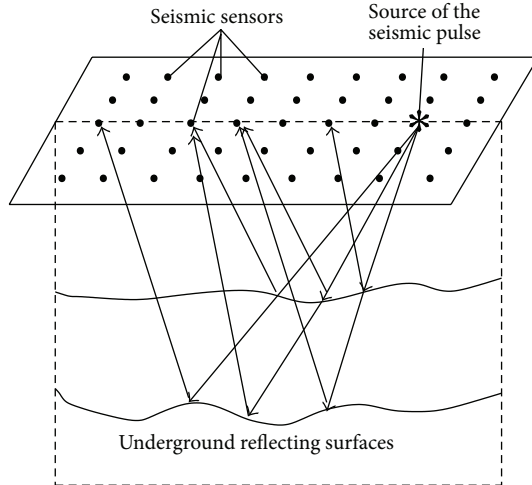
FIGURE 1: Diagram showing the formation of the reflected seismic signals.

the horizon. The location of faults is an important task for the structural interpretation of seismic data.

There are several approaches and many contributions that have been made to solve seismological problems. One approach is to perform a full 3D reconstruction of the sub-surface from the surface measurements. This type of problem is well known and is referred to as an ill-posed inverse problem which often requires substantial regularization. A general background to inverse problems can be found in Ribés and Schmitt [5] and in a mathematical discussion in Stuart [6]. Even in 2D the reconstruction process is usually mathematically and computationally challenging, and performing a full 3D reconstruction for a large scale seismic problem (possibly involving terabytes of data) may be impractical and ineffective—this is in contrast to other geophysical data collection methods (see, e.g., the review paper of [7]). Also, the reconstructed distribution would still need postprocessing to identify features, and a major drawback of this two-stage approach is that the regularization required in the first step may mask the features of interest. Instead, other types of analysis have been more widely studied looking to produce surface maps to indicate what is beneath the surface. Lendzionowski et al. [8] and van der Baan and Paul [9] use pattern recognition methods in seismology. In particular, they compare the traces against a library of patterns from known features with the aim of identifying similar features. The key to success is the construction of a varied library as otherwise features of interest could be ignored. Walden [10] uses simple summary measures of the traces followed by projection pursuit and thresholding to classify the traces into homogeneous groups. This, however, was performed on the entire trace and hence no depth information is retained. Also, there is no spatial smoothing, meaning that some locations can be classified differently to neighbouring locations simply due to the effects of noise.

The rest of the paper is organized as follows. In Section 2, a model is presented which describes the full multistage process leading to a seismic dataset. This model will be used

to generate synthetic data, but it will not be used for data analysis. Instead, the method of data analysis is described in Section 3 along with examples. Finally, there is a conclusions' section followed by references.

2. A Model for Seismic Data

In this section, the standard model for seismic data is described. In this paper, the model has been used for data generation, but it is not used in the proposed method of analysis. The use of a realistic seismic model will mean that synthetic data mimic real data. If the aim of an analysis is full 3D subsurface reconstruction, then such a model would also be used in analysis; however, in this work, the proposed method detailed in the next section interprets the data directly.

Suppose that acoustic sensors are placed on the earth's surface at n distinct locations forming a regular grid across some study regions. At time $t_0 = 0$ a seismic pulse is generated and data recording starts. The energy from the pulse spreads through the ground with some proportion being reflected back towards the surface. The sensors record this reflected energy at n equally spaced time points, $\mathbf{T} = \{t_i : i = 1, 2, \ldots, n\}$, spread over several seconds before data collection stops. Hence, at each location, a time series, $\mathbf{y} = \{y_i : i = 1, 2, \ldots, n\}$, is recorded. Let the n sensor locations be denoted by $\mathbf{S} = \{\mathbf{s}_j : j = 1, 2, \ldots, m\}$. Collecting the multiple time series together produces the full seismic dataset, $\mathbf{Y} = \{y_{i,j} : i = 1, 2, \ldots, n, j = 1, 2, \ldots, m\}$, where $y_{i,j}$ represents the reflected energy measured at time t_i and location $\mathbf{s}_j$.

The amount of reflected energy depends on the acoustic impedance distribution of the subsurface, as well as proper-ties of the initial seismic pulse. A common model used in reflection seismology is that of a stack of horizontal layers (see [4, 11]). Suppose that there are K distinct layers with the bottom of the layers at depths $\mathbf{D} = \{d_k : k = 1, 2, \ldots, K-1\}$—note that the lowest layer is assumed to stretch to infinity and hence there is no value for d_K—and that the corresponding acoustic impedances are $\mathbf{W} = \{\omega_k : k = 1, 2, \ldots, K\}$. In turn, the acoustic impedance is determined by the product of the density, ρ, and the acoustic velocity, v, of the material in the layer. Typical values for density (in g/cm^3) and acoustic velocity (km/sec) are water $\rho = 1.0$, $v = 1.4$; oil $\rho = 0.8$, $v = 1.3$; and rock $\rho = 1.5$–3.0, $v = 3.0$–6.0. This means that it is possible for two materials to have different density and acoustic velocity but to have indistinguishable acoustic impedance values.

