Denoising and Recovering Better Error Estimates For Geophysical Data

Johnathan C. Kuttai
Department of Earth, Ocean and Atmospheric Sciences
University of British Columbia University
Vancouver, BC 15213
11074580
jkutt@eoas.ubc.ca

Anran Xu
Department of Earth, Ocean and Atmospheric Sciences
University of British Columbia University
Vancouver, BC 15213
54002811
anranxu@eoas.ubc.ca

Abstract

Geophysical surveys are often used to image subsurface mineral exploration targets. Commonly, electromagnetic surveys are used. When these surveys are employed in areas near existing mine sites, external noise from the infrastructure degrades the geophysical signal of interest. Controlled source surveys such as direct-current and induced polarization derive physical properties from the measured time-series collected by a receiver spread. Typically filters are applied to remove the infrastructure noise but is difficult to do so without corrupting the induced polarization response. Another difficulty is characterizing the error that is used in geophysical inversion. In noisy environments, if the noise is not characterized properly, artifact structure and or targets are created. This project explores applying auto-encoders and variational auto-encoders to the above problems. Using these architectures estimates on error and denoising are analyzed.

1 Introduction

In mineral exploration, follow up exploration often extends from known ore bodies. These ore bodies are typically mined which are accompanied by large infrastructure. A lot of it is electrical. Most non-invasive electromagnetic geophysical surveys used in exploration are greatly affected by this infrastructure (figure 1). The electrical interference corrupts the time-series data collected from geophysical surveys. Specifically, direct current and induced polarisation (DCIP) derive subsurface physical properties from measured time-series data. To improve the derived measurements, denoising or reliable noise estimates of the signals collected near mine-sites are required.

Often digital and analog filtering is unsuccessful at fully removing the unwanted signal in time-series data. Mine-site signals can vary in frequency significantly, far beyond the typical 60 Hz power line. Some signals can be less than 1 Hz or higher than 100 Hz (figure 2). Using notch filters or band pass filters can do more harm than good and further corrupt the induced polarisation response of the time-series.

Parameters derived from the time-series data are often used in subsurface modelling known as geophysical inversion. Failing to successfully filter the unwanted signal, estimates of the error on

Course project for UBC CPSC 440/540, April 2023.

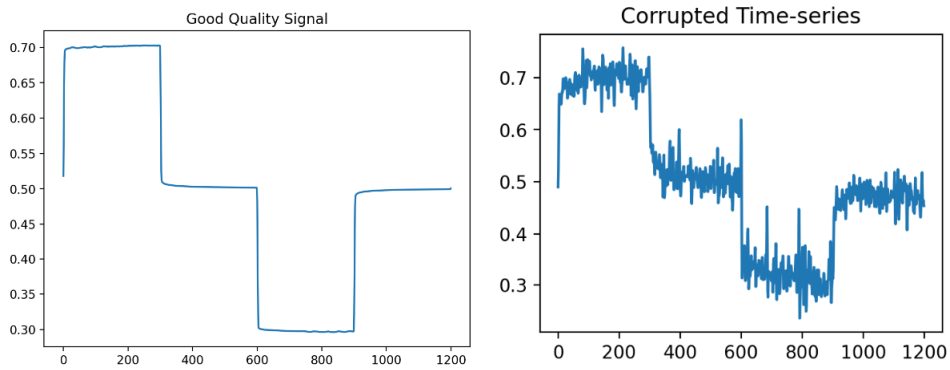

Figure 1: Examples of a good and external noise corrupted time-series portion.

