Processing of Magnetotelluric Data Using Machine Learning Techniques

Author: Paula Nocelo Sampedro.
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.[*]

Advisor: Alejandro Marcuello Pascual

Abstract: This report is based on magnetotelluric (MT) data from the electromagnetic (EM) Earth field. The main objective is to analyze, reduce, and asses EM noise, coming from man-made technologies. To evaluate how to denoise MT data, a code is developed in Python, which uses the K-Means technique from machine learning. In the end, it is obtained a preliminary method to identify noise situations.

I. INTRODUCTION

Magnetotelluric (MT) method consists of the Earth's exploration using electromagnetic (EM) natural waves. In the current technological era, the increasing of electrical man-made noise disturbs field measurements and becoming useless data. A precisely well-done processing data is required to denoise them [1].

Many researches have worked on this topic, which is booming in recent years. Some of them using Machine Learning methodologies [2]. This branch of data science develops programs that can automatically learn by themselves. Particularly, in this report, I will focus on K-Means, which can involve the use of a large volume of information.

This research aimed to study machine learning methods applied in MT data processing and to implement them in an existing processing program [3] to express and understand better magnetotelluric data.

The structure of this paper consists of a brief introduction of the theoretical terms followed by an explanation of the code developed. Next, the obtained results are presented, and finally the discussions and conclusions.

II. METHODOLOGY

A. Theoretical Framework

A few concepts are needed just before the Code Structure explanation. The first one is **magnetotelluric method**. As said before, it is a natural passive method that captures the Earth's resistivity distribution [4]. In this study, we will focus on lower frequency sources ($f < 1Hz$) generated from the interaction of the solar wind with the magnetosphere and ionosphere.

Starting from the Maxwell equations it could be seen the temporal dependence of electromagnetic fields. Because of their possibility to be expressed as a combination of sines and cosines its temporal dependence is harmonic, only participating wave amplitudes to the equations. At first assumption, it will be considered a homogeneous Earth, making it possible a solution for the quasi-static condition. The impedance is the ratio between E_x and H_y complex magnitude. From this relation is extracted apparent resistivity. The electrical distribution of each is given by the Earth's composition.

$$\rho(\omega) = \frac{|Z(\omega)|^2}{\mu\omega} \tag{1}$$

For further study, an inhomogeneous assumption is made and the expression (1) introduces the concept of apparent resistivity. Now, to extract the resistivity, it is going to be used The Impedance Equation.

$$E(\omega) = Z(\omega)H(\omega) \tag{2}$$

Therefore, to solve Z is useful to do cross and auto-spectrum. For the E_x component and being ${H_x}^*$ and ${H_y}^*$ the conjugated magnetic North-South and East-West fields, respectively, we obtain [5].

$$\begin{aligned} E_x{H_x}^* &= Z_{xx}H_x{H_x}^* + Z_{xy}H_y{H_x}^* \\ E_x{H_y}^* &= Z_{xx}H_x{H_y}^* + Z_{xy}H_y{H_y}^* \end{aligned} \tag{3}$$

Notice that E_x will have two complex components, Z_{xx} and Z_{xy}, from the two directions. Moreover, apparent resistivity (1) would also have the same behavior, ρ_{xx} and ρ_{xy}.

The second one is **K-Means**, whose aim is to cluster dispersed points, in this case Z, to find K groups in which centroids are more akin to the points than to the average.

B. Code Structure

The complete program, written in Python, is composed of 3 parts, as seen in FIG. 1. This method allows us to separate data acquisition according to the needs of each procedure. The main code is based on a previous

[*] Electronic address: pnocelsa7@alumnes.ub.edu

code developed in [3] to study geomagnetically induced currents.

In the first place, it is presented the magnetic and electric fields recorded in the Caceres area during 3 days, 11th, 12th, and 13th of January 2019, every second. Named B_x, B_y and E_x. Being x and y horizontal coordinates of North-South and East-West, respectively. The impedance could be found through the recorded time signals as indicated in (3).