Where two layers touch, there is an acoustic impedance contrast described by the reflection coefficient. This is defined as the difference between the acoustic impedances of the two layers divided by their sum. So the reflection coefficient, γ_k, at the bottom of layer k is given by

$$\gamma_k = \frac{\omega_{k+1} - \omega_k}{\omega_{k+1} + \omega_k}, \quad k = 1, 2, \ldots, K-1. \tag{1}$$

In a region with multiple layers, there will be several reflecting surfaces at different depths. A reflection coefficient series, or spike series, records the depth and value of the various
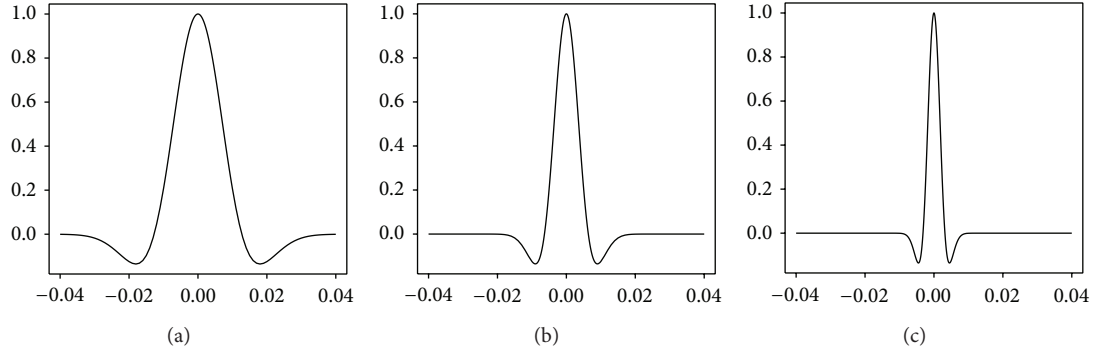
FIGURE 2: Ricker wavelet of varying dominant frequency: (a) 25 Hz; (b) 50 Hz; (c) 100 Hz.

reflection coefficients and can be written as $\boldsymbol{\gamma} = \{\gamma_k : k = 1, 2, \ldots, K-1\}$ at depths $\mathbf{D} = \{d_k : k = 1, 2, \ldots, K-1\}$. This series therefore encodes the information about the horizons, in the sense that the arrival time of a reflection is directly proportional to the location of the spike, d_k, and the amplitude of the spike represents the strength of the reflection, γ_k.

The model of recorded seismic data is completed by the convolution of the reflection coefficient series with a transfer function which is often taken as the Ricker wavelet [12]. The Ricker wavelet is defined as the scaled second derivative of the Gaussian function written as

$$w(d, t) = \left(1 - 2\pi^2 f_R^2 \left(t - \frac{d}{\overline{v}}\right)^2\right) e^{-\pi^2 f_R^2 (t - d/\overline{v})^2}, \quad (2)$$

where d is the depth of the reflecting surface, t is the time delay for the reflection to return to the surface, f_R is the dominant frequency of the sound wave, and $\overline{v}$ is the average velocity of the sound wave along its path. The transfer function relates the time taken for the energy pulse to cover a particular distance. Examples of the Ricker wavelet for three different dominant frequencies are shown in Figure 2. The Ricker wavelet is wide for low frequencies and more compact for higher dominant frequencies. Although this means that higher frequency sound waves have the advantage of greater time resolution, such waves have the overwhelming disadvantage that they penetrate less deeply into the ground. Then, the discrete convolution model can be written as

$$f(t_i) = \sum_{k=1}^{K-1} w(d_k, t_i) \times \gamma_k, \quad i = 1, 2, \ldots, n, \quad (3)$$

and including the effects of measurement error gives the final data equation

$$y_i = f(t_i) + \epsilon_i, \quad i = 1, 2, \ldots, n, \quad (4)$$

where $\{\epsilon_i, i = 1, 2, \ldots, n\}$ are independent measurement errors modelled by a Gaussian distribution. Figure 3 shows a diagrammatic representation of the formation of a seismic signal which results from the convolution of the reflection coefficient and the Ricker wavelet. The horizon locations can be seen as spikes in the reflection coefficient series
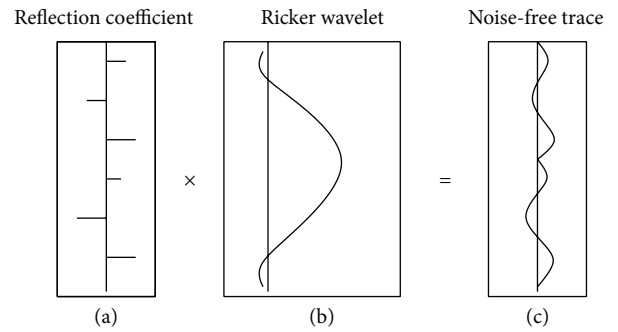

FIGURE 3: Diagram illustrating data production: (a) reflection coefficients; (b) Ricker wavelet; (c) seismic trace.

corresponding to contrasting acoustic impedance values. The convolution of this series with the Ricker wavelet leads to a blurred version of the reflection coefficient series. In practice, the reflection coefficient series will be contaminated by minor acoustic impedance contrasts due to natural variation and by random noise describing other sources of error.