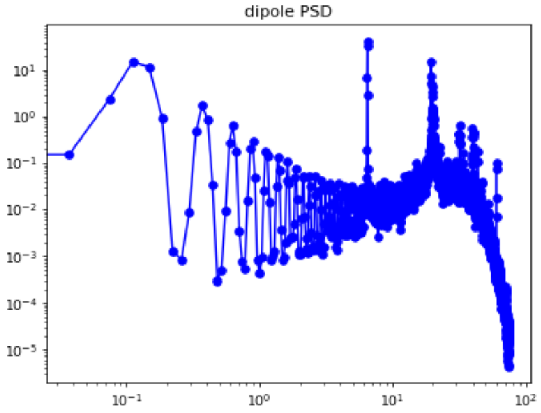

Figure 2: Example of unwanted frequencies (peaks > 10 Hz) from the power spectral density.

the data can greatly improve the resulting model. This can also be more difficult than expected. Traditionally uncertainties are defined by a single guessed standard deviation plus a floor value. This is often a gross generalisation and requires fine tuning. To better characterise the noise, is to improve the modelling results.

Two ways to improve DCIP survey data collected near mine sites is to either denoise the time-series leaving only the desired signals or provide better noise estimates to feed the geophysical inversion process. Using real data collected near or in mine sites, the idea is to apply machine learning techniques to denoise or at the very least provide better estimates of the noise. For training, good clean data collected in less noisy environments will be used to train autoencoder and variational autoencoder networks.

Since autoencoders and variational autoencoders provide a reconstruction of its inputs, there is potential to remove/characterise the noise. The reconstruction error itself can be applied to defining the error estimate/uncertainty used in geophysical inversion. In addition, this can be applied as a quality control tool. When reconstruction error is too large this is likely associated with heavily corrupted, unusable time-series. These will be removed from the overall time-series data. Auto-encoders have been applied to Electromagnetic geophysics previously but were applied to prediction of the forward model data in the inversion routine itself (**1**).

The characterization of the noise will ideally be captured by the networks and provided as a standard deviation estimate of the time-series portions that derive the physical properties. If the time-series data is likely, the error estimate used in the geophysical inversion will reflect that. This approach will be used to compare against traditional error assignments to see the improvement in overall models.

2 Training and Test data

The training data consists of 11,000 examples of 0.125Hz source full period time-series. Of these training data, 4683 examples are labeled "good" data. The training step will use these data in order to reconstruct noisy data into the cleaned "good" data of the signal. For this study, test data known to consist of external noise were compiled in order to measure how well the networks reconstruct the transmitted input signal and estimate the noise in an extreme noise environment.

Given that the processing procedure for controlled source geophysical data involves reconfiguring a recorded time-series into "stackable" chunks, the VAE network is trained on similar chunks of data. Though instead of half-period samples, full-period samples are used to fully capture the 50% duty cycle geometry.

3 Denoising and Error Estimates with Variational Autoencoders

DCIP time-series are measured by creating an artificial source with a known input signal that is often a 50% duty cycle, meaning that for an entire period, for one half period, half of it is signal "power on" and the last 50% is signal "power off") (**2**). Measurements are derived from taking the measured time-series and rearranged into half period snippets or "stacks" that get polarity corrected. They are then averaged in a process called "stacking" (**3**). Variations on the averaging exist and typically a kernel filter is applied to act as a low pass filter (or any filter for that matter) (**3**). This greatly reduces the high frequency noise in the measured time-series leaving cleaner signal to derive the resistivity and induced polarisation physical properties. The primary voltage is calculated from the measured time-series data from a window of the "power on" portion of the signal. From this the apparent resistivity physical property can be calculated knowing the current (injected signal) and the electrode geometry (**2**) .

$$\rho_{apparent} = \frac{2\pi V}{I} \frac{1}{\frac{1}{r1} - \frac{1}{r2} - (\frac{1}{r3} - \frac{1}{r4})}$$

Induced polarisation is derived from measuring the secondary voltages of the decay of the remaining power off content of the half period Figure. These secondary voltages are then combined with a defined time window and summed to get a total chargeability of the measurement (**2**).

$$M_{apparent} = \frac{1}{V_{primary}} \int_{t1}^{t2} V(t)dt$$

To improve these derived estimates of the physical properties, the time-series are partitioned into there stacks and fed into autoencoder and variational autoencoder architectures to assess the effectiveness of denoising and or providing better error estimates.

