Turning geophysical data into geological information or Why a broader range of mathematical strategies is needed to better enable discovery

The goal of exploration geophysics is to infer the nature of buried structure and, in particular, generate drill targets that lead to a mineral deposit discovery or reserve delineation. As a profession, we aim to turn geophysical data into geological information. Most geophysical techniques enable inferences to be made from airborne, ground-based or bore-hole data through a deterministic process whereby a single model ‘answer’ is generated. Well-founded algorithms include uncertainty estimates for different parameters in the model and/or some form of model validation. This approach has been successful to date, and we advocate the continued use of deterministic algorithms.

We also advocate that alternate strategies for extracting information from data are used alongside deterministic strategies. If we consider two general properties of data inference approaches, that of (1) assurance and (2) opportunity, deterministic approaches score poorly regarding opportunity: that is, useful answers may be missed. Alternate strategies can be computationally intensive, but several important classes of approach, summarised in this article are now tractable on workstation or high-specification notebook PCs. By using a range of strategies we can maximise both assurance and opportunity for a particular data inference goal and obtain extra, useful geological information from our data.

Keywords: data inference, information, inversion, modelling.

Introduction

Turning geophysical data into geological information presents us with a three-fold set of tasks. As practising geophysicists, we:

1. Carry out the modelling or inversion of geophysical data to find a geological ‘answer’;
2. Quantify, or represent in some way, the ‘strengths’ and ‘shortcomings’ of our model or answer;
3. Explain both the model and its limitations to other geophysicists, and very significantly, also to non-geophysicists.

In this article, we outline some of the strategies for geophysical modelling and inversion that will be familiar and not so familiar to practising exploration geophysicists. Firstly though, we examine a goal for geophysical modelling and inversion, that of mineral deposit target generation. Those targets will subsequently be drilled and hence the geophysical inference, that there is some structure playing host to an ore deposit, can be tested. Importantly, the veracity of the geological information that we have inferred may be tested against geological knowledge (albeit spatially limited) that we obtain from drilling. Two concepts are important here: assurance and opportunity.

Assurance embodies the extent to which a model answer is likely to be correct, in our example goal: an ore body discovery. High assurance corresponds to low risk. Of course, we wish for highest assurance answers, but constraining our drilling to high-assurance targets will mean that we overlook some targets. We also have to consider that there may be no high-assurance targets.

Opportunity is the idea that targets are generated with sufficiently open criteria that few targets are overlooked. Low opportunity implies that there is a very tight set of rules that map the data to a model. The natural variability of the mineral systems means that a target might still be prospective if one indicator parameter changes, hence a higher opportunity approach might (also) be appropriate.

Exploration geophysics data are often ‘inaccurate, insufficient and inconsistent’ (Jackson, 1972) and so solution existence, non-uniqueness and the instability of the solution process must all be addressed (Backus and Gilbert, 1970; Aster et al., 2005). At the present time, many established codes handle such difficulties to the satisfaction of their users, yet in a realm that can limit the models available. Moreover, the limitations of the model are not always evident. Fortunately, alternate strategies have become tractable, as computing power and data storage capacity improves. These strategies generally give a more complete set of outputs including alternative models and some measure of the likelihood of different models, or parts of the model. It is important to have in mind that an uncertainty estimate on a single model solution accounts only for noise. It does not account for solution non-uniqueness. Other very different solutions could exist that are equally well supported by the data. In forming single model solutions, we do not fully exploit the information content of our geophysical data.

Alternate strategies are generally more exploratory in character: suggesting answers that are constrained by the data but may not have been imagined by an operator (for example) building a forward model or by a deterministic inverse approach. Alternate strategies all supply material very naturally that assists us in task 2. Finally, so long as we are communicating with a colleague who can accept the inconvenient truth that there is more than
one possible answer, we can do a much better job at task 3, explaining the results and the limitations of the results. Alternate strategies provide us with a much richer range of information from the same set of geophysical data (Figure 1).

Strategies for modelling and inversion

Following the clear exposition provided by Aster et al. (2005) we define task 1 to be: ‘find a set of physical parameters which describe a model, $m$. We do this by collecting a set of observations, the data, $d$. In most cases, $m$ and $d$ are vectors. We assume that there is a function, $G$, that relates the model values to the data values:

$$G(m) = d$$

$G$ is an operator that can take many forms, for example an ordinary or partial differential equations, or a linear or non-linear system of algebraic equations. In practice the data contain noise and we can imagine $d$ to comprise some perfect set of noise-free data, $d_{true}$ plus a noise component, $\eta$.

$$d = G(m_{true}) + \eta$$

It is not desirable to fit our model to the noise, although it is mathematically possible: $m$ can be influenced strongly by even a small $\eta$. In addition, there may be many models aside from $m_{true}$ that fit the data $d_{true}$: We define the forward problem as calculating $d$, given $m$ with the assumption that the true answer lies within the chosen parameterisation of the model $m$. Often this corresponds to finding a set of modelled or simulated data, such as travel times, given an initial test model. Our task is an inverse problem: find a model $m$, given $d$. Many geophysical applications use a finite number of discrete data points, and the model is related to the data (exactly, or as an approximation) through $G$ taking the form of a set of linear equations. We are now solving a matrix equation.

$$d = Gm$$

This formulation is broader than it may at first seem as it may also be used to represent ‘weakly’ non-linear problems, where the data vector, $d$, becomes the perturbations in the data caused by a perturbation, $m$, in a model about some chosen reference model, $m_{ref}$.

