Development of electrical resistivity imaging methods for geological and archaeological prospecting

Biwen Xu

DEVELOPMENT OF ELECTRICAL RESISTIVITY IMAGING METHODS FOR GEOLOGICAL AND ARCHAEOLOGICAL PROSPECTING

Biwen Xu, B.Sc., M.Sc.

A thesis submitted to the University of Durham for the degree of Doctor of Philosophy.

Department of Geological Sciences August 1993.

Declaration

The content of this thesis is the original work of the author and has not previously been submitted for a degree at this or any other university. Other people's work is acknowledged by reference.

Biwen Xu

Department of Geological Sciences

University of Durham

Copyright

CONTENTS

ACKNOWLEDGEMENTS

I am grateful to my supervisor, Dr M J Noel, for his supervision, including guidance on the research, his patience and consistent encouragement, help in field data collection and his constructive criticism and correction of my manuscript. I especially enjoyed many discussions which we had during the past three years which usually sparked many new ideas.

I would like to thank Professor A F Harding, Head of Department of Archaeology, and Professor R C Searle, Head of Department of Geological Sciences for providing research accommodation and for permitting me to use departmental facilities.

Thanks go to Dr Barber, Department of Medical Physics and Clinical Engineering, University of Sheffield, for his generosity in providing copies of research reports and publications and for useful advice at the initial stage of the research. The Associate Editor of Geophysical Prospecting and Dr R D Barker in University of Birmingham gave useful comments on part of the content of Chapter 2. Dr N Goulty kindly checked Chapter 5 and gave some helpful suggestions. To them I am grateful.

I wish to thank Dr Laurence Thistlewood for his help and advice on how to adapt to the new environment when I arrived at the University, and Dr Cathy Batt and Mr Duncan Hale for their kindness when we shared the AP-5 laboratory. Thanks also go to MSc students in Geological Sciences, Mr A Cushnaghan and Ms A F Evens, for their help in field data collection.

I thank my wife and still my best friend, Mei Chen, for her patience and support. There were a few evenings and weekends I didn't come home in time when I said I would. I thank my parents and my parents-in-law for their encouragement and for looking after my son during my studies in the UK.

The research project was supported by grants from the Science and Engineering Research Council and Royal Society, for whose generosity I am grateful. I also wish to thank Dr Roger Walker of Geoscan Research for providing the PA7 multi-switching cable system.

SUMMARY

The methods of two dimensional electrical resistivity image reconstruction using an equally spaced linear surface array have been developed and applied to investigate geological, archaeological and environmental problems. The techniques explored consist of data collection methods, reconstruction algorithms and the image presentation methods. A *complete* or *primary data set* which contains all linearly independent measurements of apparent resistivity is collected by multiplexing two, three or four electrode configurations on a linear array placed on the ground surface. A two dimensional resistivity image in the subsurface is then reconstructed from the measurements by the Born procedure or by the methods of regularised least squares and presented as grey scale or pseudo-colour images. The use of a complete data set has two advantages: (1) any other data sets obtainable on the linear array can be synthesised from it through a process of superposition; (2) it is likely to produce a better image than a conventional pseudo-section data set in the presence of noise. In comparison to the regularised non-linear least squares algorithms, the Born inversion is fast, taking several minutes on an ordinary PC computer if the sensitivity matrix is pre-inverted. It is also not so sensitive to the resistivity changes along the strike direction and therefore can be applied to image the cross-section of three dimensional structures. It is, however, a perturbation method with the ability only to reconstruct the relative variations of resistivities and is limited by its lack of resolution when resistivity structure is complicated or resistivity contrast is high. Also it is relatively sensitive to noise in the data. The algorithms of the regularised non-linear least squares plus various spatial smoothers normally produce better results than Born inversion if the initial model is not far away from the true solution and a proper degree of smoothness constraints is imposed on the image for the problem at hand. The initial model may be constructed from *a priori* knowledge or estimated from the corresponding apparent resistivity pseudo-section while the smoothness constraints are introduced by smoothness functions and regulated by the regularisation parameter λ. Experiments indicate that among the various choices of smoothing functions, the zero-order regularisation function plus weighted spatial averaging smoother seem to be the most efficient in maintaining the stability of the iteration while being able to have a relatively good resolution. Experiments indicate that the choice of regularisation λ is critical for the success of the regularised non-linear least squares algorithm. Although the experiments have demonstrated that good results may be obtained by choosing λ automatically either as a minimizer of data residual or generalised cross-over functions, or even as a constant value, in general it is context dependent and thus needs in practice to be selected by experiment.

The possibility of using genetic algorithms and neural network methods to reconstruct an electrical resistivity image is also explored. Initial numerical experiments suggest that although the genetic algorithms or neural network methods may be applied to process the data over 2D resistivity structures in the future, currently their use in practice is severely limited by the high cost of computation.

CHAPTER 1 INTRODUCTION

1.1 STATEMENT OF THE PROBLEM

During recent years there has been an increasing demand for new or improved methods for imaging subsurface structures for applications where conventional methods are inadequate. Examples of these applications are water pollution monitoring, site assessment for storage of nuclear and toxic wastes, enhanced oil recovery, mineral exploration, hydrogeological investigation, guidance for tunnel construction or mining and archaeological prospecting (see Table 1.1). In archaeology, detailed subsurface images are important for planning site conservation and excavation. In environmental engineering, the knowledge of detailed distribution or movement of pollutants is essential before any cleansing or remedial procedures can be applied. In selecting sites for the disposal of high-level nuclear and toxic chemical wastes, the subsurface characteristics of the site must be established with a high degree of confidence and here the level of spatial resolution desired is substantially greater than that commonly required for routine engineering investigations (Olsson et al., 1992).

Tomography, which has been so successful in medical diagnostics, is one such technique that has the promise to satisfy the demands mentioned above. Tomography is normally referred to as the reconstruction of an image from its line integrals. More broadly, it can also refer to any reconstruction of an image of some physical property (e.g. density, velocity, resistivity, etc.) from the measurements of emanations (e.g. radiation, wave motion, static field, etc.) that have passed through the region to be imaged (Bates et al., 1983). In geophysics, seismic waves and high-frequency electromagnetic waves are such emanations and the corresponding tomography methods are termed 'seismic tomography' and 'electromagnetic tomography' respectively. Using seismic waves and varying the depths of transmitter in one borehole and receiver in another, an image of velocities may be constructed from collected data (Dines and Lytle, 1979; McMechan, 1983). Considerable research has been conducted in seismic tomography owing to its potential for commercial applications and its ability to obtain relatively high spatial resolution. The most widely applied technique is traveltime tomography based on the straight-ray or curved-ray tracing algorithm where the transmission of seismic waves is approximated as a

straight line or curved line and the data are the measurements of travel time from transmitter to receiver (Dines & Lytle, 1979; Ivansson, 1986; Dyer & Worthington, 1988; Bregman et al., 1989). Clearly seismic tomography only works where targets have adequate velocity contrasts. It may not function if the subsurface structures lack such contrasts; for example, seismic tomography cannot provide information about the salinity of ground water.

If geologic targets have an electrical conductivity contrast as often is the case in mineral prospecting and water pollution monitoring, cross borehole electromagnetic tomography is an alternative approach. The transmission energy can be the high frequency continuous electromagnetic (EM) wave or a pulsed radar signal (Daily, 1984 and Olsson et al., 1992). The disadvantage of the technique is that the decay of the signal is high if the medium is highly conductive (e.g. moist soil, water and clay, etc.) and hence the depth of penetration is very limited, in comparison with seismic tomography.

Table 1.1 Examples of these applications which may require a solution with a higher spatial resolution than that obtainable from conventional geophysical methods.

Applications	Problem Area	Nature of the Target	Material Context	Scale & Depth	Methods
Reservoir characterisation & EOR monitoring	Oil Exploration	Low velocity & low resistivity	Steam & oil, porosity of the sediments	Large & deep	Cross-hole seismic & high frequency EM tomography
Hydraulically fracture zones characterisation	Dump site evaluation & environmental monitoring, ground water exploration	Low resistivity, low velocity & high dielectric constant	Water filled fracture zones	Intermediate or small & mostly shallow	High frequency EM (radar), electrical resistivity & seismic tomography
Ground water pollution monitoring	Environmental monitoring	Variation in resistivity, dielectric constant	Various pollutes	Small to intermediate & shallow	Ground radar, electrical resistivity & EM tomography
Imaging shallow buried targets	Archaeological prospecting & civil engineering applications e.g. pipe line locating.	Variation in resistivity, dielectric constant & velocity	Walls, ditches, tunnels & pipe lines, etc..	Small to intermediate & shallow	Ground radar, electrical resistivity & seismic tomography

Note: EOR—Enhanced Oil Recovery

The development of ground-penetrating radar (GPR) technology for detection and location of buried objects has gained considerable attention in recent years. Application areas of interest include detection of underground artefacts of both civil and military significance as well as high resolution mapping of shallow soil and rock structures (Davis & Annan, 1989). Ground-penetrating radar profiling is similar in concept to sonar and seismic reflection, except that it is based on EM energy reflection. A short pulse of high frequency (e.g. ~100MHz) EM energy is transmitted into the ground, generating a wavefront that propagates downward. Some of the energy is reflected when there is a change of dielectric permittivity within the medium whereupon it returns to the surface. Most GPR data are gathered by single channel reflection and are not suitable for tomography. Only in last few years has multi-channel GPR data collection been introduced enabling conventional seismic data processing methods, such as filtering, migration and possibly, tomography, to be applied (Jol & Smith, 1991; Fisher et al., 1992). One limitation of GPR is that the reflection of a radar wave strongly depends on the water content because the dielectric constant of water is more than 10 times that of most dry geological materials (i.e. 80:4-8). The high attenuation rate of radar signals in conductive materials such as clay and water limits its depth penetration. Like seismic reflection survey, GPR works best when the subsurface structures are in layers or quasi-layers.

Electrical resistivity surveying is another simple and cost-efficient alternative for applications where resistivity contrast exists. It has better depth penetration than GPR when the subsurface moisture is high while the equipment is also cheaper and much simpler to operate. Although conventional resistivity techniques used in geophysical exploration are very crude when regarded as imaging methods, recent progress in electrical resistivity tomography, especially in medical physics, has shown that potentially a better image, in terms of spatial resolution, of subsurface structures may be achieved.

Electrical resistivity tomography or ERT is a method for producing an image of the electrical resistivity distribution in a volume from discrete measurements of voltage or current made within or on its surface by passing current through the volume. The development of ERT mainly stems from two branches: in medical diagnosis and in geophysical imaging. A review of progress in those two areas is given:

1.2 IMPEDANCE TOMOGRAPHY IN MEDICAL PHYSICS

Electrical resistivity tomography in medical diagnostics is referred as 'electrical impedance tomography', EIT, or 'applied potential tomography', APT. Although electrical resistivity surveying has been routinely employed in geophysical prospecting for a long time, the first success of resistivity tomography seems to have been in medical diagnostics where the measurements can be made around the surface of the region to be imaged. In recent years, intensive research has been carried out in this area and consequently a significant amount of progress has been made which has been documented in special issues of Clinical Physics and Physiology Measurements (Brown et al., 1987; Brown et al., 1988).

Early attempts at medical EIT imaging were based on current, rather than voltage measurements. The aim was to construct a system where current may flow in a parallel or fan beam form (e.g. Schomberg 1980; Kim et al., 1983). The underlying assumption is that the current flows along raylike paths and so reconstruction algorithms applicable to X-ray and other tomographic modalities could be applied. There are two fundamental problems with this approach. Firstly, these methods require the measurement of current or voltage at active electrodes. The effect of contact resistance will produce significant errors and may well make reconstruction impossible (Wexler et al., 1985; Barber, 1989), and secondly, the passage of current flow is strongly dependent on the distribution of resistivity and generally does not follow any straight line or predictable path. Hence it seems inappropriate to perform resistivity image reconstruction by standard algorithms which assume that the measurement is a line integral of the image like X-ray tomography, as demonstrated by Bates et al. (1980)

In recognising such problems, Barber and Brown (1984) developed a resistivity tomography system where voltage, instead of current, is measured. The system employs an encircling array of 16 electrodes. When a 50kHz, 1~5mA current is passed between two adjacent electrodes the voltages are measured between remaining adjacent pairs of electrodes (see Figure 1.1). A whole data set is collected by driving the current dipole around the object in turn. The algorithm is designed to image the relative changes of resistivity and requires collecting a set of voltage data V_{ref} before the resistivity change occurs and another set of voltage data V_{dat} after the change takes place. As shown in Figure 1.1, the

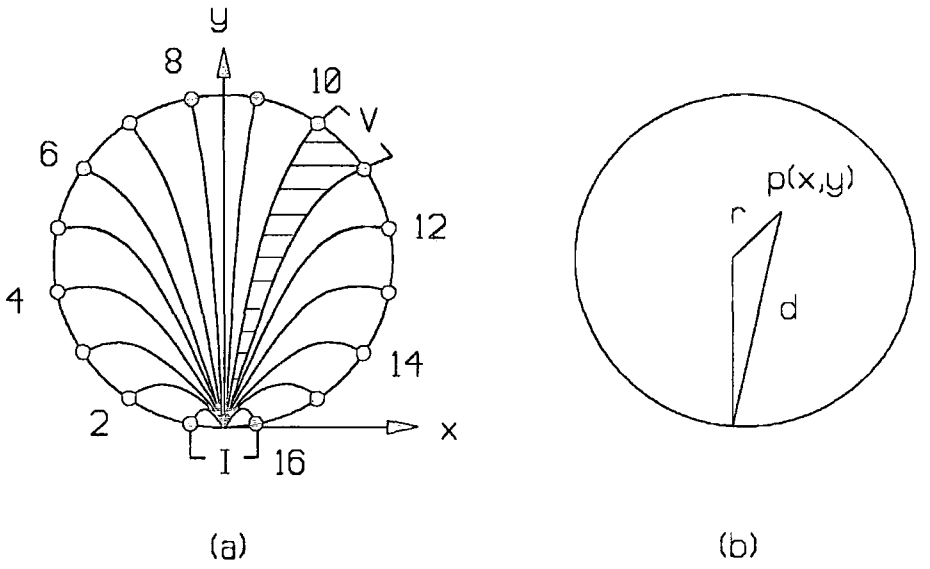


Figure 1.1 The backprojection method for an encircled array of 16 electrodes: (a) equipotentials produced by a current dipole; the shaded area represents the area into which backprojection of voltage V occurs; (b) the distance measures in the weighting function where r is the distance from the centre of circle to the pixel and d is the distance from the current dipole to the pixel (after Barber, 1990). The diameter of the circle is in 1 unit.

normalised voltage difference $g = (V_{dat} - V_{ref}) / V_{ref}$ is then weighted and backprojected along the equipotentials to form a dynamic image of resistivity in an anatomical section, i.e. if there are m equipotential lines passing through point (x, y) for the whole data set, a pixel value $c(x, y)$ is

$$c(x,y) = \frac{1}{m} \sum_{k=1}^{m} w_k(x,y) g_k \tag{1.1}$$

and

$$w_k(x,y) = \frac{1-r^2}{d^2} \tag{1.2}$$

where $w_k(x,y)$ is the weight for a pixel at (x, y) at a distance of r from the circle centre and a distance d from the current dipole, g_k is the k th normalised voltage which has an equipotential line passing through the pixel at (x, y) and $c(x,y)$ is the relative change of conductivity at (x, y). The

algorithm is fast because the weighting function can be calculated ahead of reconstruction. However, the weighting function is purely empirical and the processes of backprojection are more heuristic than being rigorously based on physical principles. For example, the changes of boundary voltage with respect to the references are not necessarily only related to the resistivity within the equipotential zones to which they belong and therefore one may ask why the backprojection is along the equipotential line and not into the entire region. The geometry of the region to be imaged is required to have circular symmetry. Despite its evident success, the rigorous physical base behind the algorithm is still not fully understood (Barber, 1989) and it may therefore be difficult to extend the algorithm to applications with other geometries. This may explain why the results were less encouraging when this algorithm was modified to image subsurface resistivities from data gathered by a linear electrode array on a flat surface (Powell et al., 1987).

A variety of reconstruction algorithms involving iterations have also been proposed. Wexler et al. (1985) and Wexler (1988) described an iterative algorithm where no sensitivity matrix (Jacobian matrix) is calculated. Firstly with an initial resistivity distribution and a given pattern of current on the surface (the Neumann boundary condition) , the distribution of current flow $\mathbb{J}$ within the object is computed by a finite element method (FEM). Then a solution for the potential distribution V within the object is computed from measured voltage data on the surface, or Dirichlet boundary condition and assumed resistivity distribution. If the initial guessed resistivity distribution is the solution, there should be

$$\mathbb{J} = -\sigma \nabla V \tag{1.3}$$

according to Ohm's Law in point form, where σ is the conductivity and ∇ is a gradient operator. Wexler et al. (1985) argued that the computed current flow and potential distribution within the object will not be consistent initially, i.e. $\mathbb{J} + \sigma \nabla V \neq \mathbf{0}$, because the initial guess of conductivity distribution is usually not correct. The resistivity image can then be estimated by minimising the difference between those two current distributions, i.e. $\mathbb{J}$ and $-\sigma \nabla V$. Wexler (1988) showed that a new estimate for the conductivity of the i th element is

$$\sigma_i = \frac{-\sum \int_{\Omega_i} \mathbb{J} \circ \nabla V d\Omega}{\sum \int_{\Omega_i} \nabla V \circ \nabla V d\Omega} \tag{1.4}$$

where the integrations are over the volume Ω_i occupied by the i th element and the summations are over all the applied current patterns. The process is repeated until it is convergent. Such an approach has shown some encouraging results (Wexler, 1988) but its convergence speed is very slow. Yorkey et al. (1987) also proposed an iterative algorithm which combined the finite element method with a modified Newton-Raphson technique where the FEM is used to compute the voltage boundary data from a model and the Newton-Raphson method is used to modify the model parameters to minimise the errors between observed and computed data. The final resistivity distribution of the model is the reconstructed image.

Several of these algorithms have been compared by Yorkey et al. (1987) on synthetic data under the assumption that both current and resistivity distributions are confined within a 2D sheet-like region. They conclude that iterative algorithms with a correct sensitivity matrix (Jacobian matrix), such as Yorkey's algorithm (Yorkey et al., 1987), can achieve the best results. Recent progress in this area can be found in Barber(1989), Ider et al. (1990), and Hua et al. (1991, 1993). In general, it appears that the algorithms in medical science are currently developing in two directions: the first is the backprojection algorithm devised initially by Barber et al. and the other is the iterative algorithm which normally involves a different kind of forward modelling and function minimisation or optimisation methods. The backprojection algorithm has been proved to be very fast because the reconstruction is simply a process of multiplying normalised voltages by pre-calculated weights so that it can be implemented in a real-time image tomography system. It also has the advantage of not being very sensitive to the variation of surface shape of the target volume compared with iterative methods (Barber, 1989). The disadvantage is that this algorithm can only reconstruct the image of resistivity changes, not the absolute values of resistivity distribution and hence cannot be used in situations where

there is no change in resistivity during data collection or when an absolute resistivity image is required. On the other hand, iterative algorithms can construct an image of an absolute resistivity distribution but require accurate forward calculation of the boundary data within 1%-2% which could only be achieved with high cost computation, for example, by FEM (Barber, 1989). The problem of ill-posedness in the inversion is another difficulty for iterative algorithms, as it will be discussed in Chapter 5.