3.1 Reconstructing Time-series with Variational Autoencoders

The VAE architecture here uses a 7 layer encoder, 4 layer decoder and 2 dimensional latent space architecture (figure 3). It is used to produce a reconstructed representation of the input trained to remove the noise parts of the signal (**4**). With this we can determine the reconstruction error as a measurement of the quality of the data.

3.2 Denoising

Since the VAE output is a reconstruction of the input source signal, an estimate of the noise can be calculated by subtracting the reconstruction from the recorded signal. What is left are ideally the noise signals (Appendix A figure 5a). Calculating the power spectral density of the raw, reconstruction and noise time-series, we see that the VAE is effective at removing the high frequency noise in relation to the base source signal (Appendix A figure 5b). The VAE also acceptably preserves the base frequency and its harmonics. Though not perfectly, it does well enough to gauge the noise added from the environment.

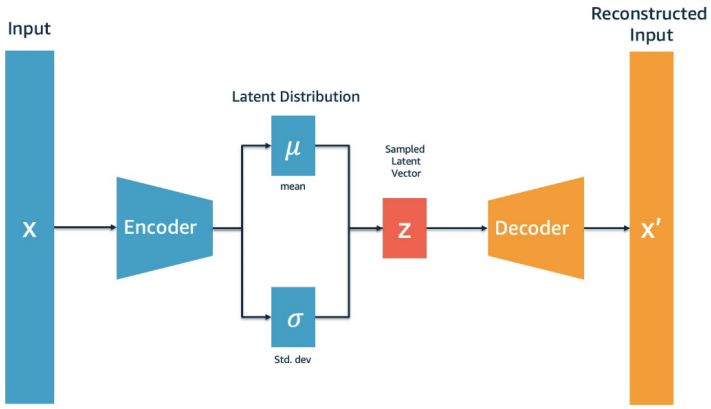

Figure 3: Variational auto-encoder Architecture (**5**).

The effect of noise on the derived physical properties can be significantly corruptive. As seen in Appendix A figure 6a, the on time signal has variable amplitude and contains a lot of high frequency noise. Although the stacking method has the added benefit of being effectively a low-pass filter, it's only marginally effective at removing powerline noise of 50-60Hz and up. Mostly due to the roll off nature of the filter response. Stacking also suffers from being incapable of dealing with other problematic noise signals such as electromagnetic coupling which is indicative of the noise spikes at the beginning and or end of the power-on portion of the signal. Appendix A figure 7 illustrates a positive electromagnetic coupling effectively increasing the estimation of the chargeable response in the starting of the off-time portion of the signal. The VAE on the contrary is particularly effective at removing (or at the very least minimising) the electromagnetic coupling response. The on-time signal is overestimated but this is likely due to its variable amplitude period to period and not the coupling.

We can see from Appendix A figure 6b that the electromagnetic coupling creates a significant difference in the derived chargeable response. So much so that this could indicate an anomalous response when in actuality, there is not. This can be costly as drilling anomalous targets in mining are expensive (millions of dollars per hole). Here we see that due to the electromagnetic coupling, the chargeable response is 7.1 mV/v higher than what the VAE estimates as the noise free signal.