Now, consider an example of the simulation of a synthetic dataset using the above model. Consider a sedimentary geological region which consists of several rock layers bedded horizontally upon each other with each layer having a characteristic, but naturally varying, acoustic impedance. This latter point will create weak reflections due to low-contrast changes in acoustic impedance within a layer creating clutter, which may partially obscure the true horizon locations. This will have the effect of contaminating the reflection coefficient series which will be transformed into correlated errors in the seismic trace—in addition to the independent measurement errors. Further, suppose there is a fault causing one part of the region to be deeper relative to the surrounding area. As a particular example, Figure 4 shows the seismogram of traces along a transect—in these traces moderate natural variation and measurement noise have been included. The dataset consists of 21 traces, each having 361 readings at equal time intervals. Three horizons can easily be identified, one curving up towards the top-right and the other two almost horizontal in the bottom half of the picture. The discontinuity in the horizons is caused by the presence of a fault towards
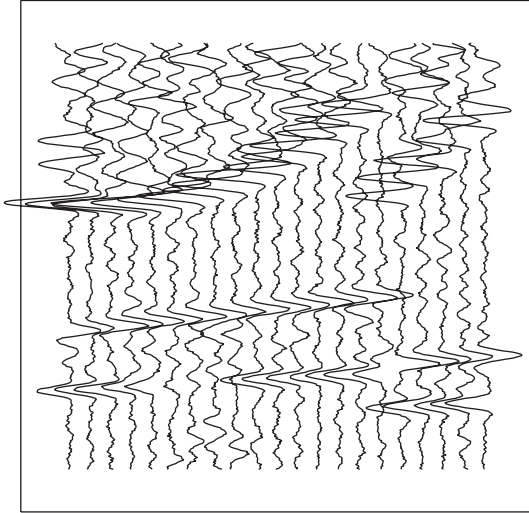
FIGURE 4: Seismogram showing multiple traces in a synthetic dataset.

the right-hand side. There is also a gap in the lowest horizon towards the left-hand side.

3. Method of Analysis

3.1. General Strategy. The method proposed here takes as its starting point the seismogram. Although this may have already been preprocessed, it is assumed that this involved little or no smoothing. Hence, the method starts by considering smoothing to remove the effects of noise, before moving on to identify key features within the data.

The overall objective of the analysis is to identify the stratigraphic structure in the seismogram, that is, long, linear features which are called horizons. These represent lines of high acoustic impedance contrast which are associated with important changes in subsurface structure. In the presence of low noise, the points along a strong contrast horizon can easily be identified as turning points in the individual traces—see Figure 4. Even in this case, however, as the number of horizons is unknown, it might not be obvious how many turning points to take. Also, when the contrast is lower, relative to the noise level, identification will be further complicated as turning points caused by the noise and feature might sometimes have similar signal values.

In the proposed method, first, each time series trace is considered separately and candidate horizon locations are identified as local maxima and minima after smoothing. Then, these candidate points are classified into groups where members of each class have similar signals. Again, noise may mean that some points are incorrectly classified. The identified classes are defined as structural points. Finally, selected points in each class are linked to identify the horizons. The diagram in Figure 5 illustrates the steps of the analysis strategy.

3.2. Extracting Candidate Point Locations. Figure 6(a) shows a single trace, from the middle of the earlier seismogram, with

local minima and maxima shown as open and filled circles, respectively. The sets of the locations of the local maxima and minima are simply determined as

$$L_0^{\max} = \{t_i : y_i - y_{i-1} > 0, y_{i+1} - y_i < 0\},$$
$$L_0^{\min} = \{t_i : y_i - y_{i-1} < 0, y_{i+1} - y_i > 0\}. \tag{5}$$

In total, there are 61 turning points and although the three horizon locations can easily be seen, the remaining 58 are caused by the noise. In situations where there are higher noise levels, it will not be so easy to separate the extreme points created by horizons from those due to noise. This will mean that some horizons are lost and that spurious points contaminate the true locations. To reduce the influence of noise the traces are smoothed.

For a given trace, consider the n paired data points, $\{(t_i, y_i) : i = 1, \ldots, n\}$, related by an unknown nonlinear relationship, $f(\cdot)$, with additive noise; that is,

$$y_i = f(t_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{6}$$

where the noise, $\{\epsilon_i, i = 1, \ldots, n\}$, are independent, and identically distributed Gaussian random variables with zero mean and constant variance, σ^2.

To reduce the effects of noise, the unknown function can be estimated using smoothing splines (see, e.g., [13, 14]), which minimizes the objective function

$$J(\lambda) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(t_i))^2 + \lambda \int (f''(t))^2 dt \tag{7}$$

subject to the constraint that the fitted curve is twice differentiable. That is,

$$\widehat{f}_\lambda(t) = \arg\min_f \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(t_i))^2 + \lambda \int (f''(t))^2 dt \right\}. \tag{8}$$

In these, the integral is over the interval $[t_{(1)}, t_{(n)}]$, that is, the minimum and maximum of the data collection times. The first term, a residual mean-square, measures the discrepancy between the data and the fitted function and the second, an integrated squared curvature, measures the roughness of the fitted function, and λ is a smoothing or bandwidth parameter. To make the residuals zero, the fitted function must interpolate the data. In general, this will lead to a quickly varying function which will have a high curvature. In contrast, to make the curvature zero, the fitted function must be linear, which is unlikely to fit the data well, making the residuals large. The smoothing parameter, λ, controls the balance between these two measures. Figures 6(b) and 6(c) show the results of spline smoothing (using the R function smooth.spline, [15]) for small and moderate values of λ, respectively. In each, the major structure is still evident and some undesirable points are removed. However, care is needed to avoid oversmoothing which might eliminate important features. In the extreme, however, even the main structural features would be lost.
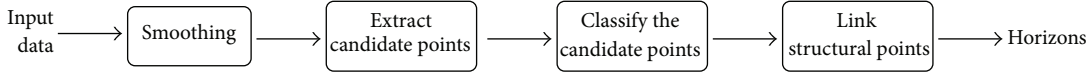
Input data → Smoothing → Extract candidate points → Classify the candidate points → Link structural points → Horizons
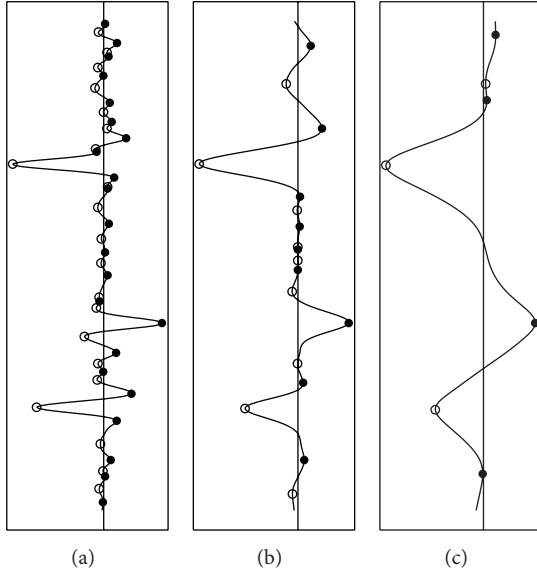

