Exploratory data analysis of the SOLA land dataset

Fantine Huot

ABSTRACT

Seismic surveys provide us with an abundance of data characteristics. Is there an informative way to visualize a survey's metadata? Using exploratory data analysis (EDA) tools, we investigate a land survey's metadata to detect trends among the observations. We derive indicators of the quality of a seismic trace.

INTRODUCTION

With the advent of multi-component sensors, we are faced with an abundance of data characteristics which may contain precious information about our survey. For example, within a seismic survey, noise levels vary by receiver location, source location, offset, frequency, time of the day, etc. While bad traces are easy to pick out, the study of data characteristics can yield criteria for assessing the quality of a seismic trace.

High noise levels degrade seismic images. Thus, significant efforts are deployed in seismic signal processing to attenuate and remove noise from the signal (Yilmaz, 2001), while stacking techniques are commonly used to improve signal to noise ratio (Claerbout and Black, 2008). Once bad traces are identified, they can be selectively left out of the imaging process using weighing and penalty functions as introduced by Claerbout (2014).

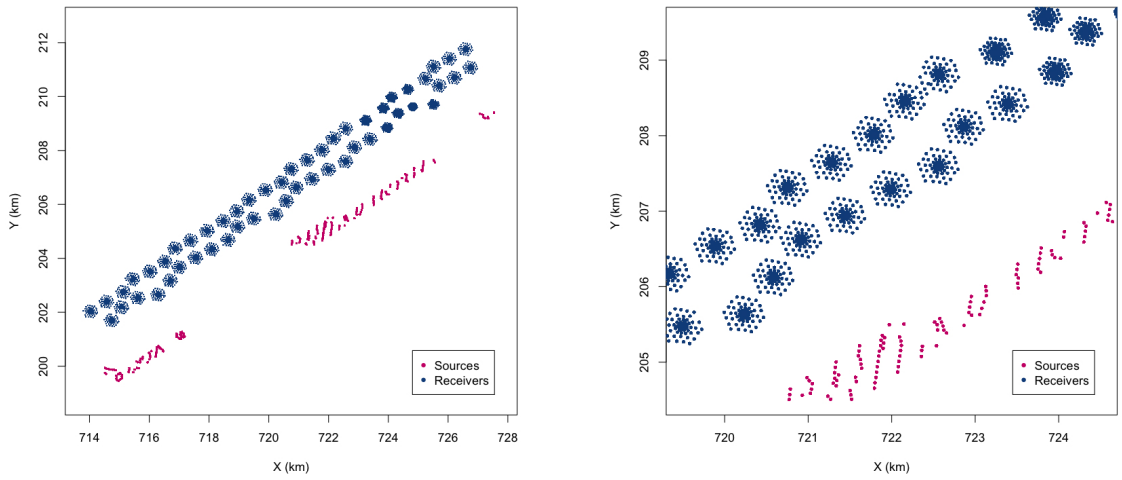
Therefore, it would be extremely useful to derive indicators of noise level or recording quality from a survey's metadata. However, what would be an informative way to visualize these metadata? How can we efficiently discover trends among the variables or among the observations? Which are the important variables?

Exploratory data analysis (EDA) refers to a diverse set of techniques for answering questions such as these. Herein we use R, a free software environment for statistical computing, to perform our analysis. We conduct our study on the metadata of the SOLA dataset.

THE SOLA LAND DATASET

Acquisition geometry

The SOLA survey is a three-component 2D array land acquisition conducted during summer 2015. It has 3,600 receivers arranged into 54 hexagonal arrays and 4,000 shots. The arrays are 200 m to 600 m wide, with 63 to 110 receivers per array. In order to record a wide range of frequencies while still providing comprehensive coverage, the arrays are irregularly sampled: the receiver spacing varies from 5 to 10 m on the inside of the array to 50 m on the outside. The survey has about 7 million traces, but for this study we were provided with the data characteristics of a subset of 350,000 traces. This subset's acquisition geometry is illustrated in Figure 1.



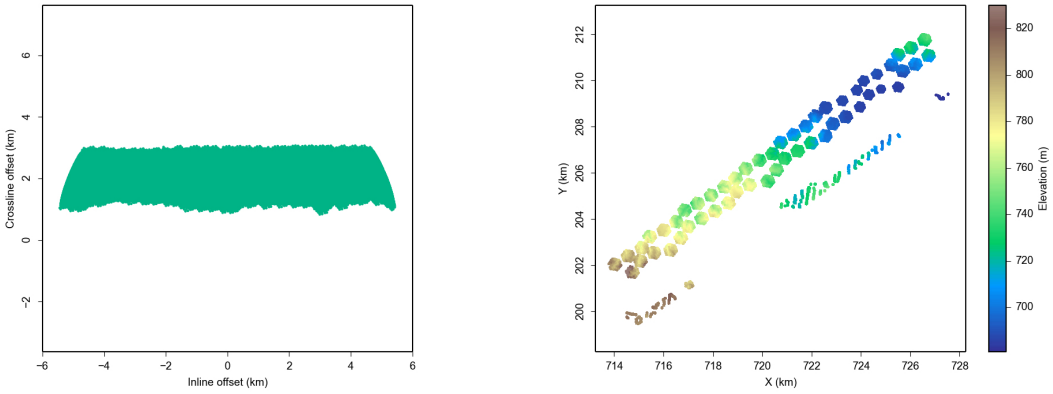
(a) (b)

Figure 1: (a) The SOLA acquisition geometry. (b) Zoomed in portion of the survey. In order to record a wide range of frequencies while still providing comprehensive coverage, the receivers are arranged in irregularly sampled hexagonal arrays. [**ER**]

The metadata we will be considering for this study contains two types of entries: descriptive headers and data characteristics.

Descriptive headers

The descriptive headers cover all acquisition parameters for each seismic trace, such as source and receiver location, shot times, offset and azimuth (Figure 2). The complete list of headers is provided for reference in the appendix. Certain headers were left out from this study for they were redundant or did not vary over the selected subset, leaving us with 25 variables.