Building from time electromagnetic series, a data pre-processing is applied. The long time series are divided into multiple equal-length segments (or windows) of N points. To smooth the responses, an overlapping between consecutive segments of one half a segment, having M segments. For each segment, all the steps the impedances are computed following the following steps.

Firstly, as said before in the Theoretical Framework, to minimize the effect of possible discontinuities, it is applied a Hanning Window. Through a cosine bell, the values will be smoothed, removing discontinuities when taking the segments which may produce Gibbs Effect. Where N is the number of the output points noticing that N must be a power of 2 in computational science.

Secondly, a detrend method is applied to balance the tendency over time. A Numpy function exists for both, Hanning Window and Detrend. At this point, B_x, B_y and E_x are graphically represented, FIG. 2 and 3, respectively.

All the previous work was in the time-domain, after applying the Fourier transform, the next analysis will be done in spectral-domain. The Nyquist Sample Theorem will limit our maximum frequency. The Spectrograms of the magnetic and electric components now can be displayed. They represent the amplitudes of the EM fields in front of the time and frequency, giving information of the field frequency changes along time, FIG. 4, 5 and 6.

The Fourier transform gives frequency the spectrum at frequency following an arithmetic progression. However, MT typically works with logarithms. Therefore, the spectrum will be distorted. To consider evenly spaced frequencies in the logarithmic scale the Parzen filter [6] is applied.

$$parzen(f) = \begin{cases} 1 & |f_t - f| = 0 \\ (\sin(u)/u)^4 & 0 < |f_t - f| \leq f_r \\ 0 & |f_t - f| \geq f_r \end{cases} \tag{4}$$

Being $u = \pi |f_t - f| / f_r$ considering f_t as the middle frequency of each window while f_r is the interval width.

In (3) it has only been considered the expression for E_x but could be exactly calculated the same way for E_y. The results might be similar for E_x than E_y.

At this point is important to remind that the goal of MT processing was to obtain the $Z(\omega)$, or the apparent resistivity and phase curves.

Finally, since the impedances Z are computed for all the segments at the same frequencies, for a given frequency there are M computed values of Z (M is the number of segments). For a given frequency, the series of the real and imaginary parts of Z of all the values obtained can be displayed, showing a cloud of data to be analyzed. It is further normalized by $\sqrt{\mu\omega}$ [7].

To analyze these data dispersion, the K-means method is applied. Two (K=2) and three (K=3) clusters had been considered. In the discussion part, the accuracy of the conjectures will be examined. A Python library called Sk-learn has been used. There are multiple commands to cluster data or to calculate the centroids.

