Theory of model-based geophysical survey and experimental design

Part A—linear problems

ANDREW CURTIS, Schlumberger Cambridge Research, Cambridge, U.K.

Enormous sums of money are invested by industry and scientific funding agencies every year in seismic, well logging, electromagnetic, earthquake monitoring and microseismic surveys, and in laboratory-based experiments. For each survey or experiment a design process must first take place. An efficient design is usually a compromise—a suitable trade-off between information that is expected to be retrieved about a model of interest and the cost of data acquisition and processing. In some fields of geophysics, advanced methods from design theory are used, not only to optimize the survey design, but also to shift this entire trade-off relationship between information and cost. In others, either crude rules of thumb are used or, indeed, expected model information is not optimized at all.

This is the first part of a two-part tutorial that provides a theoretical framework from the field of statistical experimental design (SED), within which model-based survey and experimental design problems and methods can be understood. Specifically, these two articles describe methods that are pertinent to the detection and inference of physical properties of rocks in the laboratory, or in the earth.

The choice of method to use when designing experiments depends greatly on how easily one can measure information. This in turn depends principally on whether the relationship between data that will be measured and model parameters of interest is approximately linear, or significantly nonlinear. Consequently, the first article focuses on the case where this relationship is approximately linear and the next (in next month's issue of *TLE*) deals with theory for nonlinear design.

The tutorial begins with an introduction to concepts in model-based SED theory. This is then extended to create a more general, theoretical framework of design for linear model-data relationships. The same techniques can be used to design model parameterizations that explain information contained in measured data optimally (with maximum resolution), and this is illustrated at the end of the article. Geophysical applications of the various techniques are described throughout the tutorial, drawing on examples in which the author has been involved over recent years. A discussion of profitable areas of future research and development in linear, model-based design theory is left to the end of the companion article in next month's *TLE*.

Introduction to SED. SED techniques have been used to find optimal designs in a variety of different geophysical areas: determining locations of seismometers to locate earthquakes with minimum uncertainty (Rabinowitz and Steinberg, 1990; Steinberg et al. 1995); locating receivers optimally within a well to locate induced microseismicity during production (Curtis et al., 2004); designing source/receiver geometries for acoustic tomography that optimally detects underwater velocity anomalies (Barth and Wunsch, 1990); designing surveys for electromagnetic tomography from the ground surface to constrain optimally the shallow subsurface conductivity structure (Maurer and Boerner, GJI, 1998; Maurer et al., 2000); designing the interrogation of human experts to obtain optimal information to condition geophysical surveys (Curtis and Wood, 2004); designing nonlinear AVO surveys (van den Berg et al., 2003); planning crosswell seismic tomography surveys that illuminate the inter-well structure optimally (Curtis, 1999; Curtis et al., 2004); updating shallow resistivity survey designs in real-time as new data, and hence new information are acquired (Stummer et al., 2004); creating seismic acquisition geometries that maximize resolution of the earth model (Gibson and Tzimeas, 2002).

This tutorial considers the case where we would like to perform an experiment to collect data **d** (seismic, electromagnetic, logs, core, etc.) to constrain some model of earth properties or architecture described by a vector **m**. Say we define a set of basis functions $\{\mathbf{B}_j(\mathbf{x}){:}j=1,...,P\}$ that describe elementary components of earth properties or architecture. Examples of such basis functions used in geophysics are rock properties in each of a set of mutually-exclusive spatial cells, discrete Fourier components over a finite band-width, scatterers of energy at a set of fixed locations, or statistical properties observed over a finite range of length scales. Possible models of the earth can then be expressed as:

$$\mathrm{M} = \sum_{j=1}^{P} m_j \, \mathbf{B}_j(\mathbf{x}) \tag{1}$$

The problem of estimating earth composition consists of estimating coefficients m_j.

Say data **d** are to be recorded using a survey design described by vector **S**. Vector **S** might describe, for example, locations and types of sources and receivers to be used, bandwidth, and data type (seismic, electromagnetic, logs). Clearly, expected uncertainties in the data can be controlled at least in part by varying the survey design **S** (e.g., by changing either the equipment used or the number of repeated measurements). However, as shown below, considering simple examples allows a more profound understanding of the effects of changing the design.

Let the set of surveys that can reasonably be carried out given logistical and financial constraints be Σ. Determining this set, or finding some of its members, often accounts for a large proportion of the work that goes into geophysical survey design. For example, in 3D land seismic surveys the logistical problems of deploying equipment and recording data are immense. Sophisticated modeling packages are required simply to calculate how to execute survey logistics such that any proposed source and receiver pattern is honored. This is necessary to estimate the cost of the proposed survey geometry to see whether the design, say **S**, is acceptable ($\mathbf{S} \in \Sigma$) or unacceptable ($\mathbf{S} \notin \Sigma$). However, despite the computational expense and effort required, it is important either that set Σ is completely defined, or at least that any survey design can either be assigned inside or outside of this set, since Σ bounds the space of designs that must be considered. The design problem then consists of finding $\mathbf{S} \in \Sigma$ such that information about the model is maximized, and such that any additional measures of cost (logistical, temporal, financial, effort) are minimized.

Let function $\mathbf{F}_S$ represent the relationship between the earth model and the data, such that data **d** that would be recorded if a model **m** was true are predicted by

$$\mathbf{d} = \mathbf{F}_S(\mathbf{m}) \tag{2}$$

The subscript in $\mathbf{F}_S$ indicates that the form of the model-data relationship (often referred to as the forward function) depends on the survey design **S.** This is clearly the case: if, for example, three nonrepeated measurements were collected instead of two, then in equation 2, $\mathbf{F}_S$ would have to be three- rather than two-dimensional. However, more subtle variations in $\mathbf{F}_S$ are possible by changing the survey design, and indeed designing an appropriate $\mathbf{F}_S$ is the essence of most model-based survey or experimental design techniques. Whereas, postsurvey, we are interested in translating information in recorded data **d** into information on model **m** (usually an inverse problem), prior to data collection we vary the survey design in order to change function $\mathbf{F}_S$ such that we *expect* most information about **m** to be gleaned from the data that *are likely to be* measured in the survey. As such, survey or experimental design is really a macro-optimization problem: optimizing the inverse problem (equation 2) to be solved postsurvey, while simultaneously respecting cost constraints.