(a) (b)

Figure 2: Examples of parameters contained in the survey's descriptive headers. (a) Inline and crossline offset. (b) Source and receiver elevation. [**ER**]

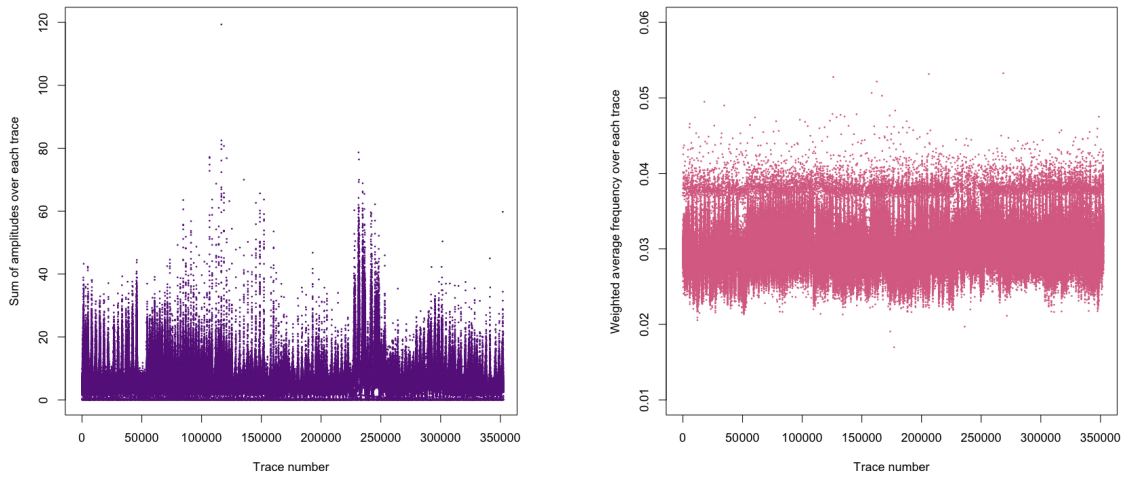
Data characteristics

The data quality varied greatly over the survey, and the acquisition recorded various levels of surface noise. To capture this variability, a certain number of data characteristics were computed for each trace. These data characteristics include variables such as first break pick, average amplitude, and average frequency or spikiness of the signal (Figure 3). When possible, they were computed both over the full trace and over various time windows, constituting a total of 52 variables. The complete list of computed data characteristics is provided in the appendix.

Using both the header information and the computed data characteristics, we herein seek to identify trends in the noise variability.

DATA VISUALIZATION

Combining the descriptive headers and data characteristics, we have a total of $n = 350,000$ observations over $p = 77$ different variables. We could visualize these data by plotting two-dimensional scatter plots, each of which contains the n observations' measurements on two of the p variables. However, there are $\binom{p}{2} = p(p-1)/2 = 2,926$ such scatter plots, which makes it prohibitive to look at all of them. Moreover, most of them would not be very informative since they each contain only a small fraction of the total information present in the dataset.



(a) (b)

Figure 3: Examples of the computed data characteristics. (a) Sum of amplitudes over each trace. (b) Weighted average frequency of each trace. [**ER**]

Principal component analysis (PCA)

When faced with a large set of correlated variables, principal components allow us to summarize the dataset with a smaller number of representative variables that collectively explain most of the variability in the original set. The idea is that each of the n observations lives in a p-dimensional space, but not all of these dimensions are equally interesting. Principal component analysis (PCA) seeks a small number of dimensions that are as interesting as possible, where the concept of interesting is measured by the amount that the observations vary along each dimension. The dimensions found by PCA are called principal components. Principal components are a sequence of linear combinations of the p variables, mutually uncorrelated and ordered in variance. They are the directions along which the original data is highly variable.

In the following, we provide a brief overview on how to compute the principal components of a dataset, based on formulations by Hastie et al. (2005). The first principal component Z_1 of a set of variables $X_1, X_2, ..., X_p$ is the normalized linear combination,

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + ... + \phi_{p1}X_p,$$

that has the largest variance. The elements $\phi_{11}, ..., \phi_{p1}$ are called the loadings of the first principal component. Together, they make up the principal component loading vector, $\phi_1 = (\phi_{11}\, \phi_{21} \ldots \phi_{p1})^T$. As an arbitrarily large value of these loadings would result in an arbitrarily large variance, the loadings are normalized such that $\sum_{j=1}^{p} \phi_{j1}^2 = 1$.

Let's consider a certain $n \times p$ dataset $\mathbf{X}$. As we are only interested in variance, we assume that each of the variables in $\mathbf{X}$ has been centered to have mean zero. The first principal component of $\mathbf{X}$ is computed by finding the linear combination of the sample variable values of the form $z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + ... + \phi_{p1}x_{ip}$, that has largest sample variance, subject to the constraint that $\sum_{j=1}^{p} \phi_{j1}^2 = 1$. In other words, the first principal component loading vector solves the following optimization problem:

$$\max_{\phi_{11},...,\phi_{p1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \text{subject to} \sum_{j=1}^{p} \phi_{j1}^2 = 1. \quad (1)$$

