RESEARCH ARTICLE

10.1002/2013JB010494

Key Points:

- Nonstationary impedance estimation for the magnetotelluric method
- Approach verified with synthetic, real, and nonstationary data

Correspondence to:

M. Neukirch,
neukirch@icm.csic.es

Citation:

Neukirch, M., and X. Garcia (2014), Nonstationary magnetotelluric data processing with instantaneous parameter, *J. Geophys. Res. Solid Earth*, 119, 1634–1654, doi:10.1002/2013JB010494.

Received 5 JUL 2013

Accepted 24 FEB 2014

Accepted article online 27 FEB 2014

Published online 21 MAR 2014

Nonstationary magnetotelluric data processing with instantaneous parameter

M. Neukirch¹ and X. Garcia¹
¹Barcelona Center for Subsurface Imaging, Institute of Marine Science, CSIC, Barcelona, Spain

Abstract Nonstationarity in electromagnetic data affects the computation of Fourier spectra and therefore the traditional estimation of the magnetotelluric (MT) transfer functions (TF). We provide a TF estimation scheme based on an emerging nonlinear, nonstationary time series analysis tool, called empirical mode decomposition (EMD) and show that this technique can handle nonstationary effects with which traditional methods encounter difficulties. In contrast to previous works that employ EMD for MT data processing, we argue the advantages of a multivariate decomposition, highlight the possibility to use instantaneous parameters, and define the homogenization of frequency discrepancies between data channels. Our scheme uses the robust statistical estimation of transfer functions based on robust principal component analysis and a robust iteratively reweighted least squares regression with a Huber weight function. The scheme can be applied with and without aid of any number of available remote reference stations. Uncertainties are estimated by iterating the complete robust regression, including the robust weight computation, with a bootstrap routine. We apply our scheme to synthetic and real data (Southern Africa) with and without nonstationary character and compare different processing techniques to the one presented here. As a conclusion, nonstationary noise can heavily affect Fourier-based MT data processing but the presented nonstationary approach is nonetheless able to extract the impedances.

1. Introduction

Natural electromagnetic (EM) field variations are caused by two major working mechanisms: lightning activity at high frequencies (> 8 Hz) and magnetospheric currents excited by solar wind at low frequencies (< 8 Hz) [e.g., Garcia and Jones, 2002; Viljanen, 2012]. Rakov and Uman [2007] summarize the electromagnetic lightning discharge to three modes: (a) fast and transient leader-return stroke sequences, (b) slow and quasi-stationary continuing currents, and (c) perturbations and surges on the continuing currents. The longest lasting and most abundant in an electromagnetic time series measurements are the perturbed continuing currents, which may be viewed as being stationary on a section with some dynamic length confined by the recurrent transient strokes. Liu and Fujimoto [2011] conclude that the magnetospheric current is nonlinearly driven by the dynamic solar wind but behaves in a static manner for high magnetospheric pressure conditions. Both of these EM sources are naturally nonstationary, since both, lightning strokes and magnetospheric pressure conditions, are very dynamic and thus strictly limit the duration of any stationary electromagnetic signal.

Practitioners argue that the magnetotelluric (MT) signal is quasi-stationary (stationary on reasonably long time windows) and, thus, justify the application of the windowed Fourier transform. In practice, this procedure works very well for data with high signal-to-noise ratios but frequently encounters problems in the presence of electromagnetic noise (clearly what is called noise here would include nonstationary signal) [cp. Junge, 1996]. A concise treatise of sophisticated MT signal processing based on the Fourier transform is given by Chave [2012] in which nonstationarity is listed as one of the problems that affect transfer function estimation.

For instance, if there would be a nonstationary electric discharge, the window (data segment) of this event would not qualify as containing stationary data and such a window would have to be considered noise in a windowed Fourier transform algorithm. Moreover, noise sources (which do not include nonstationary signal) can be of any kind and do not need to be quasi-stationary (e.g., imagine a road with irregularly passing cars near the instruments) [cp. Adam et al., 1986]. All nonstationary noise sources (may also include nonstationary signal) will affect the (windowed) Fourier transform in unpredictable ways just because the data breaks the necessary assumption for the Fourier transform at least in the relevant windows. This is not an

issue when there are few affected windows, but it would become a problem when nonstationary effects are frequent. A more concise treatment of electromagnetic noise and its characteristics is given by Szarka [1987] and Junge [1996], where both acknowledge nonstationary noise sources and the aforementioned difficulties. Therefore, we argue that even though the MT signal may behave sufficiently stationary, the contained noise in the data clearly cannot always be assumed quasi-stationary as it would be required for the application of the Fourier transform.

The isolation or separation of noise has been studied intensively since the introduction of the MT method and the two major noise counteract breakthroughs date back to the 1980s. Gamble [1979] propose the use of a remote station to apply the technique of instrumental variables [Reiersøl, 1941] in order to drastically reduce bias by uncorrelated noise. Later, Jones and Jodice [1984], Egbert and Booker [1986], and Chave and Thomson [1987] advocate robust regression procedures for transfer function estimation to reduce the influence of unlikely but highly influential data points. Besides these two milestones, there has been much effort in reducing noise influence further by either trying to estimate and remove the noise directly [e.g., Egbert, 1997; Oettinger et al., 2001] in the frequency domain or by filtering or extracting quiet data sections in the time domain by visual inspection [Garcia et al., 1997] and in the time-frequency domain [e.g., Weckmann et al., 2005, and references therein]. The latter procedures are reported to be effective for particular data sets but require intense user attention and good, detailed knowledge about the data. Moreover, noise identification, separation, and/or removal is not always successful, sometimes practitioners encounter data from which it is seemingly impossible to extract reasonable transfer functions. This could be partly due to the fact that EM data (the combination of signal and noise) are not as quasi-stationary as required for the (windowed) Fourier transform. A very simplistic example would be the presence of a spike in the data, which would compromise the particular data segment (or window) in which it is present. Clearly, the presence of a moderate number of spikes is easy to counteract (through interpolation) [e.g., Jones et al., 1989; Junge, 1996], but we argue that the same principle applies to other nonstationary effects which might not be as easily identified and mitigated.

Huang et al. [1998] introduce empirical mode decomposition (EMD) in the framework of the Hilbert-Huang Transform (HHT), a novel time series analysis tool, which is data adaptive and suitable for nonlinear and nonstationary data. The decomposition provides data modes (called intrinsic mode functions (IMFs)) which are defined such that they can be represented as a single oscillation. Thus, Huang et al. [1998] argues that the definition of the IMF allow for a meaningful computation of its instantaneous parameters, like amplitude, phase, and frequency, with the Hilbert Transform. In practice, however, Huang et al. [2009] demonstrate that the Hilbert Transform often is numerically unstable and advocate a more practical routine to obtain the instantaneous parameters, which first separates amplitude and oscillation and then acquires the instantaneous phase by direct quadrature.

EMD has been tried and applied in several fields, including geophysics and the magnetotelluric method [Battista et al., 2007; Zhang et al., 2003; Cai et al., 2009; Chen and Jegen, 2008]. In particular, for MT, Cai et al. [2009] present how EMD could be used to separate obvious noise from the signal. Later, Cai [2012] attempts to substitute the Fourier transform in favor of HHT in MT processing, but the segmentation and averaging of data in order to construct marginal spectra (comparable to Fourier spectra) are unnecessary and limit the potential strength of EMD. In the same year, Chen et al. [2012] present an estimation scheme for the transfer functions in MT data by using the instantaneous parameters (in contrast to marginal spectra). However, they conclude that the implementation of remote reference processing and robust statistics can further improve their approach, because both techniques are very often required to estimate transfer functions from regular field data.