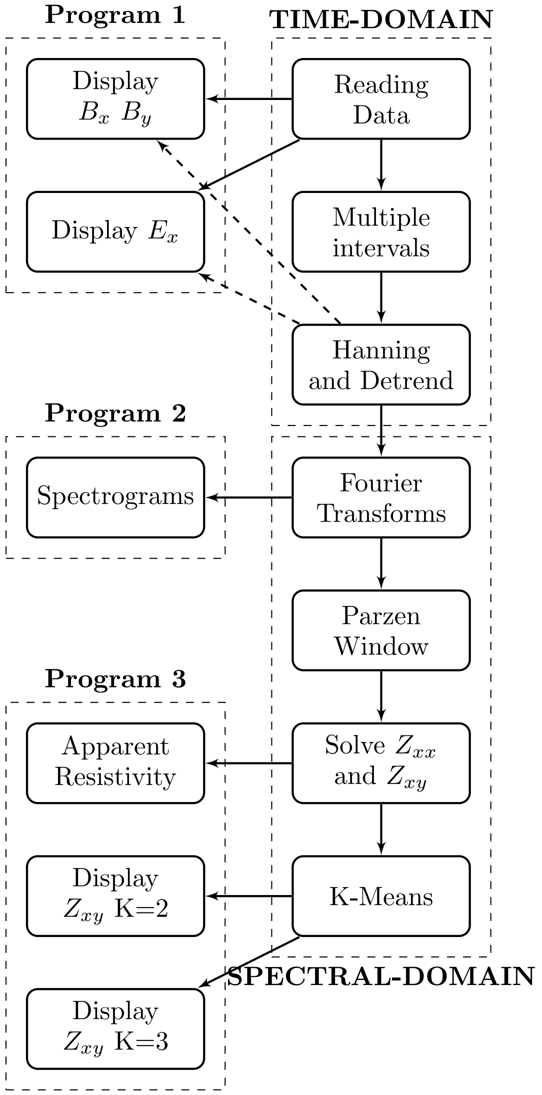

FIG. 1: Scheme of developed program.

III. RESULTS

The results that will be presented are the left panels pointed by the arrows in FIG.1. To begin with this section, B_x, B_y, and E_x time series, respectively, are displayed just after Hanning and Detrend windows.

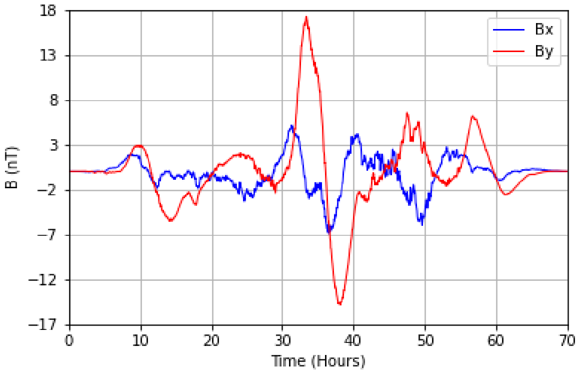

FIG. 2: Magnetic North-South field, B_x, and East-West field, B_y, in nT in front of the time, in hours.

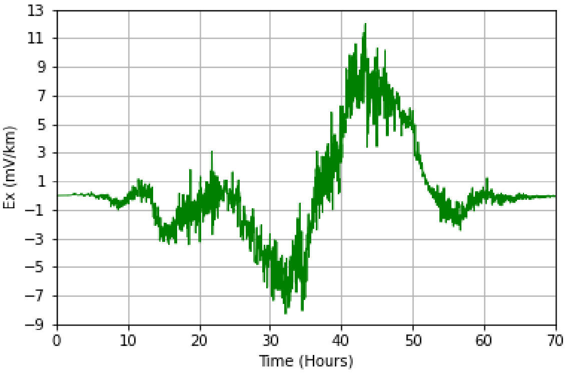

FIG. 3: Electric North-South field, E_x in mV/km, in front of the time, in hours.

To continue, it is going to be shown the three spectrograms of 3 days long.

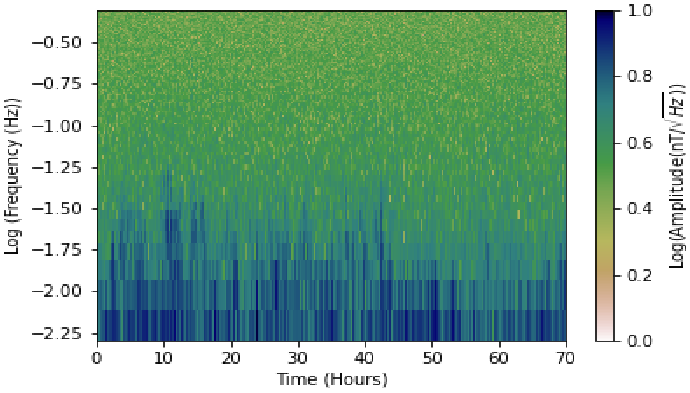

FIG. 4: B_x Spectrogram.

