Outlier detection in magnetotelluric time series from a multivariable perspective

Cortés-Arroyo, Olaf J.

Mexican Center for Innovation in Geothermal Energy (CEMIE-GEO)

Abstract

Here I propose a simple Octave/Matlab algorithm to detect outliers in magnetotelluric time series. The algorithm defines a tolerance ellipse for a MT data set using the different statistical properties of the recorded time series. This method has benn tested with very large data sets in CEMIE-GEO, proving its effectiveness in MT data sets where no coherent noise is present.

Function calling

tolerance_ellipse(S, F ,mode)

Algorithm variables

S – Matrix with dimensions *m x n,* with *n* selected channels and *n* samples.

F - Numerical vector with *n* elements, where each element is a multiplier factor of the standard deviation/MAD corresponding value for each channel.

mode – Numerical value:

- 0 - Applies standard deviation values to define the tolerance ellipse limits.
- 1 - Applies MAD values to define the tolerance ellipse limits.

Theory

The algorithm works under the assumption that for a MT recording site, every field component selected for the analysis (Ex, Ey, Hx, Hy and/or Hz) can be regarded as stationary with a gaussian distribution (Banks, 1998), i.e., no coherent noise is observed in data.

First, a matrix $\boldsymbol{S}$ with *n* rows and m columns is defined, where each column contains the time series of a particular channel (Ex, Ey, Hx, Hy or Hz) and the number of rows thus represents the total number of samples. Because of the statistical properties of the data series, a *m*-dimensional tolerance ellipsoid $\boldsymbol{E}$ can be estimated for matrix $\boldsymbol{S}$.

First, the median value for each channel is calculated and removed from the time series, to locate the mth central coordinate of the ellipse in 0.

$$p_m = 0 \tag{1.1}$$

Next, the length of the *m*th axis is calculated by multiplying the MAD (or standard deviation) of the *m*th channel by a factor $\boldsymbol{F_m}$ defined by the user:

$$A_m = F_m \, MAD(C_m) \quad \text{or} \tag{1.2}$$

$$A_m = F_m \, \sigma(C_m) \tag{1.3}$$

These properties of the ellipsoid allows the user to estimate if a set of values ($\boldsymbol{S}_{n,1}, \ldots, \boldsymbol{S}_{n,m}$) is located inside or outside the ellipse, i.e,

$$C_n = \frac{S_{n,1}}{A_{n,1}^2} + \frac{S_{n,2}}{A_{n,2}^2} \ldots + \frac{S_{n,m}}{A_{n,m}^2} \tag{1.4}$$

or

$$C_n = \sum_{i=1}^{m} \frac{S_{n,i}}{A_{n,i}^2} \tag{1.5}$$

Where $C_n > 1$ is located outside the tolerance ellipsoid and is classified as outlier.

Values $\boldsymbol{F_m} = 3\sigma$ or $\boldsymbol{F_m} = 4$ MAD are recommended, to approximate a 95% confidence interval. However, it is strongly suggested to perform a visual inspection of the series (as shown in the following example) to corroborate if this values provides a correct analysis for the dataset or if a correction is necessary. If a change in the values is needed, the user must select new values by a trial and error approach.

Although adjustment in $\boldsymbol{F_m}$ vaules needs to be perfomed by the user, making of this a semi-automatic evaluation method, the algorithm reduces the tedious, time-consuming task of outlier detection in very large data series to a simple selection of a number *m* of values.

Synthetic example

Consider two time series X, Y, both with a length of 1,000 data points and composed by random values with a normal distribution. Each time series is created separately, so they are independent one of the other (Figure 1).