(a) (b) (c)

Figure 6: A seismic trace after increasing levels of smoothing: (a) low, (b) moderate, and (c) high smoothing. Also shown are the locations of local minima and maxima.

Figure 7: A single seismic trace after increasing levels of smoothing.

Figures 6(b) and 6(c) also show the locations of the local turning points of the smoothed function estimates, which can be defined as

$$L_\lambda^{\max} = \left\{ t_i : \widehat{f}_\lambda \left(t_i \right) - \widehat{f}_\lambda \left(t_{i-1} \right) > 0, \widehat{f}_\lambda \left(t_{i+1} \right) - \widehat{f}_\lambda \left(t_i \right) < 0 \right\},$$

$$L_\lambda^{\min} = \left\{ t_i : \widehat{f}_\lambda \left(t_i \right) - \widehat{f}_\lambda \left(t_{i-1} \right) < 0, \widehat{f}_\lambda \left(t_{i+1} \right) - \widehat{f}_\lambda \left(t_i \right) > 0 \right\}. \tag{9}$$

Here, and later, a scale-space approach (see, e.g., [16, 17]) will be considered. The idea is that one should not try to focus on a single smoothing level but consider several levels simultaneously with the aim of identifying any dominant scales. For any value of the smoothing parameter, λ, minimizing (7) will yield a fitted function, $\widehat{f}_\lambda$, and hence this can be repeated for a range of values of the smoothing parameter. Figure 7 shows a set of smoothed traces. As the level of smoothing increases, from left to right, the resulting smoothed traces become smoother and smoother until they are a straight line on the far right. The best smoothed trace will be somewhere within this set.

Recall that the aim here is not to produce a "nicely smoothed trace," but, instead, to accurately locate the turning points which correspond to the horizons. Hence the choice of smoothing parameter focuses on this output rather than on a more usual global goodness-of-fit measure. To help with the analysis, a novel turning-point tree is proposed with an example shown in Figure 8 where the locations of
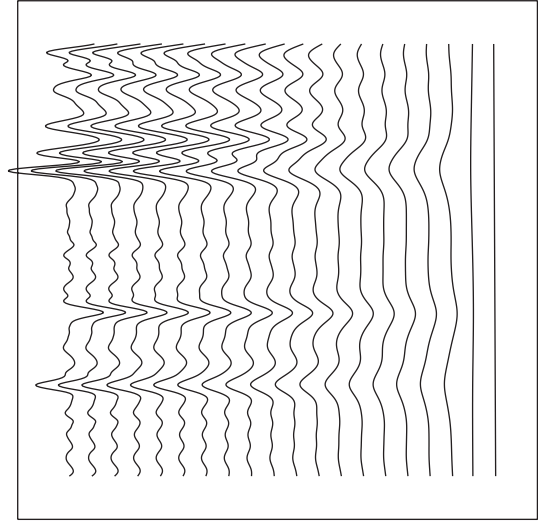
all minima and maxima are shown as the smoothing level changes. That is, the tree shows a graphical representation of the sets $L_\lambda^{\max}$ and $L_\lambda^{\min}$ as functions of λ. The bottom edge of Figure 8 corresponds to no smoothing and every turning point in the data appears. As the smoothing increases, moving upwards within the graph, the turning point locations change and sometimes turning points disappear. Note that, as the smoothing increases, pairs of neighbouring maxima and minima move towards each other before disappearing. Gradually, the number of turning points decreases until at the top of the graph only the three principal turning points are seen and eventually even these will disappear. The horizontal dotted lines correspond to the low, moderate, and high smoothing situations shown in Figure 6. Overall, it seems that turning points associated with true horizons remain for a very wide range of levels of smoothing and hence the method does not seem to be too sensitive to the exact choice.

