S1–04

MT Data Processing with Bootstrapping Followed by Constrained Smoothness Optimization

Kim Kang Sop[*], *Jo Kyong Rae*

Faculty of Resources Exploration Engineering, Kim Chaek University of Technology, Pyongyang, DPRK

[*]Corresponding Author: Email: kgs6673@star–co. net. kp

Summary

In Magnetotelluric (MT) method which generally handles noisy data, data processing is a key–procedure. A robust method with remote reference is in standard use. However, there are often situations, where no method yields useable transfer functions (TF) due to ultimately high noise level.

For such situation, we propose a new method, based on the bootstrap inference that yields many candidate estimates for MT TFs and their errorbars immediately via resampling, and on their constrained optimization for picking one smoothly varying with period and ranging within feasible region. We show field illustrations, using long–period MT datasets from EMSLAB and SAMTEX projects and MT data in Korean peninsula. We conclude that the method extracts usable impedance estimates even from heavily contaminated MT data objectively without making fabrication.

Keywords: Magnetotellurics; Transfer function; Bootstrap; Constrained Optimization;

1. Introduction

The MT deep sounding method, which is based on the measurement of natural electromagnetic field, is inevitably affected by various natural and cultural noises. Especially, sources of cultural noise such as powerline, DC electrified railway, are growing with the industrialization and urbanization. Hence, data processing is historically regarded as a paramount issue of MT method.

The traditional method for TF estimation was least squares method, which is based on Gaussian distribution assumption of measurement noise [2]. Later, robust methods were introduced in order to suppress non–Gaussian noise and downweight outliers [4, 10]. Presently, there exists many variants of robust technique, such as the bounded influence estimator [6], the robust smooth estimator [13], and the robust method for AMT dead band [11], etc.; robust methods with remote reference become standard practice in MT data processing.

Meanwhile, beyond the traditional statistics, non–parametric methods begin to be employed in MT data processing. In particular, jackknife method based on the delete–one resampling is widely used for estimating errorbar of MT TF [6, 11]. Also, boothstrap method, another non–parametirc inference, has been used to compute errorbars of MT TFs by the ELICIT–RR method [3, 7]. However, estimation of TF itself by non–parametric inference have scarcely been discussed in our view.

Now, there have been reported many failure to obtain interpretable TFs due to high noise level. In such cases, a rule of thumb is that "if in doubt throw it out [14] " or, to "put such data in trash [2]". However, it is often requested to inevitably extract useful TFs from given data as much as possible.

We propose a method to address this problem. The method is based on the bootstrap resampling to yield many candidate estimates of TFs from given data, followed by solving constrained smoothness optimization with resulted candidates. The method will be validated through some field tests, using MT datasets collected in Korean peninsula and long–period MT datasets from EMSLAB and SAMTEX projects.

2. Method

2.1. Main idea

MT data processing involves the frequency–by–frequency estimation of impedance tensor, satisfying the linear regressive model:

$$\boldsymbol{E}=\boldsymbol{Z}\boldsymbol{H}+\boldsymbol{r} \tag{1}$$

where $\boldsymbol{E}$ and $\boldsymbol{H}$ are respectively matrix of x– and y– components of measured electrical and magnetic fields, and $\boldsymbol{Z}$, $\boldsymbol{r}$ are respectively vectors of impedance and random error (or residual).

Conventional robust methods give point–estimators of impedance, assuming a certain statistical distribution of measurement error. For example, the iterative reweighted least squares method (IRLS), as a standard robust estimator, repeatedly calculates the following estimates ([5])

$$\boldsymbol{Z}(\boldsymbol{E},\ \boldsymbol{H})=(\boldsymbol{H}^{*}\boldsymbol{W}\boldsymbol{H})^{-1}(\boldsymbol{H}^{*}\boldsymbol{W}\boldsymbol{E}) \tag{2}$$

where $\boldsymbol{W}$ is a diagonal matrix composed of weights for measurements, elements of which are function of residual. Traditionally, Gaussian, Student's t–, and Fisher distributions were employed to approximate unknown statistical property of measurement error in MT data processing; [7] recently discussed that MT data are pervasively described by the alpha stable distribution family. Namely, the statistical features of MT data and its inherent noise is generally undetermined.

