Geophys. J. Int. (2003) **152,** 1–7

Magnetotelluric data processing with a robust statistical procedure having a high breakdown point

M. Yu. Smirnov

Department of Earth Sciences, Geophysics, Uppsala University, Villavagen 16, SE-752 36 *Uppsala, Sweden. E-mail: ms@geofys.uu.se*

Accepted 2002 July 9. Received 2002 July 9; in original form 2001 August 3

SUMMARY

A new robust magnetotelluric (MT) data processing algorithm is described, involving Siegel estimation on the basis of a repeated median (RM) algorithm for maximum protection against the influence of outliers and large errors. The spectral transformation is performed by means of a fast Fourier transformation followed by segment coherence sorting. To remove outliers and gaps in the time domain, an algorithm of forward autoregression prediction is applied. The processing technique is tested using two 7 day long synthetic MT time-series prepared within the framework of the COMDAT processing software comparison project. The first test contains pure MT signals, whereas in the second test the same signal is superimposed on different types of noise. To show the efficiency of the algorithm some examples of real MT data processing are also presented.

Key words: data processing, electromagnetic induction, magnetotellurics, robust statistics, spectral analysis.

1 INTRODUCTION

Over the previous decade, the instruments available for magnetotelluric (MT) measurements have been improved significantly. This requires equally advanced data analyses techniques in order to fully exploit the increasing quality of MT measurements in many cases contaminated by industrial noise.

One of the most successful ways to improve the quality of MT transfer functions estimations is to follow the principles of robust statistics. Adaptation of robust statistics to the MT data processing problem has been discussed by Egbert & Booker (1986), Chave *et al.* (1987), Chave & Thomson (1989) and Larsen *et al.* (1996). Many recent examples have shown the effectiveness of such techniques and demonstrated their advantage over standard least-squares (LS) methods (Jones *et al.* 1989).

It has been shown that, in contrast to the traditional LS solution, robust procedures produce more stable and unbiased results in the presence of large errors in the data, in both the frequency and time domains. In order for the conventional LS method to work reasonably well, data have to be examined first and appropriate data editing and rejection have to be made by hand. Robust statistics schemes make it possible to formalize and automate such hand-preparation. However, in many cases only formal robust procedures allow an adequate data treatment to be performed, especially when large data sets are available.

The basic measure of the robustness of an estimator is its breakdown point ε^*, that is, the fraction (up to 50 per cent) of outlying data points that can corrupt the estimator (Hampel *et al.* 1986). In other words, the breakdown point may be roughly defined as the smallest percentage of gross errors that may cause an estimator to take on arbitrarily large values. It is well known that the breakdown point of the LS solution is zero, which means that even a small amount of noise might have a strong influence on the final estimate. This leads to applying different kinds of robust schemes. Commonly used ones are based on *M*-estimators (Huber 1981). The stablest of them have breakdown points approaching 30 per cent in the case of a simple linear regression. However, they do not have the highest achievable breakdown point, but their efficiency in the case of outlier-free Gaussian data is comparable with that of the LS solution. Robust schemes may exist that have a higher breakdown point (Hampel *et al.* 1986). It would be promising to use these estimators for MT data processing.

In this study, such a robust scheme is utilized. The method was suggested by Siegel (1982) and his calculations are based on a repeated median algorithm. This estimator has the highest breakdown point, namely 50 per cent. This implies that nearly one-half of the data can be outliers, but the solution will still yield a reasonable result. However, methods with a very high breakdown point usually have a smaller asymptotic relative efficiency at the Gaussian distribution than LS. This means that the higher the robustness of the estimator the higher the asymptotic variance. In order to achieve the same parameter uncertainties by the robust procedure more measurements are required. For instance, the loss of efficiency of the median estimator is approximately 60 per cent relative to the L_2 norm estimator. To increase the efficiency of the final estimator for short time-series, we supplement the Siegel estimator, which serves as an initial approximation, with a reduced *M*-estimator.

To eliminate and remove obvious outliers and fill short gaps in the time domain before the main processing, the AR-prediction method is used.

The spectral transformation of the original time-series is performed by means of the fast Fourier transformation technique by subdividing the data into segments. To overcome the problem with highly noisy segments, sorting is applied using a coherence criterion. A cascade decimation technique is used to obtain results over the whole period range.

The algorithm is described here in the single-station implementation and in terms of impedance tensor estimations, but it can be applied to estimate magnetic transfer functions, as well. An adaptation for remote reference and multistation processing has also been performed with appropriate modifications, but will not be discussed here.

2 THE MAGNETOTELLURIC DATA PROCESSING ALGORITHM

2.1 Correction for outliers in the time-series and the filling of gaps with predicted values

Prior to performing a spectral analysis to derive magnetotelluric transfer functions, the time-series involved are preconditioned in the time domain. The ultimate goal is to reduce the bias of the final estimate. We proceed in the following way.

An autoregressive (AR) model is used to identify outliers and to close short gaps in a given data set. Assuming that the time-series-generating process is sufficiently well described by such a model, predictions and thereby prediction errors can be derived. If the latter exceed a preset threshold level, then the respective datum is replaced by the predicted value of the model and, by the same procedure, short data gaps are filled in.