This work follows Chen et al. [2012] by using directly the instantaneous parameters obtained from EMD but in contrast to their work; here the multivariate variant of EMD by Rehman and Mandic [2009] is discussed and applied. Robust procedures are introduced to estimate instantaneous parameters, and a data selection scheme is proposed to ensure independent data. For transfer function estimation, a robust regression is advocated, which uses regressors defined by the two major robust principle components (robust principal component analysis described by Hubert et al. [2009]) of all remote data sets or for single site processing, all the available channels. Effectively, this procedure excludes the site channels from the regressors if remote data are available in order to further reduce the risk of propagating correlated noise from between site channels into the principal components. Synthetic examples demonstrate the effect of nonstationarity of the

Figure 1. EMT workflow chart to compute spectra.

source on traditional processing schemes. Semisynthetic signals, which consist of real signal and synthetic (nonstationary) noise, present the effects due to nonstationary of noise in real data, and lastly, real-world data sets verify the power of the algorithm for regular data and most notably, data in which nonstationary noise is suspected. Additionally, a MATLAB routine is presented, which creates (non)stationary synthetic MT data (or noise).

2. Outline of the EMT Algorithm

Figure 1 outlines the workflow chart of the algorithm that we have developed to process MT data using the EMD technique. We call the following scheme empirical mode decomposition-based magnetotelluric data processing, in short EMT. Here we present the outline of the code, the following sections will describe each of the steps thoroughly:

1. Decompose time series with multivariate empirical mode decomposition (MEMD). The MEMD method is used to decompose the multivariate data of all available channels (station and remotes) into oscillatory modes.
2. Compute instantaneous Parameters. Separate amplitude and oscillatory phase functions of the modes with amplitude-phase demodulation according to Huang *et al.* [2009]. Generate the complex IMFs from amplitude and oscillatory phase for each channel to permit the computation of the instantaneous phase and the instantaneous frequency defined as time derivative of the phase.
3. Gather independent data points. We ensure linear independence of the data points by defining a time scale of data dependency.
4. Organize data in frequency domain. The data points are collected in wide bins, typically 5 to 10 bins per decade, ensuring enough estimates per decade and statistical stability of the impedance estimation for each bin by exploiting the fact that the MT transfer functions vary slowly with frequency.
5. Estimate transfer functions. (i) Compute the two major robust principal components from data to use as regressor, (ii) robust regression of each channel on principal components, and (iii) estimate confidence intervals by means of bootstrapping the robust regression.

3. Algorithm Step 1: MEMD

Huang *et al.* [1998] only present the application of their technique to univariate data, but MT data consist of at least four data channels, which depend on each other. Using a univariate EMD, each signal is sifted and

Figure 2. (a–f) The related modes for two channels of a short example signal from Southern Africa data are compared with respect to time scale consistency for different EMD algorithms and added Gaussian noise variance. The presented mode is #8 out of 11 IMFs obtained from 1000 data points; however, note that the actual mode number is irrelevant for the example being representative, since Gaussian noise is equally present in all modes. The dashed boxes emphasize where there are major differences in the time scale between the two channels for the univariate case, and then, for the multivariate algorithm, the boxes highlight the differences due to the added noise. Figure 2e shows that the time scale in the mode of E_y is affected by losing one oscillation when adding 25% noise variance, and in Figure 2f the noise variance added to E_y even begins to affect the time scale in the modes of H_x . Percentage of noise refers to the noise variance for Gaussian noise relative to the average signal amplitude.

Hilbert transformed independently, decreasing the likelihood that the signals remain dependent throughout all modes due to the possible channel-independent noise characteristics. For example, if there would be a high-frequency noise in one channel which is absent in the other three, the first IMF of the first channel would contain that noise and start with the rest of the signal from mode two on, whereas the other channels would contain signal from mode one on, resulting in different time scales for all modes (Figure 2). In this example, without any previous knowledge of this noise, the corresponding modes of different channels could never be used jointly for a linear least squares approach, since they do not contain the signal of the same frequency range. For that reason *Chen and Jegen* [2008] and *Cai* [2012] suggest to calculate the marginal spectra for each channel and use those in a similar manner as it would be done with the Fourier spectra. This approach has been shown to work very similarly to the usual Fourier approach and to provide novel noise control mechanism but does not take full advantage of the possibilities that EMD offers, namely, the instantaneous parameters. *Chen et al.* [2012] circumvent this problem by only taking into account the data points of a time instant when they find a match for the instantaneous frequency (IF) for each channel but in any mode. This procedure certainly solves the problem given in the simple example above, but a procedure that only uses data points were the IF matches (arguably within a certain limit) might run into problems as soon as the channels are more seriously distorted by noise, and hence, the frequency computation for one or more channels is rather poor, ultimately decreasing the number of valid data points. In this section we discuss a multivariate decomposition algorithm that alleviates that problem by forcing all channels to decompose into correlated IMFs or in other words into IMFs of similar time scales, so that we can attribute one common IF value to all channels.

Rehman and Mandic [2009] developed a scheme to analyze multivariate signals and compute IMFs of each of the signal's components such that they remain correlated in their time scale as much as possible. The algorithm is summarized in the following:

1. Project the multivariate signal on an orthogonal n dimensional hypersphere (basis functions defined by Hammersley sequences). The dimensions of the hypersphere represents different time scales much like the orthogonal sine functions in the Fourier transform.
2. Locate the extrema of each projection (n projections in total).
3. Interpolate the multivariate signal by using the projection extrema locations for each dimension, to obtain a distinct upper and lower envelopes of the multivariate signal for each dimension of the hypersphere.
4. Average the means of upper and lower envelopes for each channel over all dimensions.
5. Subtract the average envelope mean from the data and repeat to convergence to obtain the multivariate IMF.

MEMD provides a set of IMFs for each channel and retains the dependency in between those with respect to a most similar time scale (frequency) in all channels. It is also worth noting that for a source in EM field theory all components of the electric and magnetic field have the same frequencies present at all times, meaning that if there is an electric source of 10 Hz, it will be accompanied by a magnetic field of 10 Hz. Therefore, MEMD does not at all introduce additional assumptions on the field components but rather ensures a fundamental property inherent in EM field theory for each IMF, and thus, it decomposes the MT data into IMFs which can be conceived as independent data sets.

MEMD decomposes the data set into a number of IMFs, which have the information of instantaneous amplitude, phase, and frequency at each time step, and each IMF is a time series with a dynamic and locally narrow banded IF [Flandrin and Rilling, 2004]. Each IMF is interchannel dependent, and each time step fulfills the MT equation for its IF in the same way as narrow frequency-banded time series do [Berdichevsky and Bezruk, 1973; Chen et al., 2012; Neukirch and Garcia, 2013]. However, real data will always contain noise in all channels, and the effect of the noise on the IMFs will largely depend on the (timely) local signal-to-noise ratio and can easily span from subtle effects (e.g., some noise is present in one of many clean channels) to affecting the amplitude in (originally) clean channels (e.g., half the channels are corrupted by coherent noise and affect the clean ones) to even introduce false information in all channels (e.g., severe noise introduces new extrema). As an example for noise effects, Figures 2c to 2f illustrate data with added Gaussian noise to a single channel.