1.3 RESISTIVITY TOMOGRAPHY IN GEOPHYSICS

When compared to impedance tomography in medical science, the development of electrical resistivity tomography in geophysics has been limited in its scope. Resistivity tomography in geophysics differs from impedance tomography in medical diagnostics both in physical scale and measuring angular coverage. Geophysical applications require sampling on a larger physical scale than medical applications and hence to achieve sufficient received-signal levels over practical distances much lower frequencies such as 0.5-10Hz and stronger currents have to be used. Generally, the region to be imaged in geophysics is not accessible through 360^0 and hence the angular coverage of measurements available here are much less than in the case of anatomical studies, especially when only a surface electrode array can be deployed. However, significant progress has been made (Noel & Walker, 1990; Daily & Owen ,1991; Barker, 1992; Shima, 1992; Ramirez, et al., 1993)

Resistivity survey methods have been used to investigate subsurface structures since the beginning of this century. Applications include mineral and hydrothermal prospecting, groundwater location, contamination monitoring and archaeological evaluation (Young & Droege, 1986; Zohdy & Bisdorf, 1990; Bevc & Morrison, 1991). The early presentation of resistivity data was not much more than a tabulation of measured apparent resistivities while later the data were interpreted by curve matching (e.g. Keller & Frischknecht, 1966; Orellana & Mooney, 1966) and model fitting through trial and error or nonlinear inversion (e.g. Van Nostrand & Cook, 1966; Inman, 1975). The development of the filter method in fast forward modelling (Ghosh, 1971a, b) was significant, since it enabled 1D forward modelling or nonlinear inversion to be routinely and quickly carried out on calculators or personal computers. However, in general these techniques are only suitable for layered earth models and consequently, it was difficult to interpret any but simple subsurface structures.

Attempts to invert surface potential data to arbitrary 2D subsurface resistivity structures can be traced back to Alfano's work (1959). Later, Pelton et al. (1978) proposed an inexpensive inversion algorithm based on least-squares optimisation and a precalculated forward modelling data bank. Although this algorithm is computationally efficient, it is impossible to fit arbitrary 2D resistivity structures to information in a limited data bank. In fact the algorithm assumed that the structure to be recovered is a single 2D prism lying immediately below a single layer of overburden. Dines and Lytle (1981) developed an algorithm for imaging core samples immersed in a saline bath using an encircling electrode array. Tripp et al. (1984) formulated a more general iterative algorithm of nonlinear inversion using a subsurface model based on the resistance network for forward modelling and nonlinear least-squares as a search method. Here the geometry of the model must be specified in advance and the algorithm is only suitable for dipole-dipole electrode configurations. Tong and Yang's (1990) algorithm is similar to Tripp et al.'s but incorporates topographic modelling into their finite-elements forward modelling and is suitable for general collinear 4-electrode configurations, i.e. any collinear configuration which transmits current through the subsurface by a pair of electrodes and measures the potentials on another pair. Shima (1990, 1992) developed an algorithm where the forward modelling was conducted by the alpha centres method (Stefanescu 1970) and data were collected using a pole-pole electrode array. The algorithm is fast but may suffer from its inaccuracy in forward modelling for arbitrary 2D structures. Daily and Owen (1991) employed Yorkey et al.'s (1987) algorithm in cross-borehole resistivity inversion where resistivity and electrical currents are assumed to be distributed only within a 2D sheet. A recent progress has been to impose smoothness constraints on the reconstructed resistivity images (e.g. Sasaki, 1992). A different approach was described by Barker (1992) where the iterations were based on Zohdy's (1989) technique, not the conventional nonlinear least squares method. Although the algorithm is fast and has provided some successful results, the underlying assumption that the apparent resistivities in the Wenner pseudosection are related to the true resistivities of a forward model by the depth investigation may make the algorithm unstable for complicated 2D structures. This arises because the 'depth of investigation' is not only a function of electrode spacing but also a function of resistivity contrast.

As stated above, nearly all algorithms for resistivity imaging are iterative and follow similar strategies. An initial resistivity distribution is assumed and from this starting model the boundary data are calculated by forward modelling. The computed boundary data are then compared with the actual measurements and the differences are used to adjust the model until the boundary data eventually best fit the observed data. The final model is then taken to be the reconstructed field section. This process is normally carried out through nonlinear least-squares function minimisation (optimisation) methods. Given the complex nonlinear relationships between a resistivity distribution and the potential data on the ground surface, it seems justified to employ iterative algorithms. However, experience suggests that the success of iterative algorithms largely depends on the accuracy of both forward modelling and the initial model structure. The former is due to the inherent ill-conditioned problem of resistivity inversion, i.e. a small data error can cause a large fluctuation in the reconstructed image. The latter is due to the fact that most function minimisation techniques devised so far are local search methods which can be trapped at a local rather than global minimum. To achieve higher accuracy in forward modelling requires a finer mesh to be employed, for example, by FEM. Using such a mesh to characterise the subsurface will result in a large number of cells, rapidly leading to an increase in computing time. It is also not yet clear how to carry out the global search (Press et al., 1989; p274). Another disadvantage is that in most cases an iterative algorithm requires use of major computing facilities.

In addition to the iterative algorithms mentioned above, Noel & Walker (1990) proposed a non-iterative algorithm using a linear electrode array with dipole-dipole electrode configurations. The measured voltages are first normalised by reference voltages calculated for a assumed uniform half space and are then backprojected into the equipotential zones in a similar manner to Powell et al.'s (1987) and Barber & Brown's (1984) backprojection algorithm. The algorithm has achieved some success (Noel & Xu 1991), but it suffers the same problems as Barber & Brown (1984)'s algorithm mentioned before in relation to image artefacts arising from structures with appreciable resistivity contrast. Therefore, the algorithm may not work well with complex 2D structures. A similar backprojection algorithm was also proposed by Shima (1992) in which the weighting function was estimated from the linearized Jacobian Matrix and the same weakness as mentioned above would be expected.

1.4 AIMS OF THE STUDY

The aims of this study are to develop electrical resistivity tomography techniques for geological and archaeological investigation using a linear electrode array placed on the ground. The study includes the development of suitable field data collection methods, reconstruction algorithms and image presentation and processing methods.

Chapter 2 will describe how to collect a large dataset of linearly independent voltage measurements on a linear array of electrodes and how to transform such data sets into others. Chapter 3 discusses some problems arising from the practical implementation of field data collection methods, such as roll-along data collection mode, the effect of spacing error and topography on the measurements and some considerations on the design of multielectrode switching systems. Chapter 4 introduces a fast reconstruction algorithm based on Born inversion. The nonlinear algorithms, including the method of regularised nonlinear inversion algorithms and the potential application of genetic and neural networks are explored in Chapter 5. The image presentation and processing methods are discussed in Chapter 6 while the results of numerical and field experiments are discussed in Chapters 7 and 8. Finally Chapter 9 presents a summary of the achievements of this research and possible future work.

CHAPTER 2 DATA COLLECTION

2.1 INTRODUCTION

Design of an efficient data collection method comprises an important part of the electrical resistivity tomography method. It is known that the effective resolution in any tomographic method is largely limited by the measurement geometry and the error distribution in the available data (e.g. Natterer, 1986). The measurement geometry is also referred to as angular coverage of the data since it is a measure of the angle from which the object can be 'viewed' by transmitter and receiver. Almost all medical tomography systems have a measuring geometry of up to 360^0 because the transmitter and receiver can be placed around the object, enabling the data to be collected by scanning around the object with transmitter-receiver (e.g. Figure 1.1). In geophysical tomography, the choice of the measuring geometry is much wider, as shown by the samples in Figure 2.1. In the case of borehole to borehole and borehole to surface measuring, the angular coverage of data may approach 360^0. On the other hand, if measurements are confined to the ground surface, then the angular coverage with respect to the target may be at most up to 180^0, resulting in an information deficient system. This chapter will describe how, using a surface linear array of equally spaced electrodes, potential data can be obtained with limited measurement geometry for use in ERT imaging.

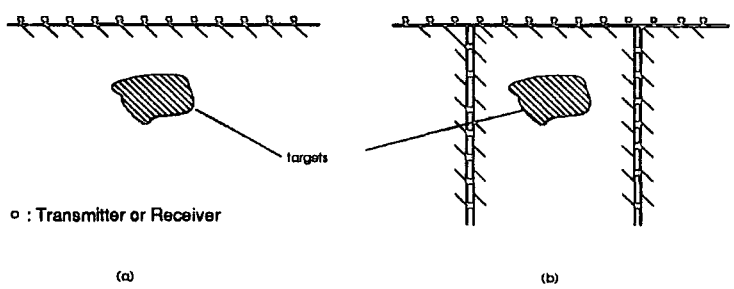


Figure 2.1 Angular coverage of measurements: (a) on the surface; (b) borehole to borehole and borehole to surface.

In electrical resistivity surveying, the current and potential electrodes are to be arranged in a certain pattern which is called an electrode configuration; for example a current dipole and a

potential dipole. Although the line source or other electrode configurations may also be suitable for ERT, the discussion here will be confined to colinear 4-electrode configurations which have wider practical application. Electrical resistivity survey using colinear 4-electrodes involves passing current through the earth between a pair of electrodes (C1 & C2) while measuring the potential difference arising between a second pair (P1 & P2). If one or two of the electrodes are fixed at a great distance from the others (which are mobile), then measurements are regarded as employing a 'three electrode' or 'two electrode' configuration, respectively. The primary aim of this chapter is to explore how to make best use of the limited measurement geometry to collect a large series of data which contains all linearly independent measurements of apparent resistivity on such an array using two, three or four electrode configurations. From this primary data set, it is shown that any other value for apparent resistivity on the array can be synthesised through a process of superposition.

Resistivity data acquisition using a colinear array of four electrodes with different configurations has been investigated by many researchers (see, Keller & Frischknecht, 1966; Telford et al., 1990). Most of the data collection methods developed so far are specifically intended for vertical sounding or horizontal profiling. Although the pseudosection method (see Edwards, 1977) can provide some information regarding the 2D distribution of resistivity, data from such conventional resistivity surveys may still be too sparse to be suitable for tomographic processing and therefore it becomes necessary to develop a routine which ensures a larger set of linearly independent data being collected.

The advent of resistivity meters electronically multiplexed to large electrode arrays has been a significant recent development in the field of resistivity surveying (Griffiths & Turnbull, 1985; Van Overmeeren & Ritsema, 1988; Noel & Walker, 1990). These systems expand the choice of electrode measuring configurations and the number of data that can be collected in a realistic time although, hitherto, their main application has been to the automation of conventional pseudosection surveys (e.g. Griffiths *et al*, 1990). However, this interesting development raises some fundamental questions:

1 What is the relationship between electrode spacing, length of array and desired resolution?

2 What is the maximum number of linearly independent apparent resistivity data that can be obtained using an array of N electrodes?

3 In which sequence should the electrodes be connected to the resistance meter to ensure that the resulting dataset is both complete and contains independent items?

4 How can one dataset (with one pattern of measurement electrodes) be transformed into another (with a different pattern) while maintaining the status of independence and completeness?

5 How can the data be collected more efficiently when the array is moved along a traversing line while ensuring completeness?

6 What are the relative merits of different datasets when used as inputs to the ERT inversion?

Questions 1-4 are addressed in this Chapter and questions 5 and 6 will be discussed in Chapters 3 and 7 respectively.

2.2 MEASUREMENT GEOMETRY

Measurement geometry (i.e. scanning or projection geometry) is one of the predominant factors which determines the number of data and the amount of information available for image reconstruction. By comparing the measurement geometry of ERT to that of ray based tomography (i.e. the tomography where the paths of emanations can be approximated as lines or beams), we may gain some insight into the data collection problem of ERT. In ray based tomography that employs X-rays or seismic waves, the importance of sampling geometry is

clearly indicated by Radon's theorem which implies that the data from all projection geometries around the object to be imaged should be collected if an image of a structure described by a continuous function is to be reconstructed (see Herman, 1980; Natterer, 1986). The maximum angular coverage of the projection is 360^0. In ERT, there is no direct analogy to the concept of projection because the path of current flow is itself a function of resistivity distribution but we may define each transmitter-receiver position around the target region as one *projection (or measurement) geometry* and the potential datum associated with each such position as one *projection.* In this way the principle of data acquisition in ray-based tomography may be applied to ERT, i.e. the maximum angular coverage of potential data is also 360^0 when the region to be imaged can be encircled by electrodes and all projections (potentials) are collected by scanning the region with measuring electrodes. Since the electrical field of a point source fills the whole space of the medium, for each source location all potential data on the boundary should be collected rather like the fan-beam configuration used in X-ray tomography.

Clearly the best measurement geometry for electrical resistivity surveying in the field can only be achieved through borehole to borehole and borehole to surface measuring schemes where the angular coverage of projections may approach 360^0. Another advantage of such a configuration is that the secondary field due to resistivity structures is strong because of the short distance between the target and transmitter or receiver. This explains why most ERT as well as other geotomography methods are developed for boreholes. However, boreholes are expensive to drill and therefore are not always available. In some archaeological investigations, drilling may be forbidden because it may destroy the underground remains. In this case, an ERT method based on a linear electrode array on the surface is attractive. The disadvantage with this method is that the projection geometry is severely limited (angular coverage $< 180^0$), resulting in an information deficient system where the reconstruction is fundamentally underdetermined. This will make an already ill-posed inversion problem even worse, as will be demonstrated in the later chapters involving tomography algorithms.

In principle, the maximum angular coverage of potential data with a linear electrode array on the surface may be achieved by injecting current through each electrode in turn and collecting potential data from the rest of the electrodes. The electrode array should extend to infinite distance and electrodes should be closely spaced although, of course, it is impossible to carry out such a data collection. Firstly, the electric field of a point source, especially the secondary field due to a resistivity anomaly, decays rapidly with distance. Thus for a specific region the useful length of the electrode array is limited. For the same reason, the choice of electrode spacing will depend on the burial depth and dimensions of objects of interest, as will be discussed in section 2.3. Secondly, the potential field produced by a point current source is known to be symmetric and additive. It is symmetric in the sense that interchange of the positions of the current and potential electrodes does not change the measured value of potential. It is additive because the total potential field from several current sources is equal to the sum of the fields generated by individual current sources. Consequently such a potential data set may contain values which are redundant since they can be superposed or they arise from interchange of the positions of current and potential electrodes and hence it may not be necessary to collect all data over a linear array for each current injection. The situation is further complicated by the fact that the data can be gathered by several different configurations of measuring electrodes, for example, a dipole-dipole or a Wenner configuration. There is then a question as how many electrode configurations have to be applied to collect all those data and if more than one of the configurations are applied, whether the data from different electrode configurations are linearly independent. This will be discussed in sections 2.4 to 2.6. It will be shown that in general, the potential data should be collected by scanning the measuring electrodes over a linear array in a way similar to X-ray tomography so that a larger number of data, or 'projections', can be gathered.

To scan over a linear array with current sources and potential electrodes is also justified by the fact that this approach can generate a more isotropic current distribution within the region of interest, as argued by Noel and Xu (1991). The currents flow through the region with different direction for each location of current electrodes and therefore the measurements may provide

some information about anisotropy properties of the resistivity distribution. This possibility has yet to be explored.

Another interesting discussion of data collection methods for resistivity survey was given by Stevenson (1934) and Vozoff (1960). They pointed out that in the determination of an underground resistivity distribution it is not sufficient simply to generate a number of data equal to the number of unknowns. They indicated that there must be a certain relation between the number of dimensions desired of the solution and the number of dimensions of the collected data. For instance, the horizontal layered earth is a 1D problem and thus only one full dimension of information is needed. The example of such a full dimension of information is the pole-pole data set collected with the current source at a fixed point while the potential electrode moves from near the current source to 'infinite' distance or a conventional vertical resistivity curve with electrode spacing expanded from relatively small to very large distance. A 2D resistivity structure requires 'two dimensions of information' which can be obtained from resistivity sounding curves for different current sources locations on a traverse line crossing the strike of the target, i.e. one dimension of information is the curve and the 2nd is the locations of current source. This argument seems to coincide with the projection and projection geometry problem mentioned above.

2.3 ELECTRODE SPACING AND RESOLUTION

As mentioned above, one of the basic problems arising in data collection using a linear electrode array is the choice of array parameters which include the interelectrode spacing and the total number of electrodes. For a given target region, important questions are:

1) What line density of the electrodes should be used?
2) How many electrodes should be applied?
3) What is the relationship between those parameters, the spatial resolution and depth of investigation?

Questions 1) and 2) are closely related to sampling theory and therefore in principle may be treated rigorously while question 3) is complicated and rather difficult to deal with.

The fundamental theory in sampling is Shannon's sampling theorem which deals with band-limited functions (see Jerry, 1977 for a review). In practice, most functions are not band-limited in the strictest sense, for example, the apparent resistivity sounding curve over a layered earth model does not have a clear cut-off frequency in its amplitude spectrum when it is transformed into the frequency domain by the Fourier transform method (see Ghosh, 1971a). Nevertheless if its amplitude spectrum is negligible for frequencies higher than a high frequency f_c, the function is said to be a f_c-band-limited and Shannon's sampling theorem can be applied. According to Shannon's sampling theorem, a set of data sampled at a rate of twice the highest frequency in the function is enough to represent that function. Data from a sampling rate more than this produces no extra information. If the band-limit value f_c of a apparent resistivity function is known, the interelectrode spacing a should be

$$a < 1/(2f_c) \tag{2.1}$$

However, it is very difficult, if not impossible, to obtain the band-limit value f_c of apparent resistivity functions for general 2D structures because there is no analytical formula available for such functions and one must anticipate that the apparent resistivity function on the ground varies with underground resistivity distribution as well as electrode configuration applied.

