A statistical learning approach for seismic QC efficiency

Vincent Belz, Cyril Dolymnyj (CGG)

Abstract

Processing large marine seismic surveys can be challenging with regards to quality control (QC) efficiency because of the large amount of data that must be validated for each processing step (it is common to reach more than 10^6 shots for one survey). Statistical and machine learning related techniques are natural tools to tackle such a challenge.

Using data from a 10,000 km² survey in the Espirito Santo basin, we implemented a statistical approach to detect problematic shots and a supervised learning approach based on logistic regression to detect the presence of swell noise with an accuracy of 100% in our validation set.

Introduction

Marine seismic acquisition is carried out in the open ocean where sources of noise are difficult to control. The most frequent source of noise observed in the data is swell. It manifests itself as noise with high amplitude and low frequency. Another common issue is screw noise, occurring when smaller boats are present close to the streamers, which pollutes the recording with low amplitude, high frequency noise. A third common source of noise is the presence of nearby shooting vessels that create seismic interferences, which often have a different incident angle and period than the recorded seismic. To remove these different kinds of noise, we use different methods. Therefore, it is important to discriminate between them during our quality control (QC).

Marine seismic acquisitions are increasing in size and complexity. Acquiring surveys on large areas (above 10, 000 km²) is now common practice. Processing this amount of data can be challenging and time consuming, increasing the need for quality control efficiency (Twigger et al, 2016).

Statistics and machine learning are powerful tools to identify important insights in data and are gaining importance across industries, especially for geoscience as described by (Addison, 2016). For example, Gradient Boosting Tree Ensemble was used by (Guillen et al, 2015) in order to detect salt bodies using a supervised learning approach. Additionally, Deep Neural Networks were used (Araya-Polo et al, 2018) in order to derive a 2D velocity model, highlighting the potential that machine learning technology could bring to geophysics. In this paper, we propose to use such tools to improve the quality control of large seismic datasets.

Between 2017 and 2018, a variable-depth streamer survey covering 10,000 km² was acquired offshore Brazil in the Espirito Santo basin. This large dataset was used to perform our analysis.

This paper is divided into three main parts. The first one describes a methodology applied to the data in order to reduce their size while keeping important information for our QC and analysis purposes. The second part shows a statistical approach to detect abnormal shots . The final part is about a supervised learning based methodology applied to swell noise detection.

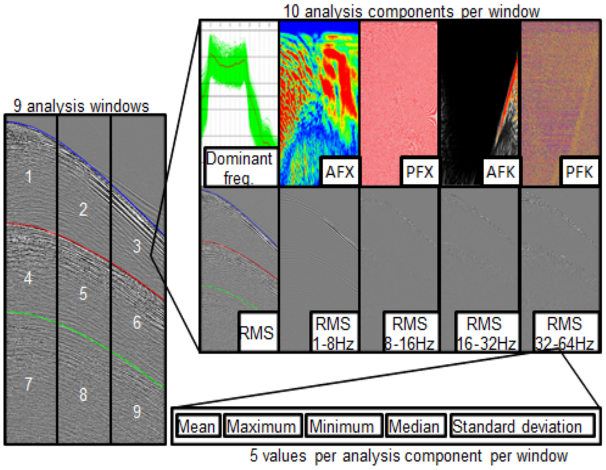

Figure 1 – *Splitting of a shot gather in 9 windows (offset and horizon based).*

Data dimension reduction using seismic attributes

A typical modern seismic shot has a length of 12 seconds sampled at 2 ms and is composed of 648 receivers per cable. It is a common practice to resample the data at 4 ms. This means that, for each shot and each cable, we have 648 * 12 / 0.004 = 1,944,000 samples that describe the data recorded. It is therefore necessary to decrease the dimension of this shot matrix while maintaining valuable information.