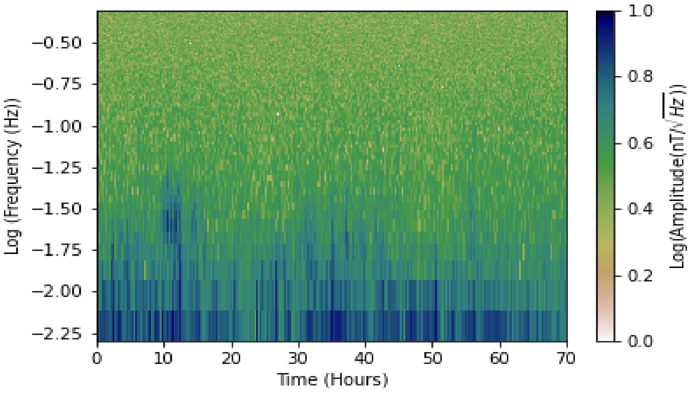

FIG. 5: B_y Spectrogram.

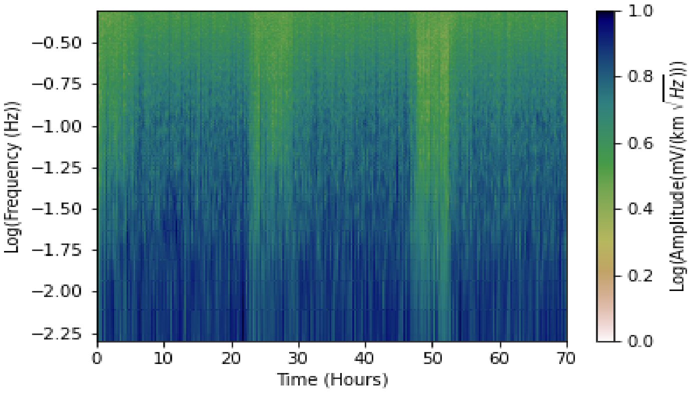

FIG. 6: E_x Spectrogram.

To apply the K-Means technique, the time series have been divided in M=123 segments of 4096 s. For each segment the value of the normalized Z_{xy} is obtained and represented in the complex plain (real and imaginary part). Now, K-Means is calculate considering 2 and 3 different distributions of component Z_{xy} for all the reference frequencies. Therefore, each figure will have 2 or 3 centroids and their corresponding groups, and the average. FIG. 7 and 8 are the f=0.0061 Hz case.

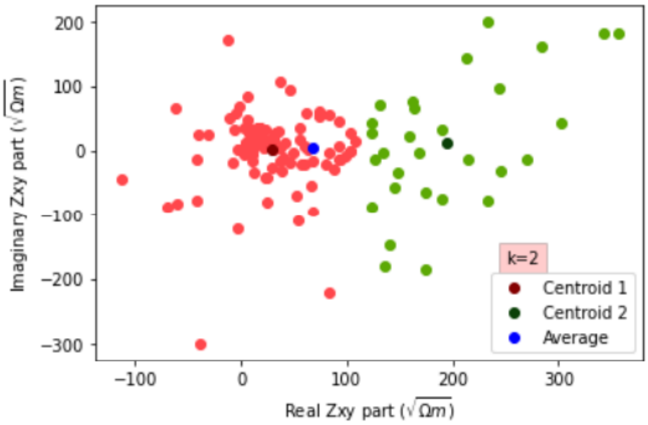

FIG. 7: Imaginary normalized Z_{xy} in front of real part in case K=2 (f=0.0061Hz).

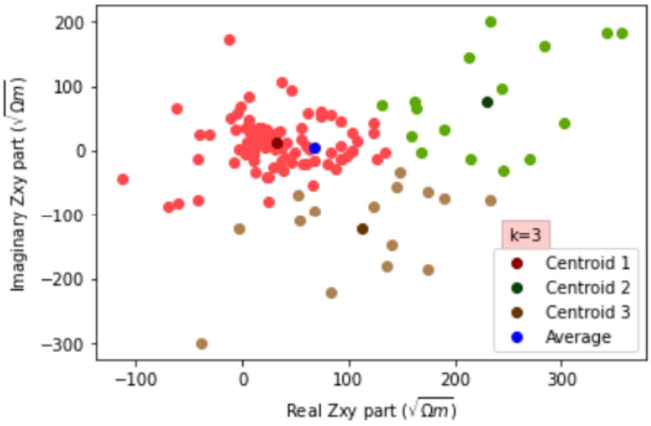

FIG. 8: Imaginary normalized Z_{xy} in front of real part in case K=3 (f=0.0061Hz).