In detail consider a time-series $x[n]$ to be the output of a causal filter, that is

$$x[n] = \sum_{k=0}^{\infty} h[k]u[n-k], \tag{1}$$

where $h[k]$ is the discrete infinite response function and $u[n]$ is the input assumed to be white noise. Then, for an AR model of order p with coefficients $a_p[1], a_p[2], \ldots, a_p[p]$, we have

$$x[n] = -\sum_{k=1}^{p} a_p[k]x[n-k] + u[n]. \tag{2}$$

Once the coefficients $\hat{a}_p[k]$ have been determined for the data segment under consideration, the resulting linear model yields a forward predicted value $\hat{x}[n]$ for the datum $x[n]$,

$$\hat{x}[n] = -\sum_{k=1}^{p} \hat{a}_p[k]x[n-k], \tag{3}$$

and thereby a forward prediction error

$$e_p[n] = x[n] - \hat{x}[n] = x[n] + \sum_{k=1}^{p} \hat{a}_p[k]x[n-k]. \tag{4}$$

Provided that the prediction errors form white noise, the derived model corresponds to the AR model, as expressed by eq. (2).

For the actual implementation, the first data segment is formed from the first N values of the time-series. From these values p coefficients $\hat{a}_p[k]$ are calculated for an AR process of a chosen order, using the modified covariance method (Marple 1987). After this, a prediction is made for the next data point according to eq. (3).

If the resulting forward prediction error $e_p[N+1]$ exceeds a specified threshold ε_e, the original datum $x[N+1]$ is replaced by $\hat{x}[N+1]$. Should the datum $x[N+1]$ be missing, the gap is filled with the prediction. Subsequently, the segment is shifted by one point forward and the prediction $x[N+2]$ is made from a new set of AR coefficients. The process is repeated until the final data point is reached. Clearly, the first segment must be without gaps and should not contain obvious outliers. Furthermore, the gaps that are to be filled should be sufficiently short compared with the order of the AR model.

It remains to define an appropriate threshold level for replacements and to consider the choices of p and N. The threshold value ε_e is determined as $\varepsilon_e = c\sqrt{D_p}$ from the dispersion of the prediction errors

$$D_p = \frac{1}{N-p-1} \sum_{p+1}^{N} |e_p[n]|^2. \tag{5}$$

In this study a c factor within the range 3–10 is used. The length, N, of the segments should be as short as possible in order to allow for changing AR processes within the time-series. For AR models in this analysis orders of p between 4 and 12 were applied. The length of the segment should be from $2p$ to $3p$ in order to have sufficient accuracy for the model parameter estimates.

2.2 Spectral transformation of magnetotelluric data

Time-series for the horizontal electromagnetic components are denoted as e_x, e_y, h_x, h_y and their Fourier transforms as E_x, E_y, H_x, H_y, respectively. The linear relations to be evaluated are

$$\begin{pmatrix} E_x \\ E_y \end{pmatrix} = \begin{bmatrix} Z_{xx} & Z_{xy} \\ Z_{yx} & Z_{yy} \end{bmatrix} \begin{pmatrix} H_x \\ H_y \end{pmatrix}, \tag{6}$$

where $[Z] = \left[\begin{smallmatrix} Z_{xx} & Z_{xy} \\ Z_{yx} & Z_{yy} \end{smallmatrix}\right]$ represents the impedance tensor.

In order to perform the spectral analysis the original time-series are subdivided into sets of segments. The procedure described below is applied in the same way in each subsequent decimation step. The decimation step involves low-pass filtering of the time-series with a recursive filter and then decimation by a factor of 2. The analysis is carried out with overlapping data segments of length N (different from N in the previous section), covering the whole time-series. The degree of overlapping ranges from zero up to 50 per cent, depending on the number of available data points.

Using $i = 0, 1, \ldots, N-1$ as the time index within a segment, the analysis proceeds as follows.

(1) Long-period trends and means are removed by a first-difference filter, yielding, in the case of e_x, the new series: $\overline{e}_x[i] = e_x[i] - e_x[i-1]$.

(2) To reduce the bias of spectral estimation each segment is tapered by a Hanning window, yielding

$$\tilde{e}_x[i] = \overline{e}_x[i]h[i]; \quad h[i] = \frac{1}{2}\left(1 - \cos\frac{2\pi \mathrm{i}}{N-1}\right). \tag{7}$$

(3) A Fourier transform of $\tilde{e}_x[i]$ is carried out, yielding the Fourier coefficients $E_x[j]$, where $j = 1, 2, \ldots, N/2$ is the frequency index.

(4) Fourier transforms X and Y, where X, Y and later Z denote any of the field components, are combined into non-smoothed auto- and cross-spectral values:

$$S^n_{XY}[j] = \frac{1}{N} X^*[j] Y^*[j], \tag{8}$$

where X^* is the complex conjugate of X.