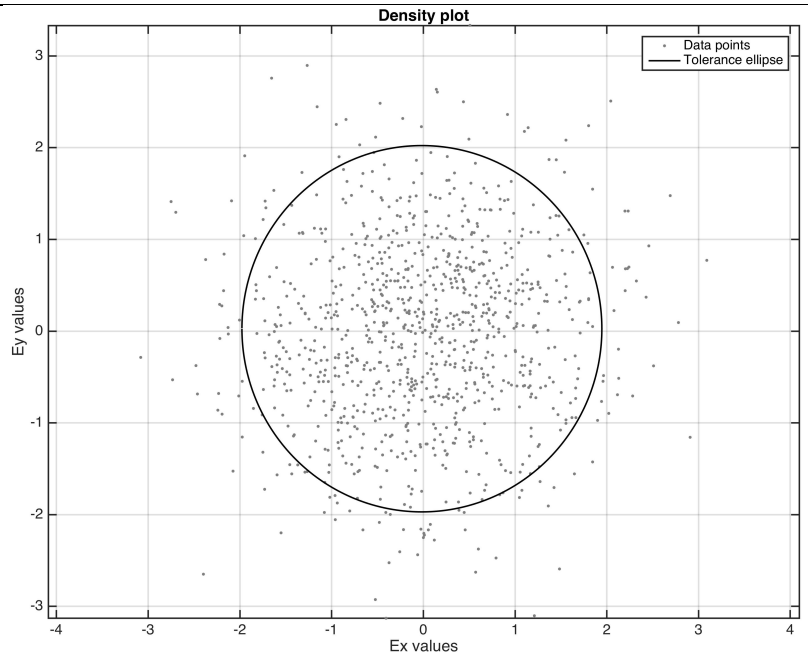
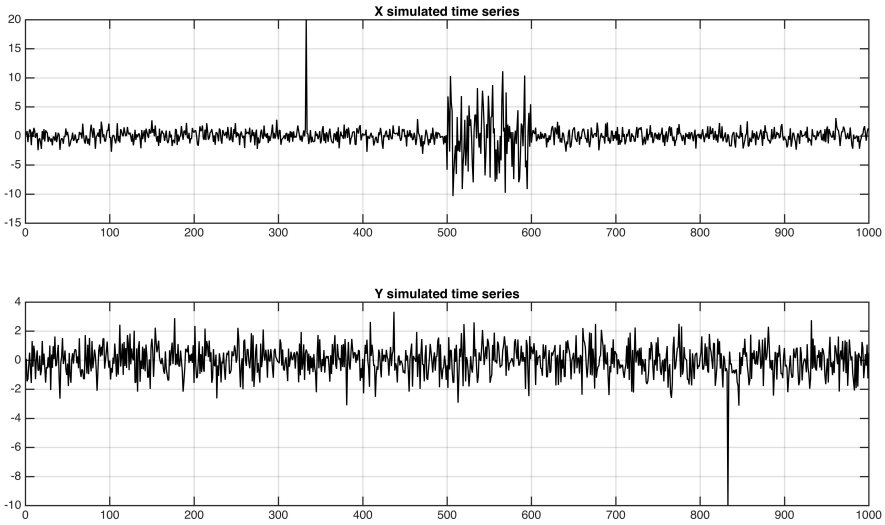


Figure 1. Tolerance ellipse representation for analysis of data series Ex and Ey, considering f_1, f_2 =2.

Lets consider now the case where noise is added in data by including an outlier on each time series and an anomalous data section between 500 and 600 seconds for the X time series (Figure 1). By making use of our proposed representation, it is observed that the outlier is located outside of the modeled ellipse, so the data point is regarded as anomalous.
Factor values are 1 for X and 2 for Y.

a)



b)

