Journal of Data Analysis & Operations Research
ISSN 2384-4671
J. Data Ana. Oper. Res. 1, No. 1 (Jun. 2014)

Using Seismic Signals to Identify Edges of Homogeneous Geological Structures

Raed A. T. Said

Al Ain University of Science and Technology; Al Ain, United Arab Emirates

Received: 15 Jun. 2013, Revised 25 Dec. 2013; Accepted: 6 Mar. 2014, Published: Jun. 2014

Abstract: Seismic signal segmentation is typically based on the statistical analysis of recorded seismic data from the exploratory stage. Quite often edges of all homogeneous regions are also captured as useful parameters for attaining optimization levels with minimal rates of misclassification. This paper investigates the applicability and limitations of selected current statistical techniques in use for the purpose of developing enhanced methods for identifying the edge of an oil field using information obtained from seismic signals. It then proposes a robust algorithm for seismic signal segmentation with its key function being to determine whether different seismic signals belong to homogeneous or heterogeneous regions. The algorithm is based on simple mechanics - identification of the most useful variables in the detection of different types of lateral change of interest in petroleum reservoirs and hence identification of the most efficient discriminating criterion. Its novelty derives from the construction of space-time models with more flexible properties than those of spatial models. This is mainly because, on the one hand, the spatial variation of the geographical position points may correlate with the pinch-out edge of the sand. On the other hand, the dependency between time and space is taken into consideration since it is not reasonable to assume that spatial locations of data occur regularly, as do most time series models, where past, present and future have no analogy in space.

Keywords: Data Visualisation, Image Segmentation, Predictive Modelling, Seismology, Spatio-Temporal modelling, Space-Time Models**.**

1. Introduction

Segmentation of a given region into homogeneous sub-regions having the same geological structure is important to estimate the capacity of new resources such as a new oil field, a new water basin or a new mineral field. The segmentation can be performed using statistical analysis of recorded seismic signals used in the exploratory phase. The identification of the edges of every homogeneous region is also useful for selecting the points at which the extraction will take place to reach optimal production levels. In seismology, for instance, seismic signals are realisations of time series and so the segmentation process reduces to classifying the signals into different groups so that each group will consist of signals that are generated by the same model. This approach is appealing in that the model coefficients of each signal depends on the geological structure of the various layers composing the underground at the point which it was recorded.

This paper proposes a robust algorithm for seismic signal segmentation with its key function being to determine whether different seismic signals belong to homogeneous or heterogeneous regions. It investigates the applicability and limitations of selected current statistical techniques in use for the purpose of developing enhanced methods for identifying the edge of an oil field using information obtained from seismic signals. The proposed novel algorithm is based on simple mechanics - identification of the most useful variables in the detection of different types of lateral change of interest in petroleum reservoirs and hence identification of the most efficient discriminating criterion. Its novelty derives from the construction of space-time models with more flexible properties than those of spatial models.

1.1 Background and motivation

In the last couple of decades, heavy investment in search of enhanced methods of oil exploration have been applied in the discovery and exploration of new oil reserves in a number of countries in Africa and South America (Eaton et al., 2003). One of the criteria applied after the discovery of a new oil field is to identify its edges in order to estimate the field capacity, drilling point and hence its feasibility. Almost a couple

E-mail: raeed.tawfeq@aau.ac.ae, raedsaeed@yahoo.com

of decades ago, Walden (1994) proposed an algorithm for summarizing the information contained in each seismic signal in a 3-dimensional point $(\alpha, \beta, \gamma)'$ where α, β and γ are physically appealing statistics and potentially useful for detecting changes in the generating process of the signals using the projection pursuit technique via the optimization of a criterion function. Projection pursuit technique helps to explore multivariate data in interesting low dimensional spaces. The purpose of projection pursuit optimization is to find all of the interesting projections. The definition of an "interesting" projection depends on the projection pursuit index and on the application or purpose (Lee et al., 2005). For the segmentation purpose in this work we used an entropy index presented in (Jones and Sibson, 1987) as a projection pursuit index.