Beyond the conventional methods, the basic idea underlying non–parametric method is to infer unknown quantity by maximally using given data through resampling; while making as few assumptions as possible. Non–parametric inference includes jackknife method of Quenouille (1949), permutation method of Fisher and Pitman (1930), and bootstrap method of Efron (1979), etc [8]. Such methods generally enable to estimate confidence intervals without any assumption about the statistical distribution of measured data. However, we focus on the other possibility that nonparametric methods can yield many candidate estimates of the quantity of interests by resampling (replication) given data.

We suppose that X={X_1, ⋯, X_N} is a given dataset, measured at random from a completely unspecified distribution F. In MT data processing, we can compose the dataset $\boldsymbol{X}$ from measured power spectral data of electrical and magnetic components as follows:

$$\boldsymbol{X}_i = \left[\overline{E_xE_x} \quad \overline{E_xE_y} \quad \overline{E_xH_x} \quad \overline{E_xH_y} \quad \cdots \quad \right]_i^T, \ i = 1, \cdots, N \quad (3)$$

When we say "at random" we mean that the X_is are independent and identically distributed (iid) random variables, each described by distribution F. Let $\boldsymbol{Z}$=$\boldsymbol{Z}$($\boldsymbol{X}_1$⋯$\boldsymbol{X}_N$) denotes an unknown parameter of F, which is the impedance in Eq. (2) in our problem.

The jackknife method, the frequently used in MT data processing, is based on the delete–one resampling of original data. So–called delete–one resampling systematically leaves out one data at a time from given dataset; repeating this resampling for $i = 1, \cdots, N$, produces N subsets X_{-i}={X_1, ⋯, X_{i-1}, X_{i+1}, ⋯, X_N} of size N–1. Thus, we can obtain N delete–one estimates from N subsets $\hat{\boldsymbol{Z}}_{-i} = \boldsymbol{Z}(\boldsymbol{X}_{-i}),\ i = 1, \cdots, N$

The bootstrap method is the other nonparametric inference, based on resampling with replacements of original data. In the former, the resampling number equals to the number of given samples; whereas in the latter, the resampling number can be chosen arbitrarily large.

Here, we suppose that all of resampling estimates are candidates of its true value, amongst which there exists an optimal one closest to the true value. In this sense, the latter can result more candidate estimates of TFs with higher freedom than the former; it probably involves a candidate closer to its true TF than the former.

On the other hand, the jackknife estimate of the variance of the median is not consistent, but its bootstrap estimate is consistent (Theorem 3.7 of Efron, 1982, [17]). So the bootstrap may result in exact confidence interval than jackknife in MT data processing. These are main idea of this paper.

2.2. Bootstrap Algorithm

It is well known that the bootstrap has been applied to a much wider class of problems than just the estimation of confidence intervals, such as discriminant analysis, regression, cluster analysis, kriging, and time series analysis.

The bootstrap method repeats the experiment a sufficient number and approximates as its empirical distribution, which puts probability 1/N at each data. This is equivalent to generate new datasets of size N with arbitrary replacement of data from the original set. This is called resampling with replacement; and the empirical distribution, based on such resampling is called bootstrap distribution, which converges to the true distribution if N increases sufficiently. Then almost any parameter of the bootstrap distribution can be used as a "bootstrap" estimate of the corresponding population parameter.