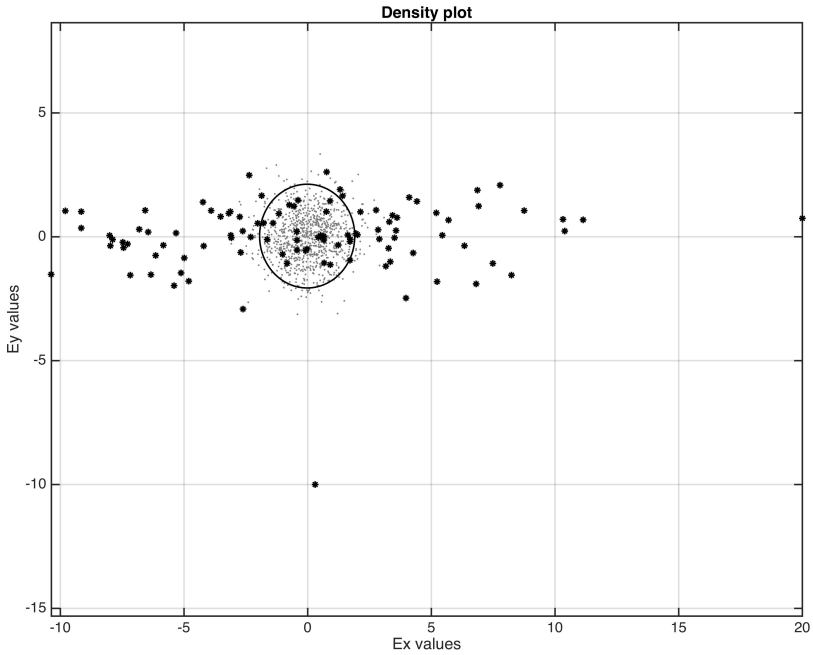


Figure 1. a) Simulated data set, with two outliers and anomalous data section between 500-600 sec. b) Application of the algorithm in presence of outliers and bad data. R1 = one standard deviation of X, R2 = two standard deviations of Y.

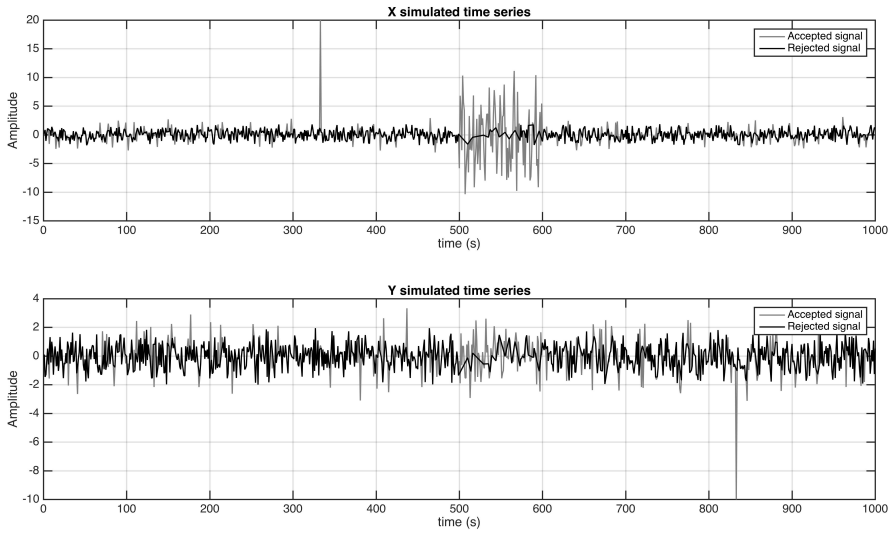


Figure 4. Classified data set. Data in gray is classified as anomalous, while data in black color has been classified as good data.

Real case example

The data set consists of electric (Ex, Ey) and magnetic (Hx, Hy) recording with a frequency sample of 18.315 Hz, registered at an unspecified location. Recovered data set consists of 6,177,600 samples per channel.

Inspection of the time series (figure 1) shows a very unusual, consistent saturation of the signal in all channels. In normal circumstances this dataset would be manually edited trough a time-consuming visual inspection, where only a fraction of the results would be useful or even the whole data set would be finally discarded.