3.3 Error assignment model

In addition to denoising, another practical use of the VAE reconstructions is to use them to map correlations of the reconstruction errors with different local geological packages within the receiver coverage. If the reconstructions cluster well, a model can then be generated to output the corresponding noise level of the area. Specific to distributed array DCIP acquisition this can be particularly useful because additional dipoles are created in post processing made up of each of the raw receiver time-series. For this study, the mapping/classification is done by fitting a Gaussian mixture model so that given an easting and northing, an output indicating the noise estimation from the assigned Gaussian. By plotting histograms of the reconstruction and standard deviations of the estimated noise done in section 2.2.1, both indicate 5 groupings (Appendix A figure 7). Furthermore, plotting the reconstructions errors and the standard deviations of the estimated noise spatially, grouping of common values are observed (Appendix A figure 8). It is observed that the northern and southern most portions of the receiver coverage contain geology that is noisier than the south central data. Compared to the how error is traditionally assigned, both the estimated error standard deviations and the reconstruction errors provide a more coherent map of the noise spatially. Comparing the standard deviations and the reconstruction errors, the latter is the most coherent.

Using the number of groupings presented in the histograms, a Gaussian mixture model with 5 components is created in order to map/classify the areas of common noise levels. The resulting model has Gaussian means and variances that classify the areas of similar reconstruction error rather well (Appendix A figure 9). Now we can apply this model to assigning errors for the geophysical inversion. Given the means of the Gaussian's determined from the model, each can be considered a different noise level. Assigning the levels can be freely chosen, be it derived from the standard deviations from

the noise estimate or any predefined level. Before data is fed into the geophysical inversion each data sent for inversion will feed its location into the Gaussian mixture model as the features and outputs the corresponding noise levels of the classified geological package. The noise level for each of the classified data points are then assigned and provided to the inversion code.

4 Uncertainty Estimation with Autoencoders

One way of estimating error is using the density network to calculate the uncertainty of the model's prediction. The density network also has the benefit of avoiding overfitting. In a plain model, the loss function is determined using the negative log likelihood with constant σ. For example, if we assume the likelihood follows the Gaussian distribution, then we have

$$p(y|x,\theta) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{(y-f_\theta(x))^2}{2\sigma^2}),$$

where f_θ is the model and θ represents the model's parameters. However, if we consider σ as a function of input X, then the σ can be an uncertainty of the prediction from the model and the negative log-likelihood will be

$$-log(p(y|x,\theta)) = -log(\frac{1}{\sigma_\theta\sqrt{2\pi}} exp(-\frac{(y-f_\theta(x))^2}{2\sigma_\theta^2})) = -log(\frac{1}{\sigma_\theta\sqrt{2\pi}}) + \frac{(y-f_\theta(x))^2}{2\sigma_\theta^2}.$$

We firstly trained an autoencoder with constant σ loss. We observed that the denoising performance of the autoencoder is similar to the variational autoencoder in the majority of scenarios. However, the variational autoencoder surpasses the autoencoder in the scenarios of very corrupted time-series.

One difference between variational autoencoder and autoencoder is that there is an extra KL divergence term in the objective function of the variational autoencoder; therefore, the latent space in the variational autoencoder is regularized to follow the prior distribution N(0,I). Appendix B figure 10 is the comparison of the latent distribution of variational autoencoder and autoencoder with the same architecture. In the cases where large noises occur, the latent representation can go to [300,300] after passing it into the encoder of the autoencoder, which makes the output of the decoder unreasonable. Thanks to the regulator in the variational autoencoder, the latent representations are less likely to have insane values, so variational autoencoder is more robust in the extremely noisy sites.

Since training time needed by an autoencoder is much shorter than a variational autoencoder with similar architecture, we choose to train autoencoders with non-constant σ loss. Sample result is shown in figure 4.

Note that the uncertainty estimation in this section is heavily depends on the training dataset, objective function, and the architecture we chose, so it's an indicator of the uncertainty of the model instead of the estimator of noise level of the input time-series.