In practice, an empirical estimate of the interelectrode spacing may be adequate. The choice of spacing is determined by the dimension of the objects to be imaged and the depth of the zone of interest. The spacing should not be so large as to 'overlook' the objects and also not so small that there is insufficient depth penetration. For an equispaced linear array with about 20 electrodes, Acworth and Griffiths (1985) suggested that the interelectrode spacing a can be chosen such that $D/3<a<2D/3$, where D is the anticipated depth to the zone of interest. This estimation is close to

the result obtained using the concept of 'depth of investigation' (e.g. Roy & Apparao, 1971; Barker, 1989)

Alternatively, the spacing a may be estimated by equation (2.1) where the band-limited or cut-off frequency f_c may be inferred through numerical modelling, i.e. the apparent resistivity sounding

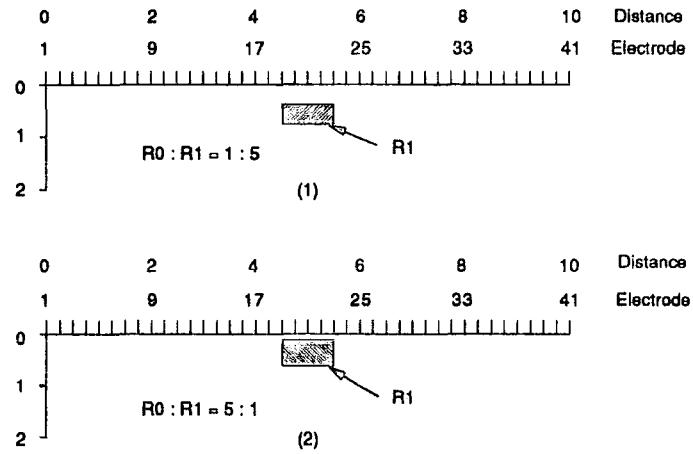
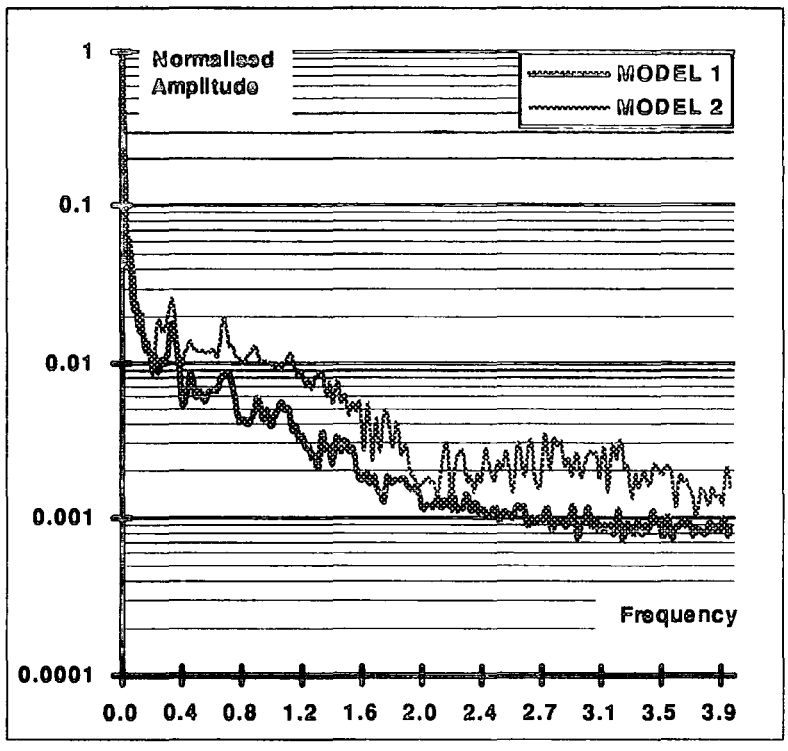


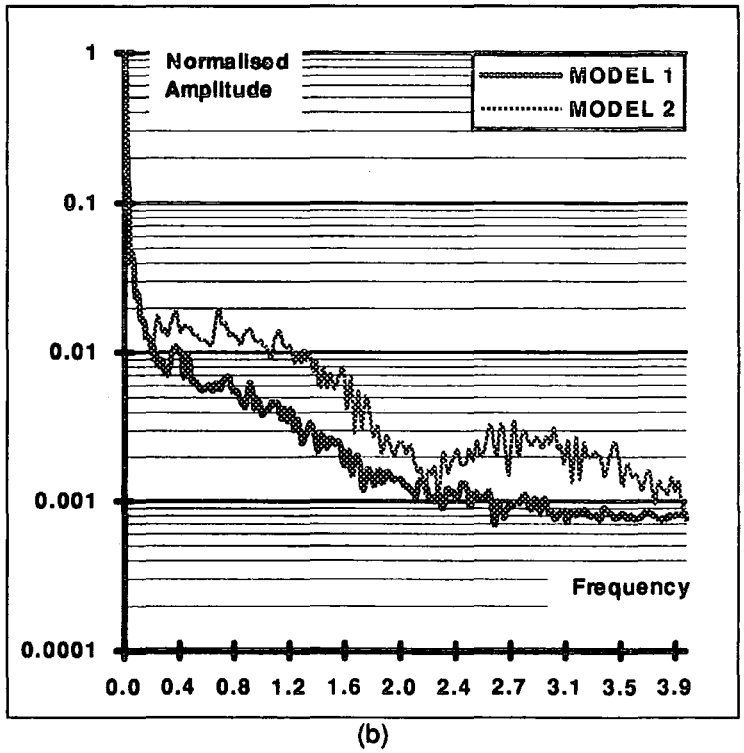
Figure 2.2 Resistivity models used to generate apparent resistivities for Fourier spectral analysis where all distances are expressed relative to the horizontal dimension (1 unit) of the object. The interelectrode spacing is 1/4 unit.

curves for 2D resistivity models are calculated by the finite difference method (or the finite element method) and their amplitude spectra are analysed by the fast Fourier transform. This was explored by the two models shown in Figure (2.2) where the horizontal dimension of the target is measured in one unit (e.g. metre or centimetre) and the 41 electrodes are equispaced at 1/4 unit intervals. The apparent resistivities from two different electrode configurations, **cCCPP** and **cCPPC** as defined in the next section, are computed and their corresponding amplitude spectra are shown in Figure (2.3). Although the spectra decay rapidly with increasing frequency, it is still difficult to determine the cut-off frequency because of the uncertainty in the choice of a proper zero level for the spectra. If the zero level of the spectra is taken as 10^{-3}, the cut-off frequency f_c is about 2.0, which implies that the interelectrode spacing should be about 4 times

smaller than the horizontal dimension of the objects to be imaged. Several similar models were explored from which it seems that the band-limit f_c is associated with several factors:



(a)



(b)

Figure 2.3 Amplitude spectra of apparent resistivities from model 1 and model 2 in **Figure 2.2.** Electrode configuration: (a)=*c***CCPP** and (b)=*c***CPPC**.

1) f_C decreases with target buried depth;

2) f_C increases with resistivity contrast;

3) f_C is related to specific electrode configurations used in the data set;

Although such modelling may never be exclusive, loosely speaking, the f_C value is about 0.5 to 2, and hence the interelectrode spacing should about 1 to 1/4 of the horizontal dimension of the smallest target of interest, given typical geological resistivity contrasts.

In its strictest sense, Shannon's theorem implies that the sampling of surface potentials is carried out from negative infinite to positive infinite distances because the Fourier transform requires to do so. In practice a limited number of sampling points within a relatively short distance may be enough because both the field of a point current source and the secondary field due to the buried objects decay rapidly with increasing distance. The useful length of the array will be limited by the measuring accuracy of the instrument and the natural sources of noise. Experience suggests that it may be adequate to have about 16 to 30 electrodes for an equispaced linear array. With a larger linear array the potential field, for instance, from a dipole-dipole configuration, will be too small to be detectable.

2.4 ELECTRODE CONFIGURATION

As mentioned in the introduction to this chapter, electrical resistivity survey using a colinear 4-electrode array employs a pair of electrodes for current injection and another pair for potential measurements. The actual electrode layouts are said to be 'three electrode' or 'two electrode' configurations if one or two of the electrodes are fixed at a great distance from the others. Given a pair of current electrodes and a pair of potential electrodes and 4 specific positions on a line, this section discuses how many electrode configurations are possible and whether the voltage data obtained from those configurations are linearly independent.

Two measuring electrode configurations

In this case, there are two electrodes placed at the area under investigation while the remaining two electrodes are sufficiently distant to be regarded as at infinity. The two possible configurations, read from left to right as 'ℂℙ' and 'ℙℂ', for such an array are shown in Figure 2.4 in which the current and potential electrodes are simply interchanged. However, according to the reciprocity theorem, the measured potentials will be identical and hence there is only one independent measurement for this 'pole-pole' configuration.

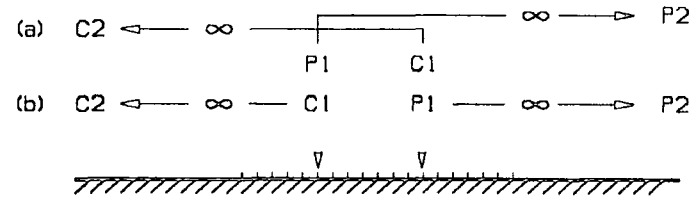


Figure 2.4 Two measuring electrode configurations in a linear array of *N* electrodes: (a) the ℂℙ configuration and (b) the ℙℂ configuration

Three and four measuring electrode configurations

Figure 2.5 shows that there are six possible electrode connections for either case of three or four

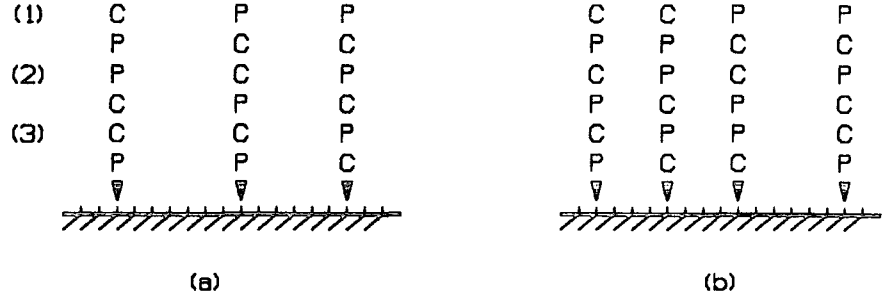


Figure 2.5 Three pairs of electrode configurations for (a) three measuring electrodes and (b) four measuring electrodes in a linear array of size *N*. If the configurations are read from left to right, they are ***CPP***(***PCC***), ***PCP***(***CPC***) and ***CCP***(***PPC***) for three measuring electrodes, and ***CCPP***(***PPCC***), ***CPCP***(***PCPC***) and ***CPPC***(***PCCP***) for four measuring electrodes where the configurations in brackets are reciprocal ones.

measuring electrodes of which half are again reciprocal and can thus be ignored. Moreover, it can be shown that only *two* of these remaining data are completely independent since the third can

always be created by superposition; for example the possible voltage data from 3-electrode configurations at positions 1, 2 and 3 shown in Figure 2.5(a) are

$$V_{cpp}=U_{12}-U_{13} \tag{2.2}$$

$$V_{pcp}=U_{21}-U_{23} \tag{2.3}$$

$$V_{ccp}=U_{13}-U_{23} \tag{2.4}$$

where U_{ij} is the potential at the j th position produced by the current source at the i th position, V_{cpp}, V_{pcp} and V_{ccp} are the voltage data measured by CPP, PCP and CCP configurations respectively. Considering that $U_{12}=U_{21}$, then

$$V_{ccp}=V_{pcp}-V_{cpp} \tag{2.5}$$

Therefore, there are two linearly independent data which can be collected using 3-electrode configurations at any location on a linear array of three. A similar relation to equation (2.5) for 4-electrode configurations was given by Carpenter (1955) and Carpenter & Habberjam (1956) who showed that there are also only two independent data from three configurations on a colinear electrode array of four.

In fact, the configurations discussed here are only the combination patterns of transmitter-receiver, without considering the intervals between electrodes as a variable. The conventional electrode configurations used for field surveys can be taken as special cases of the above configurations. For example, the Wenner array is a special case of the CPPC configuration where 4 electrodes are equally spaced. In the next section, we will discuss how to multiplex such possible patterns to a linear array with N electrodes and how to ensure that the collected data are not linearly dependent, i.e. any of them cannot be created from the data set by superposition.

2.5 DATA COLLECTION

Suppose that a linear array of N electrodes has been set up on the ground surface with equal interelectrode spacings. The questions now are

1 How many independent (not linearly correlated) data can be collected over such array?

2 How can the electrode configurations mentioned above be multiplexed to the linear array to enable such an independent data set to be gathered?

2.5.1 Two measuring electrode configurations

As shown in Figure 2.4, consider now the situation in which current electrode 'ℂ' is migrated along a array of N electrodes while potentials are recorded on the remaining electrodes. There are in total N positions for 'ℂ' and for each such 'ℂ' position there are $(N-1)$ positions for potential electrodes 'ℙ'. This provides a total of $N(N-1)$ measurements of which equal numbers arise from 'ℂℙ' and 'ℙℂ' geometries. Bearing in mind the reciprocity referred to above, we conclude that the maximum number of independent measurements $S_N^{(2)}$ is

$$S_N^{(2)} = \frac{N(N-1)}{2} \tag{2.6}$$

This dataset is listed in Table 2.1. For the convenience of later discussion, we shall introduce the

Table 2.1 The data set of pole-pole configuration with a N electrode array

Electrode	Position
C	P
1	2, 3, 4, 5, ..., N-1, N
2	3, 4, 5, ..., N-1, N
3	4, 5, ..., N-1, N
.	.
.	.
.	.
N-2	N-1, N
N-1	N

Note: C--current electrode P--potential electrode

concepts of the *Complete Dataset* and *Complete Data Space*. A Complete Dataset is defined as a series of measurements (with a given electrode configuration) which consists of only independent data and from which any other dataset (with different configurations) can be created by superposition. Otherwise the dataset is said to be *Incomplete*. By implication, such a complete dataset is not unique. In terms of linear algebra, a complete dataset with m linearly independent data is a vector in an m-dimensional linear space, $\mathbb{R}^m$. All such vectors form a space $\mathbb{R}^m$ which may be called a *Complete Data Space*. The dataset shown in Table 2.1 is one such complete data set for the pole-pole configuration over a linear array of N electrodes. A complete data set can also be referred to as a *Primary Dataset*.

It must be pointed out that a resistivity measurement is a discrete sampling of the continuous voltage function which is stimulated on the traversing line. The dataset is 'complete' in the sense that any other measurements at those discrete locations with a similar configuration can be superposed, i.e. no more independent data can be collected over the specific linear array. As mentioned above, the potential distribution on the surface is not a strictly band-limited function and extends to infinite distance. Thus the complete dataset here is in general incomplete if the interelectrode spacing is not short enough and the length of the array is small. Nevertheless the concept of such a complete dataset is useful because in practice any ERT data have to be sampled with specific electrode spacings and a limited length of array.

2.5.2 Three and four measuring electrode configurations

The extraction of complete and independent datasets from all possible systems of 3 or 4 electrodes multiplexed to a large array is a challenging and complex task. However, the problem becomes tractable when it is recalled that any 3- or 4-electrode dataset can be decomposed into a series of pole-pole measurements whose complete data space is known. Thus, by establishing and examining the matrix which transforms pole-pole measurements into the data set of 3- or 4-electrode systems it is possible to gain some insight into the complete data spaces for these electrode configurations.

The pole-pole data set in Table 2.1 can be arranged into a vector x according to rows, *i.e.*:

$$x^T = [U_{12}, U_{13}, ..., U_{1N}, U_{23}, U_{24}, ..., U_{2N}, ..., U_{ij}, ..., U_{(N-1)N}] \tag{2.7}$$

where U_{ij} is the potential at the *j th* position due to the current at the *i th* position. For any data set b, there always exists an *m*-by-*n* matrix A (denoted as $A \in \mathbb{M}_{m,n}$), such that

$$A\,x = b \qquad (x \in \mathbb{R}^n,\ b \in \mathbb{R}^m) \tag{2.8}$$

where $n = N(N-1)/2$, $\mathbb{R}^n$ is the complete data space of pole-pole array and $\mathbb{R}^m$ is a space which contains any other data set. The only possible entries for matrix A are 0, +1, -1.

Two observations can be made from equation (2.8). First, the space $\mathbb{R}^m$ is spanned by the column vector of A, *i.e.* vector b is a linear combination of the column vectors of A with the coefficients of x_i ($x_i \in x$). The vector set of all linearly independent column vectors in A is a basis which spans $\mathbb{R}^m$. To construct a matrix A, which transforms x into a complete data vector in $\mathbb{R}^m$, is to construct a basis for $\mathbb{R}^m$. Therefore, the following relation exists

$$m = rank(A) \tag{2.9}$$

This means that the maximum number of independent data from three or four electrode configurations can be obtained by checking the rank of the superposition matrix A. Secondly, the matrix A is a linear mapping operator from R^n to R^m where R^n is the domain space of A and R^m is the range space of A. For a linear mapping operator, the dimension of the range space never exceeds the dimension of domain space, *i.e.*

$$m \leq n \tag{2.10}$$

This implies that the number of independent data from three or four electrode configurations will be less than, or at most equal to, that of a pole-pole configuration. It is clear that if $A = I$, then $x=b$. Thus, the column vector set of identity matrix I is a basis which spans the complete data space of the pole-pole configuration.

Equations (2.8), (2.9) and (2.10) can serve as a general guide to constructing the complete data set, or at least to evaluate the dimensions of the complete data space R^m. For example, the following procedures can be employed to search for complete data sets for three or four electrode configurations in a multielectrode array.

(1) Design a procedure in which, when current is driven through each electrode in turn, the data are collected from all electrodes excluding those supplying current. In such a case, the number of the collected data k may be larger than the number of pole-pole data n.

(2) Construct the mapping matrix $A \in M_{k,n}$ according to the foregoing procedure and the sequence of pole-pole data in x. Matrix A can be automatically generated.

(3) Determine the rank of A to evaluate the number of independent data in b, *i.e.* the dimension of data space R^m where $m=\text{rank}(A) \leq \min(k,n)$. This may be done by the Singular Value Decomposition (SVD) method.

(4) Delete the $(k-m)$ degeneracy data in b and the corresponding rows in A. Then b is the complete data set and A contains the basis for R^m.

Special difficulties may arise in step (4) when the number of electrodes is large in the multielectrode array. In this case, the data set contains a considerable number of redundancies, making it hard to separate the independent data because the only information from step (3) is the number of redundant data, not the information on which data are actually linearly dependent. Therefore, it may be desirable to start the above procedures with a small multielectrode array, say 6 or 7 electrodes, to establish the primary relationship between $\mathbb{R}^n$ and $\mathbb{R}^m$. Then the relationship can be extended into a multielectrode array with more electrodes and verified by the above procedures. Experience also suggests that a good strategy to collect the complete data set is to

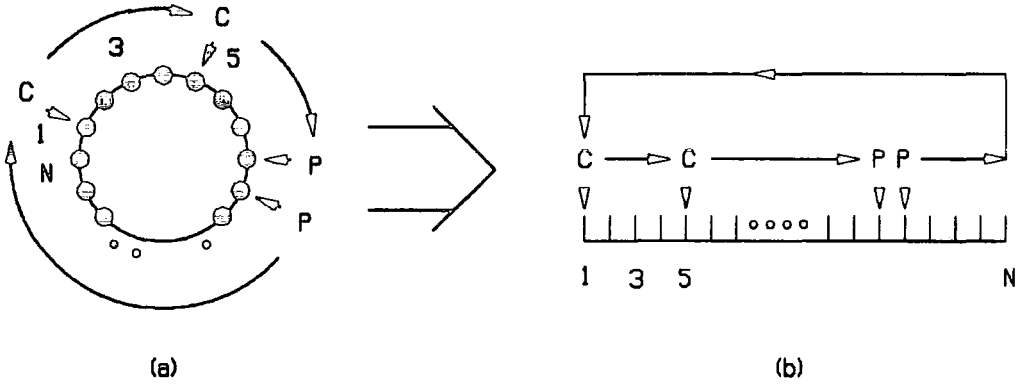


Figure 2.6 A encircled array (a) and a linear array in circulation (b).

drive a certain electrode configuration through the multielectrode array in 'circulation', *i.e.* the current and potential electrodes scan from the first electrode to the last electrode and then return to the first electrode (see Figure 2.6). In this case, the redundant data can be easily identified.

As an illustration, suppose there is a multielectrode array with 6 electrodes which are numbered 1 to 6 from left to right. A current pair CC is driven through the array and the data are collected from the rest of the electrodes as the sequence shown in Table 2.2(a). The configuration can be referred as 'circulating dipole-dipole' (or *c*CCPP) for it is exactly the dipole-dipole array in circulation. The total number of data in Table 2.2(a) is 18 which is three more than the number of independent pole-pole data. The 18-by-15 matrix **A** can be generated according to Table 2.2 (a), equations (2.7) and (2.8). The rank of **A** is 9 which indicates that half of the data are linearly

dependent. By examining Table 2.2(a) we recognise that the data are collected from CCPP, PPCC, PCCP and CPPC linear measuring electrode configurations. The redundant data are

Table 2.2 The data sets for circulating dipole-dipole configuration with a multielectrode array of six electrodes

(a)

Electrode	Position
C C	P P
1 2	3 4, 4 5, 5 6
2 3	4 5, 5 6, 6 1
3 4	5 6, 6 1, 1 2
4 5	6 1, 1 2, 2 3
5 6	1 2, 2 3, 3 4
6 1	2 3, 3 4, 4 5

(b)

Electrode	Position
C C	P P
1 2	3 4, 4 5, 5 6
2 3	4 5, 5 6, 6 1
3 4	5 6, 6 1
4 5	6 1

simply those from the reciprocal configurations (*i.e.* PPCC and CPPC). After removing the data from PPCC and CPPC electrode configurations, the rest of the data shown in Table 2.2 (b) form a complete data set. The matrix A corresponding to Table 2.2 (b) is

$$\begin{bmatrix}
0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 \\
-1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 \\
0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\
0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0
\end{bmatrix}$$

and rank(A) = 9.

The Singular Value Decomposition method is applied to evaluate the rank of the A matrices in Table 2.2 and the results are shown in Figure 2.7, which indicate that two matrices have the same rank, *i.e.* both of them contain only 9 independent row or column vectors, despite the fact that they have different dimensions.

