Journal of Applied Geophysics 135 (2016) 163–182

Contents lists available at ScienceDirect

Journal of Applied Geophysics

journal homepage: www.elsevier.com/locate/jappgeo

Data quality control methodologies for large, non-conventional DC resistivity datasets

Michael A. Mitchell *, Douglas W. Oldenburg

Geophysical Inversion Facility (GIF), Department of Earth Ocean and Atmospheric Sciences, University of British Columbia, 2020-2207 Main Mall, Vancouver, British Columbia V6T 1Z4, Canada

ARTICLE INFO

Article history:
Received 14 April 2016
Received in revised form 12 September 2016
Accepted 14 September 2016
Available online 18 September 2016

Keywords:
DC resistivity (ERT/ERI)
Data quality control
3D inversion
Data clustering

ABSTRACT

With developments in instrumentation and computational resources, the collection of large, non-conventional DC resistivity datasets has become commonplace. While the increased data content of these large datasets can significantly improve the resolution of inverse models, these datasets also present challenges for standard data quality control (QC) methodologies. Standard QC methodologies for DC resistivity datasets typically rely on our ability to decompose the dataset into 2D lines and/or reciprocal measurements. Non-conventional electrode geometries and the cost of collecting a large number of reciprocal measurements can severely limit the applicability of standard DC resistivity QC methodologies.

To address these limitations, we developed a more generalized data QC methodology which utilizes statistical analysis and classification tools. The merit of this methodology is illustrated using a field dataset collected in an underground potash mine and several synthetic examples. Results from these applications show that the methodology has the ability to identify and characterize highly noise-contaminated data from a number of different sources. The flexibility of the 4-stage methodology allows it be tailored to accommodate data from any type of DC resistivity survey and the use of statistical analysis and classification tools decreases the subjectivity of the process. Although this study focuses on the applicability of this methodology for DC resistivity data, it is potentially applicable to a variety of geophysical surveys.

1. Introduction

Despite the many advances in multi-channel instrumentation systems, survey design optimization, and inversion techniques, large DC resistivity datasets can still be difficult to work with due to limitations of standard data quality control (QC) methodologies. Identifying and dealing with highly noise-contaminated or inconsistent data is a vitally important part of the inversion process since it informs our choice of measurement uncertainties. These uncertainties define the relative importance or weight of each measurement. Standard tools for data QC include 2D pseudo-section plots of apparent resistivity and repeat or reciprocal measurements.

With 2D profiles, a pseudo-section of the apparent resistivities is typically plotted to highlight spurious or noise-contaminated data associated with a specific electrode (Deceuster et al., 2013; Edwards, 1977). Since the apparent resistivity is expected to vary smoothly in most circumstances, spurious data can be identified by small, anomalous regions of high or low apparent resistivity (Loke, 2000). Similarly, conventional 3D surveys, which consist of a regular grid of electrodes, are typically decomposed into 2D profiles so that pseudo-sections can be easily plotted (Auken et al., 2006). However, when working with large non-conventional 3D datasets, plotting a map or pseudo-section of the data is not straightforward or meaningful. Data with very different electrode geometries can plot in similar locations making it difficult to identify patterns associated with inconsistent or noisy data.

Repeat and reciprocal measurements have been shown to be a useful tool for assessing the noise levels of DC resistivity datasets by several previous studies including LaBrecque et al. (1996), Slater et al. (2000), Zhou and Dahlin (2003), LaBrecque and Daily (2008) and Wilkinson et al. (2012). Repeat measurements are made anytime that the same transmitter (TX) and receiver (RX) locations are reoccupied, while reciprocal measurements are made when the TX and RX locations are interchanged. A formal proof of the reciprocity theorem for DC resistivity, assuming an arbitrary conductivity model, is given by Parasnis (1988). Reciprocal measurements provide a more reliable estimate of measurement noise levels since they account for some systematic error sources which can go undetected by repeat measurements (LaBrecque et al., 1996).

While reciprocal measurements are certainly a valuable tool for assessing the noise level of DC resistivity datasets and identifying

* Corresponding author.

http://dx.doi.org/10.1016/j.jappgeo.2016.09.018
0009-2541/

noise-contaminated measurements a full set of reciprocal measurements is not always collected. In large distributed array systems, such as Quantec's Titan 24 system, a dense grid of RXs is laid out and the TX is moved through the grid between the RXs. Since the RX locations are never occupied by the TX, no reciprocal measurements are collected with this setup. In other scenarios, limited budgets and a desire to maximize the data content of the survey prohibits the collection of a large number of reciprocal measurements. If a large non-conventional 3D dataset lacks sufficient reciprocal measurements, other methods must be employed to QC highly noise-contaminated data.

1.1. Inversion background

After initial efforts to quality control the data and remove clear outliers, the data are input into an inversion algorithm to find a geologically reasonable conductivity model which acceptably reproduces the data. Since the inverse problem is typically very underdetermined, (i.e. there are far more model cells than data), we pose it in a regularized optimization context as the minimization of a global objective function (Φ). The inverse problem is solved by finding an estimate of the true model ($\hat{m}$), which minimizes the model objective function (Φ_m), while driving the data misfit (Φ_d) to its target level (Φ_d^*). This minimization problem can be formalized in the following manner.

$$\begin{aligned} &\min \Phi = \Phi_d + \beta\Phi_m \\ &\text{s. t. } \Phi_d \leq \Phi_d^* \end{aligned} \tag{1}$$

where β is a regularization parameter that controls the relative importance of the model objective function, which controls the complexity and smoothness of the model, and the data misfit. To obtain a meaningful solution to the inverse problem we need to choose a metric by which to measure the data misfit, assign reasonable uncertainties to each measurement, and define a target misfit to quantify an acceptable data fit (Oldenburg and Li, 2005).