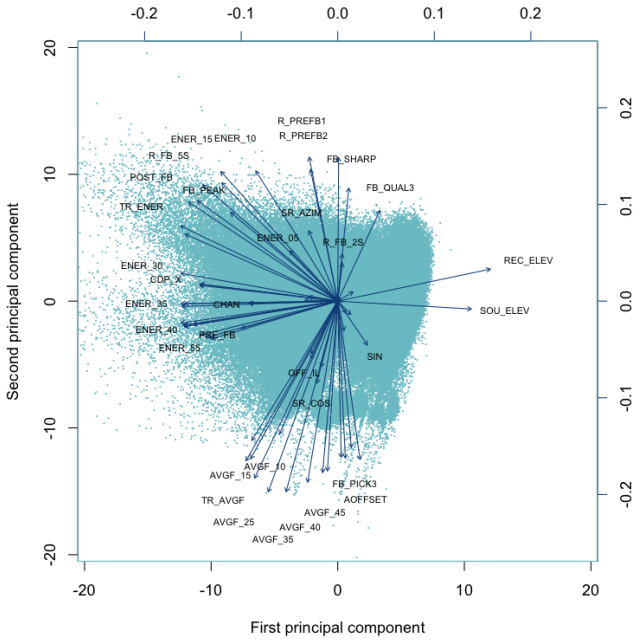
The objective in Equation 1 can be re-expressed as $\frac{1}{n}\sum_{i=1}^{n} z_{i1}^2$. Since $\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$, the average of the $z_{11}, ..., z_{n1}$ will be zero as well. Hence the objective that we are maximizing in Equation 1 is just the sample variance of the n values of z_{i1}. Equation 1 can then be solved via an eigen decomposition (Golub and Van Loan, 1983).

Once the first principal component Z_1 has been determined, the second principal component is the linear combination of $X_1, X_2, ..., X_p$ that has maximal variance out of all the linear combinations that are uncorrelated with Z_1. Constraining Z_2 to be uncorrelated with Z_1 is equivalent to constraining the loading vector ϕ_2 to be orthogonal to ϕ_1. As a consequence, to find the second principal component, we solve a problem similar to the one expressed in Equation 1, with ϕ_2 replacing ϕ_1, and with the additional constraint that ϕ_2 is orthogonal to ϕ_1.

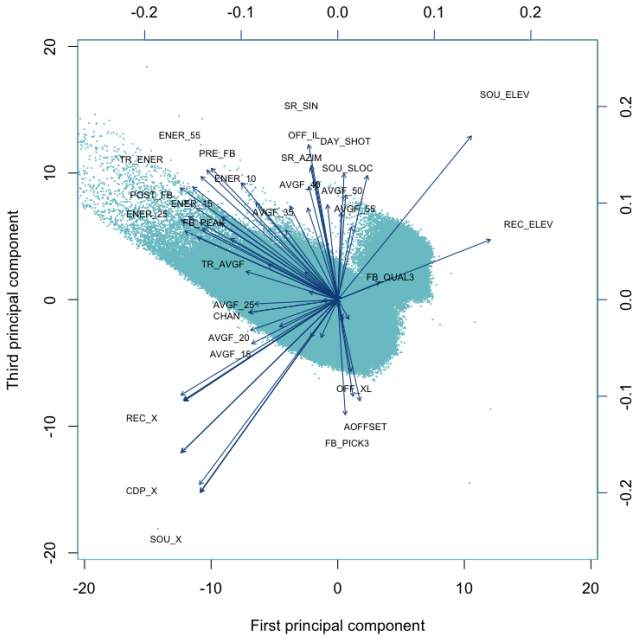
Data projection

By projecting the data along the first few principal component directions, we can build two-dimensional representations that capture most the dataset's variability. PCA was performed on the SOLA metadata after standardizing each variable to have zero mean and standard deviation one. Figure 4 represents the dataset projected along its three first principal components.

This representation allows us to visualize the nature of the first principal components. The first principal component puts weight mostly on source and receiver elevation, common depth point location, and average energy over various time windows. The second principal component accounts for absolute offset, first break pick and average frequency over various time windows. The variability over the seismic traces seems to be mostly explained in terms of source and receiver elevation and absolute offset as far as the descriptive headers are concerned, and average energy and frequency content for the data characteristics. Among the various time windows on which the summed energy and average frenquencies were computed, 2.8 to 3.2 s seems to carry most weight. The variables that accounted for the spikiness of the data have little to no impact on the first principal components, and only start carrying weight from the 11th component onwards.



(a)

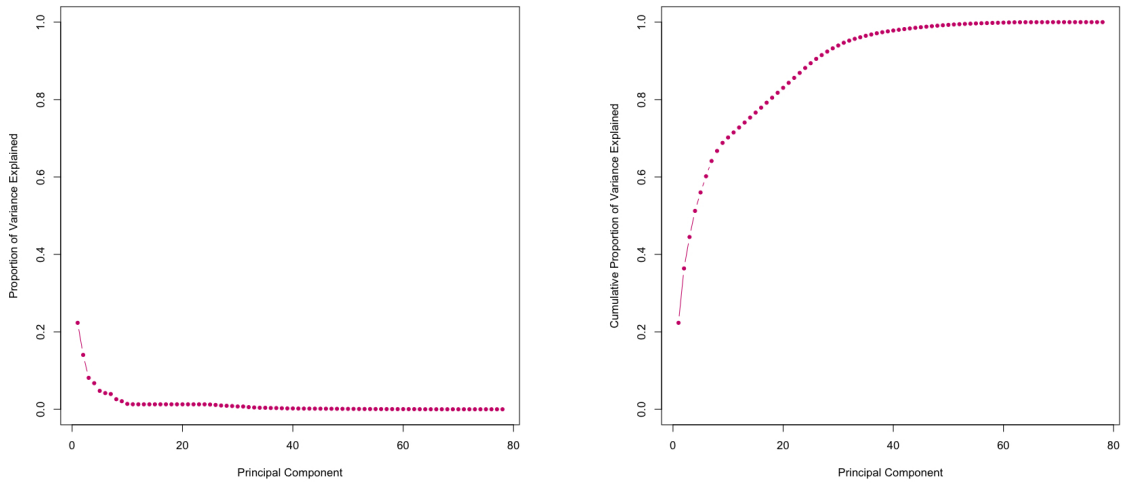


(b)

Figure 4: Projection of the SOLA metadata on its first principal component directions. The dots represent the projected metadata, while the arrows indicate the loadings associated to each variable. For readability reasons, only the loadings greater than 0.05 are labeled. The description of each label is provided in the appendix. (a) Projection on the first and second principal components. (b) Projection on the first and third principal components. [**CR**]

Proportion of variance explained

Although there are a possible $p = 77$ principal components, approximately 23 account for 90% of the total variation, while the first three ones account for 46%. Together, the first seven principal components explain around 65% of the variance in the data. This may not seem a large amount of variance. However, from the plots in Figure 5, we see that while each of the first seven principal components explain a substantial amount of variance, there is a marked decrease in the variance explained by further principal components. This suggests that there may be little benefit to examining more than seven principal components.