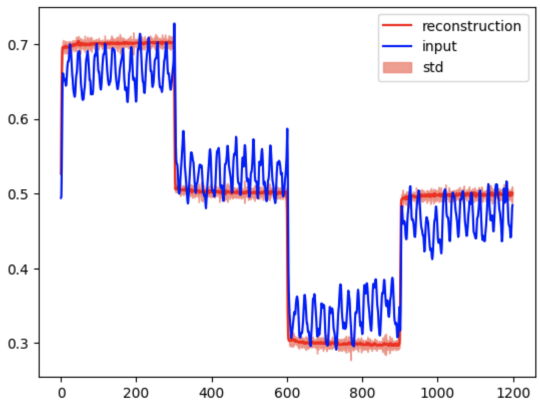

Figure 4: Auto-encoder with uncertainty estimation.

5 Discussion

By leveraging neural networks such as autoencoders and variational autoencoders, improved noise analysis can be accomplished. Since traditional filtering techniques cannot perfectly remove unwanted signals in measured time-series, characterizing the noise is beneficial for any further geophysical analysis. Traditional methods have yet to yield satisfactory estimates of noise and vary from survey to survey or flavour of the geophysical interpreter. With the onset of the recent successes in deep learning, it is only natural to apply these new tools to this specific problem. From the analysis in this study it is shown that both the autoencoders and variational autoencoders add much value characterizing noise in controlled source time-series data.

In many cases the varaitoinal autoencoder completely removes the high frequency noise while preserving most of the primary transmitted signal. In a minimal amount of cases though, effectiveness of the network is not all that beneficial. At the very least, the reconstructions can add benefit by using the reconstruction errors to classify areas within a receiver coverage that are particularly susceptible to noise. Because geophysical inversion techniques depend on error estimates, the assignment of errors can then be adjusted accordingly. Assignments can be done in a number of ways. For example, the Gaussian mixture model made up of the reconstruction error information. By using location data as features, it can classify the level of error to designate.

Error estimates alone are important in determining the level of noise in which we want to assign. Using autoencoders as density estimation networks, Uncertainty of the prediction is used to estimate the noise. These networks are also quick to train in relation to variational autoencoders. Meaning they can be added into processing workflows without much burden. The variational autoencoder though, has the benefit of fitting a time-series rather well even in most extremely corrupt cases. The reconstructions are then subtracted from the raw data ideally leaving the direct estimate of the noise signals as shown in section 3.2.

Given the costly venture to drill mineral exploration targets, focus is needed on characterizing the noise in the data used to indicate these potential targets. In this report we propose methods that can provide a complete pipeline from classifying spatial locations of varied noise to providing noise estimates for the different levels of noise within the receiver coverage. By characterizing the noise, the geophysical models produced from inversion are improved and confidence in the produced targets are higher. Essentially these methods decrease the risk required in the exploration of minable material.

References

[1] Asif, Muhammad Rizwan, Nikolaj Foged, Pradip Kumar Maurya, Denys James Grombacher, Anders Vest Christiansen, Esben Auken, and Jakob Juul Larsen. 'Integrating Neural Networks in Least-Squares Inversion of Airborne Time-Domain Electromagnetic Data'. GEOPHYSICS 87, no. 4 (1 July 2022): E177–87. https://doi.org/10.1190/geo2021-0335.1.

[2] Telford, W. M., L. P. Geldart, and Robert E. Sheriff. Applied Geophysics. 2nd ed. Cambridge [England]; New York: Cambridge University Press, 1990.

[3] Halverson M., 1967, A look at dispersion and induced polarization using the phase angle concept: Proceedings of the symposium on induced electrical polarization, February, 1967, UC Berkeley.

[4] Im, Daniel Jiwoong, et al. "Denoising criterion for variational auto-encoding framework." arXiv preprint arXiv:1511.06406 (2015).