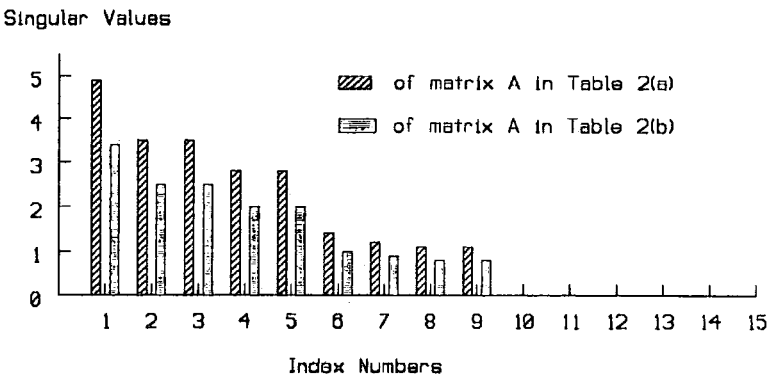


Figure 2.7. Spectra of mapping matrices

The relationship between the dimension of space $\mathbb{R}^m$ and the number of electrodes N in the multielectrode array can be derived from Table 2.2: there are a total of N pairs of positions for CC and for each CC position there are (N-3) receiver pairs (PP). The total number of data is N(N-3) and half of them are redundant, *i.e.* the dimension m = N(N-3)/2 denoted as

$$S_N^{(4)} = \frac{N(N-3)}{2} \tag{2.11}$$

Equation (2.11) is a general formula for four electrode configurations and has been verified with different N and different circulating configurations.

Similar procedures can be applied to three electrode configurations and the maximum number of independent data $S_N^{(3)}$ is found to be

$$S_N^{(3)} = \frac{(N+1)(N-2)}{2} \tag{2.12}$$

where again N is the number of electrodes in the array.

Table 2.3 Comparison of the number of independent data for different measuring configurations with a N-electrode array

Number of Measuring Electrodes	Electrode Arrangements			Number of Independent Data	No. of Data Less Than Pole-pole
	Original	Reciprocal	M_I		
2	CP	PC	1	N(N-1)/2	
3	CPP PPC PCP	PCC CCP CPC	2	(N+1)(N-2)/2	1
4	CPPC CCPP CPCP	PCCP PPCC PCPC	2	N(N-3)/2	N
4	Wenner Array			(N-1)(N-2)/6	(N+1)(N-1)/3

Note: 1) M_I is the number of independent arrangements.
2) N is the number of electrodes in the multielectrode array.
3) Number of data from Wenner array is the number of data in pseudosection (truncate to integer).
4) C—current electrode; P—potential electrode.

Table 2.3 summarises these results and presents for comparison the number of data for a Wenner array connected in pseudosection mode. It shows the pole-pole data set is largest for all N but that contains only one more datum than is possible using three electrodes. The Wenner pseudosection has the lowest number of data and the difference with respect to the number of the primary pole-pole data set increases as ~ N^2 (see Figure 2.8).

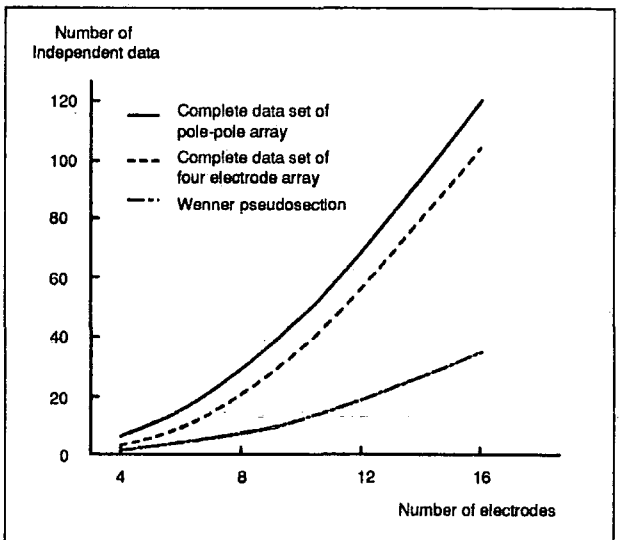


Figure 2.8. Comparison of number of independent data in different data sets

The following examples will show some possible procedures to collect complete data sets for three and four measuring electrode configurations in a linear array of N electrodes.

Example 1: Complete data set for three electrode configurations

The two scanning procedures shown in Table 2.4, namely circulating CPP and circulating pole-dipole, can be employed to collect the complete data sets for three measuring electrode configurations. The circulating pole-dipole array is similar to that used by Bristow (Bristow 1966).

Table 2.4 The complete data sets for three measuring electrode configurations in a linear array of N electrodes: (a) complete data set from circulating CPP; (b) complete data set from circulating pole-dipole.

(a)

Electrode	Position	
C1	P1	P2
1	2, 3, 4, 5, ..., N-1,	N
2	3, 4, 5, ..., N-1, 1	N
3	4, 5, ..., N-1, 1	N
.	.	.
.	.	.
.	.	.
N-3	N-2, N-1, 1	N
N-2	N-1, 1	N
N-1	1	N

(b)

Electrode	Position	
C1	P1 P2	
1	2 3, 3 4, 4 5, ..., (N-1) N,	
2	3 4, 4 5, ..., (N-1) N,	N 1
3	4 5, ..., (N-1) N,	N 1
.		.
.		.
.		.
N-2	(N-1) N,	N 1
N-1		N 1

Example 2: Complete data set for four electrode configurations

The complete data set for four measuring electrode configurations can be collected by either circulating dipole-dipole (*c*CCPP), circulating PCPC (*c*PCPC) or circulating CPPC (*c*CPPC) as shown in Table 2.5. Circulating dipole-dipole configurations have been successfully used to

collect complete data sets in both medical APT (Powell *et al.* 1987) and geophysical ERT (Noel & Walker 1990). The circulating PCPC or CPPC, however, may achieve better signal to noise ratio because the signal is stronger than that from circulating dipole-dipole when the length of the multielectrode array is large.

Table 2.5 The complete data sets for four measuring electrode configurations in a linear array of N electrodes: (a) circulating dipole-dipole (*c*CCPP); (b) circulating PCPC (*c*PCPC); (c) circulating CPPC (*c*CPPC).

(a)

Electrode	Position
C1 C2	P1 P2
1 2	3 4, 4 5, 5 6, ..., (N-1) N,
2 3	4 5, 5 6, ..., (N-1) N, N 1
3 4	5 6, ..., (N-1) N, N 1
.	.
.	.
.	.
(N-3) (N-2)	(N-1) N, N 1
(N-2) (N-1)	N 1

(b)

Electrode		Position	
P2	C1	P1	C2
1	2	3, 4, 5, 6, ..., N-1,	N
1	3	4, 5, 6, ..., N-1, 2	N
1	4	5, 6, ..., N-1, 2	N
.	.	.	.
.	.	.	.
.	.	.	.
1	N-2	N-1, 2	N
1	N-1	2	N

(c)

Electrode	Position	
C1	P1 P2	C2
1	2 3,3 4,4 5, ..., (N-2) (N-1)	N
2	3 4,4 5, ..., (N-2) (N-1),(N-1) 1	N
3	4 5, ..., (N-2) (N-1),(N-1) 1	N
.	.	.
.	.	.
.	.	.
N-3	(N-2) (N-1),(N-1) 1	N
N-2	(N-1) 1	N

The above procedures are by no means the only ones which can be employed to collect complete data sets for three or four measuring electrode configurations. The complete data sets are essentially nonunique due to the nonuniqueness of the bases of the complete data space. For example, the separations between CC in circulating dipole-dipole can be one or two or three interelectrode spacings. Nevertheless, the complete data sets are equal in the sense that they have the same number of independent data and can be mapped from one to another.

2.6 DATA TRANSFORMATION

It is desirable to transform one data set into another, such as the data set of a Wenner pseudosection, for the convenience of alternative interpretations or presentations. Apparent resistivity curves obtained from other configurations can be transformed into Schlumberger apparent resistivity curves if the data are collected over a layered earth (Patella 1974; Kumar & Das 1977; Koefoed 1977; Banerjee & Sengupta 1987). For data collected over arbitrary 2D structures, however, there is no such simple transformation. In fact the data set from conventional configurations is normally incomplete in the sense described above and hence generally it cannot be transformed into another data set. In the following discussion, we shall assume that the data set to be transformed is complete.

A complete data set can be transformed into another in the same complete data space or into a data set in its subspace. To transform a pole-pole complete data set into any other data set is fairly simple and straightforward. However, data transforms within or between the complete data spaces of four or three electrode configurations can be difficult because of their complex relationships.

If $\boldsymbol{x} \in \boldsymbol{R}^n$ is a complete pole-pole data set and $\boldsymbol{b} \in \boldsymbol{R}^{m1}$ and $\boldsymbol{c} \in \boldsymbol{R}^{m2}$ are any other two data sets, there must be two operators, $\boldsymbol{A} \in \boldsymbol{M}_{m1,n}$ and $\boldsymbol{B} \in \boldsymbol{M}_{m2,n}$, such that

$$\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b} \tag{2.13}$$

$$\mathbb{B}\,x = c \tag{2.14}$$

If there is another operator $\mathbb{K} \in \mathbb{M}_{m2,m1}$ such that

$$\mathbb{K}\,A = \mathbb{B} \tag{2.15}$$

then

$$\mathbb{K}\,b = c \tag{2.16}$$

Therefore, data set b can be transferred into data set c if there is a linear operator K which can satisfy equation(2.15).

It is obvious that if there is a operator $A' \in \mathbb{M}_{n,m1}$ such that $A\,A' = I$, then operator $\mathbb{K}$ can be constructed from equation (2.15) as

$$\mathbb{K} = \mathbb{B}\,A' \tag{2.17}$$

where A' is the pseudo-inversion of the rectangular matrix A. The pseudo-inversion can be accomplished by Singular Value Decomposition (SVD) or QR factorisation methods. In the SVD method, matrix A is decomposed as (Press *et al* 1989)

$$A = U\,\Lambda\,V^T$$

where $A \in M_{m,n}$, $U \in M_{m,n}$ is a column-orthogonal matrix and $V \in M_{n,n}$ is an orthogonal matrix, Λ=diag$\{\lambda_i\}$ (i=1,2,...,n) contains the singular values. After zeroing the small λ_i's, the pseudo-inversion can be obtained as

$$A' = \mathbb{V}\,\Lambda^{-1}\,\mathbb{U}^T \tag{2.18}$$

where $\mathbb{\Lambda}^{-1}$ =diag{$1/\lambda_1$, $1/\lambda_2$, ..., $1/\lambda_p$, 0, 0, 0,} ($\lambda_i \neq 0$, i=1, 2, .., p and p≤m)

Pseudo-inversion algorithms such as SVD have the advantage of generality, but suffer from their complexities and higher memory demands, compared with normal inversion methods, such as Gauss-Jordan elimination or LU decomposition. For all examples mentioned above, there is yet another algorithm which can construct transform operator $\mathbb{K}$ from equation (2.17) without using the SVD or QR algorithms.

Suppose the data set to be transformed is a complete dataset, *i.e.* $m1=rank(A)$ in equation (2.13), and the corresponding pole-pole data set x is arranged as equation (2.7). Then, for all complete data sets illustrated in the previous section, the last $m1$ column vectors of matrix A happen to be independent. By partitioning matrix $\mathbb{A}$ and $\mathbb{B}$ as

$$A = [F_{m1(n-m1)}, G_{m1m1}]$$

$$B = [Q_{m2(n-m1)}, R_{m2m1}]$$

then the transform matrix $\mathbb{K}$ will be

$$K = R\,G^{-1} \qquad (2.19)$$

where G^{-1} can be obtained by Gauss-Jordan elimination or LU decomposition. This method is rather heuristic, but does work with the examples mentioned above.

2.7 SUPERPOSITION EXPERIMENTS ON SYNTHETIC AND FIELD DATA

The transformation has been evaluated on both synthetic and field data. Test results with noise free and noisy synthetic data are shown in Table 2.6. The test was performed on the two layered earth model where the first layer is 1 ohm-m, 1m thickness and the second layer is 10 ohm-m. All

computations were carried out in double precision. To simulate limitations of the accuracy and precision in measuring instruments, in some data sets the data were truncated to 3 decimal places which has the effect of introducing a random error in the electrical potential of about 1~6% depending on the data set (see Table 2.6). The transformation, as shown in Table 2.6, is exact with noise-free data and is limited by inaccuracy in the source data. The noise in field data may be amplified during transformations, for example, from *c*PCPC to *c*CCPP electrode configurations.

Table 2.6. Largest relative errors (in %) in data transformation for a linear array of 10 electrodes with 1 meter interelectrode spacing.

Data Set (No.of data)	Largest Relative Errors			
Output / Input	CPP2 (44)	CCPP (35)	PCPC (35)	Wenner (12)
CPP1 (44)	5×10^{-14}	1×10^{-13}	1×10^{-13}	2×10^{-4}
*t*CCP1 (44)	1.4	6.0	4.8	0.3
*t*CCPP (35)		6.0	4.7	1.1
*t*PCPC (35)		16.0	5.0	0.7

Note: 1) CPP1: circulating pole-dipole data set.
CPP2: circulating CPP data set.
CCPP: circulating dipole-dipole data set.
PCPC: circulating PCPC data set.
2) prefix *t* means the data are truncated to 3 decimal places.

A comparison of actual Wenner pseudosection data derived from field sounding to the superposed transformations over the same structure is shown in Table 2.7. The two data sets, one Wenner dataset (57 data) and one complete dataset of *c*CCPP configuration (170 data), were collected by the same linear array over the Hett Dyke at Hamsterly near Durham, using a linear array of 20 electrodes equispaced with 5m interval. Then the dataset from *c*CCPP configuration was transformed into a Wenner dataset. The results, as indicated in Table 2.7, are quite encouraging for there is only 1.5% difference on average between measured and transformed data.

It should be pointed out that in general the transform cannot be conducted if the data set to be converted is incomplete. Thus, the transformation can be used as an indirect method to diagnose whether a data set is complete or not.

Table 2.7 A comparison of the Wenner pseudosection data from practical sounding over a dyke and those calculated from transformation of circulating dipole-dipole (cCCPP) data over the same structure.

No.	Measured	Transformed	% Error		No.	Measured	Transformed	% Error
1	14.1200	14.4948	-2.62		30	4.7300	4.7987	-1.44
2	13.5000	13.8995	-2.92		31	4.6400	4.6450	-0.11
3	12.8300	13.0805	-1.93		32	4.8200	4.8971	-1.59
4	11.5600	11.7594	-1.71		33	4.1000	4.1938	-2.26
5	10.0600	10.2472	-1.84		34	3.9200	3.9768	-1.44
6	9.3300	9.3972	-0.72		35	4.3000	4.3197	-0.46
7	9.5500	9.4364	1.20		36	4.7800	4.9046	-2.57
8	9.6600	9.4706	1.98		37	4.6800	4.7581	-1.65
9	10.5800	10.8145	-2.19		38	4.1800	4.1705	0.23
10	7.6000	8.4720	-10.85		39	3.1800	3.2731	-2.89
11	8.8500	8.7115	1.58		40	3.0800	3.1129	-1.06
12	8.6100	8.4610	1.75		41	3.3300	3.3381	-0.24
13	8.9000	8.9129	-0.14		42	3.5100	3.5371	-0.77
14	9.1800	9.3530	-1.87		43	3.1100	3.1774	-2.14
15	9.0300	9.1277	-1.08		44	3.3500	3.3632	-0.39
16	9.5700	9.5773	-0.08		45	3.5400	3.6759	-3.77
17	9.6200	9.6722	-0.54		46	3.6100	3.6542	-1.22
18	7.6300	7.7490	-1.55		47	3.5600	3.5809	-0.59
19	6.9400	7.0539	-1.63		48	3.2500	3.2156	1.06
20	6.4400	6.5276	-1.35		49	2.5600	2.6624	-3.92
21	5.8800	5.9614	-1.37		50	2.4500	2.4566	-0.27
22	5.4200	5.4936	-1.35		51	2.7900	2.9247	-4.71
23	5.8800	5.9270	-0.80		52	2.8800	2.8967	-0.58
24	6.5500	6.6411	-1.38		53	2.8600	2.8529	0.25
25	5.8200	5.8645	-0.76		54	2.8400	2.8337	0.22
26	4.6500	4.7125	-1.34		55	2.6600	2.6291	1.17
27	4.6100	4.6385	-0.62		56	2.3600	2.3494	0.45
28	4.9300	4.9959	-1.33		57	2.3500	2.3498	0.01
29	4.9200	4.9369	-0.34		Average Error			1.51

2.8 CONCLUSION AND DISCUSSION

It has been demonstrated that the complete dataset can be collected by a technique different from conventional methods with a colinear multielectrode array. This complete data set contains much more independent data than that obtained during a Wenner pseudosection survey. Those 'extra' independent data may be important for resolving complex 2D structures. The complete data sets are generally nonunique, but can be transformed from one to another or into the data set within its subspace. Although, with calculated voltage data, the errors resulting from such transformations may be governed only by machine accuracy, the error in transformations with field data is limited by noise in the data. For calculated data, one complete data set may have no significance over another if each of them can be transformed from one into another. In practice, different data sets may have different advantages and disadvantages in terms of signal to noise ratio and their suitability for hardware multiplexing or image reconstruction. In certain circumstances, it may be desirable to collect some redundant data besides the complete data set as a basis for error checking or noise reduction through stacking. For simple layered earth models (1D structure), intuitively we know that the data collection techniques described here may have no advantage over traditional methods of vertical resistivity sounding. In this case, the actual independent data only depend on electrode spacing, not the position on the ground surface because there is no lateral variation of resistivity distribution. Such a data space is degenerate.

CHAPTER 3 APPROACHES TO THE PRACTICAL IMPLEMENTATION OF ERT SURVEYING

3.1 INTRODUCTION

This chapter will address some practical problems related to ERT field data collection using a linear surface array. First, as discussed in the previous chapter, the number of data to be collected is large and hence the speed of survey may become a problem even using automated switching methods. Secondly, in a field survey, it is generally a requirement that a linear array can be rolled along the traversing line in order to cover a broader area and there is then a question of how to ensure that the data gathered from adjacent sections are complete with minimum redundancy. Thirdly, for a long traversing section the reconstructions are normally carried out separately and there may be some difficulties in joining such images to form an entire traverse section because the images may be inconsistent due to the non-uniqueness of inversion. Finally, the effects of topography and electrode spacing error on measurements need to be considered if they are significant.

The data collection could be speeded up by using a 'parallel' data collection system in which all desired potentials over a linear array are gathered simultaneously. The problem then is that the system will be expensive, complex and also not be entirely suitable for large arrays. A compromise may have to be made between the conventional multi-electrode switching system and the parallel data collection system and this will be discussed in section 3.2. The procedure for collecting complete data sets over a traversing line is presented in section 3.3 and problems arising from joining two adjacent ERT sections are discussed in section 3.4. The effects of topography and electrode spacing error on the measurements are raised in section 3.5.