Data misfit provides a quantitative measure of the difference between the observed data (d^{obs}) and the predicted data (d^{pred}) derived through numerical modelling. The data are normalized potentials or resistance values which are computed by dividing the measured potential differences (V_{MN}) by their injection currents (I). An L_p norm is typically used for this metric with the L_2 being the most commonly used and the L_1 norm sometimes employed to better handle outliers. Here we use a standard L_2 formulation for the data misfit (Φ_d).

$$\Phi_d = \sum_{i=0}^{n}\left(\frac{d_i^{obs} - d_i^{pred}}{\xi_i}\right)^2 = \left\|\mathbf{W}_d\left(\mathbf{d}^{obs} - \mathbf{d}^{pred}\right)\right\|_2^2 \tag{2}$$

where ξ are the standard deviations of the data which quantify measurement uncertainty. The data weighting matrix $\mathbf{W}_d$ is a diagonal matrix with $1/\xi$ on the main diagonal. Since these standard deviations are typically unknown, we must estimate these uncertainties, which account for discrepancies between the observed and predicted data. These uncertainties are often referred to as the noise model.

Developing a accurate noise model is challenging since many factors contribute to measurement uncertainty and each has different underlying statistical distributions. To capture the cumulative effect of the various factors in a single distribution, we assume that the noise is uncorrelated and Gaussian. Under these conditions, the target misfit is given by the expected value, which equals the number of data, provided uncertainties have been assigned reasonably. While these assumptions are almost certainly incorrect, they have served researchers well. As per common practice the uncertainties are assigned as a percentage of each measurement plus a floor value, which is often dependent on the magnitude of survey measurements and the instrument precision.

To control the complexity of the recovered model, the following formulation for the model objective function (Φ_m) is used.

$$\Phi_m = \|\mathbf{W}_s(\mathbf{m} - \mathbf{m}_o)\|_2^2 + \sum_{i=1}^{3}\|\mathbf{W}_i\mathbf{m}\|_2^2 \tag{3}$$

Where the first term controls the "smallness" of the model (i.e. how close the current model ($\mathbf{m}$) is to the reference model ($\mathbf{m}_o$)) and the summation contains directional derivative terms which control the smoothness of the model in each direction. $\mathbf{W}_s$ is a diagonal matrix containing cell weights while the three components of $\mathbf{W}_i$ ($\mathbf{W}_x$, $\mathbf{W}_y$, and $\mathbf{W}_z$) combine finite difference operators and face weight vectors in each direction.

The optimization problem is solved iteratively using a Gauss–Newton based approach, where at each β iteration the model update is estimated using an incomplete preconditioned conjugate gradient solver. The initial β is chosen to be sufficiently large so that the Φ_m dominates the objective function (Haber et al., 2004). β is then iteratively cooled with each iteration, placing more emphasis on reducing Φ_d by allowing more structure to be incorporated into the model.

For additional information on geophysical inversion we refer the reader to the following books: Menke (1989), Parker (1994), Aster et al. (2012), and Haber (2014). Oldenburg and Li (2005) provide a general overview of geophysical inversion in their tutorial while Li and Oldenburg (1994), LaBrecque et al. (1996), Loke and Barker (1996), Ramirez et al. (1996), and Loke et al. (2013) are a few of the many papers which specifically discuss the inversion of DC resistivity data.

After inverting, data misfit plots are used to examine how well the recovered model fits the data. A global misfit (Φ_d) close to the target misfit indicates that there is good overall agreement between the observed and predicted data. Looking at the distribution of individual data misfits can help identify clear outliers and provides an estimate of how many data have individual data misfits greater than the target misfit of 1. If the recovered model does a poor job of fitting the observed data then further data quality control (QC) analysis is required.

1.2. A new data QC methodology

The limitations of standard data QC tools prompted us to develop a new methodology which combines a search for correlations between high misfit data and various survey parameters with statical analysis and classification tools to identify noise sources and deal with highly noise-contaminated and inconsistent data. Graphically this search for correlations can be done using a series of survey parameter cross-plots. However, this type of manual multivariate analysis is difficult since there are many parameter spaces to explore. Boxplots, SVD analysis, and k-Means clustering (MacQueen, 1967) are used to semi-automate this procedure and reduce the subjectivity of the process.

The data quality control methodology presented, is potentially applicable to a variety of field surveys. Here we test it in an underground environment with DC resistivity data. We find that a combination of a few poor electrodes and current leakage problems within the cables, that connect electrodes, have conspired to generate a highly noise-contaminated dataset. Despite the highly contaminated nature of the dataset, we are able to use the outlined data QC methodology to identify a subset of reliable data for inversion and obtain an interpretable inversion model.

In this paper, we open with a description of the case history, which provided the impetus for this research, and then present the 4-stage data QC methodology in the context of the case history.