[5] Yi Xiang, Deploy variational autoencoders for anomaly detection with TensorFlow Serving on Amazon SageMaker, Amazon Machine Learning Solutions Lab, blog, 14 July 2021

A VAE denoising and error estimate analysis figures

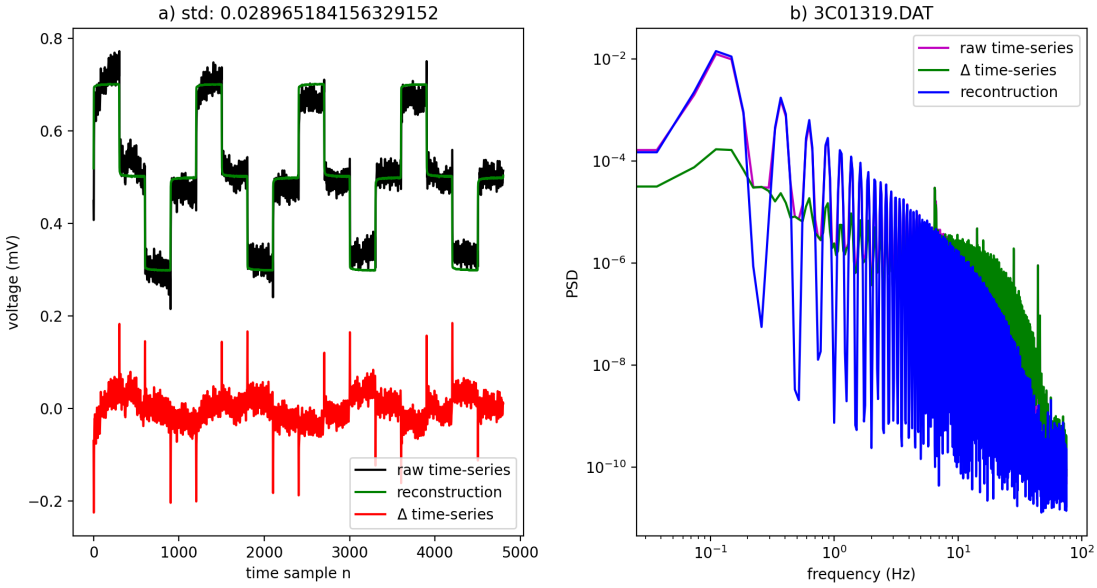

Figure 5: Example of denoising a time-series and estimating the noise. a) Partial signal plot of recorded, reconstructed and noise time-series. b) Power spectral density of each time-series make-up.

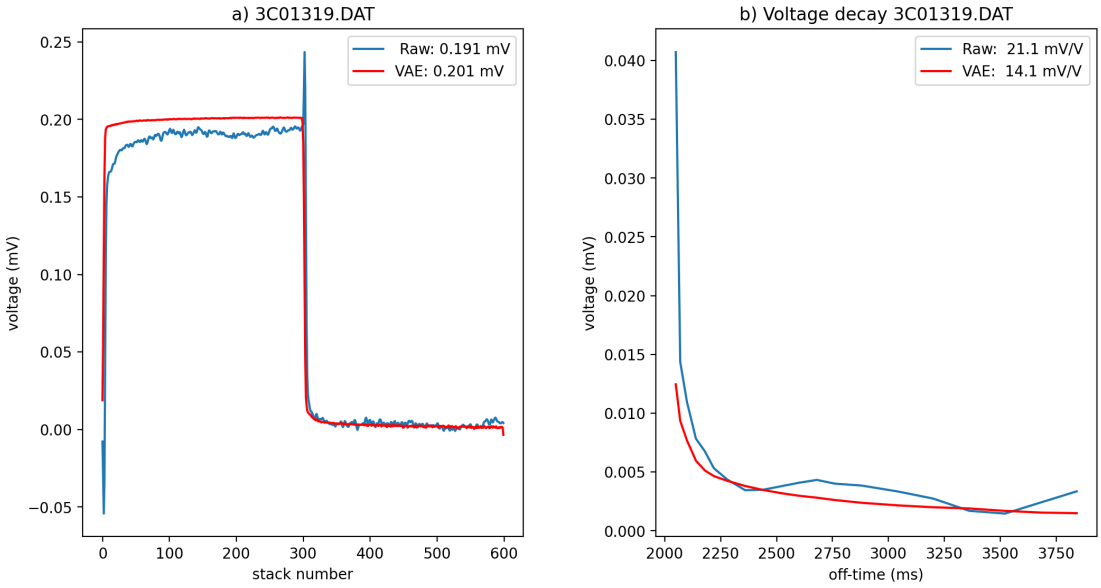

Figure 6: Example of processing the denoised and raw time-series.