To summarize the threshold tree in Figure 8, consider the number of turning points, N_λ, for each value of the smoothing parameter as shown in Figure 9(a). For small values of λ, there is little smoothing and hence there are many turning points, but as λ increases the number of turning points reduces. Turning points due to noise should be removed quickly, but true turning points due to the horizons should remain. To select an appropriate value of λ, it is proposed to use the value corresponding to the greatest change in the number of turning points. This derivative is shown in Figure 9(b) with the maximum (absolute) value shown as a vertical line at about $\lambda^* = 0.62$. Figure 6(b) shows a smoothed trace using the corresponding smoothing parameter value. Figure 10(a) shows the full dataset after
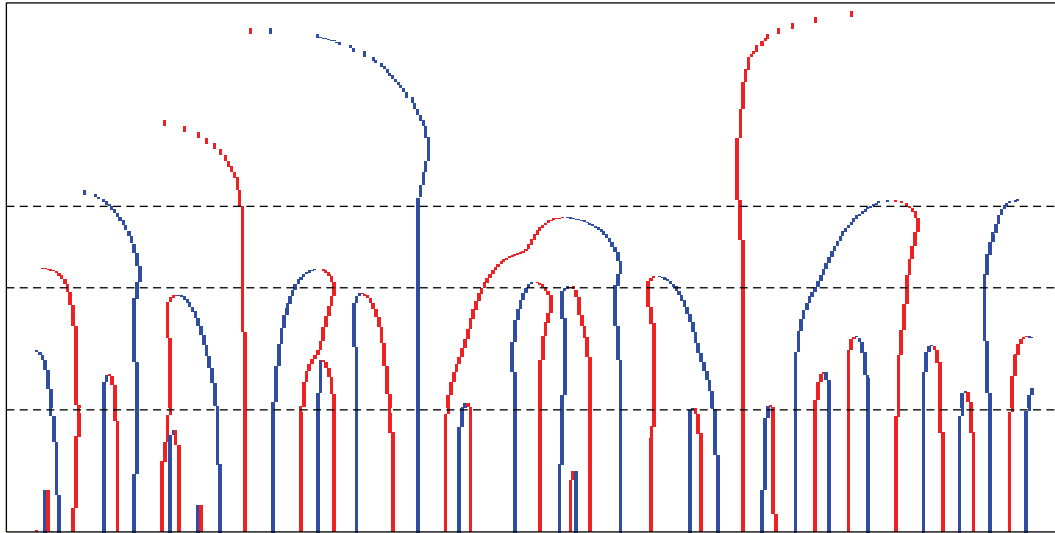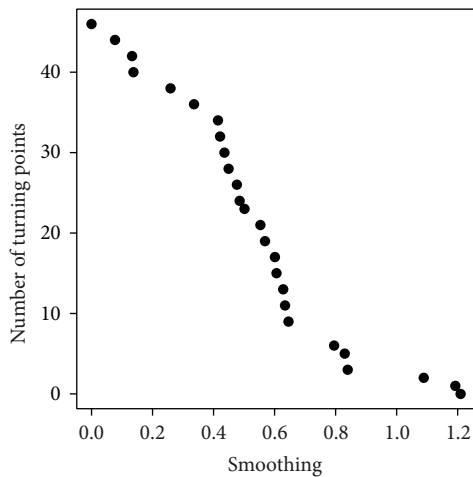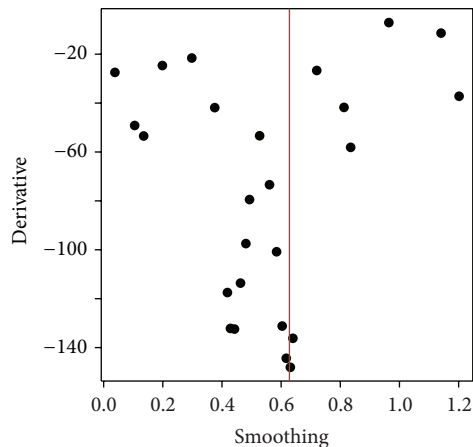
FIGURE 8: Threshold tree showing the location of the turning points (blue minima, red maxima) for varying levels of smoothing.

(a)

(b)

FIGURE 9: Graphs of (a) the number of turning points as a function of bandwidth and (b) the corresponding derivative and point of maximum absolute derivative.

smoothing using the automatically chosen bandwidth and Figure 10(b) shows the corresponding minima and maxima as open and filled circles. It is worth noting that this has located many more candidate points than the number of true turning points. This is important as, at this stage, it is vital that true horizon locations are not lost. The next stage will identify which candidate points form coherent features.

3.3. Candidate Point Classification and Structural Point Linking. The second stage in the procedure is to classify the candidate points into groups where members of a group form a single horizon. It is assumed that points along a horizon will have a distinct acoustic impedance contrast giving rise to a distinct reflectance coefficient and hence a distinct measured signal. Since each horizon is expected to have a different characteristic acoustic impedance contrast, the measured signals will be a mixture of an unknown number of distributions, each of an unknown shape. Also, the various sources of systematic and random variability may mean that measured signals from one horizon vary and may be similar to those from other horizons. It is impossible to obtain a parametric form for this distribution and so a kernel density approach is used to provide a nonparametric smoothed version of the signal strength distribution of the candidate points. For background to kernel density methods, see, for example, Wand and Jones [18] and Silverman [19].

Suppose that $\eta = N_{\lambda^*}$ turning points have been found, that their signal strengths are denoted by $\{u_i : i = 1, \ldots, \eta\}$, and that their locations along the traces are $\{v_i : i = 1, \ldots, \eta\}$; then the kernel density estimate of the signal strength distribution is