(a) (b)

Figure 5: (a) Proportion of variance explained by each principal component. (b) Cumulative proportion of variance explained. [**CR**]

IDENTIFYING TRENDS IN THE DATA

Variables that are located close to each other in the PCA projection indicate potential correlations, enabling us to identify trends in the data. Each panel of Figures 6, 7, 8 and 9 is a scatterplot for a pair of variables whose identities are given by the corresponding row and column labels.

Amplitudes seem to decrease with receiver elevation (Figure 6). The average frequency variation narrows down with receiver elevation (Figure 7). As to be expected, the time of first break is highly correlated with the absolute offset (Figure 8). However, the decrease of energy with offset only shows on the variables computed on early time windows. Unsurprisingly, the quality of first break pick is closely linked to the amplitudes before and after first break, and the first break's maximum amplitude (Figure 9).

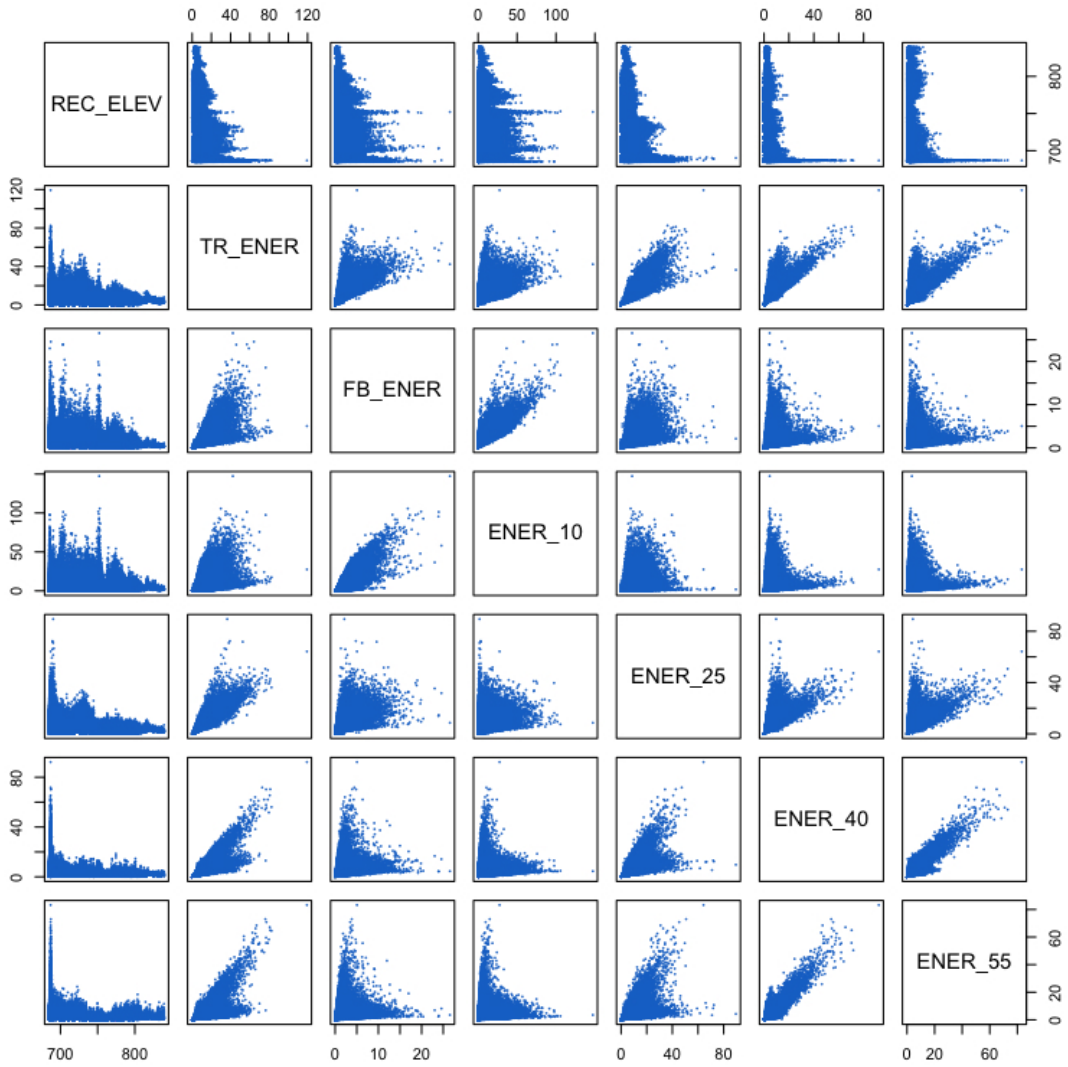


Figure 6: Scatterplots of receiver elevation (REC_ELEV) versus various measures of energy over different time windows. The full description of each variable is provided in the appendix. We notice that amplitudes seem to decrease with receiver elevation. [**CR**]