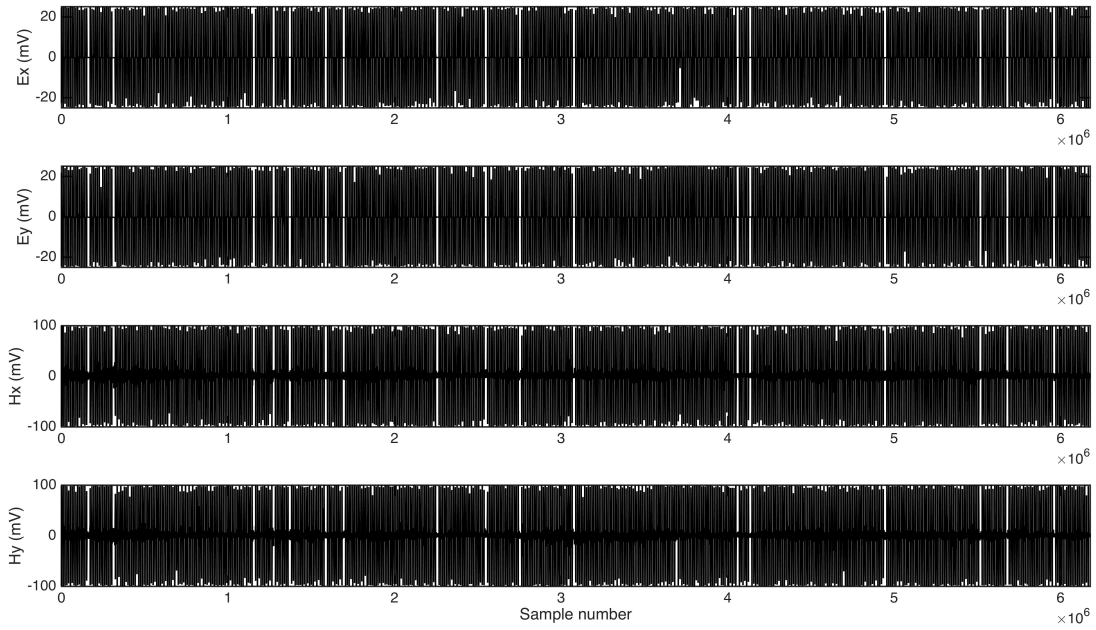


Figure 1. MT time series of the four selected channels. A great number of outliers is observed in all channels for the entire set.

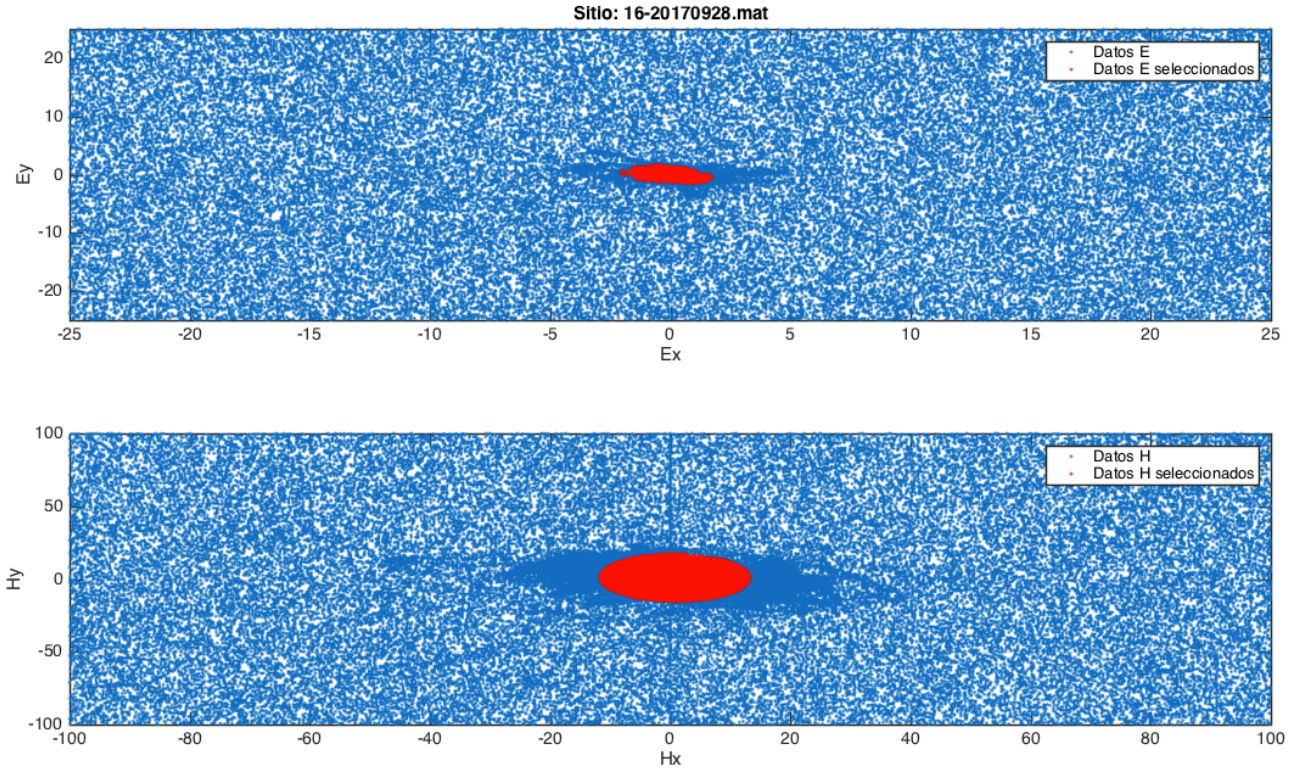


Figure 2. Data visualization of the MT data set recorded at El Mogor ranch, Guadalupe valley, Mexico. Data in red is good data, data in blue color represents anomalous data. $F_m = 1.5\sigma$ for Ex and Ey, $F_m = 2\sigma$ for Hx and Hy.