(5) In order to remove from further processing segments with a large level of noise, coherence sorting is used. Criteria are the two partial coherences:

$$Co^2_{E_x H_y \cdot H_x} = \frac{\left|S_{E_x H_y \cdot H_x}\right|^2}{S_{E_x E_x \cdot H_x} S_{H_y H_y \cdot H_x}}, \tag{9}$$

$$Co^2_{E_y H_x \cdot H_y} = \frac{\left|S_{E_y H_x \cdot H_y}\right|^2}{S_{E_y E_y \cdot H_y} S_{H_x H_x \cdot H_y}}, \tag{10}$$

where

$$S_{XY \cdot Z} = S_{XY} - S_{ZY} S_{XZ} / S_{ZZ}. \tag{11}$$

Here S_{XY} denotes smoothed spectral values

$$S_{XY}(j) = \frac{1}{k+1} \sum_{m=-k/2}^{k/2} S^n_{XY}(j+m), \quad j = \frac{k}{2} + 1, \ldots, \frac{N}{2} - \frac{k}{2}, \tag{12}$$

i.e. smoothing is carried out over $k+1$ neighbouring frequencies, where $k \leq N/2$. The whole purpose for deriving smooth spectra is to define sorting criteria in terms of coherences. Thereafter, they are not used any further.

The partial coherences introduced above refer to orthogonal electric and magnetic components. If any of them fall below a specified threshold, different for E_x and E_y, then the respective segments will be eliminated. Tests have shown that appropriate partial coherences thresholds are well defined in terms of the partial coherences of the entire time-series. The thresholds are determined during pilot analysis of the data for each decimation step separately.

As a result of these five steps, we now have *M non-smoothed* auto- and cross-spectral values from a selected set of segments for each frequency:

$$S^n_{E_x H_x i}, S^n_{E_x H_y i}, S^n_{E_y H_x i}, S^n_{E_y H_y i}, S^n_{H_x H_y i},$$
$$S^n_{E_x E_x i}, S^n_{E_y E_y i}, S^n_{H_x H_x i}, S^n_{H_y H_y i}, \quad i = 1, 2, \ldots, M.$$

These quantities are selected for the solution of the regression problem, formulated below, instead of the original Fourier coefficients because of convenience. The result of the following procedure would, however, be the same in both cases.

2.3 Robust estimation of transfer functions

In our procedure to derive robust MT transfer functions we adopt Siegel's concept of robust estimation of the repeated median algorithm. The system of equations (eq. 6) poses a linear regression problem which, in general terms, can be written as

$$y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\Theta} + e_i, \quad i = 1, \ldots, M, \tag{13}$$

where y_i is the predicted value from the ith observation of a p-dimensional vector $\mathbf{x}_i$, e_i the ith prediction error, while $\boldsymbol{\Theta}$ represents the p-dimensional vector of unknown regression parameters to be estimated.

The least-squares solution T^{LS}_M of the thus formulated regression problem can be found by minimizing the Euclidean norm of residuals e_i:

$$\Gamma(\boldsymbol{\Theta}) = \sum_{i=1}^{M} \left[\frac{\left(y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\Theta}\right)}{\sigma} \right]^2, \tag{14}$$

where σ is a scaling parameter.

The properties of the estimator are established by the Gauss–Markov theorem. The solution is optimal in the class of unbiased estimates only if errors are distributed normally. However, the normal model can never be absolutely adequate. It is well known that the LS solution is very sensitive to outliers in the data and has a breakdown point equal to zero. Huber suggested to minimize the non-quadratic loss function:

$$\Gamma(\boldsymbol{\Theta}) = \sum_{i=1}^{M} \rho \left[\frac{\left(y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\Theta}\right)}{\sigma} \right]. \tag{15}$$

Huber's M-estimator can be derived by putting $\rho(r) = \rho_{\mathrm{c}}(r)$, where $\rho_{\mathrm{c}}(r)$ is defined by the weights $w_i = \min\{1, c/|r_i|\}$, where r_i is ith residual, c is a positive constant. It is shown that, in practice, the breakdown point of such an estimator does not exceed 30 per cent (Hampel *et al.* 1986). However, M-estimators are much less sensitive to outliers than LS estimator.

In the present approach Siegel's repeated median estimator is sought with the highest possible breakdown point equal to $\varepsilon^* =$ 50 per cent, which is expressed as follows:

$$T^{(j)}_n = \underset{i_1}{\mathrm{med}} \left\{ \ldots \left\{ \underset{i_{p-1}}{\mathrm{med}} \left\{ \underset{i_p}{\mathrm{med}} \left\{ \Theta^{(j)}(i_1, \ldots, i_p) \right\} \right\} \right\} \ldots \right\}, \tag{16}$$

where $\Theta^{(j)}(i_1, \ldots, i_p)$ is the jth component of the unknown p-dimensional vector parameter, unequivocally determined by any p observations and $i = 1, \ldots, n$ is the index of observation. The estimator is described in more detail in Appendix A.