In the selection of the appropriate attributes that will define each shot, several physical considerations were made:

- Noise is not usually present on the whole shot but mostly over a specific offset range or at a given recording time (when the seismic signal becomes weaker and the signal-to-noise ratio becomes worse). Shots were therefore split into 9 different windows (3 according to the offset range and 3 according to the recording time based on the estimated water bottom event). Each of the following analyses were independently performed in these 9 windows (see Figure 1).
- Each shot is composed of valuable energy (signal) and undesired energy (noise). Often, noise can easily be identified while looking at the RMS of each trace as a strong isolated burst of amplitude could reveal the presence of noise. In some cases, the noise is frequency dependent, for instance swell has frequency content lower than 15 Hz while the noise coming from small boat engines is usually around 60 Hz. So an RMS analysis based on different frequency panels allows us to discriminate between them.
- Another aspect to consider is the frequency spectrum of the data and particularly its dominant frequency. Sometimes, a thin frequency peak, with low energy due to its sharpness, could reveal the presence of noise such as electrical noise around 50 Hz. Other characteristics thought useful include FX (amplitude (AFX) and phase (PFX)) and FK (amplitude (AFK) and phase (PFK)). These representations of the trace help to discriminate the direction of propagation of the seismic noise.

Within each window, 10 different characteristics were mapped: global RMS, RMS between 1 and 8 Hz, 8 and 16 Hz, 16 and 32 Hz, 32 and 64 Hz, dominant frequency, amplitude and phase in the FX domain, amplitude and phase in the FK domain. (Figure 1). At this point of our methodology, the data are composed of many attributes with different dimensions. Therefore, we need to further reduce our data dimension.

Data statistics offer solutions on how to condense these values. The easiest is to look at the mean and median value of the attributes defined earlier. Any difference between the two reveals the presence of a non-symmetric distribution or outliers. Other sources of useful information regarding series of values are the minimum and maximum that might also reveal outliers. And finally the standard deviation allows us to estimate the dispersion of the values relative to the mean. This data reduction allows each of the attributes described earlier to be condensed into 5 features. At the end, we split a shot into 9 windows, each window has 10 attributes and each attribute is further reduced into 5 features. Thus, we reduce each shot to 9 * 10 * 5 = 450 values. This compares favourably to the 1,944,000 samples that compose each shot. Dimension reduction allows for the manipulation of large datasets while keeping essential characteristics of the data.

Statistical method for anomalous shots detection

In a first step, we want to estimate the probability of a shot being anomalous. A natural way to do so would be to estimate a probability density function from our data.

35,000 shots were considered from a large survey offshore Brazil in the Espirito Santo Basin, with basic denoising applied (light swell removal and amplitude spike removal). We purposefully left some shots without this basic denoising application to see if they would be labelled as outliers (anomalies) by our method. The 450 parameters describing our seismic (cf Figure 1) have been normalized to make them comparable. From the 35,000 shots, we estimated the mean vector **μ** (dimension 450) and the covariance matrix ∑ (dimension 450*450), in order to derive the probability density function from the data:

$$f_{\mathbf{x}}(x_1, \dots, x_{450},) = \frac{e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}}{\sqrt{(2\pi)^{450}|\boldsymbol{\Sigma}|}}$$

Equation 1 – *Probability density function*

The shots that have the lowest probability values would be considered as anomalies.

As we cannot visualize a dataset into 450 dimensions, for visualization purposes, the dimension of our parameters should be further reduced to 2D or 3D. In order to

