Smart data selection - First insights from using machine learning for controlled-source RMT data processing

Anna Platz[1], Ute Weckmann[1, 2], and Cedric Patzer[3]

[1] *Helmholtz Centre Potsdam – German Research Centre for Geosciences (GFZ), Potsdam Germany*
[2] *University of Potsdam, Institute of Geosciences, Potsdam, Germany*
[3] *Geological Survey of Finland, Espoo, Finland; formerly GFZ*

1 Introduction

The Radio-Magnetotelluric (RMT) method is a geophysical near-surface imaging technique with a broad range of possible applications. In 2020, the German Research Centre for Geosciences (GFZ) has acquired a newly developed horizontal magnetic dipole transmitter that allows the usage of the RMT method even in regions with an insufficient coverage of radio transmitters which normally serve as source signal. First controlled-source RMT measurements were conducted in Chile in 2020. Further measurements were conducted in Ireland end of 2022. As we are able to store the raw time series, we have full control over the subsequent data processing. The processing tools at GFZ consist of the modular processing suite EMERALD (Ritter et al., 1998; Weckmann et al., 2005), which was originally designed for magnetotelluric (MT) processing, but has recently been adapted for RMT data. One main difference is that in RMT the transmitter data is considered as signal, while in natural source MT this would be regarded as (near-field) electromagnetic noise that needs to be removed using automated robust statistical approaches. However, processing the entire time series in an automated manner has a large drawback: The different emitted frequencies are transmitted in a sweep implying that only a smaller fraction of the time series contains the required signal for a particular target frequency and leading to an unfavourable signal-to-noise ratio. Since the transmitter does not have a GPS time base, synchronising the data logger and the transmitter with an accuracy of a few nanoseconds required for an automated detection scheme is difficult. Usually, several Gigabytes of raw time series are collected during field measurements, making manual editing and supervision of the time series virtually impossible. However, a careful selection of appropriate time segments is essential for the success of the data processing. To address the challenge, machine learning algorithms have a high potential to solve both problems. So far, we can only use the RMT data from Chile for the training of a suitable machine learning algorithm. Initial experience was gained with a recurrent neural network approach in order to identify suitable time segments (Patzer & Weckmann, EMTF 2021 – conference contribution and personal communication). This small feasibility study demonstrated, that machine learning algorithms are in general suitable for detection of transmitter signal in RMT time series. However, many questions remained open, e.g. the amount of training data which is necessary for a sufficient training of the machine learning algorithms and

if a trained algorithm is applicable to new data sets. We investigated both points by (i) increasing the amount of training data significantly and (ii) applying the trained network to new data measured in Ireland.

2 RMT measurements in Chile

In 2020, controlled-source RMT data were measured at three different locations in Chile (see Fig. 1) in the framework of the DFG funded EarthShape project.

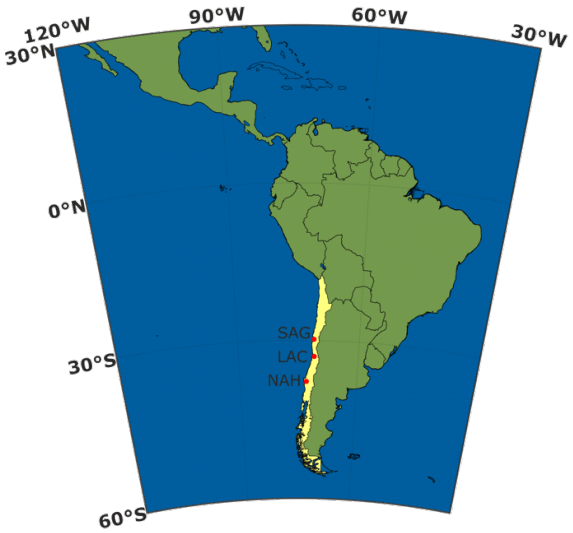

Figure 1: Map from South America displaying the three measuring locations of the controlled-source RMT field experiment in Chile in 2020. SAG = Santa Gracia; LAC = La Campana; NAH = Nahuelbuta

In total, RMT data were recorded at 166 stations along six profiles (two profiles in Santa Gracia with 34 and 32 stations, three profiles in Nahuelbuta with 33, 29 and seven stations and one profile with 31 stations in La Campana). At most stations data were recorded twice for 10 s with a sampling rate of 524 kHz using a Metronix ADU-08 data logger, the Metronix high frequency induction coil triple SHFT-02 and four spear electrodes from the Geophysical Instrument Pool Potsdam (GIPP). The horizontal magnetic dipole transmitter (see Fig. 2) was used to continuously emit eight different frequencies between 1 kHz and 128 kHz in a sweep, whereas single each frequency was emitted for 0.2 s per sweep. The two loops of the bidirectional coil antenna emit the different frequencies independent of each other.
The data were processed using the EMERALD processing suite (Ritter et al., 1998; Weckmann et al., 2005). It was originally designed for MT processing, but has recently been adapted to process RMT data. However, processing the entire time series in an automated manner leads to poor quality transfer functions due to the poor signal-to-noise ratio. As the different emitted frequencies are transmitted in a sweep, only a smaller fraction of the entire time series contains the required signal for a particular target frequency. Since it is technically very difficult to have the same time base for the data logger and the transmitter with an accuracy of a few nanoseconds, an automated detection scheme is required to find time segments that contain the transmitter signal. Due to the high amount of