This effect is conceptually comparable to how noise leaks in an ordinary Fourier transform where the signal-to-noise ratio distorts the true (noise-free) spectra, but in the EMD case the effect is local and only affects the signal at some distance around the noise occurrence, whereas the Fourier spectrum is always affected in the whole segment, since it is formulated as an integral.

The Fourier transform is a univariate algorithm, and noise in different channels cannot affect each other. Further, obviously, nonstationary effects can be reduced if the time series are broken in windows (windowed FFT). However, any nonstationary noise in a data window will affect the entire Fourier spectrum of that window, and often, robust procedures will drop exactly those spectra entirely regardless whether or not there shorter good data sections in that window. For an MEMD-based algorithm, the decision of excluding spectral information can be made for each individual time step instead of entire windows, if desired. However, care has to be taken, because even though spectral estimates are delivered at each time step, the real-time frequency resolution is much lower and depends largely on the extrema in the corresponding IMF, but let us defer discussion on this matter to section 5.

The most important point, which can be observed in Figure 2, is that channels influence each other already during the MEMD. Apparently, noise spreads throughout channels and clean channels may be affected by noise, becoming biased. This noise spreading across channels occurs because the algorithm does not assume that one of the channels can be affected by noise while the others are not; it simply finds the best correlated signal for all modes and accounts the noise as a distortion of the total electromagnetic wave field. It becomes clear that this multivariate decomposition excels with the number of provided clean channels, which aid stabilizing the mode sifting and reduce noise in noisy channels by spreading it over all channels. For this reason the mode of the E_y component in Figures 2e and 2f appears to contain less noise than one would expect from adding 25% and 100% variances of Gaussian noise, respectively. Naturally, it seems undesirable to spread noise from one channel to the others (which could be entirely avoided with a univariate EMD algorithm as Chen et al. [2012] propose), because we should preferably extract the best undistorted signal possible from our data. But since MT is an intrinsically multivariate problem, we always need the

information of all channels (of the site of interest) for the final TF estimation, and the more data points we lose due to large deviations (in, for instance, the IF, which is a data selection criterion by *Chen et al.* [2012]) in only one heavily distorted channel, the more difficult it will be to find an accurate transfer function. Using MEMD instead of EMD and enforcing a similar time scale on all channels robustifies the decomposition procedure and yields more spectral data points which can be evaluated in the regression step at the cost of spreading the multivariate noise and thus increasing noise in some channels.

Usually, a good portion of the noise is not correlated between the channels and therefore affects the channels unequally, resulting in instantaneous parameters that depart from their correct values depending on the noise. Although this is certainly not appreciated for parameters like amplitude and phase, it does come in handy for the frequency computation, which we assume to be constant between the channels. Any deviation of the IF between channels must be due to any of the following:

1. The modes do not fulfill sufficiently the definition of IMFs (having a locally zero mean).
2. The signal (channel) has been contaminated by noise (heavier contamination will result in larger deviations).
3. The frequency has been altered by nonstationary convolution with the system response of the receiver.

The first problem is a very common issue for the first modes in EMD, since the data are always sampled on some rate and the location of the extrema in the data depends much on the sampling rate (in a real nonstationary situation, the extrema can be anywhere in between the measurement directly before and directly after the recorded extrema). Routinely applied low-pass filters may alleviate much of this problem, but the exact location of the true extrema is the most crucial information for calculating the instantaneous parameter from IMFs, and this is usually not well defined for frequencies close to the sampling rate. However, in our experience the uncertainty on the location of the extrema only disperses the instantaneous parameters and does not usually introduce bias; the larger scatter in the regression is not problematic due to the larger number of data points for the higher frequencies in a data set. The second point is almost always an issue in MT, and it is broadband, meaning it is found in all frequency ranges and thus all IMFs. But since we know that the frequencies between the channels should be equal, we could use deviations between them as a selective quality marker or down weights in the later regression (similar to *Chen et al.* [2012]); however, we have not tested this idea in the present work. The last point is a rather new conclusion derived from the nonstationary convolution theorem in *Neukirch and Garcia* [2013] and will be discussed thoroughly in another work. The problem only occurs for nonstationary data convolved with a system response that varies over frequency, just like the instrument system responses for MT equipment usually do. It is not present during stationary sections and therefore a minor issue for most MT data but fairly complicated to analyze; therefore, it is out of the scope of this article. In any case, these disturbances are listed for sake of completeness as they will also affect amplitude and phase and thus can introduce undesired bias to the transfer function estimation if not removed from the data or being accounted for.

Before we continue with the subject of IF, we need to focus on the recovery of the amplitude and phase from the IMFs in the following section.

4. Algorithm Step 2: Computing Instantaneous Parameters

Huang et al. [2009] thoroughly discuss the computation of instantaneous parameters from an IMF, and *Chen et al.* [2012] continue the discussion with respect to an application in MT. We mostly follow their suggested instructions, since the IMFs of MEMD are methodically no different from the ones obtained from univariate EMD. Essentially, *Huang et al.* [2009] advise to separate amplitude and oscillatory phase with a procedure called amplitude-phase demodulation from the IMF. Then the instantaneous phase can be computed by direct quadrature from the separated, oscillatory phase function. In contrast to the original idea [*Huang et al.*, 1998] of using the Hilbert transform to obtain the phase, the direct quadrature method does not guarantee a strict analytic signal, but the routine performs well in practice and estimates the correct phase of the underlying signal more robust than the Hilbert transform.

Focusing on the differences between this work and previous studies [*Huang et al.*, 2009; *Chen et al.*, 2012], examples of instantaneous parameters are given in Figure 3, which feature two modes of a short section of a real data set from Southern Africa. Figures 3d and 3e display the instantaneous amplitude (IA), Figures 3f and 3g the instantaneous phase (IP), and Figures 3h and 3i the instantaneous frequency (IF).

Figure 3. (a) The sum of the two modes in the spirit of comparability. Examples of instantaneous parameters are displayed. (b, d, f, h, and j) High-frequency mode (IMF #2). (c, e, g, i, and k) Low-frequency mode (IMF #5). Both modes are extracted from the same data (site 072).

By definition, the direct quadrature method divides by very small numbers at the extrema of the phase modulation function leading to numerical instability at those points, which additionally amplifies uncertainties and noise. Since we apply the direct quadrature method [Huang *et al.*, 2009; Chen *et al.*, 2012], the IP usually contains small numerical errors. Especially, these numerical instabilities have a great impact on the time derivation of the phase function and are depicted by simple poles in the IF (see Figures 3h and 3i).

The poles are of first order and almost cancel each other out when summed over, which is why the phase function itself still looks smooth and the mean average over a sufficiently long time range is hardly affected. We found that a seven-point median average filter applied on the phase function before differentiating is a sufficient counteract and does not restrict the signal much more than the cubic spline interpolation already did during the sifting procedure but produces a much more stable IF (cp. Figures 3j and 3k).

In addition to the numerical instability associated with the direct quadrature method, the particular noise in each channel may cause differences in the IF between channels, where we would expect an electromagnetic field to have the very same frequency in all of its components (electric field and magnetic field) at a given time. However, we can use this fact to find a likely estimate for the common IF for all channel by using its mean or median average. The IF average is a physical meaningful representation of the true frequency of the electromagnetic signal (which is represented by all channels together) for a given time and mode. Heavy outliers from that mean average can be counteracted by using the median average and may be used to identify problematic data sections and can contribute to data quality control as mentioned in the section above. We found the median average to provide us with better estimates of the IF because of the frequent instabilities produced by the direct quadrature. These large irregularities in the IF usually do not occur in all channels at the same time, because of the impedance-related phase shift between channels (refer to Huang *et al.* [2009] for a discussion on the nature of these numerical instabilities), but occur very frequently, and thus, the median average compensates this problem, whereas the mean average would be drawn toward the outlier regularly.