3.2 MULTIELECTRODE SWITCHING SYSTEMS

Several multielectrode switching systems have been developed to reduce the manpower cost of resistivity surveys but their fitness for ERT data collection varies in terms of flexibility for various

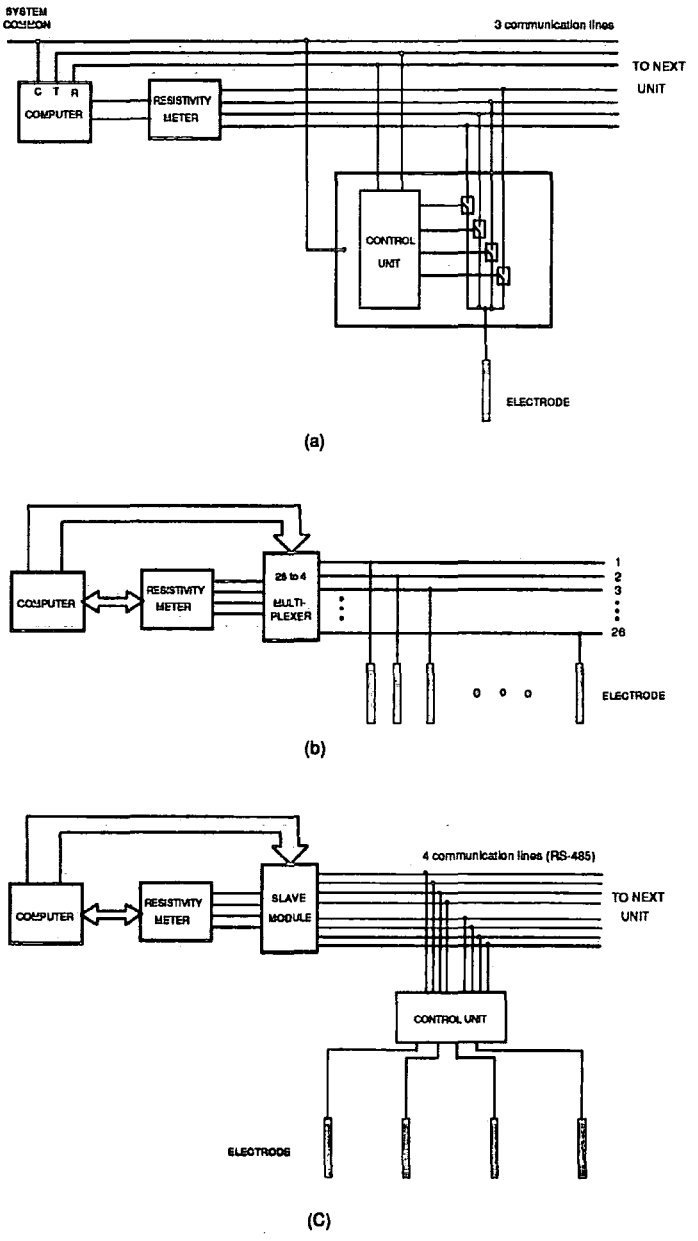


Figure 3.1: Layout of multielectrode switching systems: (a) with 7-core cables, showing one unit (after Griffiths et al 1990); (b) with 26-core cables (after Van Overmeeren & Ritsema, 1988), (c) with 8-core cables.

electrode configurations, suitability for different investigation depths and the efficiency of field operation, etc.. For example, the Wenner offset multielectrode system introduced by Barker (1981) can provide high quality data for rapid vertical sounding but may not be usable for ERT surveys because of its fixed electrode configuration and inter-electrode spacings. In contrast, a computer-controlled multielectrode array with individually addressable electrodes appears to be more flexible. One example of such a system was presented by Griffiths & Turnbull (1985) and Griffiths et al. (1990), in the form of a microprocessor-controlled resistivity traversing or MRT system (Figure 3.1(a)). The MRT employs 20 equally spaced electrodes which are linked by 50m reel-mounted sections of 7-core cable with one electrode per reel. A software-controlled relay card together with a small rechargeable battery pack is also mounted on each reel. It is therefore possible to connect any desired set of four electrodes to the resistivity meter (by sending commands to given relay card through 3 communication lines). For medium to large imaging depths, the MRT system seems to be efficient. For shallow imaging, however, the large number of reels with long sections of cable, appears to be unnecessary and inefficient. In such a case, the multielectrode systems presented by Van Overmeeren & Ritsema (1988) and by Noel & Walker (1990) seem more efficient. The former employs 26-core (seismic) cables, one electrode per line spaced at 2m intervals (Figure 3.1(b)). Through a computer-controlled relay card, the array can be scanned with any desired four electrode combination without using long communication lines. The cable is more heavy in terms of per unit length than the that of MRT system but the overall weight of the cables may be bearable because of the short length of the array. The system presented by Noel & Walker functions similarly to the MRT system mentioned above but is lighter and easier to deploy. The main array is expanded by small control boxes which are linked by 5m sections of 8-core cable. Each control box houses four addressable switch units and hence another four single cables are employed to connect them to four electrodes separately (see Figure 3.1(c)).

The use of multielectrode switching systems has increased the speed of data collection. However, published systems still operate in a *serial mode* in which only one voltage datum between a pair of

electrodes is measured in each current injection cycle. If the techniques stated in Chapter 2 are employed, the data gathering process is still rather tedious because of the large number of measurements; for example, there are about three times more measurements for a complete dataset (170-190) of a 20-electrode array than for the equivalent Wenner pseudosection (57). It is therefore desirable to explore possible methods of increasing the data collection speed.

One solution would be to use a 'parallel' data collection system in which all desired potentials over a linear array are collected simultaneously within each cycle of electrical current injection. With such an approach, a complete data set for an *N*-electrode array requires only up to *(N-1)* sequential current injection cycles, instead of *N(N-1)/2* or *N(N-3)/2* in the serial mode, and hence the data collection is about *N/2* times faster. For instance, it normally takes about 30 to 60 minutes to collect a complete data set for a 20-electrode array using an ABEM SAS 300B Terrameter resistivity meter linked to the experimental Geoscan unit PA-7 while the same task may only take about 3 to 6 minutes to finish with a parallel data collection system. This will be very significant for an array with large number of electrodes or for applications which require monitoring rapidly changing resistivity targets. The major drawbacks are: i) it requires multi-measurement channels and hence increases the complexity and cost of the instrument; ii) it requires one line per electrode and therefore the system may be feasible for shallow investigations but not suitable for intermediate to deep sounding due to the heavy weight of the multi-core cable needed. It seems that the use of a parallel system is only justified for a short array and for those applications where the speed of the data collection is critical, e.g. to monitor the dynamic changes of the ground water during a pumping test.

Alternatively, the data collection system may be implemented in a *multi-channel* or *serial scanning mode* in which a single channel resistivity meter quickly scans over the whole array through the addressable multiplexers to obtain all desired data within each current injection cycle. The hardware design of the system would be similar to a conventional multi-electrode switching system except that the A/D converter and the relay cards should have a higher speed. The time for a single scan T_s is determined by the number of voltage channels n_c (any desired combination of a pair of

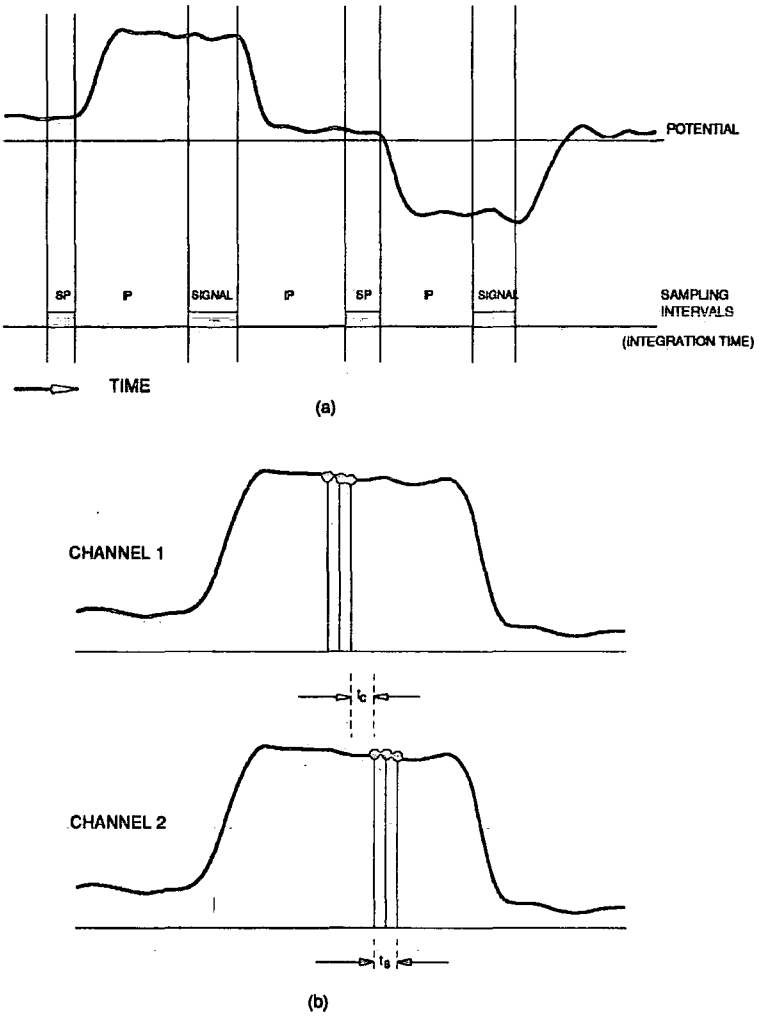


Figure 3.2: Comparison of signal sampling of resistivity surveying: (a) a sampling cycle in a conventional serial mode and (b) part of sampling cycle in a 'serial scanning' (or multi-channel scanning) mode. SP=self potential; IP= induced-polarisation.

potential electrodes forms one voltage channel), the sampling time for each channel t_s and the response time of the control unit t_c which includes the response time of the relay card and the time for command transmission through the serial communication, i.e.

$$T_s = n_c(t_c + t_s) \tag{3.1}$$

(see Figure 3.2 (b) for an illustration). The conversion time of a dual-slope type of A/D converter is about 1~50ms while the relays normally have a response time of 1ms to 4ms. Modern 'flash' A/D converters with conversion times as low as 1μs are readily available but they may not be accurate enough.

As an example, suppose that a linear array consists of 20 electrodes and hence there are 19 voltage channels. If the sampling time for each channel t_s is 10ms (the number of sampling points depends on A/D speed) and response time of control unit t_c is also 10ms. The maximum time needed for a single scan would be 380ms. Within each current injection cycle, each voltage channel has to be sampled 4 times, twice for self-potential (SP) and twice for signal (see Figure 3.2 (a)). The total scanning time for each cycle would be 1.52s. Therefore a current injection cycle of 4s to 5s would be enough for such an implementation. The total time of data collection for the whole data set is then determined by the total current injection cycles and in the 20 electrode-array case, this should be less than 10 minutes.

Perhaps one drawback of such an implementation is that the data sampling intervals may be too small to be accurate in comparison with conventional sampling which integrates the signal over a rather longer time interval (e.g. 100ms) by using voltage (or current) to frequency conversion (V-F) which averages out part of noise. Similar effects may be achieved in the multi-channel scanning mode by recursive sampling (as in seismic data acquisition), i.e. for a given signal, say SP, the sampling starts from the first channel, scanning over the whole channels, and then returns to the

first channel and the scanning is repeated again and so on with data being averaged to reduce noise. Clearly the slow response of multiplexers and the A/D converter will be the main obstacle here.

3.2 ROLL-ALONG WORK MODE.

Roll-along work mode is a field operating procedure in which the first one of transmitters or receivers of the array are placed ahead of the array on the traversing line to form a new array position once they are not required to scan the original section. It has been applied to conventional resistivity pseudo-section and seismic data collection. The idea is to keep continuous coverage of apparent resistivities for the adjacent sections with minimal redundancy of data when the whole array is traversing across the section. The same idea can be introduced to collect the complete data sets across the traversing section.

Suppose that the whole array is moved one electrode position each step along the traversing line. For the new array positions it is necessary to collect those new data required to form the next data

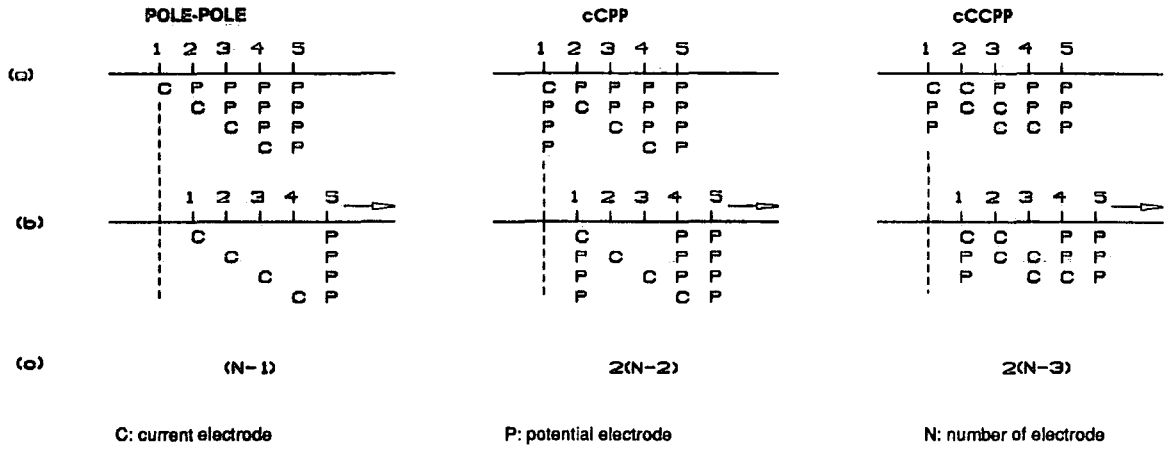


Figure 3.3: An illustration of 'roll-along' method: (a) the array before traversing and its current-potential (C-P) positions for complete data set; (b) the array after traversing one electrode spacing and its C-P positions involved with last electrode; (c) the number of new data need to be collected after each traversing step.

set. A careful examination shows that after each array movement only the linearly independent potential data involved with the new electrode position should be collected. As an illustration, Figure 3.3 presents examples of the traversing for two, three and four measuring electrode configurations, i.e. if the number of electrodes in the array is N, the number of data involved with the new electrode position is (N-1), 2(N-2), 2(N-3) for two, three and four electrode configurations respectively. The total new data arising from *m* steps of array movement should be

$$S_m^{(2)} = m(N-1) \tag{3.2}$$

$$S_m^{(3)} = 2m(N-2) \tag{3.3}$$

$$S_m^{(4)} = 2m(N-3) \tag{3.4}$$

where $S_m^{(2)}$, $S_m^{(3)}$, $S_m^{(4)}$ are the number of new data for two, three and four electrode configurations respectively.

The roll-along collection of a complete dataset can be operated as follows. After the data within the initial coverage of the array are gathered, the first electrode is moved forward, being placed ahead of the last electrode of the initial array. Then a new array is formed and the new data involving the last electrode are collected. This process is repeated again and again until the traversing line covers the whole section desired. The data set is then said to be *complete over the section* with a linear array, or referred as a *traversing complete data set.* The total number of data equals the data in the initial scan plus those obtained during the traversing:

$$S_N^{(2)} = \frac{N(N-1)}{2} + m(N-1) \tag{3.5}$$

$$S_N^{(3)} = \frac{(N+1)(N-2)}{2} + 2m(N-2) \tag{3.6}$$

$$S_N^{(4)} = \frac{N(N-3)}{2} + 2m(N-3) \tag{3.7}$$

3.4 SECTION JOINING

When the whole section cannot be recovered in a single reconstruction, for example, due to the lack of computing resources or simply because the reconstruction algorithm is not designed for the long section, then we are faced with the problem of how to join the sections together. The resistivity image sections to be joined may be adjacent or partly overlapping.

The section joining for reconstructed resistivity images could be difficult because the images to be joined may be inconsistent due to the non uniqueness of the inversion as will be discussed in the following chapters. This may give rise to the edge effects due to lack of data coverage on the two sides and hence may cause the same feature to vary from image to image. Thus a joined image may produce some artefacts which are difficult to interpret and this potential problem must be recognised in designing an image-joining algorithm.

For adjacent images the simplest joining method is to add the images side by side. If the background values of the adjacent images do not change smoothly, they should be adjusted, for example, by trend surface fitting or dithering.

It may be argued that the central zone of the image may be better determined than other parts and if such centre zones are overlapped for different image sections, a simple averaging operation within these zones could improve the quality of the joined image. Such overlaps may happen in the traversing section where data are collected by the roll-along method mentioned above. Note that for each roll-along step of array movement, there is a complete data set corresponding to the new array coverage. Using such complete data sets one can reconstruct a series of images which

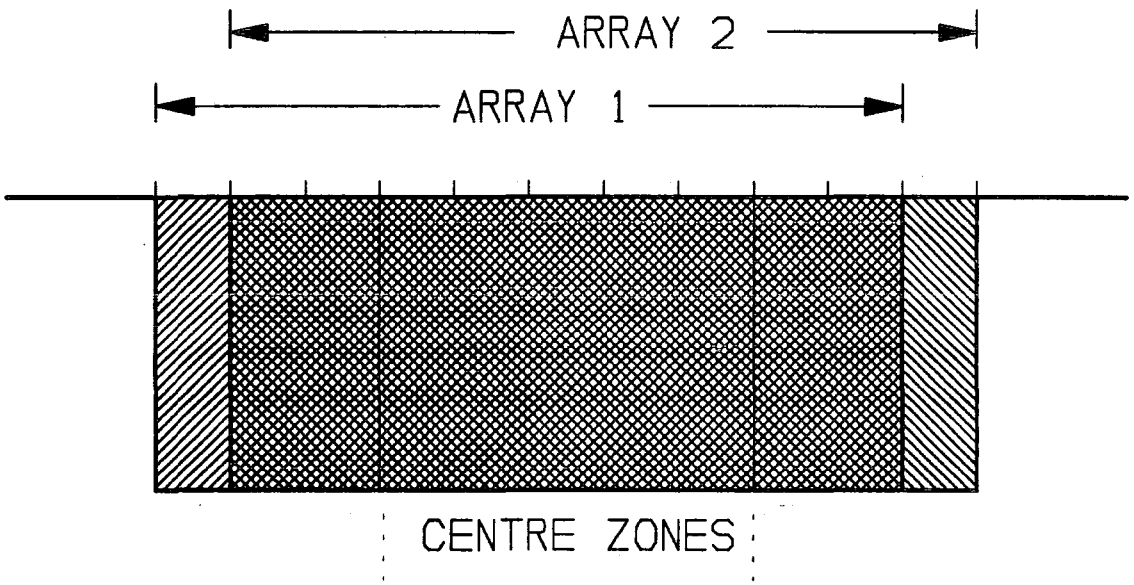


Figure 3.4 An example of two partly overlapped images joined by averaging their centre zones.

overlap with one electrode spacing interval (see Figure 3.4) and then the whole image section may be joined by averaging the overlapped central parts. Clearly the success of such an operation will depend on the stability of the reconstruction.

3.5 EFFECT OF SPACING ERROR AND TOPOGRAPHY

3.5.1 Effect of spacing error

A small perturbation of electrode position (*dx,dy,dz*) will result in a change of measured potential δV, i.e.

$$\delta V = \frac{\partial V}{\partial x}dx + \frac{\partial V}{\partial y}dy + \frac{\partial V}{\partial z}dz \tag{3.8}$$

For a colinear array on a flat-earth, both *dy* and *dz* are zero and hence if the subsurface is homogeneous, then for a pole-pole electrode configuration we have

$$\frac{\delta V}{V} = \frac{1}{V}\frac{\partial V}{\partial x}dx = -\frac{dx}{x} \tag{3.9}$$

where $x=a$ is the electrode spacing. The relative error of voltage caused by electrode spacing errors for four electrode configurations is more complicated but it has the general form

$$\begin{aligned}\frac{\delta V}{V} &= \frac{1}{V}\left[\frac{\partial V_{11}}{\partial x_{11}}dx_{11} - \frac{\partial V_{12}}{\partial x_{12}}dx_{12} - \frac{\partial V_{21}}{\partial x_{21}}dx_{21} + \frac{\partial V_{22}}{\partial x_{22}}dx_{22}\right] \\ &= -\frac{1}{G}\left[\frac{dx_{11}}{x_{11}^2} - \frac{dx_{12}}{x_{12}^2} - \frac{dx_{21}}{x_{21}^2} + \frac{dx_{22}}{x_{22}^2}\right]\end{aligned} \tag{3.10}$$

where again the subsurface is assumed to be homogeneous, $V = V_{11} - V_{12} - V_{21} + V_{22}$, V_{ij} is the potential produced at the j th position by current at the i th position and x_{ij} is its corresponding electrode spacing. The geometry coefficient $G = (1/x_{11}) - (1/x_{12}) - (1/x_{21}) + (1/x_{22})$. For

example, if a dipole-dipole configuration is planted on a linear array as shown in Figure 3.5, the relative error due to the spacing error *dx* is

$$\frac{\delta V}{V} = -\frac{(2n+1)(n+2)}{2n(n+1)}\frac{dx}{a} \Rightarrow \begin{cases} = -2.25\dfrac{dx}{a} & (n=1) \\ \approx -\dfrac{dx}{a} & (n \geq 10) \end{cases} \tag{3.11}$$

where a is the inter-electrode spacing and n is the distance (measured in terms of a) between current and potential dipoles.