Figure 2: Horizontal magnetic dipole transmitter developed and build by Radic Research and the GFZ.

raw time series normally collected during RMT field measurements, a manual selection of time segments is virtually impossible. However, a careful selection of appropriate time segments is essential for the success of the data processing. To address the challenge, machine learning algorithms have a high potential to solve both problems. During the processing the data were band-passed filtered into two different frequency bands. Both of these frequency bands contain four target frequencies ($1-8\,kHz$ and $16-128\,kHz$). Unfortunately, the quality of the data measured in La Campana is poor due to high ground contacts resistances of the electrodes with the hard ground. However, 541 data files from Santa Gracia and Nahuelbuta are available, which potentially can be used for the training of a machine learning algorithm.

3 First feasibility study

A first feasibility study was conducted using a very small training data set in order to evaluate if supervised machine learning algorithms are in general suitable for detection of transmitter signal in RMT time series. We trained a recurrent neural network, as this specific type of neural networks is very powerful for sequential or time series data. However, there exist several other possible machine learning algorithms, e.g. support vector machines, random forests and logistic regression, which could probably used for the task of automated signal and noise separation. The network was build with four bidirectional long short-term memory (LSTM) layers and one dense layer at the end using the TensorFlow library. The input layer consists of 26 nodes. The network was trained with 100 epochs and a learning rate of 0.01 using a cross-entropy loss and the accuracy as metric. As input features we used the 25 auto- and cross-spectra of the spectral density matrix for each single event as well as the frequency. This was motivated by the fact that

at least in some cases the signal of two loops of the transmitter is characterised by much higher power than the natural MT signal (see Fig. 3). The output layer consists of two nodes, representing the likelihood if an event contains transmitter signal or MT signal. The latter is regarded as noise in RMT.

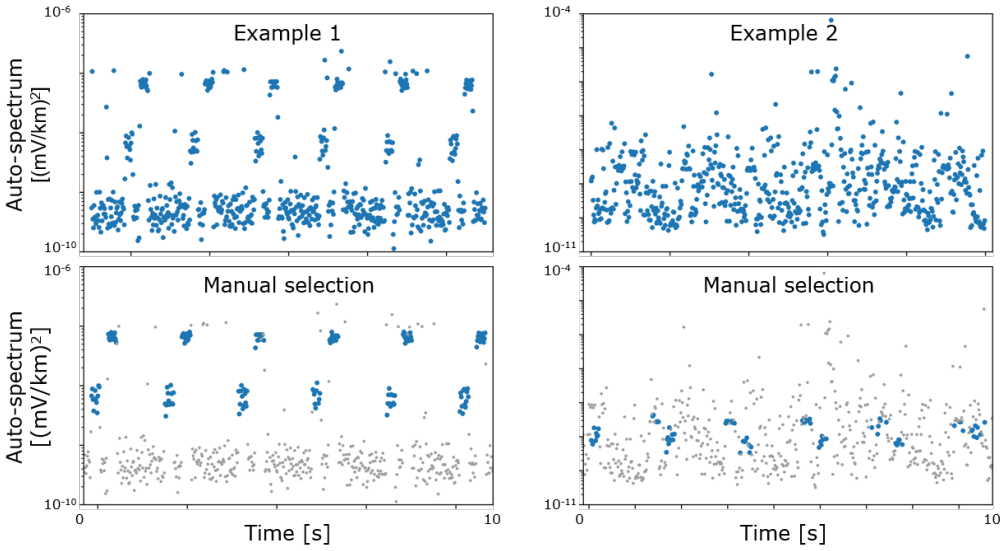

Figure 3: Manual selection of events, which contain transmitter signal for two examples. The upper images show the auto-spectrum of one electric channel for all events within the recorded $10\,s$ time series. In the lower images the manually selected events are displayed in blue, whereas events not corresponding to transmitter signal are displayed in grey.