Lastly, apparent resistivities are going to be represented next to the real data to show similarities or differences.

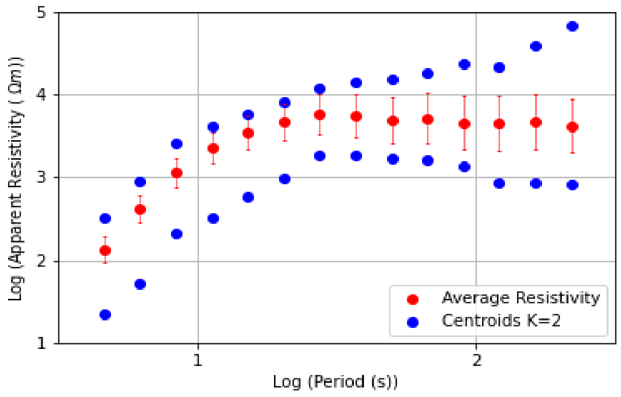

FIG. 9: Apparent resisitivity in front of period, both in logarithm scale, in case K=2.

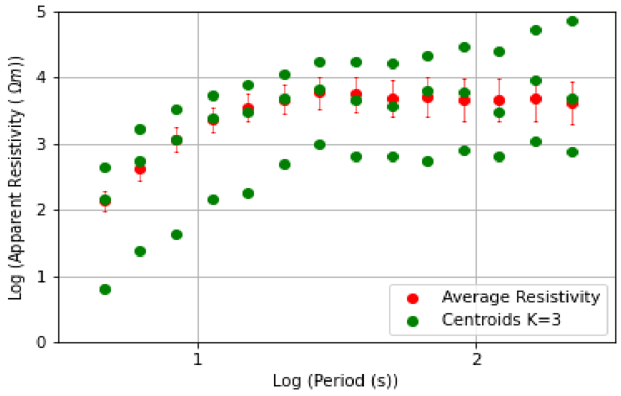

FIG. 10: Apparent resisitivity in front of period, both in logarithm scale, in case K=3.

IV. DISCUSSION

FIG. 2 and 3 let us see a significant difference between measured magnetic and electric field. Both representations had been through the same processing techniques, which they were filtering, multiple intervals, Hanning window, and Detrend. The electric field is clearly more irregular and discontinued than magnetic ones. Then, E_x presents more effect to upcoming EM noise than B_x or B_y. The scope of this report does not include E_y but, it would have the same behavior.

Moreover, in the three Spectrograms, FIG. 4, 5, and 6, it is observed that high frequencies have lower wave amplitudes.

In addition, FIG. 6 shows the EM noise and how the electric field is strongly affected, which corroborates the comment in previous discussion paragraphs. At nights, this noise greatly decreases, showing the vertical green stripes, and it could be because of the reduction of electrical activity.

On the one hand, FIG. 7 is the plot of imaginary and real part of Z_{xy} at $f = 0.0061Hz$. It becomes clear the two impedance clusters where between them lay the average, making sense of the K-Means distribution. It could be seen a larger dispersion in the green cluster, centroid 2, than the red one, centroid 1. This scattering suggests that different nature signals co-exist. If they were from the same source, all the groups would have points near to the centroid, almost overlapping like happens in the red one.

On the other hand, FIG. 8 shows exactly the same distribution impedance as FIG. 7. but contemplating three groups. Now, the third one is fitted clustering minor impedance values of the two previous groups. The average also lays between of them, supporting the method.

As long as Z_{xy} is studied, the points have to be in the first quadrant. In FIG. 7 and 8, it could be seen that some points become part of other quadrants. I would suggest directly discard these outer points. Notice that in FIG. 8, most of the brown group, centroid 3, are part of this fourth quadrant. This behavior is not just found for the presented frequency (f=0.0061 Hz), but also in other studied ones. Because of this reason, K=3 would help to automatically discard noisy points, in this case, dismissing the brown cluster in FIG. 8.