Figure 1 shows an example of how model and data information are related in a one-dimensional, linear problem. Say we would like to constrain the slowness of a medium that is assumed to be homogeneous. We could send a seismic wave from a source to a receiver at distance x from the source, and measure the traveltime datum d (Figure 1d, experiment $\mathbf{S}_1$). This is related to the slowness m by a linear, real function $d = F_S(m)$ where $F_S(m) = xm$, as shown schematically in Figure 1a. Say the datum is recorded with uniform uncertainties $d \pm e$ represented by the vertical length of the red shaded region (that is, there is an equal probability that noiseless data would lie within any subregion of fixed length within the red area). Corresponding constraints on the slowness model m are found by projecting the data uncertainties through the function F_S, producing uniform uncertainties on m spanning the blue region on the model axis. This final uncertainty is the result of the traveltime inverse problem for slowness in this simple experiment. Notice that in such linear problems, the width of the uniform model uncertainty remains the same no matter what value of the datum is recorded, as long as the datum uncertainty remains constant.

There are two common ways to redesign a survey. First, say we redesign the experiment to use a longer source-receiver distance x, then repeat the experiment (Figure 1d, experiment $\mathbf{S}_2$). In this case, the function $F_S(m)$ is steeper than in experiment $\mathbf{S}_1$, and assuming the traveltime datum is equally accurate to that in experiment $\mathbf{S}_1$, the corresponding uniform uncertainties in d and m are shown in Figure 1b. For the same data uncertainties (recording effort) as in the previous experiment, we obtain smaller uncertainties on the model m. Redesigning the survey in this case increased postexperimental information about the model, and this was directly a consequence of making the function $F_S(m)$ steeper.

A second way that one typically tries to obtain more information about the model is to repeat data measurements—e.g., increasing fold in seismic experiments (Figure 1d, experiment $\mathbf{S}_3$. In this way, the standard deviation on the datum (hence, width e above) is reduced compared to the nonrepeated case in Figure 1a, and this in turn projects into reduced uncertainty in the model parameters (Figure 1c).

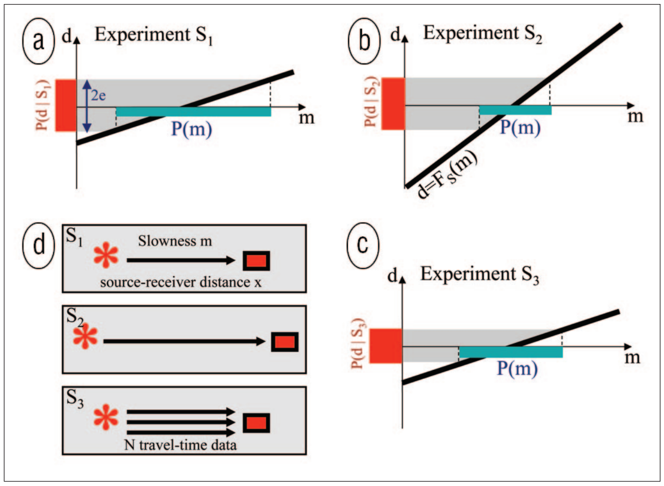

Figure 1. *(a), (b) and (c) show relationships between uncertainties in model **m** and in data **d** for experiments $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{S}_3$, respectively, shown in (d). In (a), (b) and (c) measured data uncertainties (red) are back-projected through the forward function $F_S(m)$ (black line) to produce the postexperimental model parameter uncertainty (blue).*

However, notice that none of the axes in Figure 1 have an absolute scale; the only length scales marked on each axis are the ranges of the uniform probability distributions. We could therefore rescale the data axis in Figure 1c to make the data uncertainties span the same length as that in Figure 1b. On that scale, the forward function in Figure 1c will look steeper, more similar to that in Figure 1b. Hence, both of the above ways to improve survey designs can be thought of as converting forward functions from having relatively low-gradients as shown in Figure 1a to relatively high gradients in Figure 1b, relative to the data uncertainties.

This example characterizes many design methods that increase postexperimental information. Examples include increasing fold in seismic surveys, repeating identical shots several times to reduce measurement uncertainty, and redesigning well logging tools to increase accuracy or sensitivity. In the next section it is shown how the concepts in Figure 1 generalize to higher-dimensional problems.

Maximizing expected model information is usually performed under cost and logistical constraints Σ, or may even be achieved while simultaneously attempting to minimize a cost functional. Resulting survey designs then usually represent a trade-off between increasing information about the model and reducing cost. The rest of this tutorial will concentrate on theory for maximizing model information. It will be assumed that a trade-off against cost can be affected either by using cost constraints imposed through set Σ defined above, or by using obvious cost functions or penalties attached to more "expensive" designs. Maurer and Boerner (1998) provide an example of how to incorporate costs explicitly.

General linear problems. This section shows how the concepts introduced above can be extended to practical geophysical problems with larger numbers of model parameters. Consider the simple tomographic problem in Figure 2a. Say our objective is to obtain some estimate of variations in slowness within a medium bounded by the outer square. One approach would be to assume that slowness variations

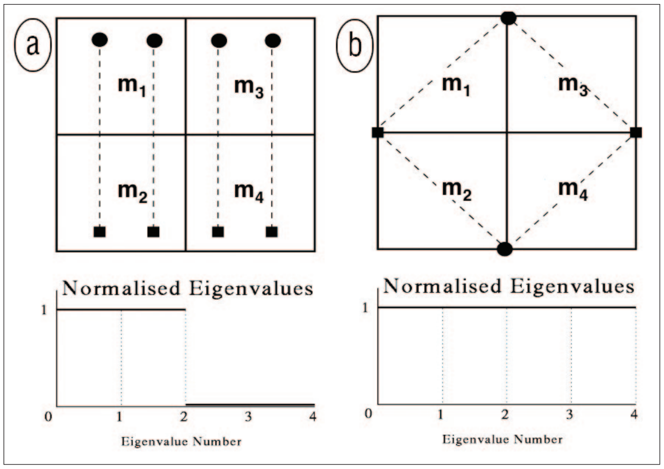

Figure 2. *Comparison of eigenvalue spectra from the inverse problem of obtaining (constant) slownesses* $\mathbf{m}_i$ *within each cell given traveltime data measured along dashed source-receiver ray paths. Twice as many independent pieces of information are available with the raypath geometry using experimental design (b) than design (a), and this is reflected in the corresponding eigenvalue spectra.*

can be represented by cells of approximately homogeneous slowness, and to divide the medium into several such cells. In Figure 2a we have arbitrarily defined four square cells with slownesses m_1 *to* m_4 which form the $P = 4$ model parameters to be estimated in equation 1.