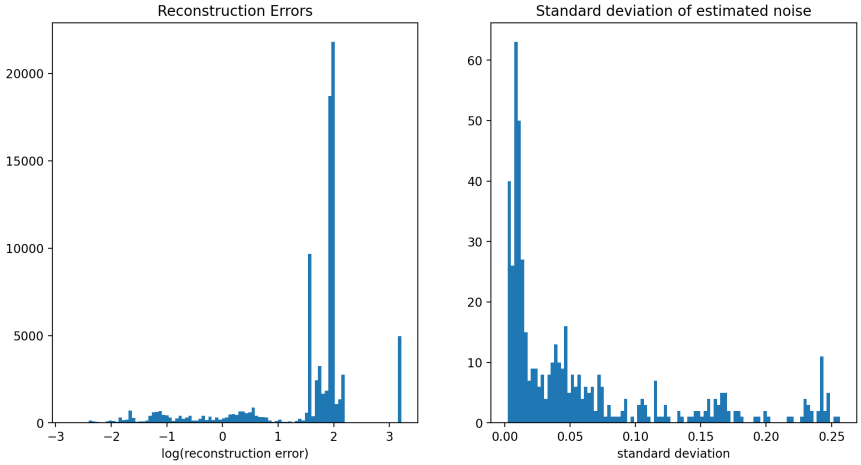

Figure 7: Distribution of reconstruction error and the standard deviation of the estimated noise.

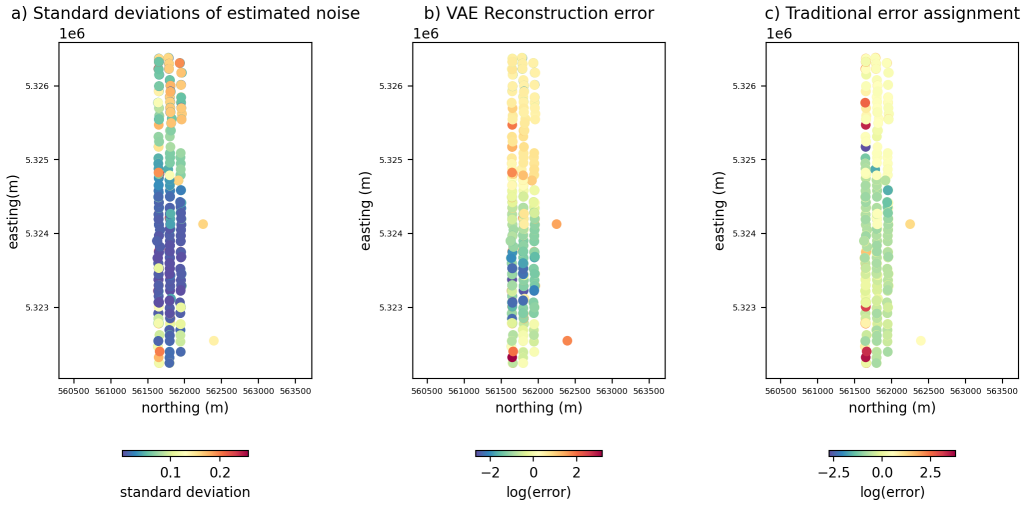

Figure 8: Mapping noise. a) error mapped using the standard deviation of the estimated noise from VAE. b) Reconstruction error map. c) The general error assignment for DCIP.