All three instantaneous parameters: IA, IP, IF, and time form data quadruples and fully describe the original data. The IA and IP can be combined to form the representation of the complex spectra for a given time and frequency. Neukirch and Garcia [2013] lay out the fundamentals for signal system convolution in a HHT context and provide proof that the convolution of complex, nonstationary IMFs with a system response in time domain can be reformulated as the multiplication of the complex, nonstationary IMFs with the system response function in the frequency domain. Therefore, when processing MT time series we can use the complex IMFs in the very same way as a Fourier spectrum and carry out a statistical analysis in order to find the spectral physical relation between the channels, known as transfer functions.

For the sake of meaningful statistics with linear regression, we should try to ensure (1) that the data errors are independent (estimation accuracy) and (2) that the errors are identically distributed (accuracy and precision of estimation). Starting with the second, since we explicitly allow for nonstationarity in our scheme, it is clear that our spectral data cannot be assumed to be drawn from a single distribution. The parameters of any distribution from which the data might start with will likely change during time; this is exactly what nonstationarity means. However, the data decomposed by EMD are represented by oscillating modes which are bound to their definition and therefore always are locally zero mean functions. Thus, the definition of the IMFs ensures that the center (location) of the data distribution is zero for whatever time-varying distribution it follows. Liu [1988] discussed the importance of data being identically and independently distributed (IID) in statistical system analysis with nonparametric methods and came to the conclusion that when the bootstrap algorithm is used, the requirement of the data being IID can be somewhat relaxed, such that it is sufficient to ensure data point independency and that the underlying distributions of the data have a common location. They argue that the nonparametric nature of the bootstrap algorithm includes a robustness toward dissimilar distributions in the data as long as the locations of the distributions are very similar (in our case even equal).

The requirement that the data points are independent is more involved and has not yet been discussed in literature for an EMD setting; therefore, we dedicate the following section to that issue, then we will return to the discussion of the statistical analysis.

5. Algorithm Step 3: Independent Data Points

Data independency is an important criterion for our statistical analysis, which if left unconsidered may bias accuracy and/or precision of the methods we use in this work. Besides, the understanding of the dependency of data points allows to draw inference about the time-frequency resolution.

In our case we need to understand how data points interact and depend on each other in the total framework of HHT. Both IA and IP derive from an analysis of the inner structure of the corresponding IMF. Each IMF is constructed by a loose sifting procedure based on the signal's extrema and guided by the required properties based on the IMF definition, a highly data adaptive procedure. The subsequent amplitude-phase demodulation and the computation of the complex IMF do not rely any more on any data characteristics. For the demodulation the amplitude function and oscillatory phase function are already defined through the IMF definition and it only strips the two apparently different signals apart. Then, the direct quadrature uses the oscillatory phase function to recover locally the argument of the assumedly complex oscillation. The demodulation procedure is comparable to calculating the argument and absolute value of a complex number, which does not change or add any data dependencies but only changes the way data are described (via the complex IMF which does not introduce information to the data). Therefore, we focus on the mode decomposition itself, when looking for dependencies in the data.

First of all, keep in mind that per definition, all IMFs of a signal are theoretically locally orthogonal, which implies that one mode to the next is linearly independent and uncorrelated. However, independency is by no means guaranteed along a mode. Since the IMFs are solely defined by a subset of points of the entire signal, namely, the extrema, the IMF itself cannot have more degrees of freedom than number of extrema. All data points of an IMF between two extrema usually share a third-order interpolation polynomial, a cubic spline, which defines these data points based on the same set of extrema. Therefore, all these points between the same two extrema are dependent, whereas points that base on different sets of extrema are independent (even if just one extremum is different). Hence, it is important to only take into account one single data value for each span between two extrema to impose independency between the final data points. Naturally, the lack of independency in the definition of an IMF compromises greatly the time domain resolution suggested by IMFs but indicates that HHT does not provide a higher spectral resolution than what would be expected by the observed frequency (thus, we still need a complete oscillation to meaningful describe spectral data). Furthermore, since the cubic spline requires the closest four extrema at each data point, the distance of influence of every extrema is about two full oscillations and represents some measure of time-frequency resolution.

Since only one interextrema data point is independent, we have to pick the one which represents the entire range. Each data point should be equally valid since they are dependent. However, noise characteristics can make some points be a poorer choice than others (be reminded on the numeric instabilities due to the direct quadrature discussed in the section above). For the moment we have not designed a selection criterion based on data quality, so we simply take one point per half oscillation defined by the location of the extrema of the function:

$$P = \sin \phi \cos \phi. \quad (1)$$

Since MT processing is multivariate, we suggest to use (1) with the integral of the common IF ω_c between the channels; thus,

$$\phi(t) = \int_{-\infty}^t \omega_c(t') dt'. \quad (2)$$

This integral is basically the inverse of the time derivative of the phase used to obtain the IF in the first place, only that now the integrand is the common IF, which results in some sort of common phase for the EM data in (2), and provides an oscillatory function in (1) according to the intrinsic oscillation of the EM data. The choice of this particular function is mainly because of its fairly random selection, if we would choose data points with certain properties (e.g., low/high amplitudes), we could easily introduce bias to the transfer functions, which is not the case for this general function. However, a more careful or sophisticated selection criterion (like a weighted average) for this point could help to reduce numeric or perhaps, even electromagnetic noise and could be discussed elsewhere.

6. Algorithm Step 4: Frequency Sorting

As noted above, EMD results in a distinct frequency value for each channel. The average of those values for a given time and mode over all channels is a physical meaningful but biased representation of the true frequency of the electromagnetic signal (which is represented by all channels together). The bias should be lower for data points which have a similar frequency value and may even be considered for data quality control as we stated before. Keeping in mind that we use a common frequency function for all channels defined by the median average between them, in the following we will assume the median frequency as the common frequency between the data channels.

Remember that the instantaneous frequency (IF) is the time derivative of the phase of the complex IMF and does not yield equidistant (as, for example, the Fourier transform) but rather continuous frequency values which vary with time and thus along a mode. For this reason, it is unlikely that we can find two data points (each with two electrical (e) and two magnetic (h) components) with the very exact frequency value (ω_0), but this would be necessary in order to find a unique estimate for the transfer function tensor (Z), which is only defined at a constant frequency:

$$\begin{pmatrix} e_x(\omega_0, t) \\ e_y(\omega_0, t) \end{pmatrix} = \begin{pmatrix} Z_{xx}(\omega_0) & Z_{xy}(\omega_0) \\ Z_{yx}(\omega_0) & Z_{yy}(\omega_0) \end{pmatrix} \cdot \begin{pmatrix} h_x(\omega_0, t) \\ h_y(\omega_0, t) \end{pmatrix}. \quad (3)$$

Note that this equation deviates from the traditional form as it includes time variance for the electromagnetic fields, since the complex IMFs of the data channels are still time series. A similar form of this time variant formula has been introduced by *Berdichevsky and Bezruk* [1973] and recycled by *Chen et al.* [2012], until this form has been proofed for the EMD context by *Neukirch and Garcia* [2013]. However, even though (3) suggests that the MT impedance equation is valid at each time instant for the IMFs of the electromagnetic field, the impedance itself cannot be solved for unless there are at least two independent measurements for the same frequency value. But since the electrical impedance only changes smoothly with frequency [Cagniard, 1953], we can group similar frequency values to increase the amount of measurements available around a certain center frequency. For this procedure, we select the independent data points based on (1) and arrange them according to the common IF, omitting time dependency of the data by considering the time axis rather as index for measurements than physical time. The data reorganization in these frequency bins follows the proposed method by *Chen et al.* [2012], only that we do not allow IMF mixture for the reasons discussed in section 3.