To adopt this estimator to the MT problem, the original system of equations eq. (6) are rewritten in terms of auto- and cross-spectral densities:

$$\begin{aligned} S_{E_x H_x} &= Z_{xx} S_{H_x H_x} + Z_{xy} S_{H_y H_x} \\ S_{E_x H_y} &= Z_{xx} S_{H_x H_y} + Z_{xy} S_{H_y H_y}, \end{aligned} \tag{17}$$

$$\begin{aligned} S_{E_y H_x} &= Z_{yx} S_{H_x H_x} + Z_{yy} S_{H_y H_x} \\ S_{E_y H_y} &= Z_{yx} S_{H_x H_y} + Z_{yy} S_{H_y H_y}, \end{aligned} \tag{18}$$

where S_{XY} denotes smoothed spectral densities, although not necessarily resulting from a smoothing process identical to that of eq. (12). It is known that the smoothing procedure leads to the LS solution that is biased by the uncorrelated noise in input channels H_x, H_y. The statistics used here allows portion of data, in theory asymptotically up to 50 per cent, to be contaminated by such noise without bias to the estimator.

For example, the Z_{xy} component of the impedance tensor is estimated as

$$Z_{xy} = \frac{S_{E_x H_y} S_{H_x H_x} - S_{E_x H_x} S_{H_x H_y}}{S_{H_x H_x} S_{H_y H_y} - S_{H_x H_y} S_{H_y H_x}}. \tag{19}$$

Taking into account that from two systems, eqs (17) and (18), the estimates of the impedance tensor components are unequivocally defined by the two independent realizations of spectral values, the repeated median estimator can be derived. For only two realizations the result of using these systems is the same as if eq. (6) (based upon Fourier coefficients) was used directly.

The Siegel estimator in terms of spectral densities for the real part of the Z_{xy} component (and the imaginary part separately in an analoguous manner) is given by

$$\mathrm{Re}[Z_{xy}]_{\mathcal{S}} = \underset{i}{\mathrm{med}}\,\underset{j\neq i}{\mathrm{med}}\,\mathrm{Re}\left[\frac{S_{E_xH_yij}S_{H_xH_xij} - S_{E_xH_xij}S_{H_xH_yij}}{S_{H_xH_xij}S_{H_yH_yij} - S_{H_xH_yij}S_{H_yH_xij}}\right], \quad (20)$$

where the indices $i, j = 1, \ldots, M$ and spectral densities S_{XY} are formed by the combination of two realizations of the non-smoothed auto- and cross-spectra:

$$S_{XYij} = \left(S^n_{XYi} + S^n_{XYj}\right)/2. \quad (21)$$

The sequence of computational steps of the Siegel estimate $[\mathbf{Z}]_{\mathcal{S}}$ of the impedance tensor is as follows (for each objective frequency it is the same).

(1) The first median operator of eq. (20) is calculated as

$$\mathrm{Re}[Z_{xy}]^M_i = \underset{j\neq i}{\mathrm{med}}\,\mathrm{Re}\left[\frac{S_{E_xH_yij}S_{H_xH_xij} - S_{E_xH_xij}S_{H_xH_yij}}{S_{H_xH_xij}S_{H_yH_yij} - S_{H_xH_yij}S_{H_yH_xij}}\right], \quad (22)$$

where the median is taken over all $M-1$ values of j. Non-smoothed spectral densities are combined here in pairs, according to eq. (21). This procedure is applied for each of the M indices i, providing us with M medians. To compute all possible combinations, it is sufficient to take for each index i a paired index j, such as $j > i$, because of symmetry $S_{XYij} = S_{XYji}$.

(2) As an option, unrealistic partial estimates $[\mathbf{Z}]^M_i$ can be omitted using the phase of the off-diagonal impedance elements as criteria, since the phases should lie within the limits (assuming an $\mathrm{e}^{-i\omega t}$ dependence)

$$-90^\circ < \arg(Z_{xy}) < 0^\circ, \quad 90^\circ < \arg(Z_{yx}) < 180^\circ. \quad (23)$$

If the estimate Z_{xy} (or Z_{xy}) is omitted, the respective estimate Z_{xx} (or Z_{yy}) is omitted as well. This option is used only in the worst cases, i.e. in the cases where the original time-series are strongly contaminated by noise, because 3-D structures or static distortions may, in some cases, cause the phase to be in another quadrant.

(3) The final median operator of the Siegel estimator is calculated separately for the real and imaginary parts of each impedance tensor component:

$$[\mathbf{Z}]_{\mathcal{S}} = \underset{i}{\mathrm{med}}\left\{[\mathbf{Z}]^M_i\right\}, \quad (24)$$

where $i = 1, 2, \ldots, M$.

(4) Confidence limits are estimated from the median of absolute deviations (MAD):

$$[\mathbf{Z}]_{\mathrm{mad}} = 1.483\ \underset{k}{\mathrm{med}}\{|[\mathbf{Z}]_k - [\mathbf{Z}]_{\mathcal{S}}|\}, \quad (25)$$

where k denotes all calculated ij-combinations. This estimate of the scale parameter is very simple to calculate and is insensitive to outliers. It also has the same features as the RM estimator, used in this study. For Gaussian errors, a 95 per cent confidence limit can be defined as

$$\Delta[\mathbf{Z}] = 1.96[\mathbf{Z}]_{\mathrm{mad}}/\sqrt{M}, \quad (26)$$

where M is the number of segments involved.