Based on algorithm for bootstrap variance estimation [18], we write an algorithm for estimating MT TFs and their variances as follows:

Step 1: Conduct the experiment and collect the random data into the dataset X={X_1 ⋯ X_N}

Step 2: Draw a random dataset of size N– $\boldsymbol{X}^* = \left\{X_1^*, \cdots, X_N^*\right\}$, with replacement, from $\boldsymbol{X}$.

Step 3: Evaluate the bootstrap estimate

$$\hat{\boldsymbol{Z}}^* = \boldsymbol{Z}(X_1^*, \cdots, X_N^*) \quad (4)$$

, based on the Eq. (2) and (3).

Step 4: Repeat Steps 2 and 3 many times to obtain a total of B bootstrap estimates $\hat{\boldsymbol{Z}}_1^*, \cdots, \hat{\boldsymbol{Z}}_B^*$.

Step 5: Estimate the mean and variance, respectively by

$$\hat{\boldsymbol{Z}}_{boot} = \frac{1}{B}\sum_i^B \hat{\boldsymbol{Z}}_i^* \quad (5)$$

$$\hat{\boldsymbol{\sigma}}_{boot}^2 = \frac{1}{B-1}\sum_i^B \left(\hat{\boldsymbol{Z}}_i^* - \hat{\boldsymbol{Z}}_{boot}\right)^2 \quad (6)$$

Step 6: Calculate the percentile confidence interval as

$$\left[\boldsymbol{Z}_{(B\alpha/2)}^*, \boldsymbol{Z}_{(B(1-\alpha)/2)}^*\right] \quad (7)$$

where α/2 and 1–α/2 are percentiles of bootstrap estimates $\hat{\boldsymbol{Z}}_1^*, \cdots, \hat{\boldsymbol{Z}}_B^*$, after sorted in ascending order.

2.3. Constrained Smoothness Optimization

Here, the problem is how to construct an optimal TF curve from bootstrap candidates $\hat{\boldsymbol{Z}}_1^*, \cdots, \hat{\boldsymbol{Z}}_B^*$, given for each frequency. For such an optimal curve, we suppose the following assumptions:

i) The optimal curve needs to possess a reasonable smoothness in the sense of norms of 1–st and 2–nd derivatives. Probably, this may become a prerequisite to obtain the smoothest model in the inversion of resulted MT TFs [9].

ii) The optimal curve is necessarily close to the standard robust estimates Eq. (2).

iii) The optimal curve should pass between the minima and maxima of given candidate estimates.

Hence, the problem of determining such an optimal curve can be established as the constrained optimization of a function with variables equal to the frequencies. We can specify the problem as

$$\underset{\hat{\boldsymbol{Z}}}{Mnimize}\left(\alpha_1\left|\hat{\boldsymbol{Z}}'\right|^2 + \alpha_2\left|\hat{\boldsymbol{Z}}''\right|^2 + \left|\boldsymbol{Z}^{\boldsymbol{0}} - \hat{\boldsymbol{Z}}\right|^2\right) \quad (8.1)$$

$$subject.\ to.\ \boldsymbol{Z}_{\min} \le \hat{\boldsymbol{Z}} \le \boldsymbol{Z}_{\max} \quad (8.2)$$

where

$$\boldsymbol{Z}^0 = \{Z^0(i)\}_{i=1}^N,\ \boldsymbol{Z}_{\min} = \left\{\min_j Z(i,\ j)\right\}_{i=1}^N$$

$$\boldsymbol{Z}_{\max} = \left\{\max_j Z(i,\ j)\right\}_{i=1}^N \quad (9)$$

respectively represent the means, minima and maxima of bootstrap candidate estimates at each frequencies. α_1, α_2 are respectively coefficients of importance of smoothness terms, represented by 1st and 2nd derivative norms (fixed as α_1=1, α_2=1 in this paper); the third term reflects the closeness of the optimal curve to the standard robust estimate; Eq. (8.2) represents the bound constraints for the solution.