Following this reorganization, we form an overdetermined system of equations that we can solve for the transfer function tensor at distinct frequency values. The estimation procedure is a bootstrapped, robust principal component regression and will be discussed in detail in the following section.

7. Algorithm Step 5: Robust Principal Component Regression

Egbert [1997] shows that MT sources are well described by two electromagnetic field polarizations. Practically, this means that the entire data vector space of all channels in a data set can be represented by the combination of two polarization vectors. Theoretically, the high-dimensional data (electric, magnetic, and all remote channels) can be described by a fundamental two-dimensional polarization space that contains all the variance of the data. Such a reduction of dimensionality of data vectors can be achieved by a (robust) principal component analysis (PCA), which provides the inherent components of the data vector, ordered by its eigenvalues. The two most dominant principal components (PCs) are the magnetotelluric source vectors since they should be present in all channels and contribute most to the variance of the data [cf. *Egbert*, 1997]. However, in practice, MT data are often contaminated by noise and source field effects, which limit this procedure [Egbert, 1997, 2002; Smirnov and Egbert, 2012] such that there are more than two dominant eigenvalues which contain a mixture of source polarization vectors and correlated noise. In order to separate the dominant principal components in such cases, a much more sophisticated multisite analysis is required and described by *Smirnov and Egbert* [2012], which should be followed for data sets with coherent noise contamination; however, the discussion or incorporation of such an analysis is beyond the scope of this work, although it could be implemented in our algorithm if desired. For this work, we assume that the first two principal components are a sufficiently good estimate of the MT source polarization vectors, but we restrict the data used for the PCA to remote channels only, if at least two are available. If not, the site channels can be used as usual.

A robust principal component analysis tool is provided by Hubert [2003] within the frame of the free LIBRARY for Robust Analysis (LIBRA) package [Verboven, 2005] for MATLAB and referred to as `rob_pca.m`. Smirnov and Egbert [2012] compare this code for consideration of its usage in the aforementioned multisite analysis of MT data and acknowledge its power but prefer a self-made solution for its flexibility. Since we do not attempt a multisite data analysis and assume two principal components to be sufficient, the algorithm from LIBRA appears the most reasonable solution at this stage of our algorithm.

After the computation of the two dominant PCs (say $r = (r_1, r_2)$), we formulate four (or five if the vertical magnetic field is provided) two-variate regression problems in order to separately deal with the noise distributions in each data channel. Assume the north-south electric field e_1 , the east-west electric field e_2 , the north-south magnetic field h_1 , the east-west magnetic field h_2 , and, if available, the vertical magnetic field h_3 as data channels. For each data channel x , the regression writes in a matrix notation:

$$x = r \cdot R_x + \sigma_x, \quad Z = \text{inv}(R_{h_1}, R_{h_2}) \cdot (R_{e_1}, R_{e_2}) \quad \text{and} \quad T = \text{inv}(R_{h_1}, R_{h_2}) \cdot (R_{h_3}). \quad (4)$$

R_x is a row vector and denotes the regression parameter for channel x on the PCs r ; σ_x represents the noise in x ; Z is the electric impedance according to (3); and T is the tipper function, which is the magnetic transfer function between the horizontal and vertical magnetic fields. The $\text{inv}()$ operator produces the inverse matrix, and the dot operator denotes the matrix multiplication. The formulation of the regression is slightly different from the one that is usually applied in MT but not as much as it seems at first. Actually, for an ordinary least squares solution for, say Z , this formulation yields exactly (3), which is the original formulation if time only indicates measurements. The idea behind this alternative formulation is that the regressors r result from a robust statistical procedure, which describe a part of the variance in the data and thus do not contain outliers that deviate from the dominant inherent information provided by the data. Originally, the regression is carried out on data channels directly, which first, contain highly influential outliers as discussed by Chave and Thomson [2004] and Chave [2012] and second, may contain correlated noise. In our solution, influential outliers in the regressor are unlikely unless they represent a repeated feature in most channels, which would only be the case for correlated noise, but if correlated noise would be present, only a careful and sophisticated data analysis (e.g., a multisite analysis [Smirnov and Egbert, 2012] or noise identification [Weckmann et al., 2005]) can mitigate the influence of this kind of noise. In any way, such noise would be removed, if possible, before any regression attempt and thus again validates the assumption that such noise is not present in the regressors.

We divide the total regression problem in substeps to separate the expected noise from all channels (compare (4)) in order to avoid a direct effect of coherent noise between channels. The regressions themselves are carried out robustly with an iteratively reweighted Huber weight function by calling the MATLAB intrinsic function `robustfit.m`, only specifying the desired weight function. Other weight functions are possible (refer to the MATLAB documentation for a discussion on the options), and we experimented with each one, concluding that the results obtained with the Huber weight function were most accurate and precise. The robust regression only accounts for outliers in the data channels and not for any possible outlier in the PCs, which have been computed robustly in the PCA and have disregarded bad influence points already.

EMT bootstraps the entire robust regression step in order to compute a data-dependent distribution of impedance values and estimate the data intrinsic errors of the procedure. Furthermore, as discussed before, the bootstrap operation also relaxes the requirement for statistical regressions for which data should be identically distributed and therefore reflect more reliably the estimates in case of nonstationary data. Empirically, we found 1000 iterations a sufficient trade-off between accuracy and computation time to estimate the uncertainty of our results.

8. Example Data Sets

In this section, we compare the processing scheme outlined above with the state-of-the-art processing algorithms Bounded Influence Remote Reference Processing (BIRRP) by Chave and Thomson [2004], EMTF by Egbert [1997], and the Long period Intelligent Magnetotelluric System (LIMS) data acquisition processing algorithm by Jones and Jodicke [1984]. The four algorithms are applied to a number of synthetic, half synthetic/half real and real data sets. We start with two synthetic data sets, one based on white noise as source signal and the other on a purely nonstationary waveform. These two examples will shed light on the differences between a quasi-stationary and nonstationary processing scheme.

Figure 4. The EMT algorithm is compared with BIRRP for synthetic stationary data based on spectral white noise as source signal. The assumed homogenous impedance model is defined as $Z_{xy} = 10^3 \cdot \exp(i\frac{\pi}{4})$ and $Z_{yx} = 3 \cdot 10^3 \cdot \exp(-i\frac{\pi}{4})$ and is plotted as a black line.

Then we present two examples of real-world data (Southern Africa) [Evans *et al.*, 2011] to compare performance of the processing algorithms on a natural problem. In order to illustrate the effect of non-stationary disturbances in the data, we add the electric fields from the second synthetic test to the electric fields of fairly good real data, which effectively introduces nonstationary noise in the electric fields but leaves the magnetic fields completely unaffected. Lastly, we process one real data set in which nonstationary noise sources are known to interfere and demonstrate the supremacy of EMT in such a situation. All plots contain the data and estimated confidence intervals for 95% of the data (doubled standard deviation).