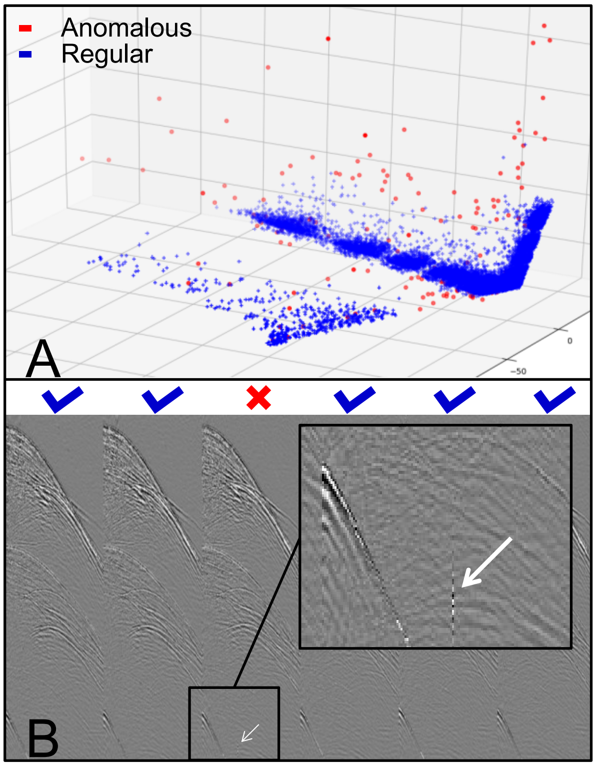

Figure 2 – *PCA 3D – 0.5% anomalies based on multidimensional Normal estimation*

visualize high dimension datasets into 2D or 3D, Principal Component Analysis (PCA) (Abdi et al, 2010) is commonly used. PCA is performed by eigenvalue decomposition of a data covariance matrix. It reveals the internal structure of data in a way that best explains the variance in that data. The first components directions have the highest impact on variance within the data. We use here the three principal components directions in order to display the shots in a 3D crossplot (Figure 2A). We highlight in red, the 0.5% of shots that have the lowest probability (outliers). Figure 2B shows some selected shots. Among these shots, one was labelled as anomalous (red cross). We can see that the abnormal detected shot contains a spike in the deep window while surrounding shots are free from such noise.

Moreover, all shots which were labelled as anomalous were detected as outliers by our method.

While this first approach is very powerful for detecting general anomalies and unexpected noise (spikes, processing artifacts, etc.), it will not help to discriminate between different types of noise. We want to know which type of noise we have in our data so we can correct for it properly. Additionally, if the dataset is composed of a majority of noisy shots, the noise will not be considered as anomalous (in an extreme case, the less noisy shots would be considered as abnormal). Hence we moved to a more specific method targeting only swell noise using supervised learning.

A statistical learning approach based on maximum likelihood to detect swell

Many marine surveys are affected by swell noise, characterized by a high amplitude and low frequency. Our Espirito Santos seismic survey was not an exception.

In order to improve the efficiency of our QC, detecting swell is very useful. To do so, we built a swell noise identification tool based on a supervised learning approach.

For the training phase of our approach, 40,000 shots containing swell noise were considered, and a light de-swell was applied to 80% of them. They were labelled as without swell (label 0). The 20% remaining shots were labeled as containing swell (label 1).

We split the dataset into a training set containing 30,000 and a validation set containing 10,000 shots.

Both training and validation set were centered and normalized with respect to the training set in order to not bias the results.

A logistic regression was applied to the training set in order to predict the swell label. It is based on the sigmoid function shown in Figure 3. It is a probability function that will tend to go to 2 different states: 0 or 1 according to a normalized value z that can be controlled (bias and scale) after a learning process.

This logistic function is a natural tool to perform binary classification (Freedman, 2009) and can be seen as the probability to detect swell noise. The predictors we used are the 450 parameters we described earlier in Figure 1.

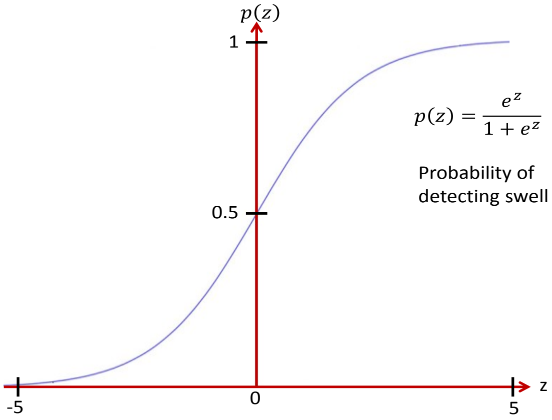

Figure 3 – *Sigmoid function*

$\beta_0, \ldots, \beta_n$ coefficients were estimated via a training approach (Equation 2).

$$p(x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}}$$

$n = 450$

$x_1, \ldots, x_n$ are the predictors

$\beta_0, \ldots, \beta_n$ are the coefficients to estimate

Equation 2 – *Logistic function*

These coefficients estimated on the training dataset were used to predict the labels for the validation dataset, where true labels are known.