$$g_\delta(u) = \sum_{i=1}^{\eta} K\left(\frac{u_j - u}{\delta}\right), \quad -\infty < u < \infty, \qquad (10)$$
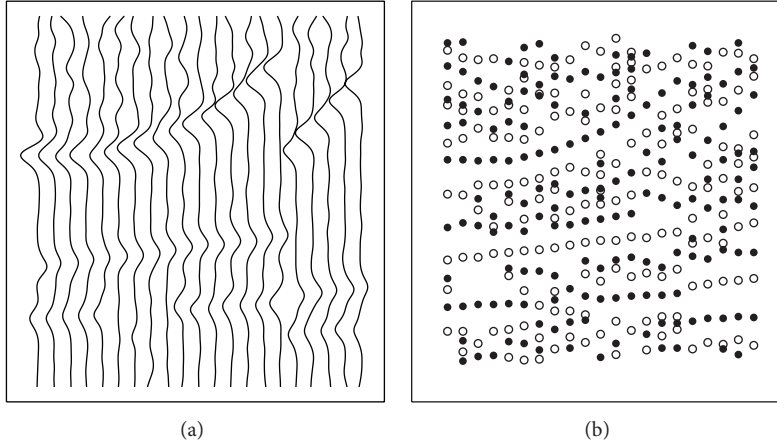
FIGURE 10: Multiple smoothed traces in (a) with corresponding candidate points in (b) where open and filled circles represent minima and maxima, respectively.
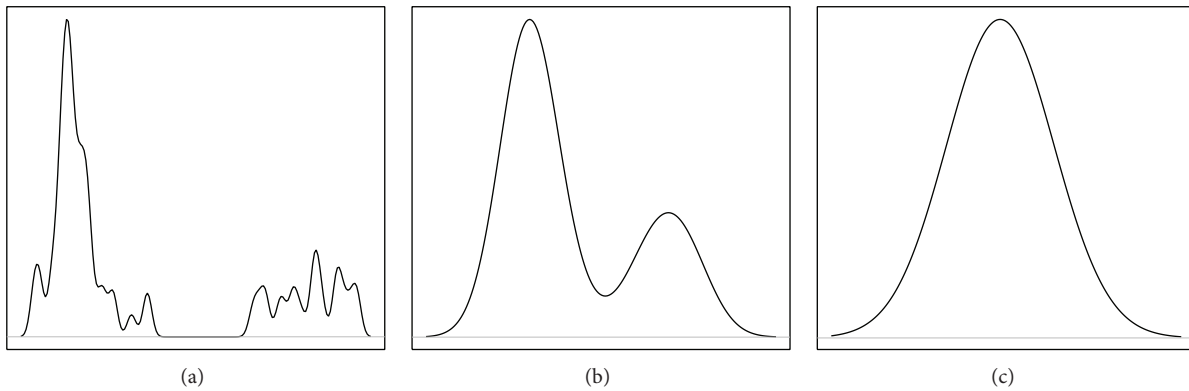

FIGURE 11: The curves represent kernel density with (a) low smoothing; (b) moderate smoothing; (c) high smoothing.

where $K(\cdot)$ is the kernel function and δ is a bandwidth parameter. A popular choice of kernel function is the standard Gaussian density. Despite the method being generally successful at finding structure, kernel density estimation can also ignore significant structure via oversmoothing or retain insignificant structure via undersmoothing, and hence the correct choice of δ is important. Figure 11 shows a simple bimodal dataset after varying levels of smoothing using the R function density [15]. We might guess that (a) is undersmoothed retaining unimportant random variation and that (c) is oversmoothed leaving little useful information. The central graph, (b), uses the default bandwidth, based on Silverman's rule of thumb [19] which seems to summarize the distribution well.

The number of modes in the kernel density estimate is regarded as the number of distinct classes and hence the number of horizons. The locations of the minima between the modes can be taken as thresholds which can then be used to classify the candidate points. Considering the three levels of smoothing used in Figure 11 would, however, lead to very different results. Rather than attempting to estimate a single level of smoothing, again a scale-space approach will be used (again see, e.g., [16, 17]). In this approach, many levels of

smoothing are considered allowing features at different scales to appear and allowing a natural scale to be identified.

Let the set of locations of the local minima, which are to be used as threshold, be defined as

$$\boldsymbol{\tau}_\delta = G_\delta^{\min}$$

$$= \{u_i : g_\delta(u_i) - g_\delta(u_{i-1}) < 0, g_\delta(u_{i+1}) - g_\delta(u_i > 0)\}. \tag{11}$$

Suppose that K thresholds are found in this way, hence defining $K + 1$ groups. For a particular bandwidth, δ, let the thresholds be denoted by $\boldsymbol{\tau}_\delta = \{\tau_k : k = 1, 2, \ldots, K\}$, which allows the members of the classes to be defined as

$$C_\delta^k = \begin{cases} \{u_i : u_i < \tau_1\} & \text{for } k = 1, \\ \{u_i : \tau_{k-1} < u_i < \tau_k\} & \text{for } 2 \le k \le K, \\ \{u_i : u_i > \tau_K\} & \text{for } k = K + 1, \end{cases} \tag{12}$$

and the number in each class as N_δ^k for $k = 1, 2, \ldots, K$.

Figure 12(a) shows the relation between the locations of the thresholds and the level of smoothing defined in terms
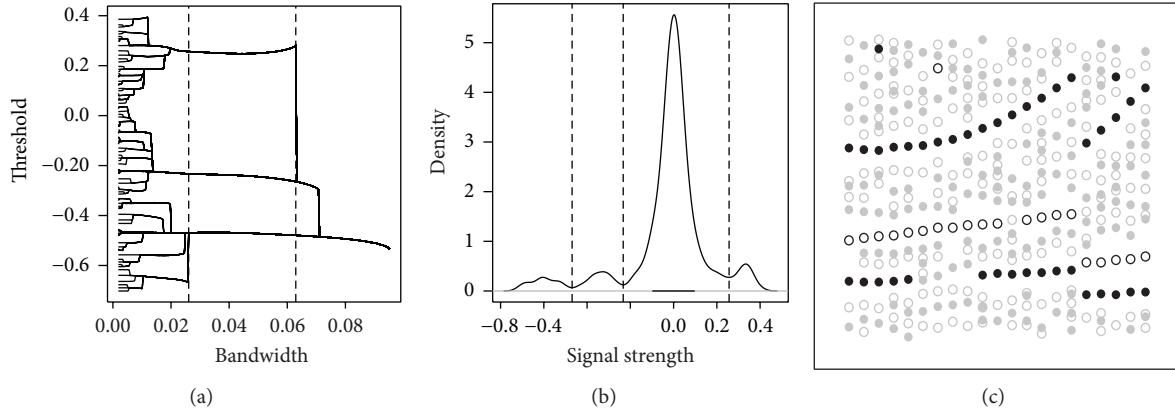
FIGURE 12: (a) Threshold tree; (b) histogram of kernel density; and (c) turning point locations.