8.1. Synthetic Data Based on White Noise

Using an auxiliary program to create MT synthetic data (see Appendix A), in this first example, we prepared two complex remote spectra $s = (s_x, s_y)$ from independent white noise:

$$s_x = n_x^{w,real} + i \cdot n_x^{w,imag} \quad \text{and} \quad s_y = n_y^{w,real} + i \cdot n_y^{w,imag}.$$

The number of frequencies is $N_f = 12,500$ with a step size of $df = 0.25$ Hz to obtain a time series of 25,000 samples with a sampling rate of $dt = 4$ s. The data $E = (E_x, E_y)$ and $H = (H_x, H_y)$ are computed in the frequency domain from $s = (s_x, s_y)$ by

$$E = s \cdot Z^{\frac{1}{2}} \quad \text{and} \quad H = s \cdot \text{inv}(Z^{\frac{1}{2}}) \quad (5)$$

with $S = M^{\frac{1}{2}}$ as the principal square root S of matrix M such that $S \cdot S = M$ in order to fulfill $E = H \cdot Z$ with the model

$$Z = \begin{pmatrix} 0 & 3000 \\ 1000 & 0 \end{pmatrix} * \exp\left(i \begin{bmatrix} 0 & -\frac{\pi}{4} \\ \frac{\pi}{4} & -\pi \end{bmatrix}\right).$$

Note that here the asterisk operator denotes the element wise multiplication of the matrices, and $\exp()$ refers to the exponential of the matrix, element by element. The results of processing this synthetic data

Figure 5. The chirp signal shown here is used as nonstationary synthetic signal. The north-south magnetic field component is illustrated with its nonstationary amplitude and frequency. Note that the frequency of the computed signal ranges from 1 to 30 mHz, and therefore, if data are processed outside of that range, it can only contain numerical noise.

are displayed in Figure 4 for the processing algorithms BIRRP and EMT. Both algorithms resemble the model fairly well, but BIRRP has the edge. We explain this by the fact that this synthetic source does not have any waveform, and therefore, the (M)EMD algorithm struggles to find correlated modes which it can relate to each other. On the other hand, BIRRP uses the spectral characteristics of the time series which are, per source definition, very well defined.

8.2. Synthetic Data Based on a Chirp

In order to clearly demonstrate the difference of the processing schemes, the synthetic data discussed here are completely nonstationary. Again, using SynDat (Appendix A), we define each of two orthogonal

Figure 6. The EMT algorithm is compared with BIRRP for synthetic nonstationary data based on a chirp signal as shown in Figure 5. The assumed homogeneous impedance model is defined as $Z_{xy} = 10^3 \cdot \exp(i\frac{\pi}{4})$ and $Z_{yx} = 3 \cdot 10^3 \cdot \exp(-i\frac{\pi}{4})$ and is plotted as a black line. Note that the frequency of the computed signal ranges from 1 to 30 mHz, and therefore, the processed data outside that range can only contain noise; however, inside the range, only EMT is successful in recovering the model.

Figure 7. Good example data from Southern Africa (site 072). The LIMS results are the original results from the Southern African Magnetotelluric Experiment (SAMTEX) [Evans et al., 2011].

magnetic source fields $s = (s_x, s_y)$ by a logarithmic frequency oscillation f and a logarithmic amplitude oscillation a :

$$s = \Re \left(a \cdot \exp \left(i \int f dt \right) \right) \quad (6)$$

$$\log(f) = A + B \cos(F_f t)$$

$$\log(a) = C + D \sin(F_a t).$$

The parameters $A, B, C = (C_x, C_y)$ and $D = (D_x, D_y)$ define frequency, and amplitude range and the parameters F_f and F_a control the degree of nonstationarity by the oscillation rate of f and a , respectively. The time axis t is sampled at a rate of 4 s for a total length of 100,000 s or 25,000 samples. Figure 5 displays the magnetic north-south component with its respective amplitude and frequency function. By design the signal is a locally zero mean function to ensure that it complies with the definition of the IMFs, even without the need to apply (M)EMD. As in the last example, the impedance Z is assumed to be homogenous with

$$Z = \begin{pmatrix} 10 & 3000 \\ 1000 & 30 \end{pmatrix} * \exp \left(i \begin{bmatrix} \frac{\pi}{4} & -\frac{\pi}{4} \\ \frac{\pi}{4} & -\frac{\pi}{4} \end{bmatrix} \right).$$

The electric and magnetic fields are computed according to (5). Figure 6 compares the results of processing the electric and magnetic data with BIRRP and EMT. Both algorithms are called with their respective default parameters to compare the results assuming no a priori knowledge about the data. EMT successfully recovers the model in the frequency range of the computed data, but BIRRP fails processing the data, which can only be addressed to the strict nonstationarity of the signal and exemplifies that Fourier transform based methods are not suitable for strictly nonstationary signals, even those that apply the windowed Fourier transform. However, this example is not a fair comparison as this kind of signal is not natural and treatises of the physics of typical MT sources [see Rakov and Uman, 2007; Liu and Fujimoto, 2011] suggest that the

Figure 8. Fair example data from Southern Africa (site 027). The LIMS results are the original results from the SAMTEX [Evans et al., 2011].

sources are not as nonstationary as this example for an extended period of time and instead can be treated as quasi-stationary. This example serves only illustrative purposes and is designed to highlight the strength of EMD, to expose the weakness of the Fourier methods and most of all, to demonstrate clearly how non-stationarity appears in the results of Fourier methods. In the following section, we present more practical examples using real data.

8.3. Fairly Good Real Data From Southern Africa

Now let us compare the algorithms using three real data sets from Southern Africa which correspond to the sites 027 and 072 with site 045 as remote reference, 042 with 027 as remote reference for long-period data, and 043 as remote reference for the short-period data [Evans et al., 2011]. The first two time series have approximately 500,000 samples on a sampling rate of 5 s, and we only consider the horizontal magnetic fields as remote reference, since they have proven to be sufficiently efficient in removing coherent noise from the local fields. The last example has up to 2 million samples for the high frequencies at 2560 Hz and around 500,000 samples on a sampling rate of 5 s. The high-frequency data only have one remote reference site, and for the long periods, we selected the best suitable one.

The first example (site 072) is considered good for MT processing purposes when processed with the available remote magnetic channels (of site 045). Figure 7 displays the processing results for the *LIMS processing algorithm* (original results) and EMT, and both algorithms agree very well.

The second example (site 027) contains somewhat more noise even when processed with the available remote magnetic channels (site 045). Figure 8 compares the LIMS processing algorithm (original results) with EMT and shows that there are only marginal differences. Both algorithms agree well with the phase, but there is a slight difference in the amplitudes. Concluding this example, EMT appears to obtain similar results but the smaller confidence intervals are less conservative or suggest higher precision.

Figure 9. Southern Africa data (site 072) jammed with synthetic nonstationary electric noise. (a) The electric field channel with and without the added noise to illustrate the impact of the noise compared to the data. (b) Section of the added noise to emphasize that both amplitude and frequency contents of this noise is clearly nonstationary.

8.4. Real Data Jammed With Synthetic Nonstationary Noise

As a semisynthetic test, we combine the good real data set (site 072) from the previous section with synthetic, nonstationary noise. The goal of this test is to learn how easily a quasi-stationary source can be compromised by nonstationary noise and to test if our algorithm is able to treat the situation correctly. We consider the nonstationary noise to be present in the electric fields only and leave the magnetic and remote channels completely unaffected. This way we can see if the computation of the spectra via Fourier transform fails or succeeds, since stationary noise in the electric channels could be cleaned by the unaffected magnetic and/or remote reference channels by the remote-referencing technique.