The heavy investment required in most geological explorations in search of enhanced methods of oil exploration implies that developing novel methods potentially adds up to the cost of explorations. It is therefore reasonable to design and develop flexible approaches which can be adapted to different situations. Olhede and Walden (2005) introduced a flexible approach based on time-frequency analysis of multicomponent signals which focuses mainly on additive noisy data. Thompson et al., (2010) seek to reduce noise using homogeneous image regions, which they refer to as "superpixels". Their main idea is to combine homogeneous image regions with "endmember extraction" to produce concise mineralogical summaries seismic signal-dependent algorithms. Segmentation of geological structures in this paper is motivated by the current practices as outlined above. The segmentation can be performed using statistical analysis of recorded seismic data that were used in the exploratory phenomena and the identification of the edges of every homogeneous region is useful for attaining optimal levels with a minimum error rate. So that, edges of homogeneous regions are captured as useful parameters for attaining optimization levels with minimal rates of misclassification. We propose a robust algorithm for seismic signal segmentation with its key function being to determine whether different seismic signals belong to homogeneous or heterogeneous regions. The paper's main objectives can be summarised as follows.

a. To highlight the interactions between the geological structures of the underground and the parameters of time series models to confirm or discard the Walden (1994) model and
b. To produce an enhanced approach to seismic signal segmentation.

The paper is organized as follows. Methods are described in Section 2, data analyses and discussions in Section 3 and conclusions and potential new research directions in Section 4.

2. Methods

Walden (1994) suggested an approach for classifying N signals into two groups so that signals of one group characterize one single homogeneous region whose edges are inferred from the corresponding pixels of the signals. As a first step, his approach consists of reducing the problem of time series classification to a classical one of a sample of N 3-dimensional observations once each signal $S_t, t = 1,2,...,N$ had been summarized in a 3-dimentional statistics $d_t = (\alpha_t,\ \beta_t,\ \gamma_t)'$ where α is its energy, β is the index / time at which half of α is reached, and γ denotes the central frequency of the signal S_t. Secondly, he transformed the obtained sample $\{d_t\ \ t = 1,..N\}$ into a univariate sample $\{x_t\ ,\ t = 1,..,N\}$ using the projection pursuit algorithm (For detailed explanations, see Jones and Sibson,1987) where x_t is a linear combination of the three summary statistics α_t , β_t , and γt i.e. :

$$x_t = a_1\alpha_t + a_2\beta_t + a_3\gamma_t \text{ for } t = 1,...,N \qquad (1)$$

where the scalars a_1, a_2, and a_3 characterize the optimal line in the space on which $\{d_t\ \ t = 1,..N\}$ are projected. Since we are concerned with clustering problem, the optimal line is such that the univariate sample $\{x_1,...,x_N\}$ is composed of two separate clusters where each cluster is the projection of a subset of the points which are close to each other. Consequently, if two points d_i and d_j are close to each other in the space, then $\alpha_i \approx \alpha_j$, $\beta_i \approx \beta_j$ and $\gamma_i \approx \gamma_j$ which could indicate that the two original signals S_i and S_j are similar. Thus, by Walden (1994) technique, the projection line is chosen to maximize the entropy of the kernel density function $f_a(.,h)$ obtained from $\{x_1,...,x_N\}$. This choice is based on the argument that the entropy measures the deviation of the density $f_a(.,h)$ from the normal density. The third step of Walden's method consists of separating the

univariate sample $\{x_1,...,x_N\}$ into two subsamples where the first one is composed of the $x_i's$ that are smaller than the anti-mode, say x^*,- located between the two modes of the bimodal kernel density estimate $f_{\hat{a}}(.,h)$ and the second consists of the $x_i's$ that are greater than x^*. Finally, the two regions are recovered from the pixels of the two classified signal groups. In this work we investigate the properties of the foregoing approach focusing on:

a. The appropriateness of the variables α, β and γ for summarizing seismic signals.
b. The efficiency of the entropy as a criterion for identifying the optimal line.
c. The application of Walden's method on a real data set.

1.2 Data source and description of a typical data grid

Our proposed method was applied on real data set described in Said (2007). The data were obtained from the Department of Geophysics at the University of Leeds. The data consist of 861 signals of 76 observations each. The signals were recorded in a 21x41 pixel region the grid points of which are graphically presented in Figure 1. The rationale is that when a region $\Re$ of a given locality is explored for the first time, geophysicists analyze the seismic signals that are recorded at grid points forming the intersecting points obtained by dividing the region $\Re$ using N_1 X-lines and N_2 Y-lines as graphically illustrated in this plot.