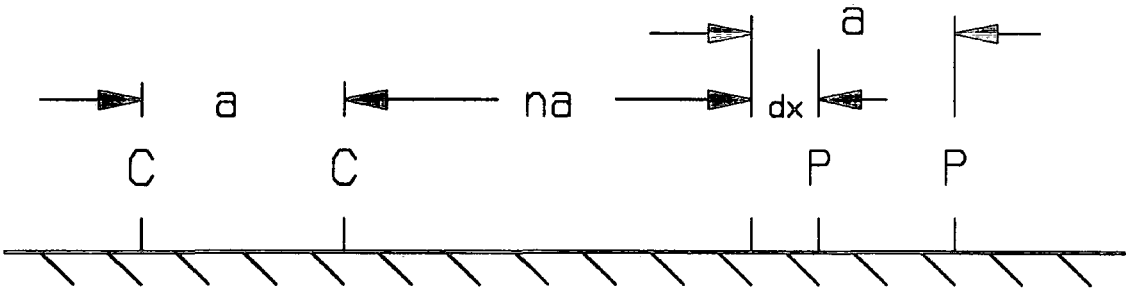


Figure 3.5 A dipole-dipole array with one electrode mis-placed.

It seems that an accuracy of 1/100 inter-electrode spacing or better is necessary for field work. For medium to large scale geological prospecting the spacing error should not be a problem because the electrode spacings are large and the accuracy of modern surveying to a precision of 1/1000 is readily attainable. For small scale problems like archaeological investigation, however, there may be practical difficulties and the electrode array must be carefully planted to keep the errors to a minimum.

3.5.2 Effect of topography

The algorithms for resistivity reconstruction normally assume that the ground surface is flat. If field resistivity survey is carried out on an irregular surface, the terrain effect may need to be taken into account. Fox et al. (1980) suggested that if the finite element method is used to compute the response of a resistivity survey then the terrain effect can be modelled with a flat-earth model by filling the area occupied by air with high resistivity materials ($10^3 - 10^5\,\Omega m$) (see Figure 3.6). The field data with topographic effects are then compensated by multiplying with the terrain-correction coefficients which are calculated as

$$c_i = \frac{p_i^f}{p_i^t} \tag{3.12}$$

where p_i^f, p_i^t are potentials of the homogeneous earth with a flat surface and the actual terrain surface, respectively. In ERT the topographic effects can be compensated before the reconstruction using the above method or, alternatively, the topography is directly incorporated into the reconstruction by a mesh shown in Figure 3.6 if finite difference or finite element forward modelling methods are employed in the reconstruction process (see Chapter 5).

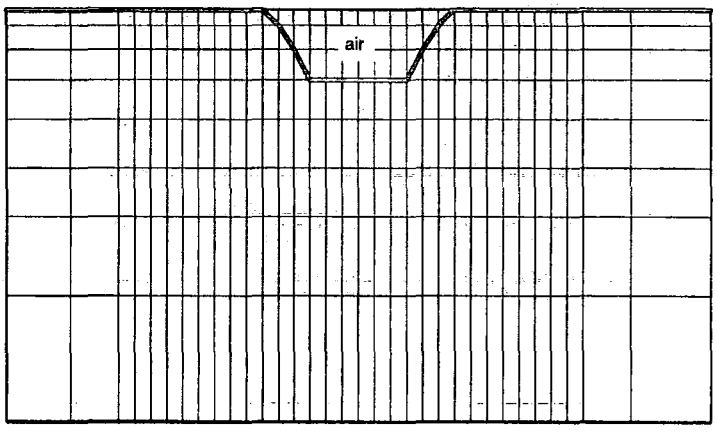


Figure 3.6 Finite elements (or finite difference) modelling of topographic effects with flat surface by assigning high resistivities to the area of air (after Fox et al., 1980).

It is known that for slopes of less than 10^0 terrain effects are insignificant and they are smaller if the spread of the array is parallel, rather than normal, to the strike of a 2D terrain structure (Fox et al. 1980). In such cases it may not be necessary to apply terrain corrections.

CHAPTER 4 LINEARIZED ALGORITHM: BORN INVERSION

4.1 INTRODUCTION

This chapter describes a Born procedure for imaging arbitrary 2D resistivity structures from the data gathered using a linear array of electrodes at the ground surface. The algorithm is based on a sensitivity relation which links the variation of resistivity to the deviation of surface potential data. In this formulation, a procedure similar to Born inversion in seismology is applied to compute the weighting coefficients for each current source-potential location. The resistivity measurements are then weighted and summed to produce an image of subsurface resistivity which can be computed within minutes on a desktop PC.

The work presented here is the further development of the early research on a simple backprojection algorithm for electrical resistivity tomography or ERT (Noel & Walker 1990, Noel & Xu 1991). Instead of computing the weighting coefficients by a simplified but inaccurate analytical weighting function, one step Born inversion is applied where the current distribution is assumed to be distributed within a homogeneous half space and resistivity anomalies are perturbations from this reference medium. Given a linear array of equally spaced electrodes on the ground surface, all independent potential data are collected using the methods described in Chapter 2 and then they are multiplied by the pre-calculated weighting coefficients, forming a subsurface resistivity image. Numerical and field trials on geological and archaeological structures have shown that this algorithm can achieve moderate resolution.

4.2 RECONSTRUCTION ALGORITHM

The resistivity image reconstruction starts from a sensitivity operator which relates the subsurface resistivity distribution to the surface potential distribution induced when current flows within the medium. The direct current field may best be described by Ohm's law and Maxwell's equations. In the conducting medium it is found experimentally that current density $\mathbb{J}$ is related to electric field $\mathbb{E}$ by the following expression (Ohm's law):

$$\mathbb{J} = \sigma \mathbb{E} \tag{4.1}$$

where σ is the conductivity, $\sigma = 1/\rho$, and ρ is the resistivity. The electric field $\mathbb{E}$ is related to the density of electric charge q by one of Maxwell's equations:

$$q = \nabla \circ (\varepsilon_0 \mathbb{E}) \tag{4.2}$$

where ε_0 is the dielectric permittivity.

Substituting equation (4.1) into equation (4.2):

$$q = \varepsilon_0 \left[\rho I \delta(\mathbb{r} - \mathbb{r}_c) + \frac{\mathbb{E} \circ \nabla \rho}{\rho} \right] \tag{4.3}$$

where $\mathbb{r}$ is any point in the medium, $\mathbb{r}_c$ is the current source point and I is current strength.

Combining equation (4.3) with the expression for the point charge potential in a half space, the total potential $U(\mathbf{r}_p)$ on the ground is

$$\begin{aligned} U(\mathbf{r}_p) &= \frac{1}{2\pi} \int_\Omega \left[\frac{\rho I \delta(\mathbf{r} - \mathbf{r}_c)}{\|\mathbf{r} - \mathbf{r}_p\|} + \frac{\mathbf{J} \circ \nabla \rho}{\|\mathbf{r} - \mathbf{r}_p\|} \right] dv \\ &= \frac{\rho_0 I}{2\pi \|\mathbf{r}_c - \mathbf{r}_p\|} + \frac{1}{2\pi} \int_\Omega \frac{\mathbf{J} \bullet \nabla \rho}{\|\mathbf{r} - \mathbf{r}_p\|} dv \\ &= U_0(\mathbf{r}_p) + U_a(\mathbf{r}_p) \end{aligned} \tag{4.4}$$

where

$\mathbf{r}_p$ = the potential electrode position.

$U_0(\mathbf{r}_p)$ = the primary potential in a homogeneous half space with resistivity ρ_0.

$U_a(\mathbf{r}_p)$ = the potential due to variation of resistivity.

dv = the variable of volume integration.

Ω= the half space.

$\|\mathbf{r}\|$= *norm* or length of a vector $\mathbf{r}$.

Equation (4.4) is the boundary integration expression of potential due to the point source of d.c. current in a half space and similar results have been obtained by others (Vozoff 1960; Keller & Frischknecht 1966; Eskola 1979; Snyder 1976; Hohmann & Raiche 1988; Li & Oldenburg 1991).

Let $\Delta U(\mathbf{r}_p) = U(\mathbf{r}_p) - U_0(\mathbf{r}_p)$ ($=U_a(\mathbf{r}_p)$). Equation (4.4) then can be rewritten as

$$\Delta U(\mathbf{r}_p) = \frac{1}{2\pi}\int_{\Omega}\frac{\mathbf{J}\ominus\nabla\rho}{\|\mathbf{r}-\mathbf{r}_p\|}dv \tag{4.5}$$

This is a nonlinear equation which relates the changes of resistivity to the deviations of potential on the ground, i.e. it is a *sensitivity* relation. If changes of resistivity are small, the current field may be approximated by $\mathbf{J}_0$, the current field produced by a point source I in a uniform half space with resistivity ρ_0:

$$\mathbf{J} \approx \mathbf{J}_0 = \frac{\rho_0 I(\mathbf{r}-\mathbf{r}_c)}{2\pi\|\mathbf{r}-\mathbf{r}_c\|^3} \tag{4.6}$$

Then equation (4.5) can be linearized by Born approximation, a well-known procedure in seismology (Cohen & Bleistein 1979).

Suppose the resistivity distribution being imaged is 2-dimensional and the electrodes are equally spaced on the ground in a line perpendicular to the 2D structures. We divide the half space into small cells (Figure 4.1) and assume the resistivity is constant within each cell and varies from cell to cell. Thus $\nabla\rho = 0$ within each cell and on the discontinuous boundaries between cells, $\nabla\rho$ becomes

$$\nabla\rho = (\rho_{ij} - \rho_{i,j-1})\mathbf{i} + (\rho_{ij} - \rho_{i-1,j})\mathbf{k} \tag{4.7}$$

where $\mathbb{i}$ and $\mathbb{k}$ are the unit vectors of the X, Z axes and there is no change of resistivity in the Y

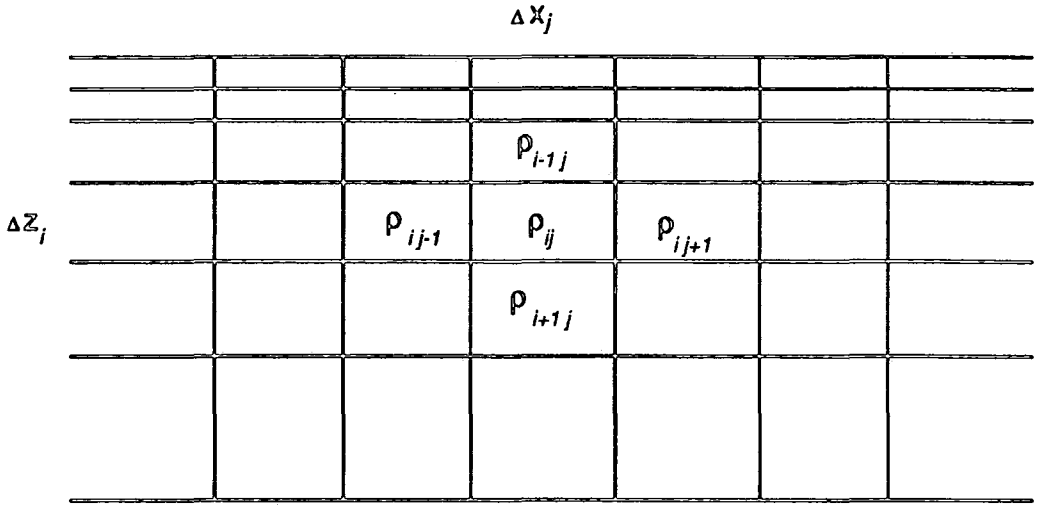


Figure 4.1 Rectangular discretization grid.

direction. We substitute equations (4.6) and (4.7) into equation (4.5) and approximate the surface integrations in the X-Z plane by summations. Equation (4.5) then becomes

$$\Delta U_n(\mathbb{r}_p) = \sum_{ij} \mathbb{S}'_{ij} \cdot \nabla \rho_{ij}$$

which can be re-arranged into

$$\Delta U_n(\mathbb{r}_p) = \sum_{ij} S_{ij} \rho_{ij} \tag{4.8}$$

where

$$\Delta U_n(\mathbb{r}_p) = \frac{\Delta U(\mathbb{r}_p)}{\mathrm{I}}$$

$$\mathbb{S}'_{ij} = S_{zij-1}\mathbb{i} + S_{xi-1j}\mathbb{k}$$

$$S_{ij} = (S_{zij-1} - S_{zij}) + (S_{xi-1j} - S_{xij})$$

$$S_{zij} = \frac{1}{2\pi^2} \int_0^\infty \frac{(x - x_c)\Delta z_{ij}}{\left[(x-x_c)^2 + y^2 + z^2\right]^{3/2} \left[(x-x_p)^2 + y^2 + z^2\right]^{1/2}} dy$$

$$S_{xij} = \frac{1}{2\pi^2} \int_0^\infty \frac{z\,\Delta x_{ij}}{\left[(x-x_c)^2 + y^2 + z^2\right]^{3/2} \left[(x-x_p)^2 + y^2 + z^2\right]^{1/2}} dy$$

$(x_c, 0, 0)$: position of current source

$(x_p, 0, 0)$: position of potential electrode

(x, y, z): point on boundary of cell

Since the sensitivity coefficients S_{ij} are calculated on a discrete mesh, exact reciprocity will not be found numerically. In this work, reciprocal elements are averaged. The above formulation is derived for a pole-pole electrode arrangement and the sensitivity relationship for other configurations can be superposed from it as explained in Chapter 2.

If there are M independent potential data and N cells in the half space, equation (4.8) can be rearranged as

$$\Delta \mathbb{u} = \mathbb{S}\,\mathbb{P} \tag{4.9}$$

where

$\Delta \mathbf{u} = [\Delta U_{ni}]$: vector of potential deviations.

$\mathbf{S} = [S_{ij}]$: sensitivity matrix.

$\mathbf{P} = [\rho_j]$: vector of resistivity.

i = 1, 2, ..., M and j = 1, 2, ..., N

If there is a matrix **B** such that $\mathbf{BS} = \mathbf{I}$ where **I** is unit matrix, the distribution of resistivity may be recovered from the surface potential data as

$$\hat{\mathbf{P}} = \mathbf{B}\,\Delta\mathbf{u} \tag{4.10}$$

where $\hat{\mathbf{P}}$ is the recovered resistivity distribution.

If we define each independent source—potential measurement as one 'projection', the inversion model can be conceived as tomographic inasmuch as the resistivity structure $\mathbb{P}$ is being 'projected' onto the surface potential $\Delta\mathbb{u}$ through a sensitivity operator $\mathbb{S}$. The question now is how to find an operator $\mathbb{B}$ to backproject $\Delta\mathbb{u}$ to form a reconstructed image of $\mathbb{P}$.

Since the sensitivity matrix $\mathbb{S}$ is a rectangular matrix in general, the operator $\mathbb{B}$ has to be a pseudo inverse of $\mathbb{S}$. This can be accomplished by the Singular Value Decomposition (SVD) method (Jackson, 1972; Jupp & Vozoff, 1975; Press et al., 1989), i.e. the *m* by *n* matrix $\mathbb{S}$ can be decomposed as

$$\mathbb{S} = \mathbb{U}\,\Lambda\,\mathbb{V}^T \tag{4.11}$$

where $\mathbb{U}$ is an *m* by *n* column-orthogonal matrix, $\mathbb{V}$ is an *n* by *n* orthogonal matrix and $\Lambda = \mathrm{diag}\{\lambda_i\}$ (i=1, 2, ... ,n) where λ_i are the singular values. After zeroing the small λ_i's, the operator $\mathbb{B}$ is

$$\mathbb{B} = \mathbb{V}\,\Lambda^{-1}\mathbb{U}^T \tag{4.12}$$

where $\Lambda^{-1} = \mathrm{diag}\{1/\lambda_1, 1/\lambda_2, 1/\lambda_3,..., 1/\lambda_t, 0, 0, ...,0\}$ ($\lambda_i \neq 0$, t<n). λ_t is the threshold where any singular value smaller than λ_t is zeroed in the process of constructing $\mathbb{B}$. The threshold is a variable which controls the distortion and resolution of the reconstructed resistivity image $\hat{\mathbb{P}}$. Normally, a higher threshold will lead to less distortion but lower resolution or vice versa and for noisy data several trials may be needed to choose it. In our work, the threshold value λ_t is taken as

$$\lambda_t = \lambda_{max}(0.5 \sim 10)\times 10^{-4}$$

depending on the noise level in the potential data.

It may be desirable to normalise the potential data and the estimated resistivity distribution in equation (4.10) by ρ_0, the resistivity of a uniform half space (i.e. reference resistivity):

$$\hat{\mathbb{P}}_n = \mathbb{B}\,\Delta\mathbb{u}_n \tag{4.13}$$

where $\hat{\mathbb{P}}_n = \hat{\mathbb{P}}/\rho_0$ and $\Delta\mathbb{u}_n = \Delta\mathbb{u}/\rho_0$. Resistivity ρ_0 may be estimated from apparent resistivities measured with small electrode spacings. In fact, numerical tests have shown that the algorithm is not highly sensitive to the choice of ρ_0 (see Chapter 7).

This algorithm in general can only reconstruct the relative variations of resistivities, not their absolute values, because equation (4.13) is underdetermined and linearized from equation (4.5). Fortunately, for most geological or archaeological applications this information may be sufficient when presented as a grey scale or pseudo-colour image. However, it is not surprising that there may be some negative elements of vector $\hat{\mathbb{P}}_n$ which have no meaning in physics although, mathematically, there is no reason why all such elements should be positive because the algorithm cannot fully recover the original image $\mathbb{P}_n$ and hence $\hat{\mathbb{P}}_n$ is not equal to $\mathbb{P}_n$. This difficulty may be overcome by formulating equations (4.5)—(4.13) with resistivity ρ in a logarithmic scale or by the following transformation which may be understood as a change of coordinate origin:

$$\hat{p}_{nti} = \hat{p}_{ni} + b \qquad (i=1,2,...,N) \tag{4.14}$$

where $\hat{p}_{ni}$ and $\hat{p}_{nti}$ are the elements of vector $\hat{\mathbb{P}}_n$ before and after transformation respectively and $b = 2\left|\min\{\hat{p}_{ni}\}\right|$. This transform will not change the visual image which can only show the relative variations of resistivities because the transform does not change its dynamic range.

The operator **B** in equation (4.13) is only related to geometry parameters of the array and the discrete mesh applied and hence can be computed in advance. Therefore the algorithm works fast even on ordinary desktop computers.

4.3 DISCUSSION AND CONCLUSION

A fast algorithm which can reconstruct subsurface resistivity images from surface potential and which can be implemented on microcomputers has been developed. The algorithm is based on the perturbation of a sensitivity relation between resistivity and surface potential data; a Born

approximation. To achieve better determinacy and resolution, all independent data over a multielectrode array may be collected as described in Chapter 2. The algorithm may fairly reveal the geometries of resistive inhomogeneities but cannot estimate their absolute values. High noise level in the potential data may result in a strong distortion of the reconstructed image and hence good quality data are required. Care should also be taken to reduce the spacing errors especially when interelectrode spacings are small, such as in the case of archaeological investigations. In a further development, iterative refinement may be necessary in order to optimize the results of what is a highly nonlinear problem. Although nonlinear iterative methods may ultimately give superior results, they currently suffer from high cost of computation and dependence on the initial model so that the relative merits of iteration and noniteration have yet to be determined.