As data, we use the data set shown in Figure 7 and add independent, purely nonstationary noise as defined in (6) to each electric field channel:

$$(e_{x, \text{ with noise}}, e_{y, \text{ with noise}}) = (e_x, e_y) + (s_1, s_2).$$

Then, we try to recover the original impedance (Figure 7) by processing the altered data with BIRRP and EMT to study the effects of the added, nonstationary noise. The test is performed thrice, first with a certain noise level, then again with the noise doubled and quadrupled. Figure 9a displays the electrical field north-south component with and without the added noise for all three tests and as an example; a section of the added noise is illustrated in Figure 9b with its parameters amplitude and frequency. The spectral range of the noise is set between 1.7 mHz and 19 mHz, respectively 52 s and 610 s, so we expect to see the biggest impact on the data processing results in that range.

Figures 10 to 12 display the estimated impedances with the increasing impact of the nonstationary noise. Where in Figure 10 the noise only raises the confidence limits for BIRRP, it camouflages the estimates in their entirety for larger noise amplitudes in Figures 11 and 12 so much that the impedance cannot be retrieved. On the other hand, EMT is barely affected by the lowest and medium amplitude noises and still provides interpretable results with the highest noise amplitude.

Figure 10. Southern Africa data (site 072) jammed with synthetic, nonstationary electric noise of low amplitude.

Figure 11. Southern Africa data (site 072) jammed with synthetic, nonstationary electric noise of medium amplitude.

Figure 12. Southern Africa data (site 072) jammed with synthetic, nonstationary electric noise of high amplitude.

8.5. Problematic Real Data From Southern Africa

This last example is a real, broadband data set and has been acquired in a region where DC trains operate and active mining takes place. *Evans et al.* [2011] report problems in processing the data in particular due to these noise sources.

We focus on site 042 with 027 as reference for long-period measurements (> 20 s) and with 043 as reference for short-period data (< 20 s). The long-period data have been collected with LIMS instruments and the short-period data were measured by band 5 of Phoenix Systems' instruments. The site contains a tremendous amount of noise which makes interpretation difficult from about 3 s in Figure 13. The data of this site were originally processed with EMTF [Egbert, 1997] for short periods (BBMT) and with the LIMS processing algorithm [Jones and Jodicke, 1984] for the long-period data (long-period magnetotelluric (LMT)). The amplitude results from LMT have been scaled by the acquisition team to account for static shift according to the interpretation of the BBMT data, whereby the results from EMT are unchanged, since it does not suggest that the measurements of the long-period data have been affected notably by static shift.

We use originally published data for this plot, because we argue that (in time of original publication) the interpretation of the data (that it is affected by static shift) was wrong due to some noise effect. The EMT result is not shifted, because it does not lead to the conclusion that the LMT data require a shift, which exemplifies the long-ranging effect of noise beyond data processing and highlights the strength of the algorithm in this situation.

Besides the apparent noise between 3 s and 20 s, the phase estimations between 1 s and 100 s obtained from EMT are consistently 5° to 10° lower than the results estimated by the other algorithms, which we cannot explain at this point. Two possible reasons for this discrepancy could be due to nonstationary spectral leakage in the other algorithms (compare processing of a purely nonstationary data set in Figure 6) and due to strong correlated noise distorting significantly the first two dominant principal components.

Figure 13. Site 042 contains a tremendous amount of noise which complicates interpretation of periods of 3 s and more. The EMTF [Egbert, 1997] (BBMT) and the LIMS [Jones and Jodice, 1984] (LMT) results are the original results from the SAMTEX [Evans et al., 2011]. The merge of BBMT and LMT responses was performed manually, and as it still is common practice, the LMT apparent resistivity were shifted to match the BBMT apparent resistivity at the overlapping periods.

9. Conclusion

In the course of this work, we outlined a robust magnetotelluric data processing scheme purely based on nonstationary methods and showed that its results compare to state-of-the-art algorithms. In contrast to other groups, our algorithm directly uses the instantaneous parameters of the measured multivariate time series and therefore naturally handles nonstationary sources. In theory, our scheme is less apt to introduce bias from spectral leakage due to this kind of noise and our synthetic and real data examples support this.

The algorithm carefully incorporates the most general and important data quality control measures like remote referencing and robust statistics as countermeasures for uncorrelated noise between occupied sites and control of highly influential but statistically unlikely data points, respectively.

This new methodology operates in a time-frequency domain and, therefore, potentially enables new data quality control measures like controlling instantaneous changes in the parameters amplitude, phase, and frequency, which could be investigated in a future work.

The function to select the independent data samples assures that the correct amount of data is selected, but the function of choice is somewhat arbitrary. On one hand it can be seen as an advantage that the samples are drawn arbitrary or random, but on the other hand alternative ways should be investigated for assessing their performance.

We demonstrated on synthetic and real data that a nonstationary approach in MT processing can be fruitful. The synthetic, nonstationary source in this work is specifically designed to disturb the Fourier transform and to break its assumptions; however, the results provide an insight in how bad real, nonstationary noise can affect MT measurements and encourage to verify the findings on more real-world data sets that are suspected to contain, in particular, nonstationary noise, e.g., data that are acquired close to train lines, mining activity, or electric fences.

Lastly, we present one such example of real data and find that at the time of original data processing, even the interpretation of the data has been affected by nonstationary noise, because the long-period data have been corrected unnecessarily for static shift by the original processing team.

We encourage to reassess more data sets that have been difficult to process in the past in order to investigate for nonstationary effects. However, we wish to stress that at this moment, our proposed algorithm is realized in MATLAB and runs rather slow (about 1 day for 10 million data points) on desktop computers. Most of the time, it delivers similar results compared to much faster and more efficient processing algorithms like BIRRP [Chave and Thomson, 2004], EMTF [Egbert, 1997], or the LIMS processing algorithm [Jones and Jodicke, 1984]. Therefore, we consider our algorithm a special purpose code for data that are suspected to be contaminated by nonstationary effects.

Appendix A: SynDat: Computing (Non)Stationary Synthetic Data for MT

Availability of synthetic data is fundamental for hypothesis testing in many areas of applied science, since it offers a simple and noise-free mean of acquiring test data, which could be expensive, difficult, or time consuming in the laboratory or in the field, and it allows to design easily custom-made properties of test data, which often help to spotlight both, important problems and findings in a hypothesis.

We use the MATLAB program SynDat to generate (non)stationary synthetic data in the course of this work. It allows to define freely the frequency and amplitude time series of numerical remote channels (as given in (6)), which are used to compute synthetic MT data (as given in (5)) by means of the nonstationary convolution theorem according to Neukirch and Garcia [2013]. Additionally, the synthetic data can be modeled for any impedance by importing the respective transfer function (TF) from files of the Electrical Data Interchange format or be computed for the impedance of a one-dimensional conductivity model [Keller and Frischknecht, 1966]. This program is freely available upon request to any of the authors.

Acknowledgments

The authors thank Alan Chave, Woods Hole Oceanographic Institution, for suggesting the use of the autocorrelation for data independency, which ultimately led to our final solution. Further, we appreciate the constructive discussions with Jin Chen and Bjoern Heincke at Geomar, Kiel. Data from Southern Africa are part of the SAM-TEX. This work has been funded by REPSOL under the framework of the CO-DOS project. And lastly, we thank Jin Chen, an anonymous reviewer, and the Associate Editor for their very constructive criticism.