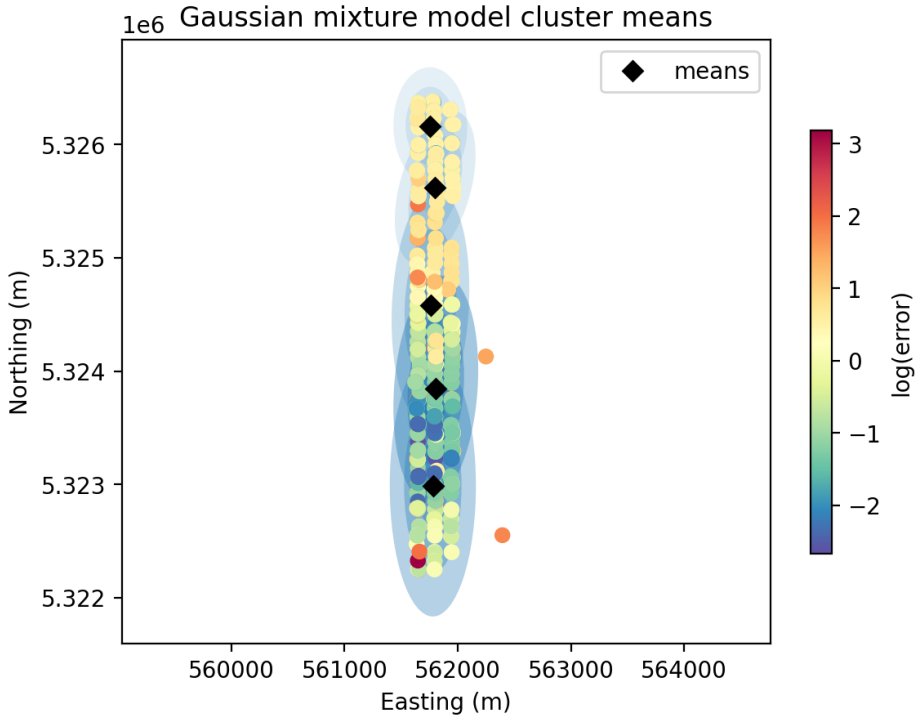

Figure 9: Noise classification/mapping.

B Latent distribution of autoencoder and variational autoencoder

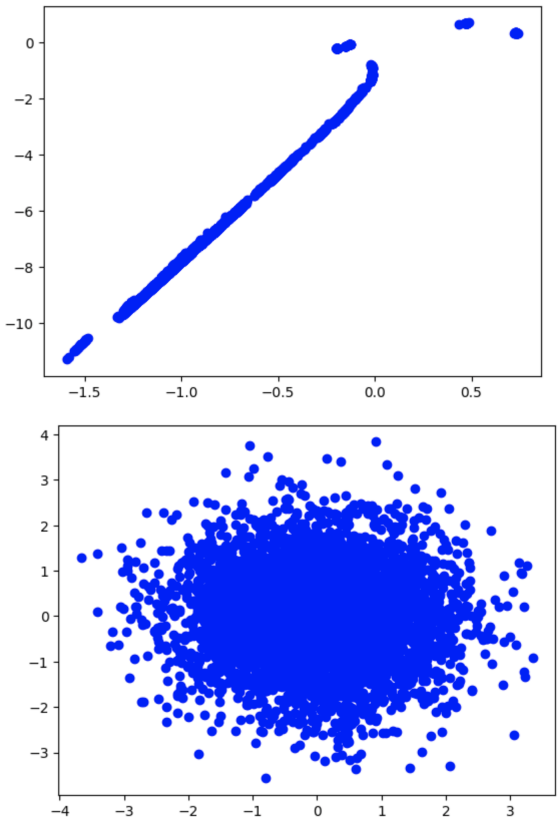

Figure 10: The top plot is the latent distribution of autoencoder, and the bottom plot is the latent distribution of variational autoencoder. They have the same architecture for encoder CNN and decoder CNN. The latent represntations are from the training set.