In an idealistic situation, EM noise does not appear, a single measurement is expected, coinciding all the Z_{xy} points, one cluster. Thus, the existence of two clusters suggests the following interpretation. One group might correspond approximately to the apparent resistivity not affected by noise, and the other one is influenced by noise.

For this reason, the most affected values by noise are considered to be part of the green cluster, FIG. 7 and 8, more dispersion that come from electrical man-made noise. If this green group is not considered acceptable, it could be discarded. As a result, removing the disturbances coming from the other groups, like green and brown ones, makes a useful denoising method.

The most suitable case, K=2 or K=3, will also be examined through apparent resistivity representations.

Finally, both FIG. 9 and 10 show an increasing trend with the period that may conclude that lower layers have a higher resistance than the upper ones. It has also been verified the phase to be around $\pi/4$.

The error bars have been calculated through the standard deviation apparent resistivity. According to this, FIG. 9 and 10, high frequencies correspond to lower error bars.

In FIG. 9 it could be seen that approximately the medium point between the two centroids series along the frequencies correspond with the average apparent resistivity values. Being the upper curves nearer to the average value than the lower one.

In the second case, FIG. 10, the previous behavior also happens, adding to the third centroid curve that mainly coincides with the red curve, the average one. In preceding paragraphs, it was said that this third cluster is the most affected by noise, brown one in FIG. 8. Now, it coincides with apparent resistivity, FIG. 10. If this K-Means treatment had not been done, we would not notice that the average is greatly affected by noise.

To sum up, K=3 gives more denoising information than K=2, and dismissing these two clusters could be considered an automatic denoise method for all the frequencies in this study.

V. CONCLUSIONS

In this report, the developed MT processing program has been presented showing the procedures and techniques of data science and learning machine. Also, it has been included the commands used during the Python code. The objectives of this work have been achieved.

On one hand, one of the most important things of this report is the code developed to implement the K-Means technique, which allows us to manage large series of information.

On the other hand, the graphical results exceed satisfactorily the capacity to separate data in clusters automatically, like a learning machine. It will separate different origin of EM signals but, further analysis will be required to identify them. This new code could proportionate a base for future projects.

Acknowledgments

I would like to thank Alex Marcuello Pascual for bringing me the opportunity to develop this project and for always being such supportive and patient even though the situation. Besides, to my colleagues and parents for encouraging me.

[1] Jin Li, Yiqun Peng, Jingtian Tang and Yong Li (2020). Denoising of magnetotelluric data using K-SVD dictionary training. doi: 10.1111/1365-2478.13058.

[2] Andreas Tzanis (2014).The Characteristic States of the Magnetotelluric Impedance Tensor: Construction, Analytic Properties and Utility in the Analysis of General Earth Conductivity Distributions. arXiv preprint arXiv:1404.1478.

[3] Maria Pifarré Prats (2020). Analysis of geomagnetically induced currents in the power networks. TFG Facultat de Física.

[4] Joan Campanyà i Llovet (2013). Innovation of the Magnetotelluric method and its application to the characterization of the Pyrenean lithosphere. PhD Thesis. http://www.tdx.cat/handle/10803/112702.

[5] Vozoff, K. (1972). The magnetotelluric method in the exploration of sedimentary basins. Geophysics, 37(1), 98-141.

[6] Pedro J. García Laencina, José Luis Sancho Gómez (2014). Estimación de densidad de probabilidad mediante ventanas de Parzen. En III Jornadas de introducción a la investigación de la UPCT, nº 3, 68-70.

[7] Ahmet T. Bas¸okur, Cemal Kaya and Emin U. Ulugergerli (1997). Direct interpretation of magnetotelluric sounding data based on the frequency-normalized impedance function, Geophysical Prospecting, 45, 21–37.