Swell is easily described by its features, and the estimator is able to predict the swell label with 100% accuracy (comparing predicted labels and true labels). We display in Figure 4 (top), the shots in a 2D PCA crossplot, highlighting in red the swell labels.We can see that swell and non-swell shots are easily separated by a hyperplane in the 2D PCA crossplot, explaining why we can reach such a high level of predictability. Figure 4 (bottom) shows a selection of gathers. Gathers with a red cross are predicted with swell while gathers with a blue check are predicted without swell.

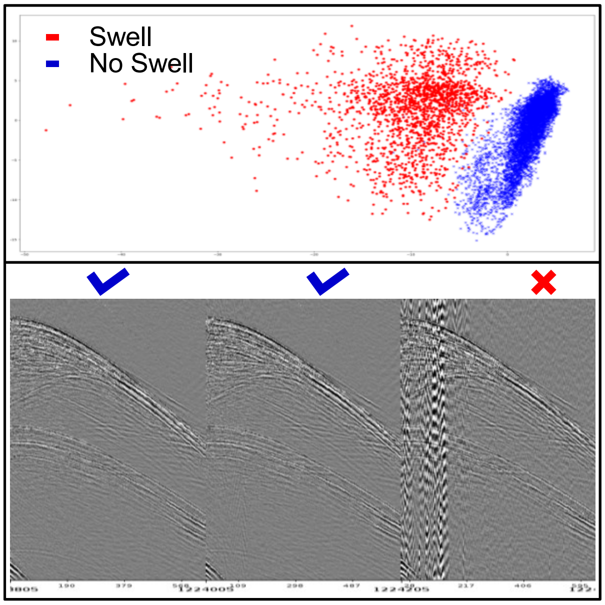

Figure 4 – *PCA 2D – Swell labels representation*

Conclusion

We presented a statistical learning approach in order to gain efficiency in the quality control of our data processing.

We presented a methodology that enabled us to reduce the dimension of a shot from 1,944,000 to 450 (reduction factor of more than 4300) while maintaining inherent information from the data. We estimated a probability density function of our data using a multidimensional normal distribution that allowed us to distinguish abnormal shots (outliers). A logistic regression estimator has been built, enabling us to predict swell with an accuracy of 100% on our validation set. In order to identify more complex noise, such as seismic interference, a more sophisticated approach, which involves for instance neural networks, is required.

Acknowledgments

We would like to thank CGG Multi-Client New Ventures for permission to publish this work.

References

Araya-Polo, M., Jennings, J., Adler, A., Dahlke, T., 2018, Deep learning tomography, The Leading Edge, Vol. 37, No. 1, pp. 58-66

Addison, V., 2016, Artificial intelligence takes shape in oil and gas sector: EPmag, April.

Twigger, L., Schouten, R., James, G., firsth, J., 2016, Seismic efficiency on a vast scale – a case study from offshore Gabon, First break volume 34.

Guillen, P., Larrazabal, G., Gonzales, G., Boumber, D. & Vilalta, R., 2015, Supervised learning to detect salt body: 85th Annual International Meeting, SEG, Expanded Abstracts,1826–1829.

Abdi. H., & Williams, L.J., *2010,* Principal component analysis: Wiley Interdisciplinary Reviews: Computational Statistics. 2 (4): 433–459.

David A. Freedman, 2009, *Statistical Models: Theory and Practice*. Cambridge University Press. p. 128.