If we had four energy sources and four receivers, we might decide to design an experiment where traveltime data are collected along raypaths defined by the dashed lines in Figure 2a. Traveltime d_i on raypath i is related to slowness according to,

$$d_i = \int_{ray\,path\,i} slowness(\mathbf{u})\, d\mathbf{u} = \sum_{k=1}^{P} l_{ik} m_k \qquad (3)$$

where in the intregral $\mathbf{u}$ runs along the raypath, and l_{ik} is the length of raypath i in cell k. Hence, traveltimes from the two left raypaths constrain the sum of slownesses m_1+m_2. However, they do not constrain the difference m_1-m_2 because m_1 can be increased and m_2 decreased such that the sum remains constant and hence so do the traveltimes in equation 3 on the two left paths. Similarly the two right paths constrain the sum m_3+m_4 but not the difference m_3-m_4. Hence, using this design, from a total of four traveltime measurements we obtain two independent pieces of information about the model (the sums of slownesses) while two pieces of information about the model remain unknown (the slowness differences).

Although raypaths in general depend on the slowness structure itself leading to nonlinearity in equation 3, let us assume that the slownesses lie within a range such that the raypaths between source and receivers remain as shown in Figure 2. The inverse problem to be solved postexperiment (equation 3) can then be represented by a linearized system of equations,

$$\mathbf{d} = A_S\mathbf{m} \qquad (4)$$

where the ijth element of matrix A_S is,

$$[A_S]_{ij} = l_{ij} \qquad (5)$$

The solution to equation 4 is,

$$\mathbf{m} = [A_S^T A_S]^{-1}\, A_S^T \mathbf{d} \qquad (6)$$

If matrix A_S^T is singular then the matrix inverse can be replaced by a generalized inverse. Matrix A_S^T can be decomposed numerically into its eigenvalues and eigenvectors. In the survey design context, eigenvalues have the following property: positive eigenvalues correspond to independent pieces of information obtainable from the survey, and zero eigenvalues to pieces of information that are unobtainable, where each corresponding piece of information is some linear combination of the model parameters. Eigenvalues of A_S^T are always nonnegative, and the magnitude of each eigenvalue relates directly to how well each piece of information can be estimated from the data to be collected. The actual pieces of information related to each eigenvalue are exactly the eigenvectors of matrix A_S^T, and these pieces of information are always linearly independent (i.e., knowledge about one eigenvector tells us nothing about any other eigenvector).

As an example, the (normalized) eigenvalues of A_S^T for the above experiment (plotted in the lower part of Figure 2a) show two positive and two zero eigenvalues. The positive eigenvalues relate to the two linear combinations of model parameters (eigenvectors) that are obtainable from the experiment (sums of slownesses m_1+m_2 and m_3+m_4); the zero eigenvalues correspond to those combinations that can not be obtained (differences m_1-m_2 and m_3-m_4). Hence, this decomposition tells us immediately how many independent pieces of information can be obtained from the survey design, and exactly what these pieces of information are.

If we had put more thought into the problem, we might have designed the survey as shown in Figure 2b. The two sources each fire to the same two receivers, again resulting in traveltime measurements along four raypaths. In this case, however, each traveltime constrains the slowness in exactly one cell, and hence all four cell slownesses can be estimated. Correspondingly, there are four positive eigenvalues and each eigenvector is simply a single model parameter (m_1, m_2, m_3, and m_4).

Five important general points about survey design in linearized problems are demonstrated by the examples in Figure 2:

1) Notice that no actual traveltime data were necessary for this analysis. Only matrices A_S or $A_S^T A_S$ were analyzed, and these matrices depend only on the survey design, not on the actual data obtained during the survey. Hence, the analysis above can be carried out before any survey has been executed.
2) In both experiments, four traveltime measurements on nonrepeating paths were recorded, and hence the amount of information in the data is expected to be equal in both designs. However, half as many pieces of information about the model would be obtained using design A than using design B in Figure 2. From the point of view of constraining the model, we would usually prefer design B.
3) The eigenvalue and eigenvector decomposition of matrix $A_S^T A_S$ (or equivalently, the singular value decomposition of matrix A_S) analyzes exactly how many, and which pieces of information are expected to be resolved by a survey, and hence can be used to compare the quality of any set of candidate survey designs.
4) In design B, half as many sources and receivers were used as in design A, and yet twice as many pieces of information were obtained about the model space. Using survey design techniques, it is almost always possible to obtain more information, and at lower cost, compared to surveys designed heuristically (using rules of thumb).

5) Finally, notice that the raypath coverage (sum of the proportions of each ray in each cell) is identical in every cell for both designs A and B (every cell contains on average one complete raypath—two half raypaths in each cell in design A). Yet, twice as many pieces of information would be recovered using design B than design A. This illustrates that designing surveys using the criterion of maximizing raypath (or generally data) coverage does not necessarily maximize information about the model space; data coverage is a bad design criterion that generally should not be used for survey design if it can be avoided (Curtis and Snieder, 1997).

Linear design measures. Linear SED methods concern the maximization of *quality measures*. These are measures of the amount of information expected to be transferred from measured data into information about model parameters of interest, possibly traded off against the cost of data acquisition and processing. From the above discussion it is clear why quality measures are usually sensitive to the eigenvalues in linear problems, since these indicate how much information is transferred between data and model parameters. Let $\{\lambda_i : i=1,...,P\}$ be the P eigenvalues of matrix $A_S^T A_S$ in order of decreasing magnitude, i.e., λ_1 is the largest eigenvalue. Commonly used quality measures are:

$$\begin{aligned}
\Theta_0 &= \sum_{i=1}^{P} \frac{-1}{\lambda_i + \delta} \\
\Theta_1 &= \sum_{i=1}^{P} \lambda_i \qquad \left[= trace\,(A_S^T A_S)\right] \\
\Theta_2 &= \sum_{i=1}^{P} \frac{\lambda_i}{\lambda_1} \qquad \left[= \frac{1}{\hat{\lambda}_1} trace\,(A_S^T A_S)\right] \\
\Theta_3 &= \prod_{i=1}^{P} \lambda_i \qquad \left[= \det\,(A_S^T A_S)\right]
\end{aligned} \tag{7}$$

Three of the four quality measures above can be calculated for any particular survey design *without* explicitly calculating the eigenvalues of matrix $A_S^T A_S$(the squared singular values of matrix A_S). This reduces the computation required. Where available, mathematical tricks to calculate the quality measures are shown in square brackets in equation 7; for Θ_1 and Θ_3 it is only necessary to calculate the trace or determinant of matrix $A_S^T A_S$, respectively. For Θ_2 it is also necessary to estimate λ_1, the largest eigenvalue. The estimate, $\hat{\lambda}_1$, can be obtained using the power method (Curtis and Snieder, 1997; Curtis, 1999). To the best of my knowledge, Θ_0 requires calculation of the complete eigenvalue spectrum.