of the bandwidth δ. The change in the bandwidth leads to a change in the number and value of the thresholds and hence in the number of classes. As the bandwidth increases, the number of thresholds decreases, and when the curves combine into a single line, there is only a single threshold, which means in turn that the number of groups is two. The procedure is stopped when the smoothed distribution becomes unimodal. In the proposed method, the best classification is defined to be when the number of thresholds is stable for the longest.

In this example, three thresholds persist the longest, a bandwidth interval indicated by the vertical dotted lines, and hence suggests four clusters. The optimum smoothing bandwidth, δ^*, is then taken as the smallest value in this interval. The corresponding kernel density estimate is plotted in Figure 12(b); the vertical dotted lines show the position of the thresholds. It is now important to realise that the cluster surrounding signal strength zero includes turning points generated by the random noise and should be excluded. This will be considered now, before moving to the final linking step.

The observed signal strength distribution will be a mixture of background values and values due to true horizons. Of course the various means and the variance of these components are unknown. To describe the distribution of the signal values from the background, the following approach is proposed. The centre of the background signal strength distribution is estimated using the median of the full data

$$\tilde{\mu} = \operatorname*{median}_{j} u_j, \tag{13}$$

and the variance estimate is based on the median absolute deviation (MAD)

$$\tilde{\sigma}^2 = \frac{\text{MAD}}{\Phi^{-1}(3/4)} \quad \text{with MAD} = \operatorname*{median}_{j} \left| u_j - \tilde{\mu} \right|, \tag{14}$$

where the value $1/\Phi^{-1}(3/4) \approx 1.48$ is chosen so that when the values follow a Gaussian distribution, then $E[\tilde{\sigma}^2] = \sigma^2$. These estimators are chosen to be robust to outliers

caused by the signal strengths from the horizons. The solid horizontal line near the bottom of Figure 12(b) represents a 95% confidence interval for background signal strength. The calculation assumes a Gaussian distribution and uses the above mean and variance estimates. This interval encloses the majority of this zero-centred mode and confirms that it is reasonable to exclude the whole group. Before moving on, it is worth noting that, for datasets containing no horizons, all turning points will form a single group which will be identified as a background group, and hence the procedure would not identify any horizons.

Figure 12(c) shows all the minima and maxima as open and filled circles, respectively. The grey circles correspond to the background group and the black circles to the horizon groups. The algorithm achieves good results in extracting and classifying the structural points which make the horizons, but notice that, although the horizons can be clearly seen, there are three extra points near the top and one missing along the central horizon. Also, although not visible, there are six points misclassified to the wrong group. This means that linking would not be perfect with some crossing of the proposed horizons. This suggests that further criteria must be included in this final linking process. As a simple example, here, isolated structural point will be excluded from the linking. In particular, if the distance to the nearest other point in the group is greater than, say, twice the average within-group interpoint distance then is excluded.

Let the ordered set of all locations along the traces of member of class k be denoted by $\{v_i : 1, 2, \ldots, N_{\delta^*}^k\}$. Then consider the median distance between neighbouring points in the class

$$d^k = \operatorname*{median}_{i} \left| v_{i+1} - v_i \right| \tag{15}$$

and the distance from a point to the nearest other point in the class

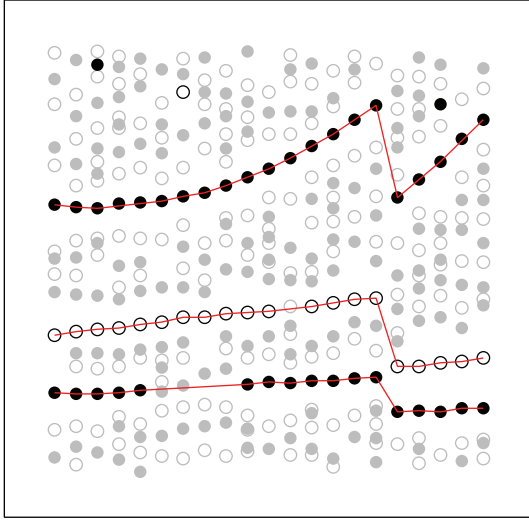$$m_i^k = \min_{j} \left| v_i - v_j \right|. \tag{16}$$

FIGURE 13: Linked structural points with modified method.

This allows the following modified rule for class membership to be defined:

$$C_{\delta^*}^k = \begin{cases} \left\{v_i : v_i < \tau_1, m_i^k < 2 \times d^k\right\} \\ \qquad \text{for } k = 1, \\ \left\{v_i : \tau_{k-1} < v_i < \tau_k, m_i^k < 2 \times d^k\right\} \\ \qquad \text{for } 2 \le k \le K, \\ \left\{v_i : v_i > \tau_K, m_i^k < 2 \times d^k\right\} \\ \qquad \text{for } k = K + 1. \end{cases} \qquad (17)$$

Applying this final rule to the previous example produces three groups of structural points forming clear and coherent horizons as shown in Figure 13. Although not shown here, it has also been applied to examples with higher noise levels with equal success.