We solve the optimization problem of Eq. (8) by the following algorithm:

i) Determine sets of the means, minima and maxima in Eq. (9).

ii) Obtain an optimal curve by starting at initial value $\boldsymbol{Z}^0$ to search a solution $\hat{\boldsymbol{Z}}$ which minimizes Eq. (8.1), subject to the inequality constraints (8.2). Here, we use the interior–point algorithm [1].

iii) Pick out candidates, nearest to the optimal ones frequency by frequency. We call the set of selected candidates as suboptimal curve. This will converge to the optimal curve, when the resampling number increases sufficiently.

Although this method is implemented via optimization algorithm, it is essentially 'a picking–out procedure' from given candidate estimates. Hence, it is strictly distinguished from polynomial smoothing which can produce distortion of original information of TFs.

3. Examples

3.1. EMSLAB Data

First, we demonstrate the proposed method for long–period MT data at site EMS02 (44°53.5'N, 123°46.5'W) from EMSLAB project [2], which is freely available from Mtnet homepage. According to the data explanation, data quality are bad at longer periods, particularly in the E_x channels [5].

We show the results of bootstrap optimization in Figure 1. Here, candidate TFs are denoted as points, their means as triangles, and optimal and suboptimal TFs as black lines and triangles. Bootstrap errorbars and the corresponding theoretical TFs resulted by ρ+ algorithm [16] are shown in Figure 1 (C). For comparison, each panel shows the TFs (black circles), resulted by using LiMS (Long–period Intelligent Magnetotelluric System processing) code with remotely referenced to site EMS01, too.

As shown from Fig. 1(A), bootstrap candidates have wide distribution, mean TFs show oscillation. However, optimal and suboptimal TFs show good consistency with the LiMS–reference results and the corresponding theoretical TFs. Such improvements is possibly achieved by the fact that while bootstrap method has higher freedom in generating candidates. This example shows that the bootstrap optimization has revealed sufficiently useable TFs.