The measures above are sensitive to different properties of the eigenvalue spectrum. Θ_0 is relatively expensive to calculate, but is only sensitive to the magnitude of eigenvalues approximately around or above the magnitude of δ—hence, value δ can be set to create a lower threshold of sensitivity (based on expected data noise levels—Maurer and Boerner, GJI, 1998). If we plot the eigenvalues λ_1 on a graph as a function of eigenvalue number i, then measure Θ_1 is simply the area under the eigenvalue curve. Maximizing Θ_1 can therefore be achieved by increasing the largest eigenvalues at the expense of the small ones, and this is commonly what occurs with designs found using this measure. Hence, Θ_1 often gives extremely reliable constraints on relatively few pieces of information.

Measure Θ_2, on the other hand, only measures the area under the normalized eigenvalue curve (where the eigenvalues have all been rescaled by λ_1 such that the maximum normalized eigenvalue is one). The only way to increase this measure is to increase the value of smaller eigenvalues relative to the largest one, hence this measure gives a more even spread of constraints over a larger number of pieces of information than does Θ_1.

Measure Θ_3 is the most evenly sensitive to the magnitude of all eigenvalues. It also has a particular interpretation in the case of Gaussian errors: consider the case where the data uncertainties are Gaussian with uniform variance. Since the problem is linear, model parameter uncertainties will also be Gaussian. Then $(\Theta_3)^{-1}$ is the factor of contraction of data uncertainties as they are translated into the model space. Maximizing Θ_3 is therefore directly equivalent to minimizing expected postsurvey model parameter uncertainties.

You may be wondering how maximizing eigenvalue-based measures relates to the schematic linear problem illustrated in Figure 1. There is a simple relationship: in that experiment, the eigenvalue is the gradient of $F_S(m)$ squared. Maximizing the gradient of this function is therefore equivalent to maximizing any of the measures Θ_0, Θ_1 or Θ_3 in equation 7.

It is often the case that we are primarily interested in obtaining information on a subspace of the model space (for example, properties of the reservoir interval are usually of greater interest than those of the overburden). In this case we may design surveys to focus information only on that subspace by using a quality measure that is most sensitive to such information. Let $\{\mathbf{e}_i : i=1,...,P\}$ be the eigenvectors of matrix $A_S^T A_S$ such that eigenvector $\mathbf{e}_i$ corresponds to eigenvalue λ_i. Also, let $\{\mathbf{v}_i : i=1,...,Q\}$ where $Q < P$ be a basis for the subspace of interest (for example, this could be a subset of model basis vectors $\mathbf{B}_j(x)$ in equation 1). Then focused quality measures are,

$$\begin{aligned}
\Theta_4 &= \sum_{i=1}^{P} \sum_{j=1}^{Q} \lambda_i^2 \, (\mathbf{e}_i \cdot \mathbf{v}_j)^2 \\
\Theta_5 &= \sum_{i=1}^{P} \frac{\sum_{j=1}^{Q} \lambda_i^2 \, (\mathbf{e}_i \cdot \mathbf{v}_j)^2}{\sum_{j=1}^{Q} \lambda_1^2 \, (\mathbf{e}_1 \cdot \mathbf{v}_j)^2}
\end{aligned} \tag{8}$$

Measures Θ_4 and Θ_5 are the focused equivalents of Θ_1 and Θ_2, respectively, except that each term in the summation is squared for computational convenience (see below). The extra summation over dot products with each subspace basis vector $\mathbf{v}_j$ ensures that only information in the subspace of interest counts towards the original summation over i in equation 7. Measure Θ_5 is a normalized version of Θ_4.

In 1999, I showed that Θ_4 is related to more usual measures of resolution and covariance within the model subspace of interest. In fact, it differs only in the power to which the eigenvalues are raised in equation 8. However, it was also shown that there are mathematical tricks to calculate Θ_4 without calculating eigenvalues and eigenvectors (as would be required in order to calculate resolution and covariance); hence the measures in equation 8 are computationally more efficient to use. Other measures based on what is called the reduced information matrix are used in statistical literature (e.g., Atkinson and Donev, 1992; references therein), but again my 1999 article shows that in large geophysical problems calculation of this matrix is also likely to be less efficient than using the above measures.

Examples: Focused and unfocused crosswell tomography. I presented examples of designing crosswell tomography surveys by maximizing most of these measures in two 1999