For this feasibility study a very small training data set was used. It consists of only 44 data files using data from Santa Gracia and Nahuelbuta. The trained recurrent neural network was applied to other data from Chile, which were not included in the training data set. Despite the very limited size of the training data set, the trained network often performed quite well reaching accuracy values greater than 90 %.
Fig. 4 shows exemplary the comparison of the selection of events done by the trained recurrent neural network and the manual selection for one station. For the target frequency of $128\,kHz$ the network reached an accuracy of 98.1 % with only a small amount of false positives (red rectangle) and false negatives (red dashed rectangle). We conclude from this small feasibility study that machine learning algorithms can be used to automatically identify transmitter signal in controlled-source RMT data. However, more work has to be done to improve the quality of the trained network as the training data set was extremely small and none of the hyperparameters has been properly tested. Furthermore, we neither have tested other possible machine learning algorithms nor other possible input features, as e.g. coherences, eigenvalues of the spectral density matrix, single event transfer functions or polarisation values, yet. This will be done in the near future. The second example in Fig. 4b demonstrate that the performance of the trained network can be much lower (35.4 % in this case) if e.g. the power of the transmitter signal is in a similar range as the background noise. A first attempt to increase the performance of the network would be to increase the amount of training data.

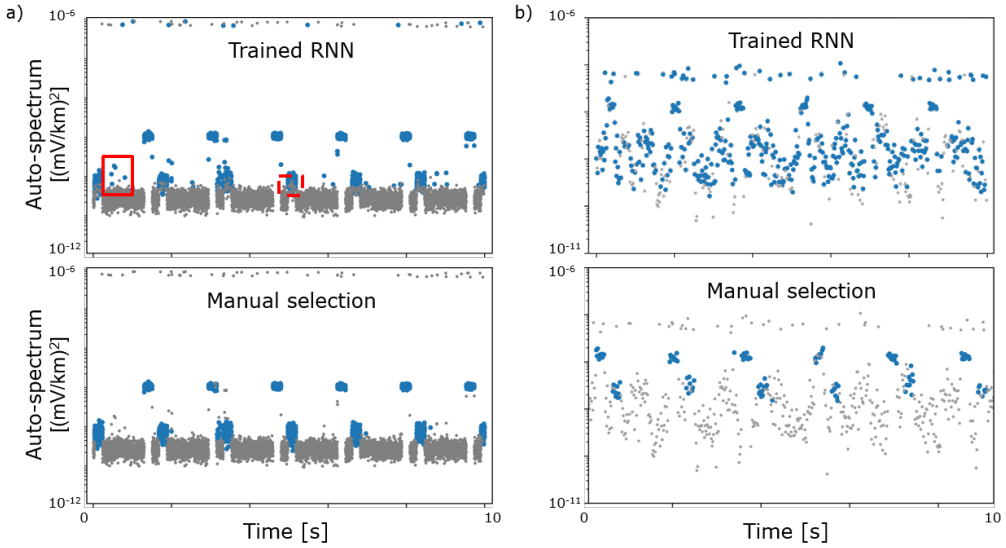

Figure 4: Comparison of the selection of events done by the trained network of the feasibility study (upper images) and manual selection (lower images) for one Chilean station, which was not included in the training data set for the target frequencies of a) $128\,kHz$ and b) $4\,kHz$. The selected events are displayed in blue. The red rectangles show exemplary false positives and false negatives.

4 Results with increased training data set

To increase the size of the training data set, we labelled all of the 541 data files from Santa Gracia and Nahuelbuta by manually selecting the events, which correspond to transmitter signal. As we needed on average $40 - 45$ minutes per station, this process was very time consuming. Especially, as in 19 % of all files we could not identify the transmitter signal at all and in another 22 % we could identify the signal only for some of the four target frequencies. However, for 269 data files (50 %) we were able to identify the transmitter signal for all four target frequencies with a very high accuracy. These 269 data files were used as a new training data set. In contrast to the 44 files used for the first feasibility study, the size of the training data set was increased by a factor of six. Furthermore, the new data set is more balanced in terms of station locations and used frequency bands as the originally training data set. 182 data files (67.7 %) were used for the training, 78 files (30 %) were used for the validation of the model and 9 files were used to compare the performance of the trained network from the feasibility study and the retrained model using the new data set. The retrained model was trained with the same parameters as the original model, only the training data set was increased.
Fig. 5 shows the comparison of the selection done by the model of the feasibility study and the retrained model for the same station as in Fig. 4 plotting the auto-spectrum of one magnetic channel, the bivariate coherence and the magnetic polarisation direction. For this extreme example the accuracy of the model could be increased from 35.4 % (model from the feasibility study) to 92 % (retrained model). The average accuracy for the test data set were increased from 81.8 % for the model from the feasibility study to 89.8 % for the retrained model. Furthermore, the retrained model has for 33 of the 36 frequencies and accuracy greater than 80 %. This is a quite good result, especially as the training