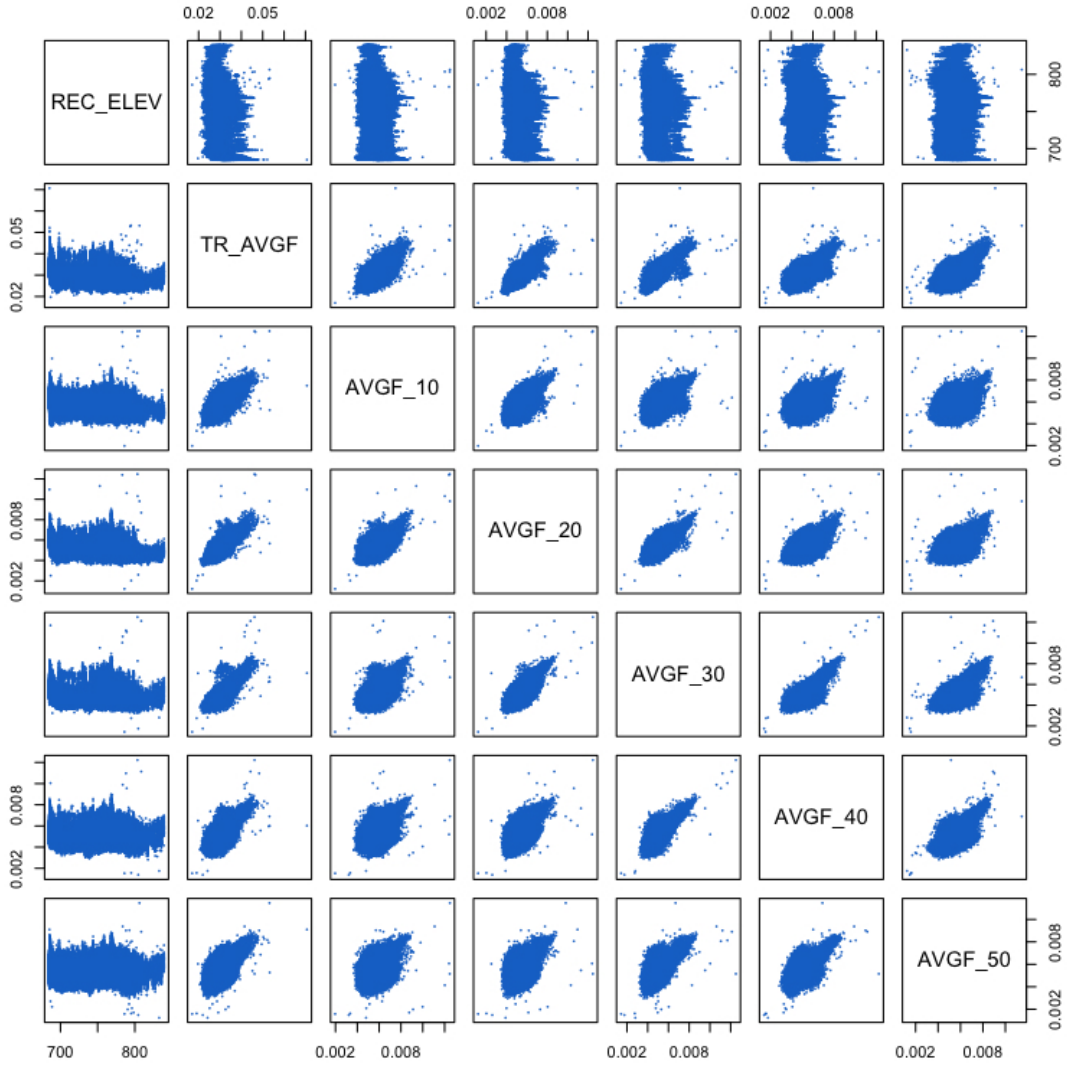


Figure 7: Scatterplots of receiver elevation (REC_ELEV) versus various measures of average frequency over different time windows. The full description of each variable is provided in the appendix. We notice that average frequency variation narrows down with receiver elevation. [**CR**]