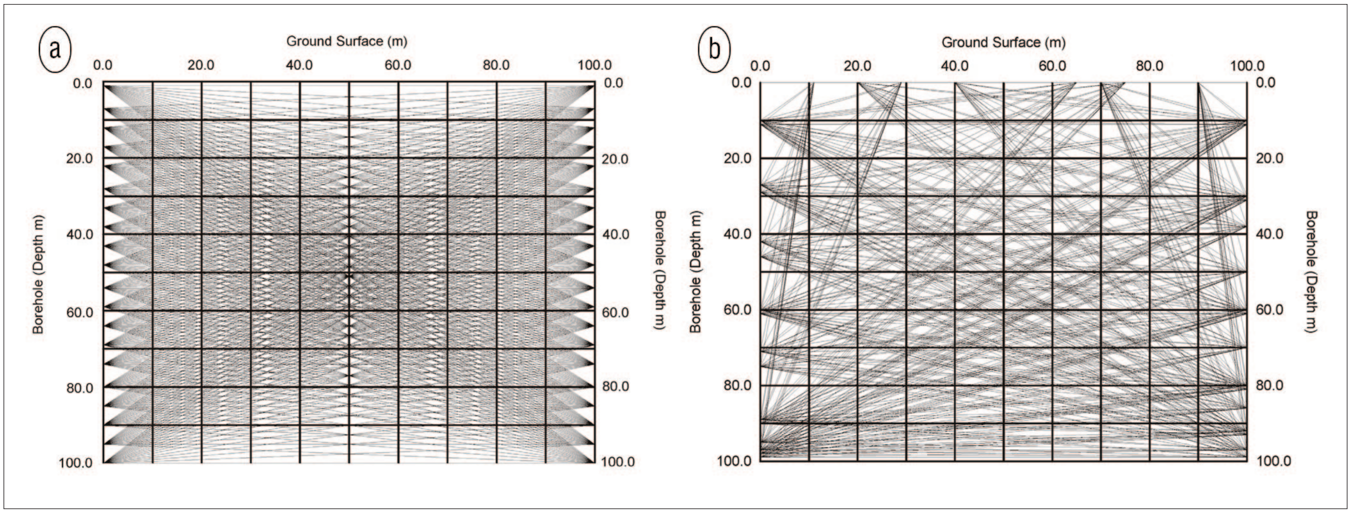

Figure 3. A crosswell tomography experiment. Left and right sides of each plot represent wells, the top represents the ground surface. The slowness model parameterisation consists of regularly spaced, homogeneous, square cells. Thin lines show predicted raypaths between 20 seismic sources and 20 receivers (hence 400 raypaths). The velocity model is assumed to be approximately constant over this 100 m interval. (a) Raypaths when sources and receivers are distributed evenly within each well. (b) An optimized geometry of sources and receivers.

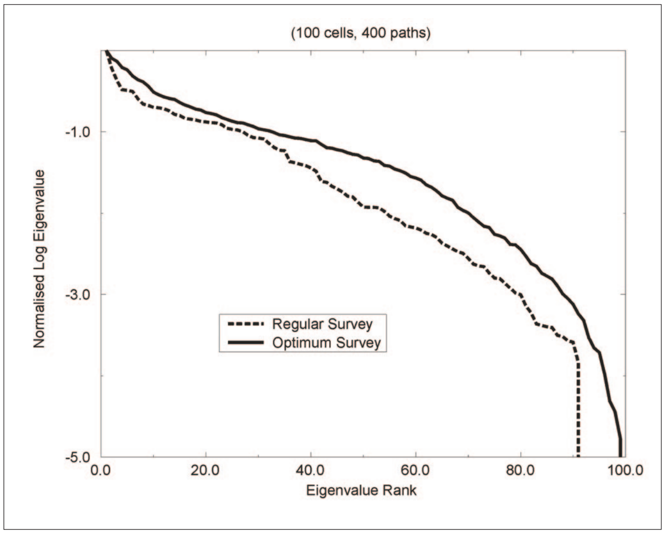

Figure 4. Logarithm of eigenvalue spectra of the tomographic inverse problems corresponding to the experimental designs depicted in Figure 3a (dashed line) and 3b (solid line), normalized by the largest eigenvalue in each case.

articles. Maurer and Boerner, and Maurer et al. (2000) use measure Θ_0 to design electromagnetic surveys for shallow conductivity tomography. Rabinowitz and Steinberg (1990) maximized measure Θ_0 to design an optimal earthquake-monitoring network. Barth and Wunsch (1990) used a variation on this theme in which they designed ship-based tomography geometries by maximizing the value of the *k*th largest eigenvalue for some prespecified constant *k*.

Figure 3 shows an example of designing a simple crosswell tomography survey by maximizing Θ_2. The left and right boundaries represent vertical wells; the top boundary is the ground surface. Twenty seismic sources and receivers are available (hence, 400 traveltime measurements along source-receiver paths), and can be placed in either well or on the ground surface. The background velocity is assumed constant, and the aim is to find the design that maximizes information across the 100 cells spanning the interwell space.

Figure 3a shows a design using regularly spaced sources and receivers down each well. This is the geometry most commonly adopted in crosswell studies. Fans of raypaths are shown emanating from each source and receiver location (sources and receivers are not discriminated or even marked in the figure since only raypath geometries are important in this design problem—equation 3). The dashed line in Figure 4 shows the normalized log eigenvalue spectrum for this geometry.

Using a genetic algorithm tuned towards optimization (Sambridge and Drijkoningen, 1992), sources and receivers were then re-arranged within the wells or on the ground surface in order to maximize Θ_2, the area under the normalized eigenvalue spectrum. Figure 3b shows the best geometry found, and the solid line in Figure 4 shows the corresponding log eigenvalue spectrum. This spectrum is clearly improved compared to the regular design, and indeed places (weak) constraints on nine additional pieces of information (eigenvalues 91-99).

The revised design has several intuitive features. First, four sources and four receivers were placed on the ground surface. This makes sense since it both provides shorter raypaths and increases the aperture of raypaths passing through many cells in the model (angular coverage is known to be generally required for high resolution). Second, the density of sources and receivers increases towards the base of each well. This also makes sense: cells lower-center in the interwell space are likely to be most poorly constrained tomographically as they can have no short raypaths and no significant aperture of raypath angles passing through them. Raypaths between sources and receivers towards the base of the wells provide the shortest possible such paths, and the dense source and receiver spacing provides the largest possible aperture through such cells.

Hence, although the design algorithms explained above are mathematical in nature, the designs produced are intuitively understandable. However, it was not possible to design this survey using intuition alone: exactly how densely would receivers be spaced towards the base of the well? Why four receivers on the surface and not five or six? Such quantitative questions escape resolution using intuition, and this is why the quantitative techniques in this paper are so important in survey design studies.