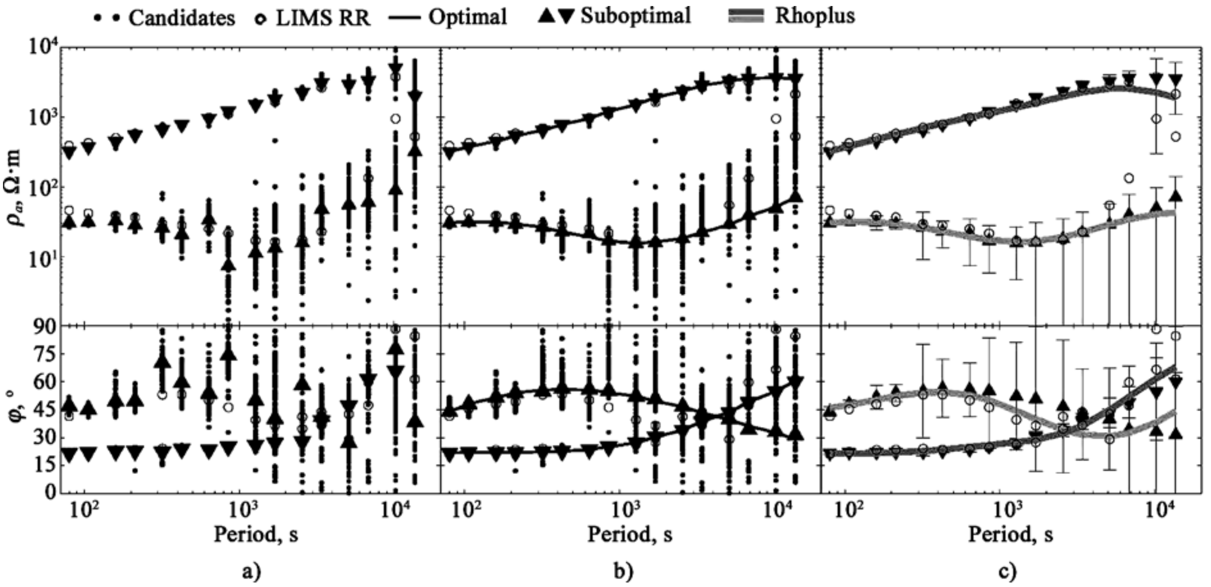

Figure 1. Bootstrap optimization at site EMS02

a) candidate estimates (points) and their means (triangles), b) optimal (lines) and suboptimal TFs (triangles) via constrained smoothness optimization, c) suboptimal TFs (triangles) and the corresponding ρ+ theoretical curves (lines); On each panel, black circles show the results of LiMS code remotely referenced to site EMS01

3.2. SAMTEX Data

Next, we validate the method for the SAMTEX long–period data, available from MTnet, too [12]. We select a site KAP145 (26°.19′05″S, 26°05′02″E), heavily contaminated by DC electrified railway and diamond mines.

Fig. 2 has similar structure with Figure 1, but each panel contains the TFs, offered in given EDI file. As shown, EDI file does not provide TFs at periods about 2 000~7 000s due to high noise level. Furthermore, when comparing with our results, it gives slightly upshifted *xy*– apparent resistivity and downshifted *yx*–one, the shifts increase with periods. Although bootstrap candidates are heavily scattered, the optimal and suboptimal TFs show smooth behavior with the change of period. Also, suboptimal TFs have good consistency to the corresponding ρ+ theoretical ones, excluding phase curves at longest periods.

Overall, we achieve our goal to have usable TFs, more improved than those offered in EDI data.

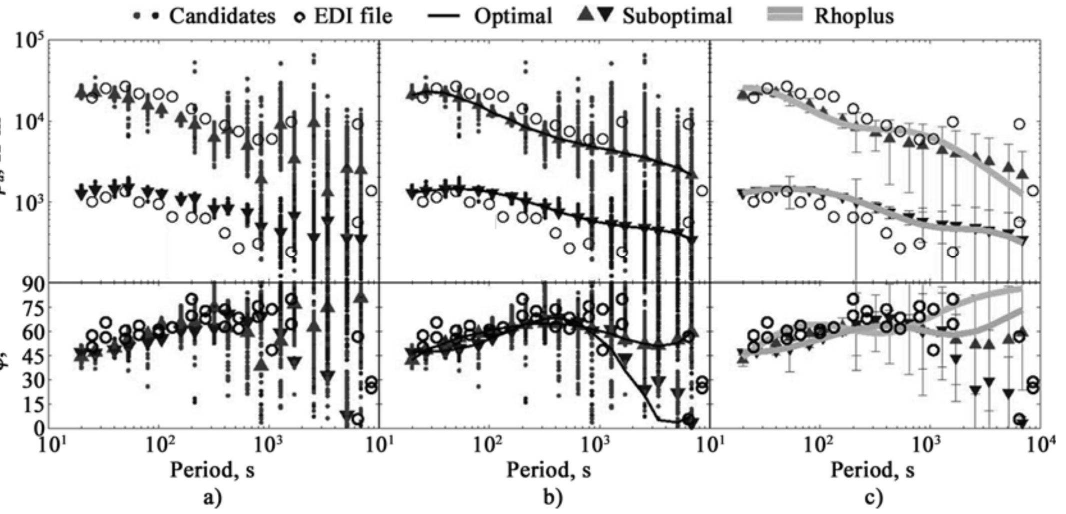

Figure 2. Bootstrap optimization at site KAP145 of SAMTEX project
a) candidate estimates (point) and their means (triangles), b) optimal (lines) and suboptimal (triangles) TFs via constrained smoothness optimization, c) comparison of suboptimal (triangles) TFs and the corresponding ρ+ theoretical curves (thick lines)
On each panel, black circles show the TFs, given in the EDI file

3.3. MT Data in Korean Peninsula

Finally, we use MT Data at site 04–05 (41°00′54″N, 128°07′17″E), North–Eastern part, Korean Peninsula.