(5) In order to obtain a more effective estimate for short time-series, the Siegel estimator is supplemented by a one-step reduced M-estimator involving neighbouring frequencies. The procedure is applied only when the final number of spectra, obtained after spectral analysis (with coherence sorting), is less than a specified threshold. In this study we used a threshold of 50. In the calculations, l neighbouring frequencies are used to define the final estimate for a particular central frequency. Here l is selected to have approximately six independent transfer function estimates per decade. All individual estimates $\{[\mathbf{Z}]_{ij}\}$, obtained from all ij-combinations in the previous step, for partial l frequencies, are used giving, in total, n values. Denoting the estimates of this set by $U_k (k = 1, 2, \ldots, n)$ and $U_{\mathcal{S}}$ for RM estimates $[\mathbf{Z}]_{\mathcal{S}}$ of the central frequency, where U denotes the real or imaginary parts of any of the four impedance tensor elements, the reduced M-estimator is expressed as

$$\overline{U[j]} = \sum_{k=1}^{n} U_k w_k \Big/ \sum_{k=1}^{n} w_k, \quad (27)$$

where the weights are

$$w_k = \begin{cases} 1 & \text{if } |r_k| \le c, \\ c/|r_k| & \text{if } c \le |r_k| \le b, \\ 0 & \text{if } |r_k| \ge b, \end{cases} \quad (28)$$

and the kth residual is

$$r_k = \frac{U_k - U_{\mathcal{S}}}{S_{\mathrm{mad}}}. \quad (29)$$

The constant c lies within the range 1–2 and b is within the range 4–5, while the scaling parameter S_{mad} is derived from the median of absolute deviations. This parameter is also used to derive confidence limits according to eq. (26).

3 TESTS AND EXAMPLES

In the first test, synthetic COMDAT data were processed and compared with a standard LS solution. Each magnetotelluric COMDAT data set is 7 d long and has a time sampling interval of 20 s. The spectra of components have both a regular part with a power dependence on frequency and a random part. The second test contains different types of noise, including normal random noise for each field component ranging from 4 to 45 per cent of the signal spectral amplitudes, outliers in the frequency domain, time domain outliers (pulses) with random amplitudes and duration. A more detailed description of the COMDAT data sets can be found in Ernst *et al.* (2001).

For noise-free data, the results are very close to the model values for both the LS and robust Siegel estimators, indicating that the developed robust procedure leads asymptotically to the true result. Test results with noisy synthetic data are shown in Fig. 1. The amplitude of the transfer function is fitted quite well by both estimators. The misfit for the Z_{xy} and the Z_{xx} components is greater than for Z_{yx} and Z_{yy}, which is caused by stronger contamination of the E_x component by frequency domain outliers in this test (Ernst *et al.* 2001). It is possible to distinguish problems with harmonic noise for periods of approximately 250 and 1000 s. The average misfit for Z_{yx} is only slightly larger than for the noise-free data. The fit for the corresponding phase is worse for all components. The diagonal elements of the impedance tensor are also estimated quite well. A comparison of two different techniques on the same data set is presented by Ernst *et al.* (2001), who observed similar features in the estimates. It is shown there that for the dominant off-diagonal impedances the accuracy is comparable to that for the noise-free test.

In the current realization of the algorithm 64-point segments were used to perform the spectral analysis. The disadvantage of such a