An example of focused crosswell tomography design is given in Figures 5 and 6. Again, left and right sides of plots in Figure 5 are two wells, but this time the top does not represent the ground surface, hence sources and receivers can only be placed within the wells. Only six sources and receivers are available (hence, 36 source-receiver traveltime

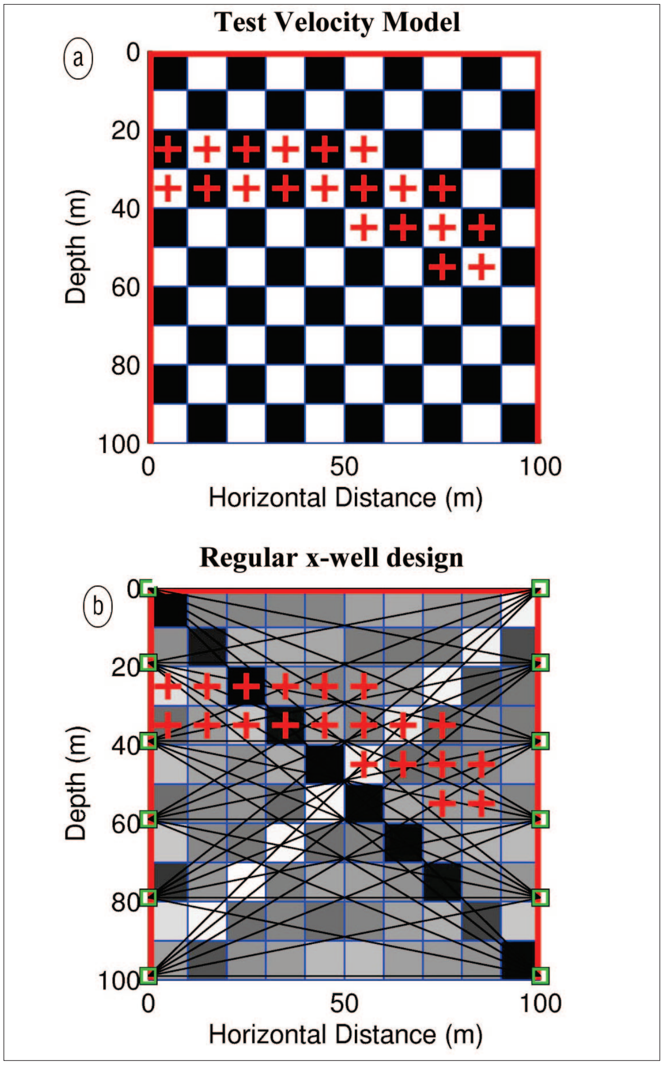

Figure 5. Left and right sides of each plot represent wells; the top does not represent the ground surface. Otherwise the plots are similar to those in Figure 3 with the following additional exceptions: no sources or receivers are shown in (a), whereas six of each are shown, regularly distributed, in (b). The checkerboard in (a) represents a model consisting of positive and negative slowness anomalies in consecutive cells, used to test designs; shading in (b) represents the best reconstruction of this checkerboard using the design shown. Red crosses mark cells of particular interest.

measurements), and the goal is to focus tomographic information on the 20 interwell cells marked with red crosses.

Figure 5a shows a checkerboard of slowness perturbations (dimensionless). This is a test structure in the interwell space: traveltime data measured on source-receiver paths will be weighted averages of the checkerboard slownesses in cells traversed by each path according to equations 4 and 5. The goal is to construct a design such that the traveltime measurements are sufficient to reconstruct this checkerboard pattern tomographically using the solution given in equation 6. Figure 5b shows the poor reconstruction of the checkerboard in the interwell space that is possible using the conventional design of regularly-spaced sources and receivers—clearly this design is not suitable for retrieving information on the 20 cells of interest.

Again, a genetic algorithm was used to reposition the sources and receivers, but this time a weighted average of measures Θ_0 (10% weight) and Θ_4 (90% weight) was maximized. The basis functions $\mathbf{v}_j$ in Θ_4 simply comprised unit vectors in the cells of interest. Hence the survey design procedure primarily optimized information in these cells. The best design found is shown in Figure 6a. This design is irregular, not at all intuitive, and hence could only have been found using mathematical procedures such as those explained herein. The increase in information about the cells of interest compared to the regular design is clear: the checkerboard in those cells is recovered almost everywhere (compare Figure 5b).

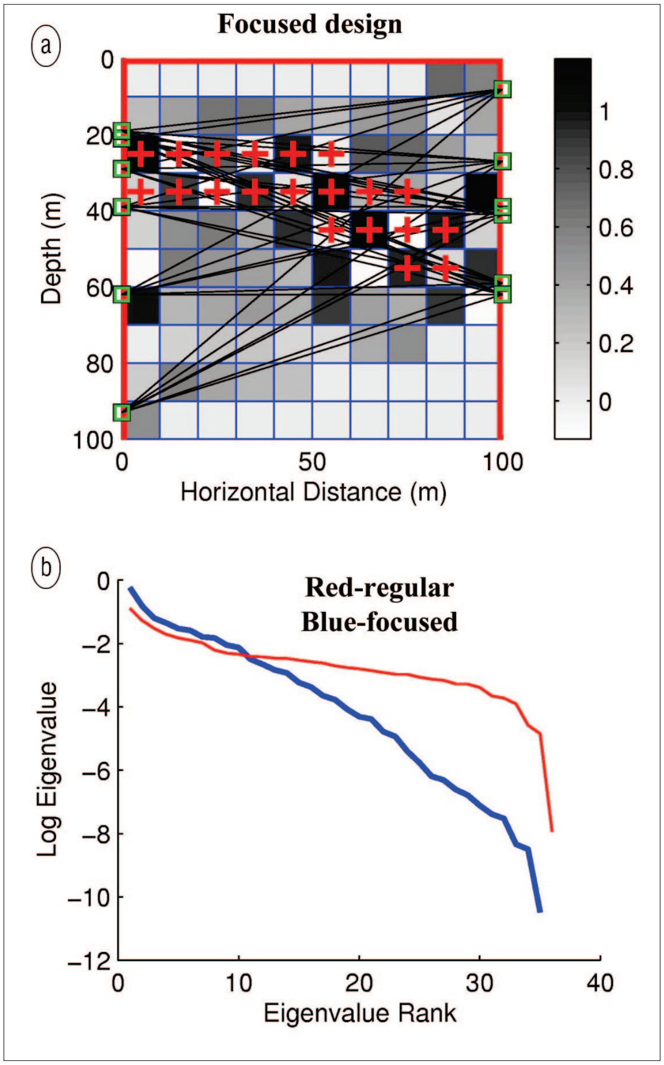

Figure 6. (a) is similar to Figure 5b, but with an optimized geometry of sources and receivers in order to focus on the cells of particular interest (red crosses). (b) shows eigenvalue spectra for the design in (a), blue, and the design in Figure 5b, red.