Fig. 3 shows the TFs, resulted by the standard robust method (Fig. 3(A)) in Eq. (2) and the proposed bootstrap optimization (Fig. 3. b). As shown, apparent resistivity curves are likely to have HK–type, but they are heavily distorted by 50Hz powerline noise and particularly at longer period than 1s by low frequency noise.

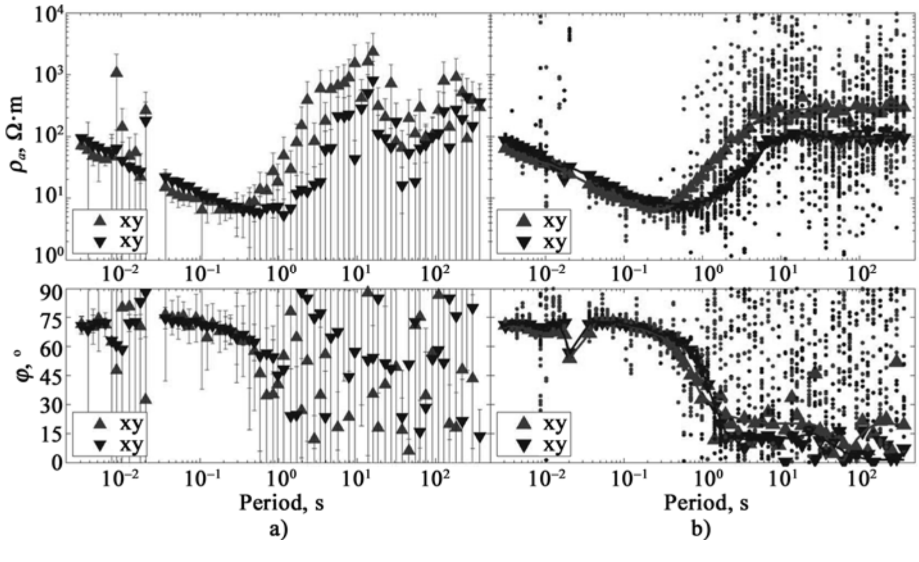

Figure 3. Processing of MT data at site 04–05 in north–eastern part, Korean Peninsula
a) standard robust method, b) bootstrap smoothness optimization (points–candidates, lines–optimal estimates, triangles–suboptimal estimates)

This area is apart as far as more than 70km from the nearest DC electrified railway, which is oriented along NW–SE direction. Therefore, rise of apparent resistivity with 45° slope, and phase close to zero at longer than 1 s periods, may be caused due to the influence of strong ground motion by wind and microseism, rather than so–called correlated noise concerned with railway [15].

Principally, one needs to throw out such bad data and re–measure new data, after relocating remote sites. Although three times of such attempts have been made, we have not achieved essential improvements of data quality, so that we inevitably reuse the original data.

In constrained smoothness optimization, thus, we omit the third term from the objective function in Eq. (8.1) with considering the worst robust estimates in Figure 1(A), instead, involve an inequality constraint into Eq. (8.2), in order to restrict the 45° rising of apparent resistivity curves.

As shown in Fig. 3. b, the ranges of bootstrap candidates of apparent resistivity are relatively narrow at periods shorter than 1s; in contrast, at longer periods, they are very widely scattered over two decades in logarithmic scale. Nevertheless, optimal and suboptimal TFs, resulted from bootstrap candidates, yield considerably smooth curves of apparent resistivity and phase. Especially, the main trend of phase curves is well recovered at long periods.

In conclusion, we have achieved our goal to yield usable TFs even from extremely bad data.

4. Conclusion

The proposed method can yield useable and interpretable TFs without producing over−smoothing distortion even from extremely contaminated data and can be used for common MT data and long period MT data.