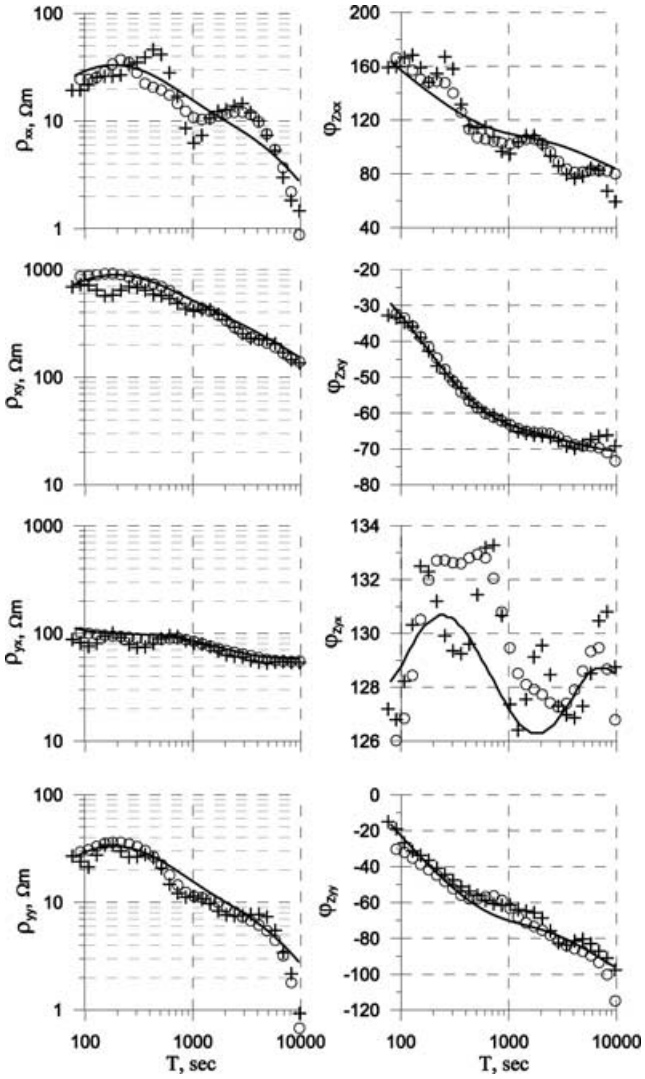

Figure 1. COMDAT data processing results. Left-hand column, amplitudes of all impedance tensor components; right-hand column, corresponding phases; solid line, true model; circles, robust estimate; crosses, LS solution.

short segment length is that the final spectral resolution is quite limited. In the case of strong harmonic noise, such as power line harmonics in audio-magnetotelluric (AMT) data, the estimates might already be biased during spectral transformation of the original data. Hence, under such circumstances an appropriate adaptation of the algorithm should be made. Short data segments are used to try to eliminate possible noise and distortions in the time domain, where contaminated segments will be removed based on coherence sorting and subsequent statistical analysis.

The problems with frequency outliers, as mentioned above, are most probably caused by insufficient spectral resolution in the Fourier analysis. In general, the robust procedure produces more stable results with a smaller average misfit for all components, but the difference in estimates for this test is not large.

The real MT data used in this study were acquired in Russian Karelia by the MT group of St Petersburg University. The duration of the time-series is approximately 3 d. Measurements were carried out in two frequency bands (4 Hz, 50 s and 30 s, DC). In the second band data were recorded continuously, while in the first band recording was done in segments of 2048 points each with coherence sorting.

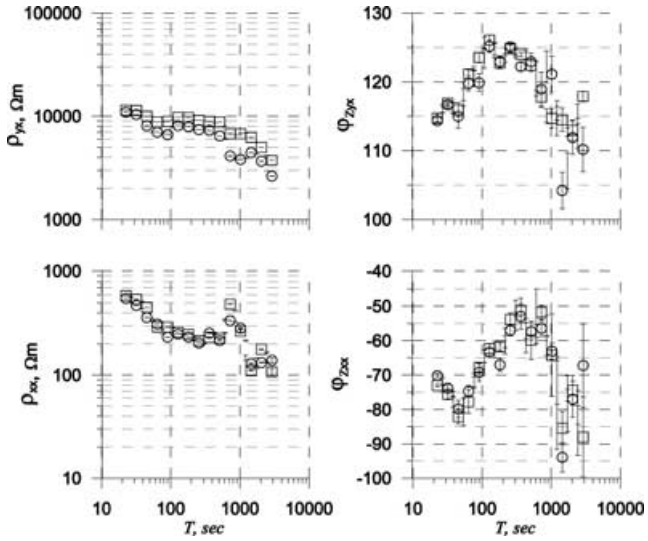

Figure 2. Processing results of real MT data from a 'noise-free' site. Z_{yx} and Z_{xx} are the impedance tensor components estimated using the LS solution and robust procedure for the 'noise-free' site. The site is located far from sources of industrial noise. Estimates agree well. Circles, LS estimate with coherence sorting; squares, robust estimate.

Only segments with average coherence exceeding some threshold were stored for further processing.

The data from the first site were collected quite far from the sources of industrial noise in the central part of Karelia, close to the southern part of Lake Topozero. Another site was approximately 50 km to the south from the first one. Both sites are located in highly resistive Archaean crust with no sedimentary cover. Sites are located at approximately latitude 65.5°, where the source field may have a complicated non-plane-wave structure caused by the closeness of the auroral zone. The second site was clearly contaminated by noise, caused by a power line approximately 10 km away.

The comparison of the present robust procedure and the LS method for the first 'noise-free' site is presented in Fig. 2, where the amplitudes and phases of impedance elements Z_{yx} and Z_{xx} are shown. The other two components (not shown) have a smaller average misfit between the LS and robust solutions. The processing results using the LS solution and the Siegel estimator coincide quite well. Distortions at the long-period tails of the apparent resistivities (approximately 1000 s) might be caused by non-uniformity of the source field. Similarly, the long-period phases have a larger misfit than the short-period ones, which is also caused by the same effects. The results indicate that even for good quality data, the robust processing produces more stable results than the LS solution. Thus the robust estimation in use is quite effective for this type of data.