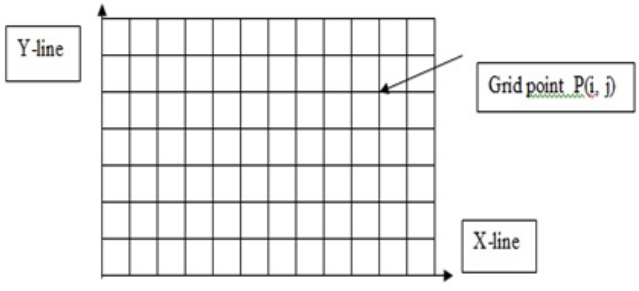

Fig.1: Standard partition of the region of interest

At each grid point $S(i,j)$ for $i=1,..N_1$ and $j=1,..,N_2$, a generated acoustic wave X_t is sent down underground. When X_t arrives to the kth interface characterizing a change in the geological structure at a depth y_k, a portion of it, $w_t = r_k X_t$, is reflected back to the ground and recorded on geophones at its arrival time. The recorded waves w_t, assumed to differ from each other only by the coefficient of reflections r_k and the arrival time, are processed to produce the seismic signal S_t. That is,

$$S_t = \sum_{j=0}^{+\infty} a_j r_{t-j} + e_t \quad \text{for} \quad t = 1,..., N \qquad (2)$$

where the sequence $\{a_t \ \ t = 0,1,.....\}$ is a sampled/discrete version of the input acoustic wave X_t which is itself a pulse of high energy and of short duration used to approximate the theoretical impulse signal (the dirac function) defined as follows.

$$\partial(t) = \begin{cases} +\infty & \text{if} \ \ t = 0 \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

Since X_t is of short duration, the value of the signal at time t depends principally only on a small number of reflection coefficients $r_i's$. In fact, one of the important concerns of the geophysicists is the estimation of the reflection coefficients $r_i's$ which are used in turn to determine the type of the rock, its depth and its width. The estimates are obtained by applying inverse filtering procedures assuming that the errors $e_t's$ are a realization of a white noise.

From this brief discussion on seismic signals theory, we deduce that the identification of the homogeneous regions should be carried out using only portions of the seismic signals rather than the complete signal (Walden, 1994). This is mainly to reduce the effect of the reflection coefficients of the other layers, which are not of concern on the values of the signals. The rationale here lies in the need for monitoring differences in relatively short portions of seismic signals which arise due to differences in the reflection coefficients of the concerned layers and its neighbors. In other words, the approach cross validates the model to balance out the potential randomness in (St) likely to arise from random training and testing samples. It therefore allows for the estimates to be obtained by

applying inverse filtering procedures assuming that the errors are a realization of a white noise.

1.3 Seismic signal variables

As noted above, Walden (1994) summarized the information contained in each signal S_t in a 3-dimensional point $(\alpha_t, \beta_t, \gamma_t)'$ where α_t, β_t, and γ_t are physically appealing statistics and potentially useful for detecting changes in the generating process of the signal S_t. The first variable α_t is the energy of the recorded signal $S_t = \{S_1,...,S_n\}$, and it is defined as the sum of the squared of its amplitudes as follows

$$\alpha_t = \sum_{t=1}^{N} S_t^{\,2} = \sum_{t=1}^{N}\left(\sum_{j=0}^{n} a_j r_{t-j}\right)^2 \qquad (4)$$

From equation (4) we see that α depends on the reflection coefficients $r_k's$. Further, any difference between the energies of two signals is mainly due to the differences that exist between the two sequences of reflection coefficients that characterize the geological structure at two distinct points since the sequence $\{a_j \quad j = 0,...,n\}$ characterizes the common input pulse that is sent in the underground at the grid points. The variable α is appealing since the reflection coefficient differs from one type of rock to another. Thus, if the type of rocks constituting the underground at a point $S(i, j)$ differs from the existing rocks at another point $S(k,l)$, we expect that the energies of the two corresponding signals will be different. Unfortunately, this is not true in general since it is well known in time series that to each ARMA model with minimum phase, there is an ARMA model with maximum phase which has the same energy as the first one (Lii and Rosenblatt, 1982). Thus, Walden (1994) considered the second variable β which differs according to whether the signal is of minimum phase or maximum phase. He defined β as the index m at which half of the energy is reached. That is,

$$\beta_t = m \text{ such that } \sum_{i=1}^{m} S_{ti}^2 = \frac{1}{2}\alpha_t \qquad (5)$$

This variable is appealing because the build-up of the energy of signal with minimum phase is rapid whereas the build-up is slow for the energy of signals with maximum phase. In other words, β_t is a measure of the discrepancy between the phase spectra of the signal S_t whereas α_t measures the discrepancy between the amplitude spectrum of the signal.