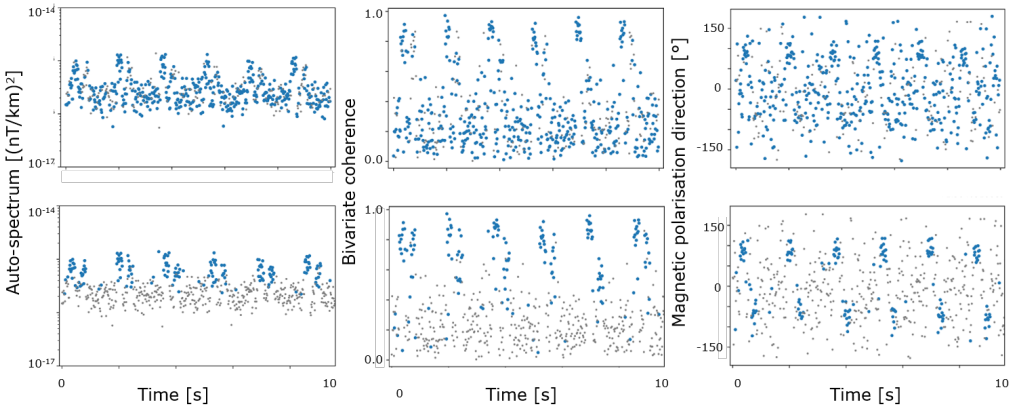

Figure 5: Comparison of the selection of events done by the trained network of the feasibility study (upper images) and the retrained network (lower images) for the same Chilean station as in Fig. 4, which was not included in the training data set for the target frequency of $4\,kHz$ displayed for three different physical parameters: auto-spectrum of one magnetic channel (left), bivariate coherence (middle), magnetic polarisation direction (right). The selected events are displayed in blue.

data set of the retrained model is still very limited. This test clearly demonstrates, that the performance of the trained model can be increased by using more training data. To increase the amount even further new data have to be measured and can be combined with synthetic data.

5 Application to new RMT data from Ireland

The retrained model was applied to RMT data measured in Ireland end of 2022 to test if a trained model can be successfully applied to new data sets. Almost all measuring parameters are different in comparison to the data from Chile as the Ireland field campaign was focused on an MT study. In theory, this should not be a problem as the trained model should be independent of these parameters. Besides the new location, a different data logger, magnetic sensors and other electrodes have been used. Furthermore a different sampling rate and a completely different transmitting scheme were applied. In contrast to Chile, where the transmitter emitted continuously, the transmitter made a short break after each cycle in Ireland and each frequency was emitted for $2 - 2.5\,s$ instead of the $0.2\,s$ used in Chile. Nonetheless, the retrained network reached accuracy up to $85.4\,\%$. This indicates that a trained model can probably applied to new data sets. One the other side, it becomes visible, that the model has a bias. Many false positives during times, where the transmitter did not emit, hint that the model "learned" the continuously transmitting scheme. This is highly undesirable as the model should work independent of the used transmitting scheme. As the model was trained with data, which all has the same transmitting scheme, this behaviour can be explained. To avoid such bias, new training data are needed, which have a high variability in the different parameters as e.g. the transmitting scheme, used sensors and sampling rates.

6 Summary and outlook

We trained a recurrent neural network with a very small training data set from Chile to identify events which contain transmitter signal for controlled-source RMT data. A first feasibility study showed that machine learning algorithms (models) are in general suitable to perform this task. We got acceptable results even with a very limited training data set of 44 files. Furthermore, we demonstrated that the accuracy of the trained model can be increased by increasing the amount of training data. The application of the trained model to new data measured under complete different conditions showed, that a trained model can probably successfully applied to new data. However, this test also demonstrates that the model has to be trained with data, which cover a higher variability in e.g. the transmitting scheme, to avoid bias. Our next steps would be to measure new data to increase the amount and the variability of the training data. Furthermore the training data set can be complemented by synthetic data. Another import step will be the identification of appropriate input features. So far, we used the single event auto- and cross-spectra. However, there exist many other possible input features as e.g. the transfer functions, coherences, polarisation directions or the eigenvalues of the spectral density matrix. We will use unsupervised clustering methods to detect an optimal set of input features. These input features will then be used to train and evaluate different supervised machine learning algorithms as support vector machines, logistic regression, neural networks and random forests in order to find one algorithm with a general high accuracy to perform the identification of transmitter signal in an automated manner in future. The final model will be tested by applying it to several new data sets.

References

Ritter, O., Junge, A., & Dawes, G.(1998). New equipment and processing for magnetotelluric remote reference observations. *Geophysical Journal International*, *132*, 535–548, doi:10.1046/j.1365-246X.1998.00440.x.

Weckmann, U., Magunia, A., & Ritter, O.(2005). Effective noise separation for magnetotelluric single site data processing using a frequency domain selection scheme. *Geophysical Journal International*, *161(3)*, 635-652, doi:10.1111/j.1365-246X.2005.02621.x.