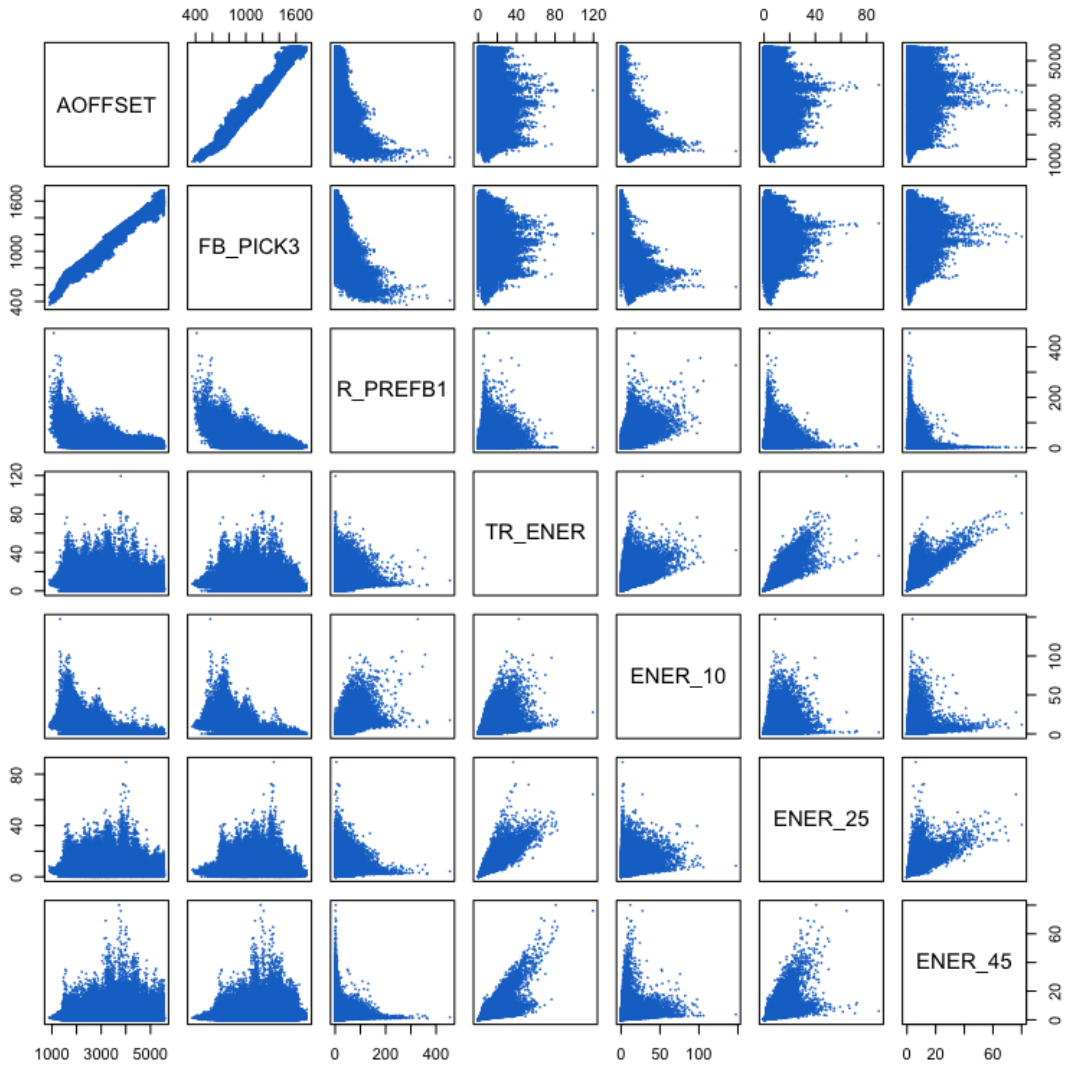


Figure 8: Scatterplots of absolute offset (AOFFSET) versus first break pick (FB_PICK3) and various measures of energy over different time windows. The full description of each variable is provided in the appendix. Unsurprisingly, we notice that first break pick is highly correlated with absolute offset. However, the decrease of energy with offset only shows on the variables computed on early time windows. [**CR**]

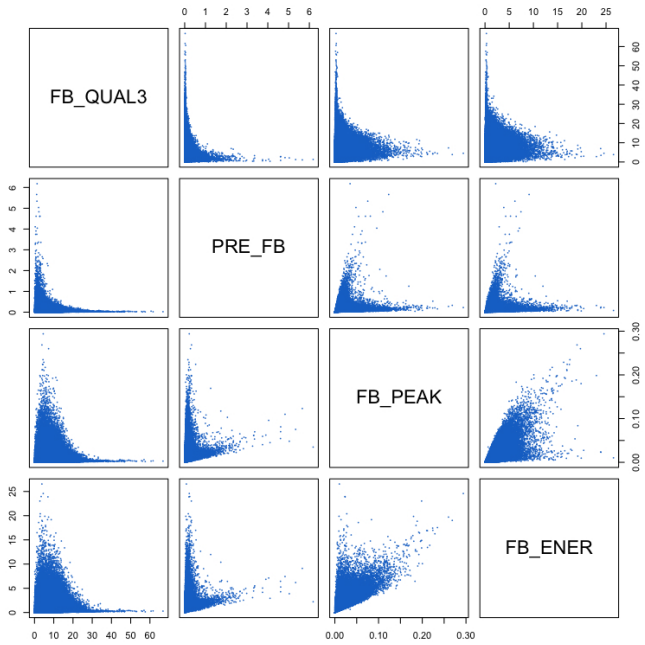


Figure 9: Scatterplots of quality of first break pick (FB_QUAL3) versus amplitudes before (PRE_FB) and after (FB_ENER) first break, and the first break's maximum amplitude (FB_PEAK). Unsurprisingly, all these variables are closely linked. The full description of each variable is provided in the appendix. [**CR**]

IDENTIFYING NOISY TRACES

Using the data characteristics computed on various time windows, we can identify portions of the signal which deviate from the remainder of the signal by unacceptable amounts, as these are likely to correspond to noise bursts.

In order to flag outliers, we use the interquartile range (IQR), which is the difference between the upper and lower quartiles of the data ($IQR = Q_3 - Q_1$), as a measure of statistical dispersion. According to Tukey's range test (Tukey, 1977), outliers are observations that fall outside the range:

$$\big[Q_1 - kIQR, Q_3 + kIQR\big],$$

where k is a positive constant. In our study, we use $k = 1.5$, a value commonly used in statistics. On a normal distribution, this value flags less than 1% of the data as outliers.

Figure 10 plots the outliers on the data characteristics both in PCA projection and over the survey's acquisition geometry. It appears that certain receiver locations accounted for noisy measurements. For comparison, Figure 11 represents the same plot, but only marks the outliers associated with spikiness of the signal.

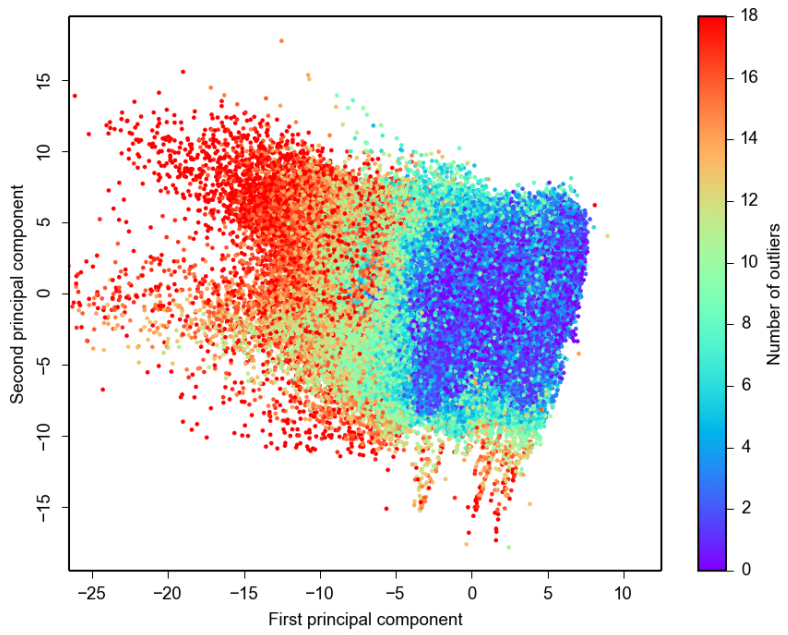
STATISTICAL ROBUSTNESS

For statistical soundness, each operation presented in this study was performed on ten different subsets of the data, where each subset contained 90% of the original data sampled at random. The different subsets did not show any significant change in the trends presented here.