CHAPTER 5 NONLINEAR ALGORITHMS

5.1 INTRODUCTION

In the previous chapter I presented a linearized, one-step algorithm which can recover 2D resistivity structures with moderate resolution. The algorithm is fast but can only recover the relative variations of a resistivity distribution. However, for such a highly nonlinear problem it may be necessary in some cases to employ an iterative method to recover the absolute resistivity values or to enhance the resolution of reconstructed images. This chapter describes several nonlinear algorithms which may be employed to reconstruct subsurface resistivity structures from surface potential data. The algorithms considered here are of two types: i) iterative algorithms involving model fitting by nonlinear function minimisation or optimisation and ii) neural network methods in which the nonlinear properties of the reconstruction are learned from 'experience', i.e. examples and memorised by the network. In the iterative algorithms, I will discuss the fundamental problem of nonuniqueness in resistivity image reconstruction and hence demonstrate the need to incorporate extra information, such as smoothness constraints, in order to stabilise the iterations. The main focus will be the nonlinear least squares method with various smoothing constraints but the possibilities of using genetic algorithms are also discussed. The initial model from which iteration starts can be constructed from *a priori* knowledge or from the output of the algorithm described in Chapter 4. For the neural network methods, only an introduction is given simply to show its potential for resistivity image reconstruction.

The use of non-linear least squares methods to invert resistivity data over a 2D structure has been investigated by several researchers. In the early development the subsurface structures were only inverted as a small number of resistivity cylinders, probably due to the limitation of computer resources at that time (e.g. Pelton et al., 1978; Trip et al., 1984). Consequently it was difficult to image any but simple structures. More recent developments are to invert the data over a more dense mesh, aiming to produce an image which is comparable to a tomographic image in terms of

spatial resolution (e.g. Shima, 1990; Barker, 1992; Sasaki, 1992; Shima, 1992). Such inversion is known to be unstable and the method of regularisation is then introduced (Sasaki, 1992). However, most of the investigations were focused on the borehole-to-borehole geometry and the limitations of using surface data alone have not yet been fully investigated. In the method of regularisation the influence of different smoothing or regularisation functions and the regularisation parameter on the reconstructed image needs to be investigated. In the following sections, the regularised non-linear least squares methods for inverting surface potential data over a 2D structure are presented and then various choices of smoothing functions are discussed. The strategies for selection of a suitable regularisation parameter, including the automated selection methods, are described. The effects of the smoothing functions and regularisation parameter will be evaluated in Chapter 7. Numerical and field experiments indicate that the regularised non-linear least squares methods normally produce better results than the Born procedure if certain conditions are satisfied (see Chapters 7 & 8).

The genetic algorithms and neural network methods have recently been introduced to process geophysical data, mainly seismic data (e.g. Stoffa & Sen, 1991; Murat & Rudman, 1992). Some of their advantages are: i) they are fully non-linear algorithms; ii) they may have better approaches for treating the local convergence problem which occurs in conventional non-linear least squares methods. It is therefore interesting to explore their usefulness in electrical resistivity image reconstruction.

5.2 FORMULATION OF THE ITERATIVE PROBLEM

5.2.1 Forward Problem

Iteration algorithms require to perform *forward* calculation of voltages or apparent resistivities from an assumed resistivity model and current source locations. There are several numerical techniques for modelling the direct current response of a 2D resistivity distribution, including finite element methods (Coggon, 1971), network analogies (Madden, 1972; Tripp, et al., 1984), finite different methods (Dey & Morrison, 1979), the alpha centre method (Stefanescu, 1970), and the integration equation or finite boundary element method (Vozoff, 1960; Okabe, 1981). Each technique has its own relative advantages and disadvantages. The finite difference method (FDM) implemented by Dey and Morrison (1979) is applied here on the account of its speed and low memory demands.

The mathematical problem to be solved is Poisson's equation in the region of interest, subject to the Neuman boundary conditions. Taking the divergence of equation (4.1) we have

$$-\nabla \cdot (\rho^{-1}\nabla U) = I\delta(x-x')\delta(y-y')\delta(z-z') \tag{5.1}$$

where

$\nabla = \frac{\partial}{\partial x}\mathbf{i} + \frac{\partial}{\partial y}\mathbf{j} + \frac{\partial}{\partial z}\mathbf{k}$, the gradient operator where **i**, **j** and **k** are unit vectors along the Cartesian axes;

$\rho\ (x,z)$ = the resistivity distribution where the strike direction is along y;

$U(x,y,z)$ = the potential distribution;

I = current strength of point source at the position of (x',y',z');

$\delta(x - x')$ = delta function.

The electric field generated by a point source in a 2D resistivity structure is 3D. To avoid the difficulty of solving a 3D problem numerically, equation (5.1) is reduced to 2D in the wave number domain by Fourier transformation along the strike direction (Dey & Morrison, 1979; Hohmann, 1988):

$$-\nabla \circ [\frac{1}{\rho(x,z)} \nabla \hat{U}(x,k_y,z)] + \frac{k_y^2}{\rho(x,z)} \hat{U}(x,k_y,z) = I\delta(x-x')\delta(z-z') \qquad (5.2)$$

where k_y is wave number and $\hat{U}(x,k_y,z)$ is the potential in the Fourier transform domain. Approximating the partial derivatives by partial differences over the x-z plane (see the discrete mesh Figure 4.1), equation (5.2) becomes

$$\hat{\mathbb{G}}\hat{\mathbb{U}} = \mathbb{S} \qquad (5.3)$$

where $\hat{\mathbb{G}}$ is an m-by-m matrix (m=number of nodes) and is a function of the geometry and the resistivity distribution $\rho(x_i,z_j)$, vector$\hat{\mathbb{U}}$ contains the potential at each node (in transform domain) corresponding to the current strength and the position of point source $\mathbb{S}$ (e.g. Dey & Morrison, 1979). Therefore

$$\hat{\mathbb{U}} = \hat{\mathbb{G}}^{-1}\mathbb{S} \qquad (5.4a)$$

After solving equation (5.4a) for several different wave numbers, the potential $U(x,y,z)$ is obtained by inverse Fourier transformation of $\hat{U}(x,k_y,z)$. Mathematically, (5.4a) can be rewritten as

$$\mathbb{U} = \mathbb{G}^{-1}\mathbb{S} \qquad (5.4b)$$

where $\mathbb{U} = F^{-1}\hat{\mathbb{U}}$, $\mathbb{G}^{-1} = F^{-1}\hat{\mathbb{G}}^{-1}$ and F^{-1} is inverse Fourier transform operator.

There are two main factors which affect the accuracy of FDM: the choice of wave numbers k_y and the number of cells between electrodes in the mesh (see Figure 4.1). With a higher density mesh and fine sampling of k_y, FDM produces more accurate potentials but also demands more computing resources.

5.2.2 Inversion

The forward modelling equation (5.4b) which relates resistivity distribution ρ (x,z) to potentials can be explicitly rewritten as

$$\mathbb{A}\mathbb{p} = \mathbb{U} \tag{5.5}$$

where $\mathbb{A}$ is an operator and vector $\mathbb{p}$ contains the discrete model parameters (i.e. resistivities) ρ_i, (i=1,2,...,M). The forward problem is, as stated in equation (5.5), a process of computation of unknown potentials from a known resistivity distribution $\rho(x,z)$. The inversion is the reverse process of forward modelling, i.e. for a known set of potential data, the question is how to relate them to unknown resistivity distribution. The process can be stated as

$$\mathbf{p} = \mathbf{B}\mathbf{U} \tag{5.6}$$

where $\mathbf{B}$ is the 'inverse' operator of $\mathbf{A}$. It seems that the resistivity image $\mathbf{p}$ can be reconstructed from boundary potentials $\mathbf{U}$ if operator $\mathbf{B}$ can be constructed. However, the nonlinear relationship between resistivity distribution and potentials (see equation 5.1) means that operator $\mathbf{A}$ (and hence $\mathbf{B}$) is a function of the resistivity image $\mathbf{p}$. Therefore the inverse operator $\mathbf{B}$ normally cannot be

constructed explicitly without approximation. To deal with this difficulty, the following approaches may be applied:

(1) *Perturbation method*

The nonlinear relationship can be linearized by the perturbation method and hence operator $\mathbb{A}$ is not a function of $\mathbb{p}$. Then operator $\mathbb{B}$ can then be constructed from the (pseudo) inversion of $\mathbb{A}$. The algorithm stated in Chapter 4 is one example of such an approach.

(2) *Model fitting by nonlinear function minimisation or optimisation*

In this approach the inverse operator $\mathbb{B}$ is not constructed explicitly. Instead, the field data are said to be inverted if one can find a resistivity model which fits the data. For simple 2D structures with a few parameters, model fitting can be carried out by a trial-and-error method (e.g. interactive modelling). If the 2D structures are complex and the model has tens or hundreds of parameters, as in the case of electrical resistivity image reconstruction, it can be very difficult to perform the trial-and-error method because of the complex coupling between parameters. Thus a more systematic procedure such as optimisation has to be applied.

Suppose that there are N measurements

$$\mathbf{U}^{\text{obs}} = [u_1^{obs}, u_2^{obs}, u_3^{obs}, \ldots, u_N^{obs}]^{\text{T}} \tag{5.7}$$

where T signifies the transpose. One can also calculate the N potentials $\mathbf{U}^{\text{pre}}$ (predicted data) from an estimated resistivity distribution $\mathbf{p}^{\text{est}} = [\rho_1^{est}, \rho_2^{est}, \rho_3^{est}, \ldots, \rho_M^{est}]^{\text{T}}$ through forward modelling.

$$\mathbf{U}^{\text{pre}} = [u_1^{pre}, u_2^{pre}, u_3^{pre}, \ldots, u_N^{pre}]^{\text{T}} \tag{5.8}$$

If the observed data $\mathbb{U}^{obs}$ are closely matched by the predicted data $\mathbb{U}^{pre}$, the resistivity distribution $\mathbb{p}^{est}$ is said to be the reconstructed image. The match normally will be not exact because of the noise in the data and also because the model used is only an approximation to the real earth structures. The *closeness*, or *goodness* of the match can be measured by the distance of the two data sets, i.e. the length (or *norm L*, denoted as $\|\cdots\|$) of the misfit, or error, vector $\mathbf{e}$ ($e_i = u_i^{obs} - u_i^{pre}$, $i=1,2,3,\ldots,N$):

$$L_1 \text{ norm:} \quad E_1 = \|\mathbf{e}\| = \left[\sum_{i=1}^{N} |e_i|\right] = \left[\sum_{i=1}^{N} |u_i^{obs} - u_i^{pre}|\right] \tag{5.9}$$

or

$$L_2 \text{ norm:} \quad E_2 = \|\mathbf{e}\| = \left[\sum_{i=1}^{N} |e_i|^2\right]^{1/2} = \left[\sum_{i=1}^{N} |u_i^{obs} - u_i^{pre}|^2\right]^{1/2} \tag{5.10}$$

where the L_2 *norm* is the familiar Euclidean length. Therefore the resistivity reconstruction now is a function minimisation or optimisation problem. In general, the model parameter values that minimise E_1 will not be the same as those which minimise E_2. The former is less affected by outlying data points than the latter because the E_2 function puts relatively more weight on larger errors than E_1 does. However, the least squares norm, E_2, is used most frequently because it implies that the data obey Gaussian statistics and hence a whole range of statistical procedures are available. The standard procedure for minimising the E_2 function (5.10) is the method of nonlinear least squares.

There are at least two problems with such approaches. First, the predicted errors or *data residual* alone are not a good measure of fitting goodness because the resistivity tomography is underdetermined (and also ill-posed), i.e. there are more unknown parameters than the independent measurements. Therefore many resistivity distributions may fit into equation (5.10) equally well

(non uniqueness). Some extra constraints must be applied to restrain the possible solution. Second, the error function E_2 may have many minima but the optimisation methods, such as the nonlinear least squares search, can only find the local minimum in the vicinity of the initial model, being unable to 'see' the whole error surface. The reconstruction will be successful if the initial guess is close to the true solution. Global optimisation methods, such as genetic algorithms, may be needed when the error function E_2 is complicated. These problems will be discussed in the following sections.

(3) *Neural networks*

The progress in artificial intelligence, for example involving neural networks, may provide a different approach to the electrical resistivity reconstruction problem. Several types of neural network may be applied to the ERT inversion. A simple method is that the nonlinear properties of the reconstruction are learned (from 'experience') and memorised in a pre-designed neural-like network. The reconstruction is then a one-step computation where the measured data is the input and the image is the output of the network. Only a simple discussion will be given to show the possibility of such an approach.

5.3 NONLINEAR LEAST SQUARES

5.3.1 Formulation

Nonlinear problems can be linearized successively and be solved by linear equation methods through successive iterations. Let $\mathbf{p}_k = [\rho_{k1}, \rho_{k2}, \ldots, \rho_{kM}]^T$ be an estimate of the model resistivities (parameters) for the k th iteration and let $\mathbf{U}_k^{pre} = [u_{k1}^{pre}, u_{k2}^{pre}, \ldots, u_{kN}^{pre}]^T$ be the corresponding potential data. The model response $\mathbf{U}^{pre}$ can be linearized by perturbing it about $\mathbf{p}_k$ with the first-order Taylor expansion:

$$\begin{aligned}\mathbf{U}(\mathbf{p})^{pre} &= \mathbf{U}_k^{pre} + \sum_{i=1}^{M} \frac{\partial \mathbf{U}^{pre}}{\partial \rho_i}\Big|_{\mathbf{p}=\mathbf{p}_k} (\rho_i - \rho_{ki}) \\ &= \mathbf{U}_k^{pre} + \mathbf{J}(\mathbf{p} - \mathbf{p}_k) \\ &= \mathbf{U}_k^{pre} + \mathbf{J}\,\delta\mathbf{p}\end{aligned} \tag{5.11}$$

where $\delta\mathbf{p} = \mathbf{p} - \mathbf{p}_k$ is the parameter correction vector and $\mathbf{J}$ is the N-by-M Jacobian matrix of partial derivatives with elements:

$$J_{ij} = \frac{\partial u_i^{pre}}{\partial \rho_j} \tag{5.12}$$

Then the error vector $\mathbf{e}$ is

$$\begin{aligned}\mathbf{e} &= \mathbf{U}^{obs} - \mathbf{U}^{pre} \\ &= \mathbf{U}^{obs} - (\mathbf{U}_k^{pre} + \mathbf{J}\,\delta\mathbf{p}) \\ &= \mathbf{e}_k - \mathbf{J}\,\delta\mathbf{p}\end{aligned} \tag{5.13}$$

where $\mathbf{e}_k = \mathbf{U}^{obs} - \mathbf{U}_k^{pre}$ is the misfit or error vector at the k th iteration.

If $\mathbf{p}$ is a model parameter vector which minimises equation (5.10), the gradient of E_2 with respect to model parameters will be

$$\nabla E_2(\mathbf{p}) = \mathbf{0}$$

that is

$$\nabla\left[\mathbf{e}^T\mathbf{e}\right] = \mathbf{0} \tag{5.14}$$

Substituting equation (5.13) into (5.14) and carrying out the gradient operation gives

$$-\mathbb{J}^T(\mathbf{e}_k - \mathbb{J}\delta\mathbb{p}) = \mathbb{0} \tag{5.15}$$

If $(\mathbb{J}^T\mathbb{J})^{-1}$ exists, the correction step is

$$\delta\mathbb{p} = (\mathbb{J}^T\mathbb{J})^{-1}\mathbb{J}^T\mathbf{e}_k \tag{5.16}$$

Notice that $\delta\mathbb{p} = \mathbb{p} - \mathbb{p}_k$ and let $\mathbb{p} = \mathbb{p}_{k+1}$. The equation (5.16) can then be rewritten as

$$\mathbb{p}_{k+1} = \mathbb{p}_k + (\mathbb{J}^T\mathbb{J})^{-1}\mathbb{J}^T\mathbf{e}_k \tag{5.17}$$

where k=0, 1, 2, 3, ..., K. This is the iteration procedure for solving the nonlinear problem of (5.10). Starting from an initial estimate $\mathbb{p}_0$ of the resistivity distribution, one can apply forward modelling described in section 5.2.1 to compute model response $\mathbb{U}_0^{pre}$. Then the error $\mathbf{e}_0 = \mathbb{U}^{obs} - \mathbb{U}_0^{pre}$ is computed and used to generate correction steps for each of the model parameters. The new estimate $\mathbb{p}_1$ of model parameters is obtained through equation (5.17). From this new estimate of model parameters, one can again compute another new estimate of model parameter in the same fashion and so on. This process will produce a model parameter series $\{\mathbf{p}_k\}$ which may converge to the nonlinear least squares problem (5.10). The nonlinear least squares method is also referred as a search method because it searches through the *model space* that contains all possible resistivity models to find one which best fits the data.

One difficulty in carrying out such an iteration is to compute the partial derivatives in Jacobian matrix $\mathbf{J}$ since the model response $\mathbf{U}^{pre}$ is not an analytic function of model parameters. A technique first proposed by Rodi (1976), which incorporates the computation of partial derivatives

into the forward modelling procedure, is used here. Because the source $\mathbb{S}$ in equation (5.3) is independent of model parameters, differentiating the equation gives

$$\hat{\mathbb{G}}\frac{\partial\hat{\mathbb{U}}}{\partial\rho_i}=-\frac{\partial\hat{\mathbb{G}}}{\partial\rho_i}\hat{\mathbb{U}} \tag{5.18}$$

Therefore

$$\frac{\partial\mathbb{U}}{\partial\rho_i}=-F^{-1}\left[\hat{\mathbb{G}}^{-1}\frac{\partial\hat{\mathbb{G}}}{\partial\rho_i}\hat{\mathbb{U}}\right] \tag{5.19a}$$

where F^{-1} is the inverse Fourier transform operator. Such an approach has two advantages. First, the computation is relatively fast compared with the method of difference approximation because matrix $\hat{\mathbb{G}}^{-1}$ has already been calculated by forward modelling. Second, the partial derivatives are computed as accurately as the electric potential since equation (5.19a) is an exact analytic formulation with the exception of the finite difference approximation.

In practice, the fitting errors and image parameters are usually scaled (or weighted); for example, the misfit can be scaled as $e_k=(u^{obs}-u_k^{pre})/u^{obs}$ and the resistivity can be represented in terms of a logarithm to ensure a non-negative value, i.e. $p_i=\ln\rho_i$ and $\mathbb{p}=[p_1,p_2,\ldots,p_M]^{\mathrm{T}}$. Then equation (5.19a) becomes

$$\frac{\partial\mathbf{U}}{\partial p_i}=\rho_i\frac{\partial\mathbf{U}}{\partial\rho_i}=-\rho_i F^{-1}\left[\hat{\mathbf{G}}^{-1}\frac{\partial\hat{\mathbf{G}}}{\partial\rho_i}\hat{\mathbf{U}}\right] \tag{5.19b}$$

and the elements of matrix $\mathbb{J}$ will be

$$J_{ij} = \frac{1}{(u^{obs})^2}\frac{\partial u_i^{pre}}{\partial p_j} \tag{5.20}$$

5.3.2 The Instability

The iteration in procedure (5.17) is very unstable in practice because ERT is an ill-posed and underdetermined problem. The ill-posedness implies that small changes in the observed data may cause correspondingly large variations in the reconstructed image. Consequently, the images may exhibit fluctuations with a roughness that is physically implausible. The underdetermined nature of the problem, for instance due to the overparameterization, means that the image reconstructed from procedure (5.17) is not unique, as mentioned in section 5.2.2. A closer examination may gain some insights into the problem so that certain remedial procedures can be applied to stabilise the iteration.

The *N*-by-*M* Jacobian matrix $\mathbb{J}$ in equation (5.17) can be decomposed by the method of singular value decomposition (SVD) (see chapter 2)

$$\mathbb{J} = \mathbb{U}\Lambda\mathbb{V}^T \tag{5.21}$$

where $\mathbb{U}^T\mathbb{U} = \mathbb{V}^T\mathbb{V} = \mathbb{I}$, $\Lambda = diag(\eta_i)$ and η_i is the singular value of $\mathbb{J}$ (i=1,2,3,...,M). Substitute equation (5.21) into procedure (5.17)

$$\mathbf{p}_{k+1} = \mathbf{p}_k + (\mathbf{V}\Lambda\mathbf{U}^T\mathbf{U}\Lambda\mathbf{V}^T)^{-1}\mathbf{V}\Lambda\mathbf{U}^T\mathbf{e}_k \tag{5.22a}$$

Notice that $(\mathbf{V}\Lambda\mathbf{U}^T\mathbf{U}\Lambda\mathbf{V}^T)^{-1} = (\mathbf{V}\Lambda^2\mathbf{V}^T)^{-1} = \mathbf{V}\Lambda^{-2}\mathbf{V}^T$, i.e.:

$$\delta\mathbf{p} = \mathbf{V}\Lambda^{-1}\mathbf{U}^T\mathbf{e}_k \tag{5.22b}$$

where $\Lambda^{-1} = diag(1/\eta_i)$. Then the effect of underdeterminacy and ill-posedness on the stability of iterations can be seen more clearly through the analysis of equation (5.22b).