To increase the efficiency of the classification procedure based on the summary variables α and β, a third variable, γ, reflecting the frequency content of the signal is included. The variable is appealing for seismic signals due to the fact that it presents a rectangular power spectrum of width B and centered at frequency f_c. That is, the difference between two seismic signals may be detected by the differences between their central frequencies f_c and their bandwidths. Walden (1994) defines γ as a nonparametric estimate of the central frequency f_c of the signal. It can be shown that the estimate of f_c is a function of the autocorrelation of lag j and to eliminate the effect of the lag, he estimates it by the median of M estimates $\hat{f}_c(j)$, j = 1, 2, …, M, where M is typically 4 or 5 and $\hat{f}_c(j)$ is defined as

$$\hat{f}_c(j) = \frac{\cos^{-1}\left(\frac{\pi j B \hat{\rho}_j}{\sin(\pi B j)}\right)}{2\pi . j} \quad \text{for a given } j \qquad (6)$$

where B is the bandwidth, to be estimated, of the rectangular spectrum, and $\hat{\rho}_j$ is the sample autocorrelation of lag j of the time series. Estimating frequency is however a more difficult problem than estimating amplitude and phase (Bloomfield, 2000). Walden and White (1990) and Walden (1994) gave a brief description to how they derived the formula $\hat{f}_c(j)$ and the unbiased estimator of the bandwidth B.

1.4 An optimization technique

The main issue in understanding the signal is how to determine the bandwidth or the smoothing parameter. Various approaches have been proposed in the literature to address this issue. Typically, an estimated

power spectrum is smoothed using one of the smoothing windows to eliminate the fluctuations. Projection pursuit – an approach proposed by Friedman and Tukey (1974) – performs exploratory data analysis as a way of by-passing the "curse of dimensionality" in multivariate data. For more information see Jones and Sibson(1987).Thus, the method can be used to transform the 3-dimensional data set $\{(\alpha_t, \beta_t, \gamma_t)' \quad t = 1,..N\}$ into a univariate data set $\{x_t\}$ which is obtained by projecting the initial data set of R^3 on a line of the space. The selection of the optimal projection line L_a depends on the problem of concern. For classification problems, L_a is chosen to preserve clustering property of the initial data – that is, close points in the space will be transformed into close univariate observations. The different k clusters in R^3 are transformed into k clusters on L_a as much separated as possible to reduce the percentage of misclassified observations. Silverman (1981) defines the optimal window width for the normally distributed data with a known standard deviation as $h = 1.06\sigma N^{-\frac{1}{5}}$. The optimal line is determined by trying various values and selecting one with the largest entropy. The main ideas here is that: departure of the entropy indicates departure of the density from normality.

3. Classification of real data

For classification purposes, the pixel signals data described above were recorded in a 21x41 pixel region as illustrated in Figure 2 based on grid points $S(16,25), S(17,34), \text{and } S(20,21)$. The question which we are interested in is whether they are generated by the same geological structure or not. The main difference, if it exists, is due to their shapes for $i > 60$. The form of the $S(20,21)$ signal doesn't differ from the other two signals for $i < 30$. The respective energies of the three signals are 81342, 92404, and 96153indicating that the $S(17,34)$ signal doesn't differ from the $S(16,25)$ signal. This remark leads us to conclude that the variable energy could be misleading since it does not take into account the number of zero-crossings which is important in comparing signals.