Figure 6b shows the eigenvalue spectra of the regular and the focused surveys. The focused survey gives more information about the eigenvectors with largest eigenvalues since these values are larger than those for the regular surveys; this trades off with a reduction in the magnitude of the smallest eigenvalues. What is happening is that the design places all of the largest eigenvectors dominantly in the subspace of interest, and then maximizes the information about those eigenvectors at the expense of the others (since the latter mainly represent information about slownesses elsewhere in the model).

An alternative approach. A different approach (Curtis et al., 2004) uses the fact that singularity (zero eigenvalues of matrix $A_S^T A_S$, or equivalently zero singular values of matrix A_S), occurs only if there are rows of A_S that are linear combinations of other rows (assuming A_S has more rows than columns). Equation 4 shows that each row of A_S corresponds

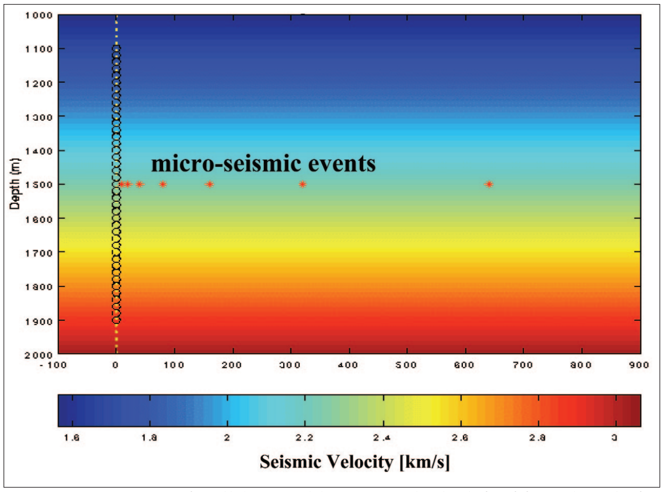

Figure 7. *Location of well (vertical line towards the left of figure) and of representative micro-seismic events (stars) in a reservoir at 1500 m depth. Circles in the well mark potential positions for seismic receivers. The color scale shows a seismic velocity gradient with depth.*

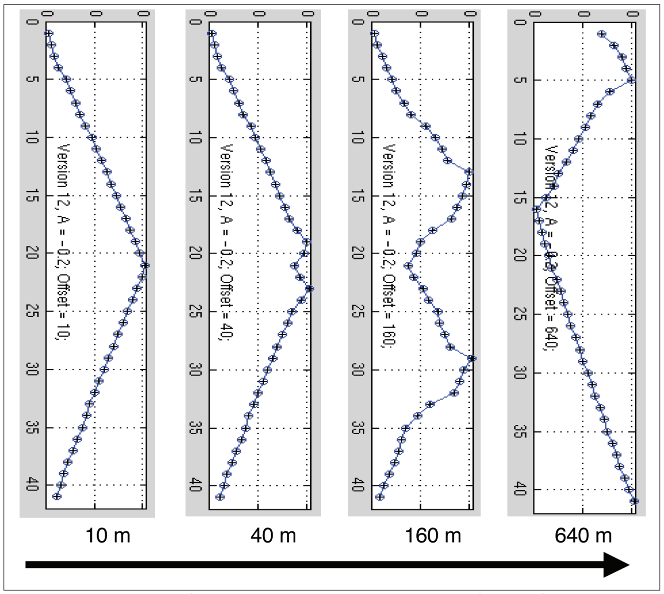

Figure 8. *Ranking of receivers in Figure 7 in terms of the information provided about event locations at increasing offset from the well (event offset is shown beneath each plot). Each plot shows the receiver that is removed first on the left, and the one removed last on the right. Intervening positions show the order in which all receivers were removed from the set of receiver locations shown in Figure 7.*

to a single datum. Hence, singularity of A_S implies that there must be redundancy in the data set.

Sabatier (1977) noticed that this property could be used to reduce the size of large data sets by removing data that are effectively repetitions of combinations of the other data, and Curtis et al. used this idea to design surveys. The algorithm used was the following: begin with a matrix A_S that is created for a design **S** that includes all possible source and receiver locations and hence all possible data that could be recorded in **d**. Then, calculate the dot product of each row of A_S with every other row of A_S and sum them (weighted by expected data uncertainties, and if desired by weights to focus on a model subspace). The result is a measure of the (weighted) angle between each row and the space spanned by all other rows, treating each row as a vector. If the row corresponds to a datum that is a linear combination of other rows then it will lie completely within that space, and the angle will be zero. If the datum is truly adding new information, the angle will be nonzero. Hence, the magnitude of the angle can be used as a quality measure for each row of A_S, and hence each datum in **d**.

The algorithm proceeds by removing from **S** the source and/or receiver corresponding to the datum d_i whose row makes the smallest angle with other rows, and this in turn removes the row from A_S. It then updates all dot products to reflect this removal, then removes the datum with the next smallest angle; it repeats this process until an acceptable (affordable) number of sources and receivers are left within the design **S**. Stummer et al. (2004) also used a similar criterion (together with one based on resolution) to *add* data that provide the most additional information to an initial base survey.

Example: Microseismic location. Figure 7 shows a vertical well with 40 possible seismic receiver locations within it. The design problem is to choose the subset of these locations that will optimally locate microseismic events occurring in a "reservoir" at a depth of 1500 m due, for example, to production-related fluid pressure changes, where the events occur at various offsets from the well. There is a vertical gradient in seismic velocity shown in the figure, and some attenuation in the medium (the exact attenuation structure is given in Curtis et al.).

First, the matrix A_S was calculated for all 40 possible receiver locations in the well, and for a model **m** consisting only of the event at 10 m from the well. The algorithm above was used to remove redundant receivers sequentially from the total possible array. Figure 8 (left) shows the order in which the receivers were removed (working from left to right in that plot). Reading this plot from right to left, we see that if we were to use a design consisting of only two receivers, we would choose the central receiver (at the depth of the seismicity) plus the one just below that level.

This process was repeated for events at 40 m, 160 m, and 640 m offset from the well, with results shown in successive plots in Figure 8. As the array is focused on events at increasing offsets, two patterns emerge: first, the optimal array of two receivers spreads out in the well. This increases the otherwise decreasing angle of aperture between the two event-receiver paths, which makes sense intuitively since better event locations require larger angles of aperture. The actual distance of the two receivers from the reservoir interval is a trade-off between increasing this aperture and increasing data uncertainty due to attenuation as event-receiver paths increase in length. Second, asymmetry about the reservoir depth emerges. This also makes sense because of the velocity gradient, which bends raypaths between event and receiver.