In Fig. 3 processing results of the data from the noisy site are shown. Several high-amplitude noisy events are present in the original time-series. The LS estimate of resistivity is obviously downward biased. The corresponding impedance phase is also distorted. This may be explained by the presence of coherent events in the original time-series that are not eliminated by the coherence sorting procedure. These distortions of the amplitude of the transfer functions might be observed when significant uncorrelated noise is present in input channels, because, in this case, auto-spectral densities are biased. However, the phases are usually not distorted. The distortion of the phase here is an indication of correlated noise and it applies, in particular, to short periods around 10–100 s, where the noise is

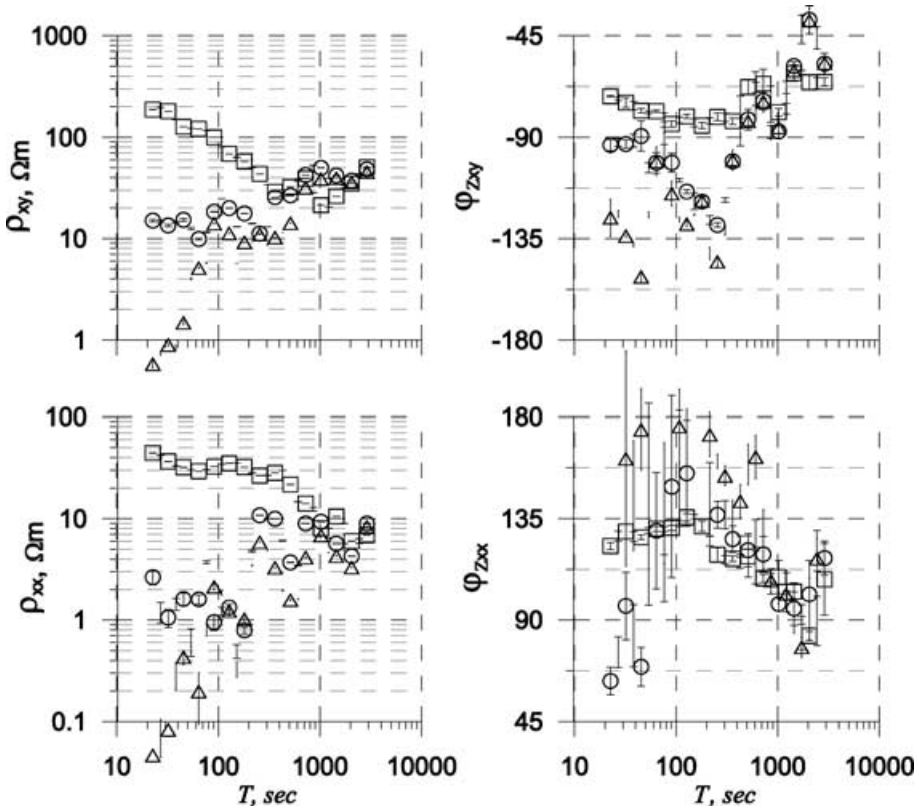

Figure 3. Processing result of real data from the 'noisy' site. The LS estimate is obviously downward biased, but the robust procedure still produces a reliable estimate. Triangles, LS solution without coherence sorting; circles, LS estimate with coherence sorting; squares, robust estimate.

concentrated. At longer periods the LS and robust estimators produce more or less similar results.

4 CONCLUSIONS

A new technique for magnetotelluric data processing has been developed, using a robust estimation procedure with the highest breakdown point. The procedure is based on the estimator suggested by Siegel, calculated using an algorithm of repeated medians. The program has been tested and compared with the standard LS solution.

Tests using the synthetic COMDAT data set show that the algorithm gives reasonably stable results, although for this test the difference between the LS solution and the robust estimate presented here is not very large. Estimates for this noise-free synthetic data set agree very well. Some problems with noisy synthetic data are observed for the Z_{xx} and Z_{xy} impedance tensor components, caused by the larger contamination of the E_x field by frequency domain outliers. The problem might be explained by the insufficient spectral resolution of the FFT, because short data segments were used, thereby frequency domain outliers have a stronger influence on the result. Time domain outliers were successfully eliminated by this technique. The robust procedure results in only a slightly smaller misfit from the true model than the LS solution.

The method has also been tested using real 'noise-free' and 'noisy' MT data. The results for the 'noise-free' site agree well for both the LS and robust methods. Agreement is good over the whole period range of interest. The advantage of the new algorithm is shown when the data from a 'noisy' site are processed. The data were contaminated by correlated noise, and, by uncorrelated time domain pulses. Coherence sorting helped to improve the data quality. The following robust procedure then removed most of the effects of the remaining noise, thus, providing a realistic estimate of the impedance tensor.

ACKNOWLEDGMENTS

I would like to thank Aida Kovtun and Stanislav Vagin for discussions on the material used in this paper. In particular, I would like to thank Ulrich Schmucker for his help with the manuscript. I also thank Laust Pedersen and Toivo Korja for useful comments and advice.

REFERENCES

Chave, A.D., Thomson, D.J. & Ander, M.E., 1987. On the robust estimation of power spectra, coherences and transfer functions, *J. geophys. Res.,* **92,** 633–648.

Chave, A.D. & Thomson, D.J., 1989. Some comments on magnetotelluric response function estimation, *J. geophys. Res.,* **94,** 14 215–14 225.