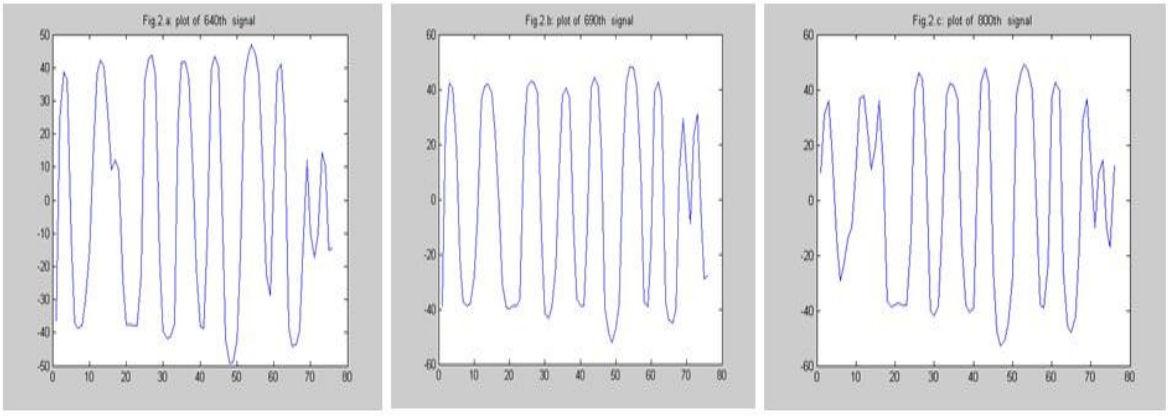

Figure 2. Plots of three different signals from the sampled data

On the other hand, the halftime variable,β of the three signals is 36, 39 and 42 respectively from which we notice that the build-up energy of 800th signal is slower than for the other two signals. This confirms what can be deduced from visual comparison of the three signals - the values of the 800th signal are relatively smaller than those of the two others at the beginning and then they become comparable to the values of the two others at the end. Also, this variable does not take into account the number of zero- crossings. The values of the third variable- that is, estimate of the central frequency, the observed values were 0.1058 , 0.1086and 0.1086 respectively which are very similar indicating that the center of the three power spectrums are close to each other. This leads us to raise the fact that differences in rectangular power spectrums depend on the signal bandwidth and the height (scaled-energy) of the rectangles. It is therefore reasonable to also include extra variables such as the number of zero-crossings and the bandwidth of the signals in summarizing the set of signals.

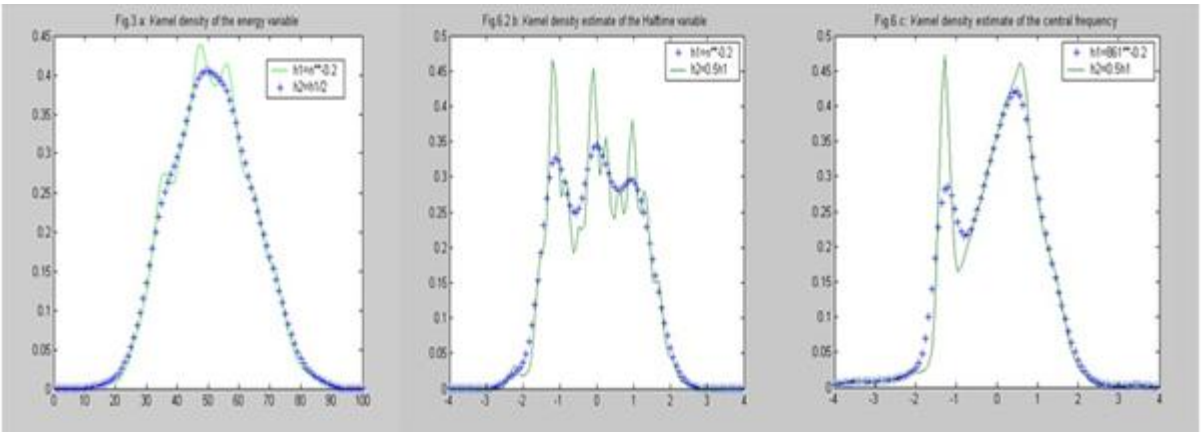

Figure 3 Plots of the kemel density functions on the principle axes using the band with $h = 0.5N^{-1/5}$

Figure 3 provides the properties of the kernel density estimates of the three variables after centering and scaling to get unit-variance using $h = N^{-1/5}$ and $h = 0.5N^{-1/5}$. This plot indicates that each of the first two variables α and β suggest that there are three different groups but the energy detects the three groups when a smaller value of the smoothing parameter is used whereas the frequency variable γ detects only two groups. We also note that the antimodes (cut-off points) are detected in the same vicinities due to centering and scaling which indicates that the first signals differ from other signals in the β and γ variables. On the other hand γ did not discriminate signals of type 2 and type 3.