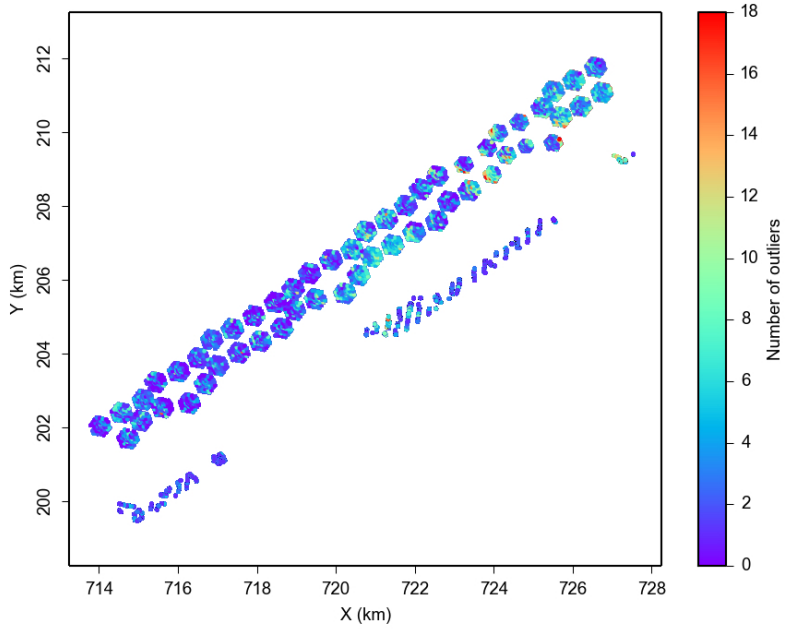
DISCUSSION

Statistical computing provides useful tools for exploring new datasets and searching for outliers and trends. Free software environments such as R ease the implementation burden for scientists and practitioners. The operations conducted in this study reflect some basic steps commonly used in exploratory data analysis. While the results obtained may not seem surprising to the trained geoscientist, a deeper study may yield more intricate trends. Therefore, it would be of interest to extend this study to the full scope of the original 7 million traces.

Moreover, a statistical approach may help identify and quantify different types of noise by flagging traces that deviate significantly from the remainder of the signal. By extending this study to a larger set of traces, it would be possible to visualize whether noise levels vary in this survey by time of the day, or day of the week. Clustering techniques may help identifying different type of noise.

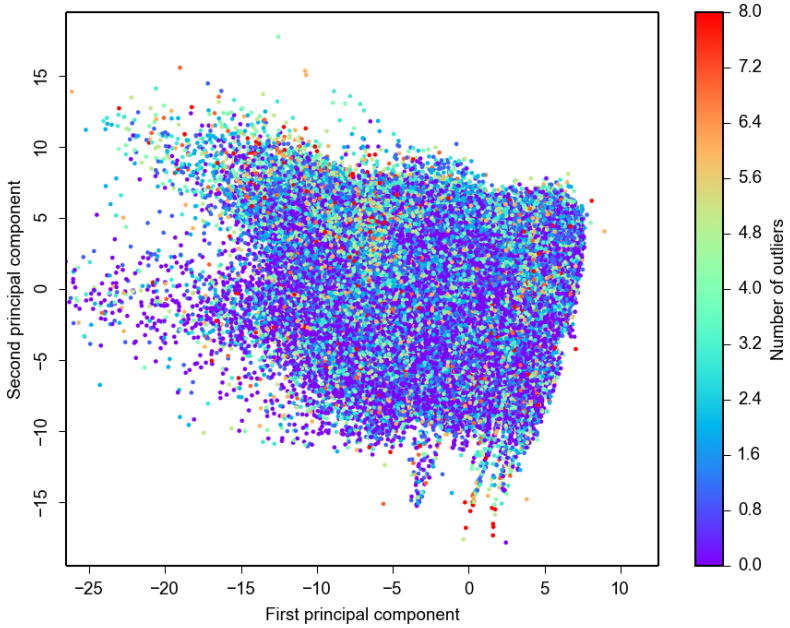


(a)

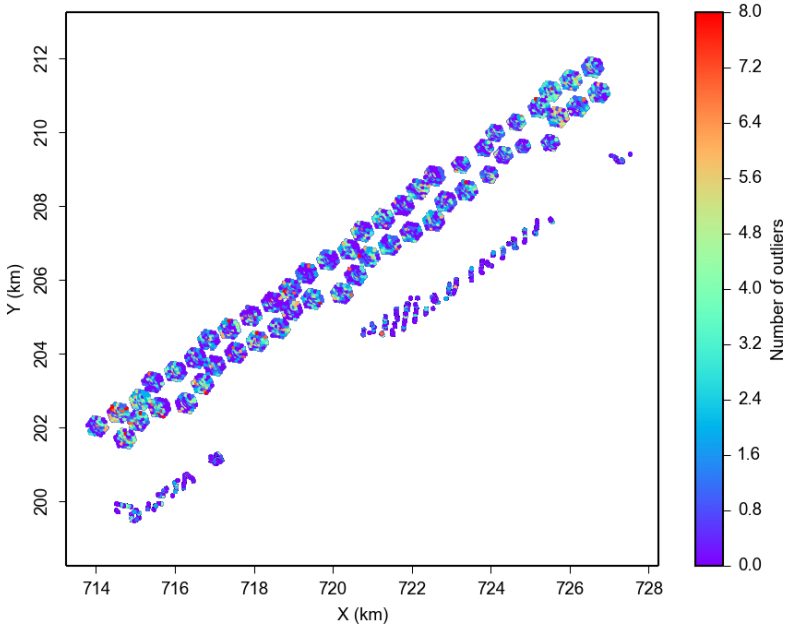


(b)

Figure 10: Outliers identified by Tukey's range test for $k = 1.5$. (a) In PCA projection. (b) Over the survey's acquisition geometry. Certain receiver locations accounted for noisy measurements. [**CR**]



(a)



(b)

Figure 11: Outliers identified by Tukey's range test for $k = 1.5$ for all the variables related to spikiness of the signal. (a) In PCA projection. (b) Over the survey's acquisition geometry. [**CR**]

ACKNOWLEDGMENTS

We would like to thank Christof Stork for providing the data characteristics used in this study.