However, we need to be careful in the practical use. The smoothest curve maybe not the best one reflecting true subsurface structure in any case. Also, if all of given data were noisy for some frequencies, this method cannot extract good TFs at the frequencies. Anyhow, the method can be used in situations when one unavoidably offer the interpretation result from given bad data.

The Earth environment is urbanized steadily and the condition for MT observation becomes more and more arduous continuously, the proposed idea can give a powerful alternative for data processing along with the existing standard techniques.

Acknowledgements

We thank all the editors and reviewers, who helped to raise the quality of our paper. Also, we truthfully thank Prof. Alan, G. Jones for offering EMSLAB and SAMTEX datasets, and LiMS processing code from the MTnet homepage

References

1. Berdichevsky M. N., V., I., Dmitriev, 2008. Models and Methods of Magnetotellurics, pp. 37−44: 508~540, Springer.

2. Campanyà J, J., Ledo, P., Queralt, A., Marcuello, A. G., Jones, 2014. A new methodology to estimate magnetotelluric (MT) tensor relationships: Estimation of Local transfer−functions by Combining Interstation Transfer−functions (ELICIT): *Geophys. J. Int.*, 198, 484−494

3. Chave A. D., 2014, Magnetotelluric data, stable distributions and impropriety: an existential combination: *Geophys. J. Int.*, 198, 622−636

4. Chave A. D., D. J., Thomson, 1989, Some comments on magnetotelluric response function estimation: *J. Geophys. Res.*, 94 (B2), 14 215−14 226.

5. Chave, A. D., D. J., Thomson, 2004. Bounded influence magnetotelluric response function estimation: *Geophys. J. Int.*, *157*, 988−1 006.

6. Chave A. D., D. J., Thomson, M. E., Ander, 1987. On the robust estimation of power spectra, coherences, and transfer functions, *J. Geophys. Res.*, 92(B1), 633−648.

7. Chernick M., R., 2008, Bootstrap Methods: A Guide for Practitioners and Researchers, 388, John Wiley & Sons, Inc.

8. Constable S. C., R. L., Parker, C. G., Constable, 1987, Occam's Inversion: A practical algorithm for generating smooth models from electromagnetic sounding data: Geophysics, 52(3): 289−300.

9. Egbert G. D., J. R., Booker, 1986. Robust estimation of geomagnetic transfer functions, *Geophys. J. R. astron. Soc.*, 87, 173−194.

10. Garcia X., A. G., Jones, 2008. Robust processing of magnetotelluric data in the AMT dead band using the continuous wavelet transform: Geophysics, 73(6), F223−F234.

11. Khoza, D., A. G., Jones, M.R., Muller, R.L., Evans, S.J., Webb, M., Miensopust, the SAMTEX team, 2013. Tectonic model of the Limpopo belt: Precambrian Research, 226, 143−156.

12. Larsen J. C., R. L., Mackie, A., Mandella, A., Fiordelisi, S., Rieven, 1996. Robust smooth magnetotelluric transfer functions: *Geophys. J. Int.*, 124, 801−819.

13. Manoj C., N. Nagarajan, 2003, The application of artificial neural networks to magnetotelluric time−series analysis. *Geophys. J. Int.*, 153, 409−423.

14. Neska A. J., Nesk, M., Sumaruk, Y., 2013, On the influence of DC railway noise on variation data from Belsk and Lviv geomagnetic observatories, Acta Geophys., 61(2), 385−403.

15. Parker R. L., J. R., Booker, 1996. Optimal one−dimensional inversion and bounding of magnetotelluric apparent resistivity and phase measurements: *Physics of the Earth and Planetary Interiors*, 98, 269−282.

16. Wasserman L., 2006, All of Non−parametric inference, Carnegie Mellon University, 271, Springer.

17. Zoubir A., M., D. R., Iskander, 2004, Bootstrap Techniques for Signal Processing, 233, Cambridge University Press.