4. Conclusions

Seismic surveys play an important role in locating underground features, but the partial reflection and natural variation in acoustic properties mean that the recorded acoustic signals provide a confusing representation of the study area. Full 3D datasets can be enormous, covering several square kilometres on the surface and representing several kilometres in depth with high frequency temporal sampling. This suggests that rapid analysis is needed even if interesting examples require further analysis or examination by experts. This also suggests that simple techniques are preferable with full reconstruction reserved for follow-up analysis. Raw data contain systematic distortions and significant noise and typically they undergo substantial preprocessing before analysis is started. The preprocessing aims at reducing the distortion and reducing noise but can introduce artefacts.

The approach proposed here is based on simple statistical procedures which aim at smoothing out noise and using contextual information to identify coherent horizons. In

particular, smoothing splines are used to smooth individual traces before local minima and maxima are located. It appears that greater smoothing can be applied to the traces, without overwhelming the major features, than would be the case if the aim was to estimate the signal. Once a smaller set of candidate points is found, they can be classified into groups where each group represents points along a horizon. Classification is performed by thresholding the signal strength distribution so that each mode represents a class. Again smoothing is needed to help select the correct number of modes, and hence classes, in the distribution. This time kernel density methods are used. Here, the choice of bandwidth is much more important as even small changes can lead to a substantial change in the smoothed density. A scale-space approach is adopted, which first allows all levels of smoothing to be considered. The best level of smoothing, and hence the number of clusters and the number of horizons, is then determined as the value for which the number of clusters persists the longest. The cluster associated with the random variation in the background is identified and excluded from horizon linkage. Locations within each cluster are then linked to form horizons. In some high noise examples, further information is used in this linking stage to exclude points which are distant from the main cluster and hence unlikely to be part of the horizon. This does not prevent, however, horizons from being linked across gaps and across faults. Clearly, it would be possible to supplement this fine tuning with other criteria as necessary. Based on the experiments conducted here, the proposed method shows great promise. It provides a simple and quick approach to locate and link horizons even in relatively noisy datasets. It is not affected by gaps in horizons nor by discontinuities caused by faults.

Finally, it is important to note that the proposed method is not specific to seismic studies but can easily be applied to other problems where the aim is to identify coherent structure within 2D or 3D data. Also, the use of smoothing and threshold trees gives a simple graphical tool for any classification problem, and the use of the most persistent thresholding is applicable to many other situations.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

References

[1] P. M. Shearer, *Introduction to Seismology*, Cambridge University Press, Cambridge, UK, 1999.

[2] R. E. Sheriff and L. P. Geldart, *Exploration Seismology*, Cambridge University Press, Cambridge, Mass, USA, 2nd edition, 1995.

[3] K. H. Waters, *Reflection Seismology*, John Wiley & Sons, New York, NY, USA, 1978.

[4] L. Hatton, M. Worthington, and J. Makin, *Seismic Data Processing: Theory and Practice*, Blackwell Scientific Publications, London, UK, 1986.

[5] A. Ribés and F. Schmitt, "Linear inverse problems in imaging: an introductory survey," *IEEE Signal Processing Magazine*, vol. 25, no. 4, pp. 84–99, 2008.

[6] A. M. Stuart, "Inverse problems: a Bayesian perspective," *Acta Numerica*, vol. 19, pp. 451–559, 2010.

[7] W. W. Symes, "The seismic reflection inverse problem," *Inverse Problems*, vol. 25, no. 12, Article ID 123008, 39 pages, 2009.

[8] V. Lendzionowski, A. T. Walden, and R. E. White, "Seismic character mapping over reservoir intervals," *Geophysical Prospecting*, vol. 38, no. 8, pp. 951–969, 1990.

[9] M. van der Baan and A. Paul, "Recognition and reconstruction of coherent energy with application to deep seismic reflection data," *Geophysics*, vol. 65, no. 2, pp. 656–667, 2000.

[10] A. T. Walden, "Spatial clustering: using simple summaries of data to find the edge of an oil-field," *Applied Statistics*, vol. 43, pp. 385–395, 1994.

[11] D. W. Oldenburg, T. Scheuer, and S. Levy, "Recovery of the acoustic impedance from reflection seismograms.," *Geophysics*, vol. 48, no. 10, pp. 1318–1337, 1983.

[12] N. Ricker, "The form and nature of seismic waves and the structure of seismograms," *Geophysics*, vol. 5, pp. 348–366, 1940.

[13] C. de Boor, *A Practical Guide to Splines*, Springer, New York, NY, USA, 2001.

[14] P. J. Green and B. W. Silverman, *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman and Hall, 1994.

[15] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013, http://www.R-project.org.

[16] T. Lindeberg, "Scale-space theory: a basic tool for analyzing structures at different scales," *Journal of Applied Statistics*, vol. 21, pp. 225–270, 1994.

[17] A. P. Witkin, "Scale-space filtering," in *Proceedings of 8th International Joint Conference on Artificial Intelligence*, pp. 1019–1022, Karlsruhe, Germany, 1983.

[18] M. P. Wand and M. C. Jones, *Kernel Smoothing*, Chapman & Hall/CRC, 1995.

[19] B. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall/CRC, 1998.