REFERENCES

Claerbout, J., 2014, Geophysical image estimation by example: Lulu.com.

Claerbout, J. and J. L. Black, 2008, Basic earth imaging: Citeseer.

Golub, G. H. and C. F. Van Loan, 1983, Matrix computations.

Hastie, T., R. Tibshirani, J. Friedman, and J. Franklin, 2005, The elements of statistical learning: Data mining, inference and prediction: 534–552.

Tukey, J. W., 1977, Exploratory data analysis.

Yilmaz, Ö., 2001, Seismic data analysis: Processing, inversion, and interpretation of seismic data: Society of Exploration Geophysicists.

APPENDIX

The complete list of descriptive headers is provided below. The headers marked with a * symbol were left out from this study for they were redundant or did not vary over the selected subset, leaving us with 25 header variables.

SOU_X	Source X coordinate
SOU_Y	Source Y coordinate
SOU_ELEV	Source elevation
DEPTH*	Source depth
UPHOLE*	Source uphole time
SIN	Source internal index number
SOU_SLOC	External source location number
FFID*	Field file index number
SOURCE*	Live source number
S_LINE	Swath or sail line number
REC_X	Receiver X coordinate
REC_Y	Receiver Y coordinate
REC_ELEV	Receiver elevation
GEO_COMP*	Geophone component (x,y,z)
CHAN	Recording channel number
REC_SLOC	Receiver index number
SRF_SLOC	External receiver location number
R_LINE	Receiver line number
CDP_X	X coordinate of common depth point
CDP_Y	Y coordinate of common depth point
OFFSET	Signed source-receiver offset
AOFFSET	Absolute value of offset
OFF_IL	Inline offset
OFF_XL	Crossline offset
SR_AZIM	Source-receiver azimuth
SR_COS	Cosine of source-receiver azimuth
SR_SIN	Sine of source-receiver azimuth
ARRAY_ID	Array index number
YER_SHOT*	Year
DAY_SHOT	Day
TIM_SHOT	Time of the day
TRC_TYPE*	Trace type

The complete list of the 52 computed data characteristics is provided below.

FB_PICK3	First break pick
FB_QUAL3	Quality of fine tuned first break
FB_NDIFF	FB_PICK3 - FB_NAVG
TR_ENER	Sum of amplitudes of trace
TR_AVGF	Weighted average frequency of trace
PRE_FB	Pre-first break energy, normalized
FB_PEAK	Max amplitude of first break
FB_ENER	Energy from first break to 0.5 s afterwards
FB_SHARP	First break sharpness
POST_FB	Energy from 0.5 to 2.0 s after first break
R_PREFB1	FB_ENER / PRE_FB
R_PREFB2	POST_FB / PRE_FB
SPIKT_AL	Spikiness of time data
SPIKF_AL	Spikiness of frequencies
ENER_05	Average amplitude from 0.2 to 0.8 s
ENER_10	Average amplitude from 0.8 to 1.2 s
ENER_15	Average amplitude from 1.2 to 1.8 s
ENER_20	Average amplitude from 1.8 to 2.2 s
ENER_25	Average amplitude from 2.2 to 2.8 s
ENER_30	Average amplitude from 2.8 to 3.2 s
ENER_35	Average amplitude from 3.2 to 3.8 s
ENER_40	Average amplitude from 3.8 to 4.2 s
ENER_45	Average amplitude from 4.2 to 4.8 s
ENER_50	Average amplitude from 4.8 to 5.2 s
ENER_55	Average amplitude from 5.2 to 5.8 s
AVGF_05	Average frequency from 0.2 to 0.8 s
AVGF_10	Average frequency from 0.8 to 1.2 s
AVGF_15	Average frequency from 1.2 to 1.8 s
AVGF_20	Average frequency from 1.8 to 2.2 s
AVGF_25	Average frequency from 2.2 to 2.8 s
AVGF_30	Average frequency from 2.8 to 3.2 s
AVGF_35	Average frequency from 3.2 to 3.8 s
AVGF_40	Average frequency from 3.8 to 4.2 s
AVGF_45	Average frequency from 4.2 to 4.8 s
AVGF_50	Average frequency from 4.8 to 5.2 s
AVGF_55	Average frequency from 5.2 to 5.8 s

SPIKF05	Spikiness of frequencies from 0.2 to 0.8 s
SPIKF10	Spikiness of frequencies from 0.8 to 1.2 s
SPIKF15	Spikiness of frequencies from 1.2 to 1.8 s
SPIKF20	Spikiness of frequencies from 1.8 to 2.2 s
SPIKF25	Spikiness of frequencies from 2.2 to 2.8 s
SPIKF30	Spikiness of frequencies from 2.8 to 3.2 s
SPIKF35	Spikiness of frequencies from 3.2 to 3.8 s
SPIKF40	Spikiness of frequencies from 3.8 to 4.2 s
SPIKF45	Spikiness of frequencies from 4.2 to 4.8 s
SPIKF50	Spikiness of frequencies from 4.8 to 5.2 s
SPIKF55	Spikiness of frequencies from 5.2 to 5.8 s
R_FB_1S	Ratio of energy from 0.5-1.0 s after first break to 1.5-2.0 s
R_FB_2S	Ratio of energy from 0.5-1.0 s after first break to 2.5-3.0 s
R_FB_3S	Ratio of energy from 0.5-1.0 s after first break to 3.5-4.0 s
R_FB_4S	Ratio of energy from 0.5-1.0 s after first break to 4.5-5.0 s
R_FB_5S	Ratio of energy from 0.5-1.0 s after first break to 5.5-6.0 s