It becomes interesting therefore to ask the question of whether the variables will lead to the same groups because we notice that the overlap of the individual densities in the left panel in Figure 3 is less important than those in the other two figures. In other words, will there be less misclassified signals of type 1 when the frequency is used alone? To answer this question we need to investigate an alternative method which consists of classifying the signals using single variables separately, associate probabilities and then combine the different results and neighboring property to obtain the final classification groups– which relates to the choice of the smoothing parameter. Walden's method revealed that the optimal linear transformation of the three variables is xi= -0.309 α^*_i + 0.951 β^*_i for $i = 1,..N$ which means that the frequency variable was judged as a non-discriminating one although when it was considered on its own it detected nicely the first group. On the other hand, the energy variable seems to affect the discriminating. The kernel density estimate of the resulting optimal univariate data set is shown in Figure 4. As expected, the mixture density estimate is similar to that of the 2nd variable shown in the middle panel of Figure 3 and it suggests the existence of three types of regions. The antimodes m_1 and m_2 were found to be equal to -0.5591 and 0.5746 respectively. However, the densities of the 2nd and 3rd groups are not well separated when the density is regarded as a mixture of 3 normal densities (Figure 4) – indicating that the number of misclassifications between group 2 and group 3 will be important.

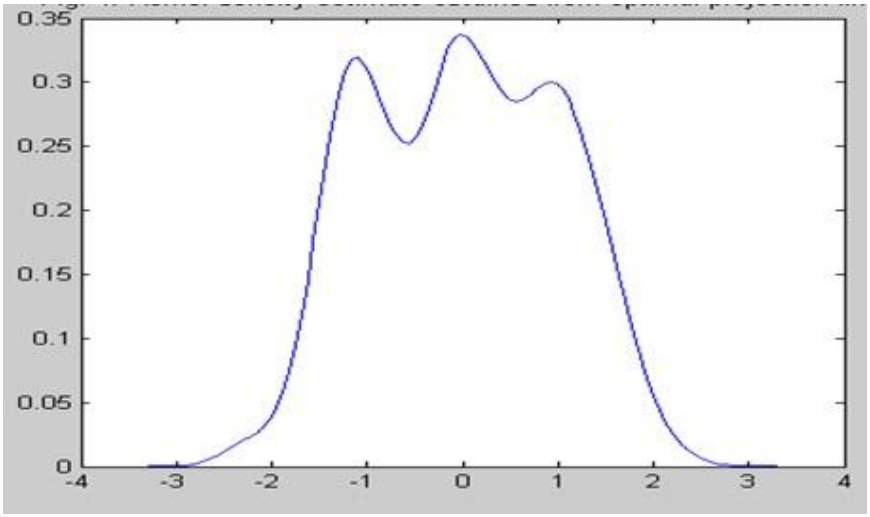

Figure 4. Kernel density estimate obtained from optimal projection line

Both panels in Figure 5 derive from the typical data grid in Figure 1. The mapping of the original signals on the 21x41 pixel rectangular region is shown in Figure 5A in which it can be seen that there are many misclassified signals, especially for signals of type 2 and type 3 (the green and blue regions). On the other hand, Figure 5B exhibits alternative results obtained using the neighboring principle in which the mode filter - a noise reduction filter that, for each pixel, outputs the mode value in its local neighborhood. The filtered results exhibit greater accuracy and consistency than those in Figure 5A.