Again, this example shows that the design algorithms introduced herein produce designs that are intuitively reasonable, but they could not have been designed by intuition alone. In this case, intuition would not allow one to define how asymmetric the design should be around the reservoir, or how one should trade off increasing the angle of aperture at the event with increasing data uncertainty due to attenuation as receivers further away from the reservoir are used.

Optimal model parameterization. Before we wrap up this part of the tutorial, notice that there is an alternative method to increase the amount of information about the model parameters from that obtainable in Figures 2a or 3a: in addition to changing the data collected (Figure 9b), one could change the model parameterization (Figure 9c). In the original experimental goals given above we wished to obtain information about slowness variations across a region. We chose the

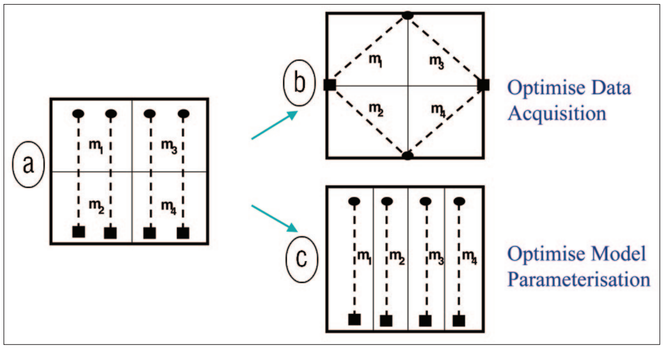

Figure 9. *Similar to Figure 2, but without the eigenvalue spectra shown. (c) shows a redesigned model parameterization. Design (b) and revised parameterization (c) both result in twice as many pieces of information about the model parameters as design (a).*

parameterization of four square cells shown in Figure 2, but this was arbitrary. Figure 9c shows an alternative cell geometry (model parameterization) in which each traveltime datum in the original experiment constrains exactly one slowness parameter.

Thus, both the survey design in Figure 9b and the alternative parameterization in Figure 9c result in four positive eigenvalues, and given only our experimental goals above, neither option is to be preferred. In fact, it is possible to design optimal model parameterizations in exactly the same way as designing surveys, by changing the parameterization to maximize the quality measures in equation 7. Curtis and Snieder (1997) demonstrated this by designing optimal irregular cell geometries to capture information in an irregular, asymmetric crosswell tomography survey.

Again, the fractional raypath coverage in every cell in all three cases in Figures 9a-c is one, yet cases b and c provide twice as many pieces of information as case a. In addition to being a poor criterion for designing surveys (see above), raypath coverage is thus demonstrated to be a poor criterion for designing model parameterizations.

Discussion. In this article I have attempted to provide a tutorial on the theory of linear, model-based, statistical experimental design (SED) techniques as they have been, and could be applied in geophysics. Unfortunately a tutorial of this length cannot provide comprehensive information, so the reader is recommended to look at the references below for sufficient additional information to implement any of these techniques in practice.

In many areas of geophysics, the majority of practitioners carry out no such sophisticated design exercises as are proposed and illustrated herein. Most use standard designs, which are often in some way based around the principle of regular sampling. While this may seem intuitively appealing, the examples presented earlier have shown that, at least in some situations, such strategies waste both money and information: adapting survey designs for the specific goals of each experiment can both reduce cost and increase acquired information. A little forethought and prior effort can go a long way.

Linear design theory is well understood and has been applied in many different fields of study. In situations where such theory is applicable it can be expected that significant improvements can be made to survey or experimental results by performing some formal design optimization process as described above.

In situations in which the model-data relationship is nonlinear, it is often possible to linearize this relationship without undue loss of accuracy if sufficient prior information about the model exists. Such situations may occur in monitoring problems for example, where baseline surveys and full processing have already been undertaken and provide much information about the earth model; subsequent surveys need only be designed to monitor changes that occur over time to that initial model, and if these changes are relatively small then linear approximations may be justified.

With very nonlinear model-data relationships, the techniques in Part A of this tutorial will not necessarily provide good results, and may indeed produce worse results than using tried and tested heuristics. For such cases the reader is referred to Part B of this tutorial, in next month's *TLE*.

Suggested reading. *Optimum Experimental Designs* by Atkinson and Donev (Clarendon Press, Oxford, 1992). "Oceanographic experiment design by simulated annealing" by Barth and Wunsch (*Journal of the Physics of the Ocean*, 1990). "Optimal experiment design: Cross-borehole tomographic examples" by Curtis (*Geophysical Journal International*, 1999). "Optimal design of focussed experiments and surveys" by Curtis (*Geophysical Journal International*, 1999). "A deterministic algorithm for experimental design applied to tomographic and microseismic monitoring surveys" by Curtis et al. (*Geophysical Journal International*, 2004. A JAVA version of the design algorithm is available at: *http://alomax.free.fr/projects/expdesign*). "Reconditioning inverse problems using the genetic algorithm and revised parameterization" by Curtis and Snieder (*Geophysics*, 1997). "Optimal elicitation of probabilistic information from experts" by Curtis and Wood (in Geological Prior Information, *Geological Society of London Special Publication*, in press). "Quantitative measures of image resolution for seismic survey design" by Gibson and Tzimeas (*Geophysics*, 2002). "Optimized and robust experimental design: a non-linear application to EM sounding" by Maurer and Boerner (*Geophysical Journal International*, 1998). "Geophysical survey design: get the most for the least!" by Maurer and Boerner (SEG 1998 *Expanded Abstracts*). "Design strategies for electromagnetic geophysical surveys" by Maurer et al. (*Inverse Problems*, 2000). "Optimal configuration of a seismographic network: a statistical approach" by Rabinowitz and Steinberg (*Bulletin of the Seismological Society of America*, 1990). "On geophysical inverse problems and constraints" by Sabatier (*Journal of Geophysics,* 1977). "Genetic algorithms in seismic wave-form inversion" by Sambridge and Drijkoningen (*Geophysical Journal International*, 1992). "Configuring a seismographic network for optimal monitoring of fault lines and multiple sources" by Steinberg et al. (*Bulletin of the Seismological Society of America*, 1995). "Experimental design: Electrical resistivity data sets that provide optimum subsurface information" by Stummer et al. (*Geophysics*, 2004).

Corresponding author: curtis@cambridge.oilfield.slb.com