In this case data plotting would correspond to a four-dimensional ellipse, being difficult to visualize. Instead, we split visualization in two graphics, one for Ex, Ey and another for Hx, Hy (Figure 2). Despite difficulties of visual inspection of the time series, our analysis shows that 97.65% of the dataset was classified as non outlier, using a F_m=1.5σ for Ex and Ey and F_m=2σ for Hx and Hy.

Figure 3 shows the results in the classical visualization. The data classified as good is plotted in red color, while data classified as anomalous is plotted in gray color. It can be observed that by choosing the good data points, the entire series can be recovered for all channels.

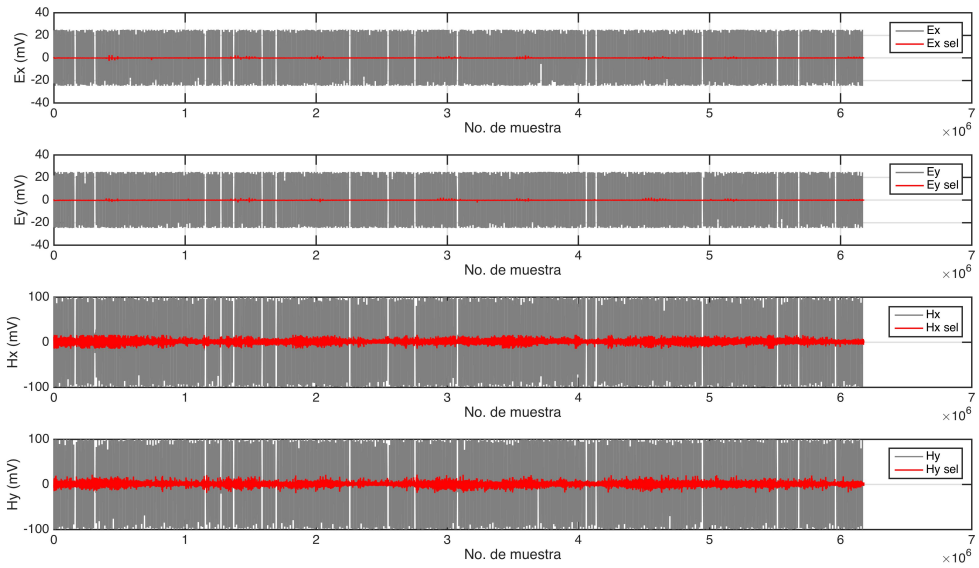


Figure 3. Time series after the analysis. Data points in red represents good data, and data points in gray represent anomalous data.