Egbert, G.D. & Booker, J.R., 1986. Robust estimation of geomagnetic transfer functions, *Geophys. J. R. astr. Soc.,* **87,** 173–194.

Egbert, G.D. & Livelybrooks, D.W., 1996. Single station magnetotelluric impedance estimation: coherence weighting and regression *M*-estimation, *Geophysics,* **61,** 964–970.

Ernst, T., Sokolova, E. Yu., Varentsov, I.M. & Golubev N.G., 2001. Comparison of two techniques for magnetotelluric data processing using synthetic data sets, *Acta Geophys. Polonica,* **XLIX,** 213–243.

Hampel, R.F., Ronchetti, M.E., Rousseeuw, P.J. & Stahel, W.A., 1986. *Robust Statistics. The Approach Based on Influence Functions,* Wiley, New York.

Huber, P.J., 1981. *Robust Statistics,* Wiley, New York.

Jones, A.G., Chave, A.D., Egbert, G.D., Auld, D. & Bahr, K., 1989. A comparison of techniques for magnetotelluric response function estimation, *J. geophys. Res.,* **94,** 14 201–14 213.

Larsen, J.C., Mackie, R.L., Manzella, A., Fiordelisi, A. & Rieven, S., 1996. Robust smooth magnetotelluric transfer functions, *Geophys. J. Int.* **124,** 801–819.

Marple, S.L. Jr., 1987. *Digital Spectral Analysis with applications,* Prentice-Hall, New Jersey.

Siegel, A.F., 1982. Robust regression using repeated medians, *Biometrika,* **69,** 242–244.

Stein, A. & Werman, M., 1992. Finding the repeated median regression line, *Proc. 3rd ACM-SIAM Symp.,* pp. 409–413, SIAM, Philadelphia, PA

APPENDIX A: SIEGEL'S REPEATED MEDIAN ESTIMATOR

Let us consider the p-component vector parameter $\mathbf{\Theta}(i_1, \ldots, i_p)$ to be estimated, which is unequivocally determined by any p observations $(\mathbf{x}_i, y_i) \ldots, (\mathbf{x}_{i_p}, y_{i_p})$. Siegel's repeated median estimator of a set of n observations $(\mathbf{x}_i, y_i) \ldots, (\mathbf{x}_{i_n}, y_{i_n})$ is defined as follows. The jth component of $\mathbf{\Theta}$ is

$$T_n^{(j)} = \underset{i_1}{\text{med}} \left\{ \ldots \left\{ \underset{i_{p-1}}{\text{med}} \left\{ \underset{i_p}{\text{med}} \left\{ \mathbf{\Theta}^{(j)}(i_1, \ldots, i_p) \right\} \right\} \right\} \ldots \right\}, \tag{A1}$$

where the median is taken over all indices $i_m = 1, \ldots, n$.

It is helpful to consider the simple linear regression model $y_i = \Theta_1 + \Theta_2 x_i + e_i$, to explain the estimator in detail. For each point (x_i, y_i), let us denote the median Θ_{2i} of the $n - 1$ slopes of the lines passing through this point and each other point of the set. The repeated median slope estimate Θ_2^* is defined to be the median of the multiset $\{\Theta_{2i}\}$:

$$\Theta_2^* = \underset{i}{\text{med}} \underset{j \neq i}{\text{med}} \frac{y_i - y_j}{x_i - x_j}. \tag{A2}$$

The intercept Θ_1 can be estimated then either separately from Θ_2, as

$$\Theta_1^* = \underset{i}{\text{med}} \underset{j \neq i}{\text{med}} \frac{y_i x_j - y_j x_i}{x_j - x_i}, \tag{A3}$$

or else hierarchically, as $\Theta_1^* = \text{med}_i \{y_i - \Theta_1^* x_i\}$.

In the bivariate linear regression model, that is used to solve the impedance linear system

$$y_i = \Theta_1 x_{1i} + \Theta_2 x_{2i} + e_i, \quad \text{where } p = 2 \tag{A4}$$

the repeated median estimate is determined in the same way. If the unknown parameter is a complex vector, then the equation is split into two independent equations for real and imaginary parts that are solved separately. The components of the vector parameter $\mathbf{\Theta}$ are estimated separately, for instance, for Θ_1 we have

$$\Theta_1^* = \underset{i}{\text{med}} \underset{j \neq i}{\text{med}} \frac{y_i x_{2j} - y_j x_{2i}}{x_{1i} x_{2j} - x_{1j} x_{2i}}. \tag{A5}$$

It means that for each ith observation first the median of combinations with all j observations is calculated and then finally the median of those $n - 1$ medians form the final estimation.

There are several algorithms to define the Siegel repeated median estimator faster than the brute method required $O(n^2)$ time, where n is the number of given points (for a simple line estimation). Some of them reduce the calculation time to $O(n \log^2 n)$, such as the randomized algorithm, however, they rely on sophisticated data structures, which make them quite difficult to program (Stein & Werman 1992).

Here, special algorithms were not applied to accelerate the calculations, because the required computer time was comparable or even less than the time needed for Fourier transformation of the original time-series.