**Deterministic methods**

Relating $m$ to $d$ through $G$, gives the sense of these procedures as deterministic. There is a pre-defined physical relationship between $m$ and $d$. For example, synthetic seismic travel time data values are determined by the values of the appropriate seismic wavespeed model parameters in $m$. A popular strategy for finding a best fitting solution is then to sum the square of the differences between the observed data values and the synthetic (modelled) data values. Subsequent iterations of the algorithm then seek to find the model which minimises this sum of squared residuals or some other objective function. As developers of deterministic modelling and inversion software know only too well, the necessary matrix inversion operations are usually unstable, hence many approaches to deterministic inversion exist (Rawlinson and Sambridge, 2003; Aster et al., 2005) and underlie most of the geophysical modelling software currently in use. At the present time, we advocate the continued use of these methods alongside one or more of the alternate strategies that follow.

Note that, although we refer to ‘deterministic’ and ‘alternate’ strategies, there is often a deterministic component inherent in some part of an alternate strategy: multiple model data inference strategies increase opportunity by suggesting a number of models that are appropriate to the data. A discussion of assurance and opportunity is given in the main text.
Innovative data inference

between data and model are either not available or are of little use due to the problem being strongly non-linear. An example of an infrasound beam tuning problem (Kennett et al., 2003) is shown in Figure 3.

**Machine learning**

Machine learning strategies for inferring useful information from geophysical and geophysics-related data take a different, empirical approach. Machine learning algorithms use sets of related observations that do not necessarily have an obvious physical relationship between each other. They may be described as ‘disparate’ datasets, related only in that they are observed at the same point on the surface of the Earth. Predictive relationships are then extracted by means of the patterns occurring between the disparate observables. These techniques have been well used across the wider information technology community (Witten and Frank, 2005) and although there are a few examples of such strategies being applied to geoscience data inference problems, there is much scope for wider use in this area.

Machine learning is a 3 stage process: (i) the input data must be prepared to produce matching sets of observations, (ii) the predictive relationships must be deduced or induced, and (iii) the output must be evaluated. Input data preparation (stage i) can involve both geographic and numeric transformations, ensuring that observations from disparate data are spatially coincident. Forming the predictive relationship (stage ii) can take the form of supervised or unsupervised learning. Supervised learning is usually carried out using a training dataset which contains both predictive (input) and dependent (output) variables (a priori information). The machine learning scheme is then allowed to deduce the appropriate predictive relationship. Evaluation of the output (stage iii) of supervised learning schemes uses methods such as cross-validation (Witten and Frank, 2005). However, as we may wish to make predictions in areas where observations have not been made, cross validation or other estimates may not provide adequate insight into the ability of a deduced relationship to maximise opportunity when applied to unseen data. This is because the training data may be overly restrictive in its representation of the inference target and, as such, high assurance may be misleading. Conversely, unsupervised learning does not include a priori information and predictive relationships are induced from interactions between predictive variables. It is therefore difficult to provide adequate measures of assurance from the outputs of these machine learning schemes. Computational cost incurred when deducing or inducing predictive relationships must be also be considered. For example, a slight increase in assurance due to a large increase in the number of predictive variables may not be practical. Development of these methods in the context of geoscience data inference problems is the subject of ongoing research by the authors.

**Model parameter sampling**

Data inference approaches that use model parameter sampling produce a solution that consists of a probability distribution for each of the model parameters, rather than a single model. Those probability distributions are found by sampling the multi-dimensional posterior model space and a best fit model is constructed at the end of the sampling process in full appreciation of the probability of the occurrence and value of each model parameter. To clarify, the modelling process results, first and foremost, in a full set of probabilistic information which is then used to construct the best fit model. This
approach falls into the category of ‘Bayesian’ techniques. Such techniques have met with failure in past usage in exploration applications, and consequent criticism, owing to the lack of understanding of the underlying fundamentals, but used with insight, they facilitate extremely well founded, high opportunity algorithms which are now sufficiently efficient to be of use in a desk-top computing environment. A recent implementation of this approach, illustrated using a geoscience dataset, was developed by Bodin et al. (2009). They used a Monte Carlo approach to sample the solution space, guided by previous samples using a Markov Chain (MCMC). This work includes an important innovation: that the number of model parameters is allowed to vary. Hence, the model parameterisation is flexible and adapts in the course of the data inference.

A further innovation is that uncertainty in the data need not be known in advance (Bodin, 2010; Sambridge et al., 2010). The data uncertainty is parameterised in the form of one hyperparameter for each of the datasets being analysed together.