References

- Adam, A., L. Szarka, J. Verö, A. Wallner, and R. Gutdeutsch (1986), Magnetotellurics (MT) in mountains—Noise, topographic and crustal inhomogeneity effects, *Phys. Earth Planet. Inter.*, **42**(3), 165–177.
- Battista, B., C. Knapp, T. McGee, and V. Goebel (2007), Application of the empirical mode decomposition and Hilbert–Huang transform to seismic reflection data, *Geophysics*, **72**(2), H29–H37, doi:10.1190/1.2437700.
- Berdichevsky, M. N., and I. A. Bezruk (1973), *Magnetotelluric Sounding With the use of Mathematical Filters*. Izv. Akad. Nauk SSSR.
- Cagniard, L. (1953), Basic theory of the magnetotelluric method of geophysical prospecting, *Geophysics*, **18**, 605–635.
- Cai, J. H. (2012), Magnetotelluric response function estimation based on Hilbert–Huang transform, *Pure Appl. Geophys.*, **170**, 1899–1911, doi:10.1007/s00024-012-0620-3.
- Cai, J. H., J. T. Tang, X. R. Hua, and Y. R. Gong (2009), An analysis method for magnetotelluric data based on the Hilbert–Huang Transform, *Explor. Geophys.*, **40**(2), 197.
- Chave, A. (2012), The magnetotelluric method: Theory and practice, in *Estimation of the Magnetotelluric Response Function*, edited by A. D. Chave and A. G. Jones, pp. 165–218, Cambridge Univ. Press, Cambridge, U. K.
- Chave, A. D., and D. J. Thomson (1987), On the robust estimation of power spectra, coherences, and transfer functions, *J. Geophys. Res.*, **92**(B1), 633–648.
- Chave, A. D., and D. J. Thomson (2004), Bounded influence magnetotelluric response function estimation, *Geophys. J. Int.*, **157**(3), 988–1006, doi:10.1111/j.1365-246X.2004.02203.x.
- Chen, J., and M. Jegen (2008), *Empirical Mode Decomposition and Hilbert–Huang transform (HHT) in MT Data Processing*. 19. International Workshop on EM Induction in the Earth.
- Chen, J., B. Heincke, M. Jegen, and M. Moorkamp (2012), Using empirical mode decomposition to process marine magnetotelluric data, *Geophys. J. Int.*, **190**(1), 293–309, doi:10.1111/j.1365-246X.2012.05470.x.
- Egbert, G. D. (1997), Robust multiple-station magnetotelluric data processing, *Geophys. J. Int.*, **130**, 475–496.
- Egbert, G. D. (2002), Processing and interpretation of electromagnetic induction array data, *Surv. Geophys.*, **23**, 207–249.
- Egbert, G. D., and J. R. Booker (1986), Robust estimation of geomagnetic transfer functions, *Geophys. J. Int.*, **87**(1), 173–194.
- Evans, R. L., et al. (2011), Electrical lithosphere beneath the Kaapvaal craton, southern Africa, *J. Geophys. Res.*, **116**(B4), B04105, doi:10.1029/2010JB007883.
- Flandrin, P., and G. Rilling (2004), Empirical mode decomposition as a filter bank, *Signal Process Lett.*, **11**, 112–114.
- Gamble, T. D. (1979), Magnetotellurics with a remote magnetic reference, *Geophysics*, **44**, 53–68.
- Garcia, X., and A. G. Jones (2002), Atmospheric sources for audio-magnetotelluric (AMT) sounding, *Geophysics*, **67**(2), 448–458.
- Garcia, X., A. D. Chave, and A. G. Jones (1997), Robust processing of magnetotelluric data from the auroral zone, *J. Geomagn. Geoelec.*, **49**(11–12), 1451–1468.
- Huang, N. E., Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu (1998), The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, *Proc. Phys. Soc. A*, **454**(1971), 903–995.
- Huang, N. E., Z. Wu, S. R. Long, K. C. Arnold, and X. Chen (2009), On instantaneous frequency, *Adv. Adapt. Data Anal.*, **01**, 77.
- Hubert, M. (2003), A robust PCR method for high-dimensional regressors, *J. Chemom.*, **17**(8–9), 438–452, doi:10.1002/cem.783.
- Hubert, M., P. J. Rousseeuw, and T. Verdonck (2009), Robust PCA for skewed data and its outlier map, *Comput. Stat. Data Anal.*, **53**, 2264–2274.

- Jones, A. G., and H. G. Jodicke (1984), Magnetotelluric transfer function estimation improvement by a coherence-based rejection technique, paper EMU, 271, presented at 54th Annl. Int. Mtg. Soc. Excl. Geophys., Atlanta, Ga.
- Jones, A. G., A. D. Chave, G. D. Egbert, and D. Auld (1989), A comparison of techniques for magnetotelluric response function estimation, *J. Geophys. Res.*, *94*, 14,201–14,213.
- Junge, A. (1996), Characterization of and correction for cultural noise, *Surv. Geophys.*, *17*, 361–391.
- Keller, G. V., and F. C. Frischknecht (1966), *Electrical Methods in Geophysical Prospecting*, 519 pp., Pergamon Press, Oxford, U. K.
- Liu, R. Y. (1988), Bootstrap procedures under some non-IID models, *Ann. Statist.*, *16*, 1696–1708.
- Liu, W., and M. Fujimoto (2011), *The Dynamic Magnetosphere*, Springer, Netherlands.
- Neukirch, M., and X. Garcia (2013), Non stationary time series convolution: On the relation between the Hilbert-Huang and Fourier transforms, *Adv. Adapt. Data Anal.*, *5*, 1,350,004.
- Oettinger, G., V. Haak, and J. C. Larsen (2001), Noise reduction in magnetotelluric time series with a new signal-noise separation method and its application to a field experiment in the Saxonian Granulite Massif, *Geophys. J. Int.*, *146*, 659–669.
- Rakov, V. A., and M. A. Uman (2007), *Lightning, Physics and Effects*, Cambridge Univ. Press, Cambridge, U. K.
- Rehman, N., and D. P. Mandic (2009), Multivariate empirical mode decomposition, *Proc. Phys. Soc. London, Sect. A*, *466*(2117), 1291–1302.
- Reiersøl, O. (1941), Confluence analysis by means of lag moments and other methods of confluence analysis, *Econometrica*, *9*, 1–24.
- Smirnov, M. Y., and G. D. Egbert (2012), Robust principal component analysis of electromagnetic arrays with missing data, *Geophys. J. Int.*, *190*, 1423–1438, doi:10.1111/j.1365-246X.2012.05569.x.
- Szarka, L. (1987), Geophysical aspects of man-made electromagnetic noise in the earth—A review, *Surv. Geophys.*, *9*, 287–318.
- Verboven, S. (2005), LIBRA: A Matlab library for robust analysis, *Chemometr. Intell. Lab.*, *75*, 127–136.
- Viljanen, A. (2012), The magnetotelluric method: Theory and practice, in *Description of the Magnetospheric/Ionospheric Sources*, edited by A. D. Chave and A. G. Jones, pp. 96–121, Cambridge Univ. Press, Cambridge, U. K.
- Weckmann, U., A. Magunia, and O. Ritter (2005), Effective noise separation for magnetotelluric single site data processing using a frequency domain selection scheme, *Geophys. J. Int.*, *161*(3), 635–652.
- Zhang, R. R., S. Ma, E. Safak, and S. Hartzell (2003), Hilbert-Huang transform analysis of dynamic and earthquake motion recordings, *J. Eng. Mech.*, *129*, 861–875.