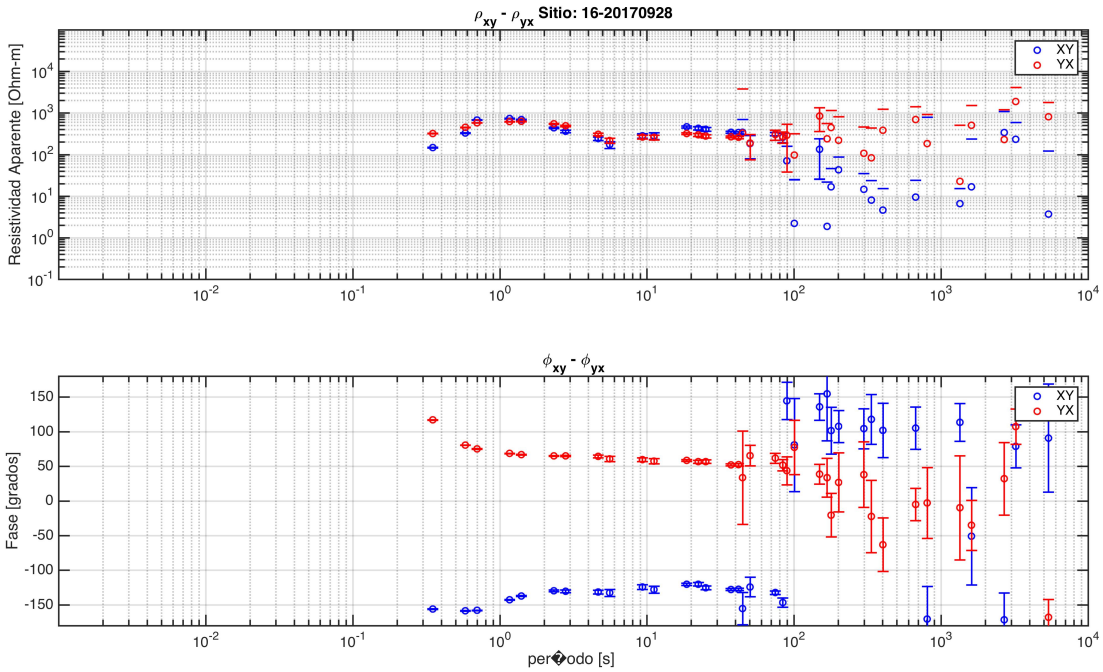


Figure 4. Apparent resistivity and phase curves obtained from time series data in figure 3. Processing was performed with BIRRP algorithm.

For comparison purposes, the original data set and the new version consisting only of selected data points (cleaned data set) were processed using the BIRRP algorithm (Chave & Thomson, 2004). The same parameters were used in both cases. Figure 4 shows the apparent resistivity and phase curves for the original data. It is observed that well-defined curves are obtained up to a 100 seconds, however the effect of the multiple outliers for larger periods could not be removed successfully by the software.

Figure 5 shows the results for the clean data set. A much better curve was obtained, with data curves better defined than in the previous case up to 1,000 seconds. This shows the practicality and efficacy of the algorithm.

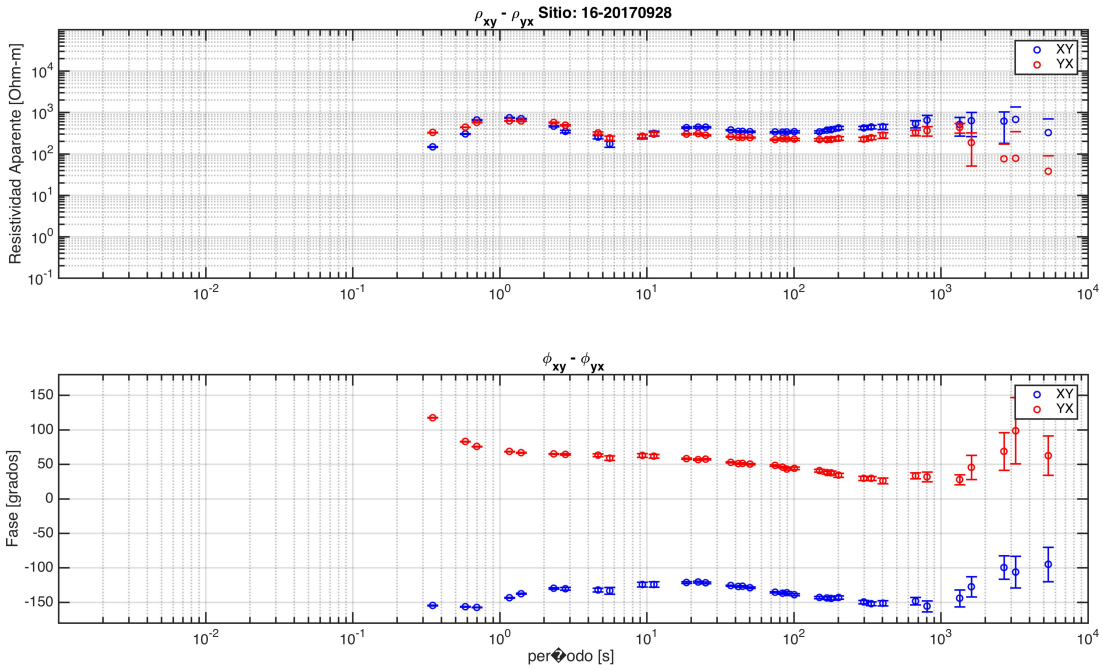


Figure 5. Apparent resistivity and phase curves obtained from selected data (red data points in figure 3).

References

-Banks, R, J. 1998. The effects of non-stationary noise on electromagnetic-response estimates. Geophysical Journal International, 135, pp. 553-563.