Fig. 3. An example illustrating a model parameter search using the Neighbourhood Algorithm. (a) Here, a single new random walk is carried out within the neighbourhood (black lines) of the previous best fitting set of parameters. After a few iterations sampling, while locally random, is concentrated in the region of parameter space where models provide a best fit to the data. (b) A misfit surface arising in an infrasound beam tuning problem (Kennett et al., 2003) and (c) the samples produced by the Neighbourhood Algorithm sampling this misfit surface. The peak of the function is located by the algorithm and the complete set of samples can be used for probabilistic uncertainty assessment (Sambridge, 1999a).
The algorithm solves for these values in the same way as for the other model parameters, i.e. by providing a posterior distribution for each. This is made possible by a Hierarchical Bayes regression formulation, HB-MCMC (Bodin, 2010; Reading et al., 2010). In the example we show here (Figure 4a), we investigate change points in wireline data using four logs. It is worth re-stating at this point, that we are not necessarily intending to replace existing wireline software. We find the probability of there being a change in layer character from the patterns inherent in the data and hence allow high opportunity in the data inference. Importantly, Figure 4b shows the relative probability of change points existing at different depths. Some depths are highly probable and tightly constrained, some are also highly probable, but their depths are less tightly constrained.

**Different strategies play different roles**

**Deterministic** strategies in this discussion are those for which a single ‘answer’ is found by matching observed to synthetic data where the synthetic data are calculated by the action of an operator, representing an underlying physical process, on a model defined by a set of parameters. The model parameters are modified to find a best fitting model. The advantage of these strategies is that very large and complex models can be accommodated. Where the data inference task is associated with a large amount of well known geological knowledge (for example, borehole intersections), this knowledge may be incorporated at a high level of detail.

In contrast, we define **alternate** strategies as those which provide a deeper insight into the solution space through a variety of methodologies which consider a range of possible ‘answers’ or show the relative probabilities of various features of the model. These might be provided through the calculation of a large number of models and a subsequent appraisal of the ensemble of possible answers (multiple model ensembles), or by an evaluation of the output (machine learning) or through the statistics of the model parameter space sampling (model parameter sampling).

A frequently misunderstood point in the presentation of alternate strategies is their usage. We do not wish to imply that (for example) a model parameter sampling approach with a comprehensive display of relative probabilities for many tens of thousands of parameters is appropriate in every case. Rather we suggest that, if a constrained, deterministic approach is the primary geophysical modelling strategy for a task, then a simplified version of the model, or particular aspects of the model is/are tested using an alternate strategy to more comprehensively investigate the solution space and quantitatively compare the extent to which important features of the model are constrained by the data. In addition, probabilistic strategies can be used to relax assumptions made in deterministic algorithms. For example, deterministic approaches often require decisions by the user as to the type of parameterisation used (i.e. definition of the model parameters, $m$) and knowledge of the statistics of the noise in the data. In the past ten years, probabilistic algorithms have been developed which allow the model parameterisation and the level of data noise to become part of the inference problem and be constrained by the data. Ultimately, each strategy has its advantages and weaknesses.

**A vision for effective exploration**

We envision procedures for effective exploration will make use of good deterministic software to produce constrained inversions with suitable error bounds. As well as this approach, we suggest the use of one of the alternate strategies outlined in this article, the investigation of multiple model ensembles, a machine learning approach or a model parameter sampling approach, to enable an increased or maximum opportunity appraisal of the

---

**Fig. 4.** An illustration of results from a model parameter sampling data inference example using the HB-MCMC implementation described in the text (Reading et al., 2010). Int. = interval, ar. = array. (a) Input data are points from a wireline log through varied lithologies (blue dots). The final result (see main text for explanation of how this is constructed) is shown by the red solid line with the green dotted lines giving the confidence limits on the final result. (b) The relative probability of change points existing at different depths is indicated by the number of models showing a change point at that depth. While the final result is given as a single ‘answer’ it is clear from the probability plot that the existence and/or location of some change points is better constrained by the data than for other change points.
solution space to be made. The alternate approach will be run as an additional inversion and may be conducted on a subset of model parameters or on a simplified model structure. There would be a requirement for this work to be conducted by skilled practitioners with insight into the theory underlying the alternate methods. In this way, we can make use of the advantages of more than one approach.

In closing, we anticipate that as geophysical practitioners become more accustomed to handling output from alternate approaches, these techniques will begin to take the place of purely deterministic methods as their advantages become more widely appreciated.

Summary

We advocate the use of a wider range of strategies for data inference in geophysics. Adding the use of an alternate strategy or strategies to the processing of geophysical data has the following benefits:

1. Alternative geological scenarios are highlighted which are consistent with the geophysical data: increasing the geological information that we extract.
2. Well constrained and poorly constrained parts of the model are clearly presented.
3. Quantitative information is provided that aids the explanation of the model and its limitations to colleagues.

Our ultimate aim is to enable geophysicists to use their data to infer geological information more effectively, thereby providing results with a combination of geological assurance, and enhanced target generation opportunity.

References


Witten, I.H., and Frank, E., 2005, Data Mining: Elsevier, San Francisco, USA.