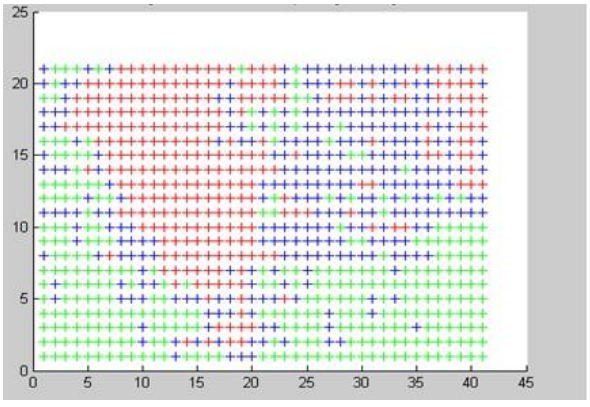

Figure 5A: Original mapping on 21x41 grid points

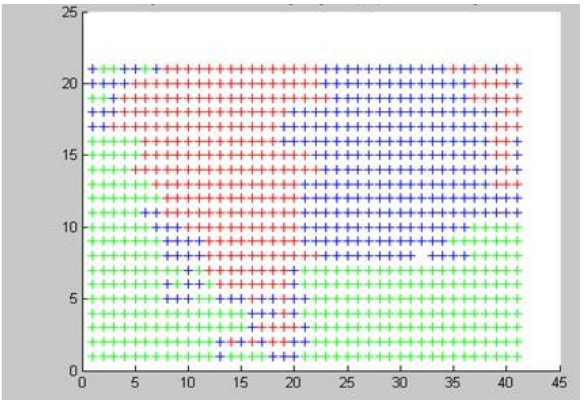

Figure 5B: Original mapping based on the mode filter

4. Concluding remarks

Motivated by the need for designing flexible and adaptive methods for applications in typically high cost geological structure applications, the paper investigated the applicability of Walden's approach and its enhancement potentials. We proposed a robust algorithm for seismic signal segmentation seeking to determine the edges of geological regions based on information captured from seismic signals. Its main idea was to develop optimizing tools based on straightforward extensions of the analysis tools initially developed by Walden (1994). Determining the homogeneity and/or heterogeneity of the regions under study was crucial and although implementation was confined to a single real data set, its multiple simulations provide a good starting point. Further testing of the optimization approach used in this work will need to be investigated further with a variety of data, in order to determine their full capabilities. Whereas the paper was confined to enhancing Walden's time based parameters, further research directions may focus on broader aspects geological structures such as data dimensionality and parameters generated by different models. Finally, it is expected that this paper will lead to better classification of spatio temporal structures of geological applications. It may also highlight new paths in other spatio-temporal applications.

References

Bloomfield, P. (2000). *Fourier Analysis of Time Series: An Introduction*, second edition. A Wiley-Interscience Publication.

Bolton, R. J., & Krzanowski, W. J.(2003). Projection pursuit clustering for exploratory data analysis. *Journal of Computational and Graphical Statistics, 12*(1).

Eaton, D.W., Milkereit, B., & Salisbury, M. (2003). Seismic methods for deep mineral exploration: Mature technologies adapted to new targets; *The Leading Edge, 22*, 580-585; GeoScienceWorld.

Friedman, J. H., & Tukey, J. W. (1974). *A projection pursuit algorithm for exploring data analysis. IEEE Transactions on Computers*, C-23, pp. 881-889.

Jones, M. C., & Sibson, R. (1987). "What is Projection Pursuit?" (with discussion) *Journal of the Royal Statistical Society, 150*(1), 1-36.

Lawrence, M. L., & Stephen, M. P. (1990) A Multi-resolution Hierarchical approach to Image Segmentation based on Intensity Extrema; *IEEE Transactions on pattern and machine intelligence.* 12, 6.

Lee, E. K., Cook, D., Klinke, S., & Lumley, T. (2005). Projection pursuit for exploratory supervised classification. *Journal of Computational and Graphical Statistics, 14*(4):831-846.

Lii, K. S., & Rosenblatt, M. (1982). Deconvolution and estimation of transfer function phase and coefficients for non-Gaussian linear processes. *Ann. Statist., 10*, 115-1208.

Olhede, S., & Walden, A. T. (2005). A generalized demodulation approach to time-frequency projections for multicomponent signals; Proceedings of the Royal Society, A; *Mathematical, Physical and Engineering Sciences, 461*, 2059, 2159-2179

Silverman, B. W. (1981) Algorithm AS 176: Kernel Density Estimation using Fast Fourier Transform. *Appl. Statist., 31*, 93-97.

Thompson, D., Mandrake, L., Gilmore, M., & Castano, R. (2010). Superpixel Endmember Detection. *Geoscience and Remote Sensing, 48*, (11), 4023-4033.

Walden, A. T. (1994). Spatial Clustering: Using Simple Summaries of Seismic Data to Find the Edge of an Oil-field. Applied statistics *Journal of the Royal Statistical Society Series C, 43*, 2, pp. 385-398.

Walden, A. T., and White, R. E. (1990). Estimating the statistical bandwidth of s time series. *Biometrika, 77*, 4, 699-707.