The overparameterization means that the image is divided into more cells M than there are degrees of freedom in the measured data N (i.e. the number of independent data). In other words, the linearized iterative equation (5.15) is underdetermined and the N-by-M matrix $\mathbb{J}$ is singular, i.e. there are at least M-N singular values which are zero (or negligible). The solution to equation (5.15) is not unique because there are $\|\delta\mathbf{p}_i^{null}\| \neq 0$ such that $(\mathbb{J}^T\mathbb{J})\delta\mathbf{p}_i^{null} = \mathbb{0}$ (e.g. Press *et al*, 1989, p54). The general solution is

$$\delta\mathbf{p}^{gen} = \delta\mathbf{p} + \sum_{i=1}^{M-N} \alpha_i \delta\mathbf{p}_i^{null} \tag{5.23}$$

where $\delta\mathbf{p}$ is a *non-null* solution constructed by the pseudo inverse (i.e. setting $1/\eta_i = 0$ for each zero singular value in equation (5.22b)), and α_i is any scalar. One particular solution may be chosen by setting $\alpha_i = 0$ for all i. If so, the iteration will result in an image solution vector $\mathbf{p}$ with minimum length $\|\mathbf{p}\|$.

The ill-posedness of electric resistivity inversion implies that there may be some very small, but not zero, singular values besides those near zero η_i's due to the overparameterization. Those small singular values will lead to the very unstable iterations because if a singular value approaches zero, its reciprocal in equation (5.22b) approaches infinity. A small change in $\mathbf{e}_k$, for example the deviation due to data noise, will produce a large correction step $\delta\mathbf{p}$. This will be reflected in the reconstructed image as geologically unrealistic, high wave number fluctuations.

It may be argued that whilst the ill-posedness is the intrinsic problem of resistivity inversion, the under determinacy can be overcome by re-meshing the image space with a coarser mesh. It is partly true that the reduction of the number of unknowns may make iteration (5.22b) more stable because it may remove the singularity of $\mathbb{J}$ due to overparameterization. However, the question is that usually little is known for the features under the reconstruction and hence a coarse mesh may suppress the significant structure which the image contained. Furthermore, for the nonlinear problem of ERT, it cannot be assumed that N measurements imply that N local values of the image can be recovered by the inversion. This can be shown evidently if the integral equation (4.5) is discretised over a 2D (i.e. x-z plane) mesh with exactly N cells and rewritten as

$$\delta U(x_{pi}) = \sum_{j=1}^{N} \iint_{\Delta S_j} K(x,z,x_{p_i}) \delta\rho(x,z) dxdz \qquad (i=1,2,3,...,N) \qquad (5.24)$$

where

ΔS_j: the integration area of jth cell.

$K(x,z,x_{pi}) = \dfrac{1}{2\pi} \int_{-\infty}^{+\infty} \dfrac{J(x,y,z)\cos\theta}{\|\mathbf{r}-\mathbf{r}_{pi}\|} dy$, for $\mathbf{r}$, $\mathbf{r}_{pi}$ and J see equation (4.5).

θ : the angle between $J(x,y,z)$ and $\nabla\rho$.

Equation (5.24) is also an iterative procedure with the current distribution $J(x,y,z)$ and resistivity $\rho(x,z)$ being updated at each iteration. According to the integral mean value theorem, equation (5.24) may be rewritten as

$$\delta U(x_{pi}) = \sum_{j=1}^{N} \overline{\delta\rho_j} \iint_{\Delta S_j} K(x,z,x_{p_i}) dxdz \qquad (i=1,2,3,...,N) \qquad (5.25)$$

where the $\overline{\delta\rho_j}$ are clearly mean values of resistivity weighted with respect to the kernel $K(x,z,x_{pi})$ over the cell ΔS_j namely

$$\overline{\delta\rho_j} = \frac{\iint_{\Delta S_j} K(x,z,x_{p_i})\delta\rho(x,z)dxdz}{\iint_{\Delta S_j} K(x,z,x_{p_i})dxdz} \qquad (j=1,2,3,..,N) \qquad (5.26)$$

Therefore, under the best conditions where there are no degeneracies in equation (5.24), the N measurements can determine N mean values of the image in this sense. Clearly, this mean value representation is difficult to interpret in terms of the usual point to point, or spatial (algebraic) averaging description of an image.

However, if it is known *a priori* that the true resistivity distribution is smooth enough, in principle the data may have contained sufficient information to reconstruct the image. More specifically, one may design a mesh such that the term $\delta\rho(x,z)$ in equation (5.26) may be taken as one value $\delta\rho_j$ for each cell and hence be removed from the integration, i.e. $\overline{\delta\rho_j} = \delta\rho_j$. The iterative procedure (5.24) may converge to the true image. In practice, it is difficult to design such a mesh without *a priori* knowledge of subsurface structures.

Alternatively, the image may be represented by overparameterizing and the reconstruction is solved by the method of least squares mentioned above. The solution image will be nonunique but one may single out a solution by imposing extra constraints or *a priori* knowledge on it. For example it may be argued that the image should be 'simple' or 'smooth', or comply in part with borehole data. The smoothness constraints will also suppress the oscillations due to the ill-posedness and hence stabilise the iteration as will be shown in the next section.

5.3.3 Regularisation and Smoothing

As stated in the above section, the ill-posedness and the under determinacy of electric resistivity tomography will cause the iterative reconstruction to be unstable and the solution image to be non-unique. The underlying problem is that although the potential data provide information about the resistivities, the information is not sufficient to determine them uniquely. Some extra information must be added to constrain the image being reconstructed. This extra information, which quantifies the expectations about the character of the image that are not based on the actual data, is called *a priori* information (Jackson, 1979). For instance, one would expect the resistivities recovered to be non-negative.

One of the commonly accepted expectations about the reconstructed image (or solution) is that it should be 'simple' or 'smooth', especially for the overparameterised inversion problem (deGroot-Hedlin & Constable, 1990). The smoothness constraint can be incorporated into the definition of the object function to be minimised (as it will be stated later) , or be applied as a simply 2D low-pass filtering on the image within each iteration. The two methods can be applied collectively. The former is known as a *regularisation* method and the latter can be referred to as the spatial smoothing filter method.

Regularisation

In the method of regularisation, the misfit or error function has the general form (Phillips, 1962; Tikhonov, 1963)

$$E = \|\mathbf{e}\|^2 + \lambda\|\mathbf{S}(\mathbf{p})\|^2 \qquad (5.27)$$

where $\mathbf{S}(\mathbf{p})$ is the roughness (or the reciprocal of the smoothness) constraint imposed on image $\mathbf{p}$, for instance, the first derivative of $\mathbf{p}$. The regularisation parameter λ (>0) which regulates the balance between the data residual and image roughness, emerges as a Lagrange multiplier in the

misfit function. The inversion process is then to minimise the misfit function subject to the roughness $\|\mathbb{S}(\mathbb{p})\|^2$ being also minimised. Following the proposal by Tikhonov (1963), the roughness can be constructed by taking a linear superposition of spatial derivatives, i.e.:

$$\mathbb{S}(\mathbb{p}) = \alpha_0\mathbb{p} + \alpha_1\mathbb{p}' + \alpha_2\mathbb{p}'' + \ldots \tag{5.28a}$$

or in matrix notation

$$\mathbb{S}(\mathbb{p}) = \alpha_0\mathbb{p} + \alpha_1\mathbb{D}_1\mathbb{p} + \alpha_2\mathbb{D}_2\mathbb{p} + \ldots \tag{5.28b}$$

where $\mathbb{D}_1$ and $\mathbb{D}_2$ represent the first- and the second-order spatial difference operators, respectively. For simplicity suppose the constraint function $\mathbb{S}(\mathbf{p})$ contains only one term, i.e.

$$\mathbb{S}(\mathbb{p}) = \mathbb{D}\mathbb{p} \tag{5.29}$$

where $\mathbb{D}$ can be a difference operator of zero-, first-order or second-order. Following the same steps described in above section, it is shown that the iteration procedure for problem (5.14) in such case is

$$\mathbb{p}_{k+1} = (\mathbb{J}^T\mathbb{J} + \lambda\,\mathbb{D}^T\mathbb{D})^{-1}\mathbb{J}^T(\mathbf{e}_k + \mathbb{J}\mathbb{P}_k) \tag{5.30}$$

Different choices of operator $\mathbf{D}$ lead to different regularising characteristics. Higher order operators imply stronger constraints on the roughness of the reconstructed image. Most frequently used operators are zero-, first- or second-order.

Alternatively, the function $\mathbb{S}(\mathbf{p})$ may take the form

$$\mathbb{S}(\mathbb{p}) = \mathbb{D}\delta\mathbb{p} \tag{5.31}$$

where $\delta\mathbb{p}$ is the step size of parameters mentioned in the above sections. The corresponding iteration procedure will be

$$\mathbb{p}_{k+1} = \mathbb{p}_k + (\mathbb{J}^T\mathbb{J} + \lambda\,\mathbb{D}^T\mathbb{D})^{-1}\mathbb{J}^T\mathbf{e}_k \tag{5.32}$$

In the special case in which $\mathbb{D}$ is a zero order difference operator, equation (5.32) becomes

$$\mathbb{p}_{k+1} = \mathbb{p}_k + (\mathbb{J}^T\mathbb{J} + \lambda\,\mathbb{I})^{-1}\mathbb{J}^T\mathbf{e}_k \tag{5.33}$$

where $\mathbb{I}$ is the identity matrix. The procedure (5.33) is well known as the damped least squares method (DLS) or 'Marquardt-Levenberg' method. Clearly the DLS method takes the total length of the step size of each iteration as the constraints, i.e. minimising data error subject to the 'size' of the $\|\delta\mathbb{p}\|^2$ being also minimised.

The first-order spatial difference operator for a 2D image can be constructed in several ways. One method is to decompose it into two operators:

$$\mathbf{D} = \mathbf{D}_x + \mathbf{D}_z \tag{5.34}$$

where the operators $\mathbf{D}_x$ and $\mathbf{D}_z$ difference the image of laterally adjacent cells and of vertically adjacent cells, respectively. For a $l_x \times l_z$ mesh with l_x cells in x-direction and l_z cells in z-direction and if cells are numbered in columns first, the row vectors of the two operators may take the form respectively:

$$\mathbf{d}_x^i = \left[0,\ldots,0,-\alpha_i,0,\ldots,0,\alpha_{i+l_z},0,\ldots,0\right] \tag{5.35}$$

$$\mathbb{d}_z^j = [0, \ldots, 0, \ -\beta_j, \ \beta_{j+1}, \ 0, \ \ldots, 0] \tag{5.36}$$

where $i, j = 1,2,3,\ldots, l_x \times l_z$. The coefficients α_i and β_j can simply be set to 1. Alternatively, as suggested by deGroot-Hedlin & Constable (1990), one may set $\beta_j = 1$ but $\alpha_i = \Delta z_i / \Delta x$ (i.e. depth-to width ratio of the cell) and this is equivalent to increasing the roughness penalty as a function of depth because although the horizontal scale of the cells is the same, the vertical scale of the cells is increased along the z-axis (see figure 4.1).

The underlying reason that the regularisation method works is that it increases the values of diagonal elements of matrix $(\mathbb{J}^T\mathbb{J})$ and hence removes the effect of the zero or very small singular values. As an example, let us evaluate the simplest case of zero-order regularisation. Because $(\mathbb{J}^T\mathbb{J} + \lambda \mathbb{I}) = (\mathbb{V}\Lambda^2\mathbb{V}^T + \lambda \mathbb{I}) = \mathbb{V}(\Lambda^2 + \lambda \mathbb{I})\mathbb{V}^T$, equation (5.33) becomes

$$\mathbb{p}_{k+1} = \mathbb{p}_k + \mathbb{V} diag\left(\frac{\eta_i}{\eta_i^2 + \lambda}\right)\mathbb{U}^T\mathbb{e}_k \tag{5.37}$$

i.e. even when $\eta_i \to 0$ there is no danger of division by zero. A similar reason also holds for equations (5.30) and (5.32) where the matrix $(\mathbb{D}^T\mathbb{D})$ usually has non-zero elements in off-diagonal positions as well.

Spatial Smoother

A spatial smoother (or filter) also can be applied to suppress the oscillation of the reconstructed image. The smoothing operation can be carried out after the image has been constructed or during each iteration. The two types of spatial smoothers which will be considered are: the spatial average operator (i.e. localised average) and median filter.

The spatial average operator $\mathbb{S}_a$ is linear and may be different in terms of the size of window and the weighting method applied—a given cell value is replaced with a weighted average of the surrounding cells and itself. A 3-by-3 window, weighted by the areas of cells, will be tested. At the edges and the corners of the image, the size of averaging window will be reduced where the unsampled cells are excluded.

The median smoother $\mathbb{S}_m$ is similar to the averaging smoother, except that it is nonlinear in the sense that the value of the central cell is replaced by the median (not any linear average) of the neighbouring values and itself. A median filter smoother has two distinguishing advantages: it does not destroy high-contrast boundaries as an averaging smoother would and it is not affected by isolated spikes that may appear in the reconstructed image due to the ill-posedness and under determinacy. The 2D median smoother tested here is accomplished by a two-pass, 1D (i.e. row or column) median filter which is easy to implement.

Incorporating a spatial smoother $\mathbb{S}$ into the iterative procedures (5.30) and (5.32) results:

$$\mathbf{p}_{k+1} = \mathbb{S}(\mathbf{J}^T\mathbf{J} + \lambda\,\mathbf{D}^T\mathbf{D})^{-1}\mathbf{J}^T(\mathbf{e}_k + \mathbf{J}\mathbf{p}_k) \tag{5.38}$$

$$\mathbf{p}_{k+1} = \mathbb{S}\left[\mathbf{p}_k + (\mathbf{J}^T\mathbf{J} + \lambda\,\mathbf{D}^T\mathbf{D})^{-1}\mathbf{J}^T\mathbf{e}_k\right] \tag{5.39}$$

where the spatial smoother $\mathbb{S}$ can be either $\mathbb{S}_a$ or $\mathbb{S}_m$. The differential operator $\mathbf{D}$ can be zero, first or second-order and the corresponding iterative procedures are referred as *zero-order, first-order* or *second-order regularisation*, respectively. The zero-order regularisation in equation (5.39) is also known as the damped least squares method, as has been mentioned above.

Regularisation Parameter λ

The success of *regularised* (or *smoothness-constrained*) least squares methods depends not only on the choice of smoothing function but also on the choice of regularisation (or damping) parameter λ which determines the amount of smoothness constraints to be incorporated in each iteration. A large λ may lead to a stable but also excessively smoothed image whilst a too small λ may cause the iteration to be unstable. Phillips (1962) argued that any λ which makes the misfit smaller than or equal to the data noise should be acceptable. The difficulty is that the exact amount of noise in the data is usually not known in practice. For the method of damped least squares, Marquardt (1963) proposed an elegant method to change λ in each iteration, i.e. in the *k* th iteration, if the fitting error is decreasing compared with the *k-1* th iteration, the λ is decreased by a factor α and otherwise it is increased by a factor of β where $\alpha<1$ and $\beta>1$. The success of Marquardt method depends on the initial λ as well as the factors α and β. It is difficult to select suitable values for these parameters without experiments because they are context dependent.

Intuitively one may choose the regularisation parameter λ which leads to reduce the data residual in each iteration. This can be achieved by a 1D search operation after each iteration (Constable et al., 1987). The λ is selected as a minimizer of the misfit function $E(\lambda)$:

$$E(\lambda) = \left\| \mathbb{U}^{obs} - \mathbb{U}^{pre}(\lambda) \right\|^2 \tag{5.40}$$

where $\mathbb{U}^{obs}$ is the measuring data vector and $\mathbb{U}^{pre}(\lambda)$ is the calculated data in current iteration. This method will be evaluated in Chapter 7.

The disadvantage of above strategy is that it does not take into account the roughness penalty in selecting λ . A more general approach is the method of Generalised Cross-Validation (GCV) due

to Golub et. al (1979). In this method, the parameter λ is selected as the minimizer of the GCV function $V(\lambda)$, defined by

$$V(\lambda) = \frac{\left\| Ax_\lambda - b \right\|^2}{\left[trace(I - A(A^T A + \lambda L^T L)^{-1} A^T) \right]^2}$$

where I is an identity matrix and x_λ is the regularised least squares solution to the linear least squares problem

$$\min_x \left\{ \left\| \mathbb{A}x - \mathbb{b} \right\|^2 + \lambda \left\| \mathbb{L}x \right\|^2 \right\}$$

Accordingly, the GCV function for regularised non-linear least squares procedures (5.30) or (5.32) may be defined as

$$V(\lambda) = \frac{\left\| \mathbb{U}^{obs} - \mathbb{U}^{pre}(\lambda) \right\|^2}{\left[trace(\mathbb{I} - \mathbb{J}(\mathbb{J}^T \mathbb{J} + \lambda \mathbb{D}^T \mathbb{D})^{-1} \mathbb{J}^T) \right]^2} \tag{5.41}$$

where $\mathbb{J}$ is the Jacobian matrix and $\mathbb{D}$ is the difference operator.

The function $V(\lambda)$ can be viewed as a weighted version of equation (5.40) where the weight function is a measure of complexity of model structure: if λ is too small, the resultant model may not be smooth enough and this reflects on the $V(\lambda)$ function as its weight function $1/trace(\mathbb{I} - \mathbb{J}(\mathbb{J}^T \mathbb{J} + \lambda \mathbb{D}^T \mathbb{D})^{-1} \mathbb{J}^T)$ approaches infinity, and vice versa. Although the GCV method is most frequently used in the linear least squares method, it may also be applicable to non-linear least squares problems (e.g. O'Sullivan & Wahba, 1985) and this will be tested in Chapter 7.

5.3.4 Discussion

One may question whether the regularisation method is justified for electrical resistivity tomography. For example, one may argue that oscillatory structures may actually be present in the target structures and hence the smoothness constraints may remove desired features in the reconstructed image. However, it may also be argued that the signature of any such structures would be close to data noise because of the low-pass filtering characteristics of the earth. This can be evidently shown in the 2D apparent resistivity pseudosection where the data change is usually much smaller than the parameter change in the model. In other words, the oscillatory structures may exist but the observed data are simply not accurate enough to resolve them. To ensure that only real structure is recovered, therefore, suppression of oscillatory components by smoothing is essential.

Another problem is that there is no guarantee that the iterative method described in the sections above will converge to the 'true' image. As mentioned above, the least squares method is a linearized, local search algorithm and hence has the potential danger of being trapped into a local minimum or even a maximum (c.f. Menke, 1989, p157). The global search strategy may need to be considered.

During recent years, considerable progress has been made in developing global optimisation algorithms, for example, simulated annealing, and currently in artificial intelligence, genetic algorithms and neural network studies. These methods are truly nonlinear (i.e. no successive linearized steps as the methods are based on local gradient information). Although the algorithms are still under development, they may have some potential for ERT detection problems in the future.

5.4 GENETIC ALGORITHM

Genetic algorithms (GA) are optimisation algorithms based on Darwinian models of natural selection, evolution, and the mechanics of genetics. The initial development was made by Holland (1975) and recent progress has been summarised by Davis (1987) and Goldberg (1989). Some pioneer experiments in applying GA to solve seismic inversion have been carried out recently (e.g. Stoffa and Sen, 1991). Genetic algorithms start with a population of likely solutions (or models, or images) to the problem to be solved. The basic idea behind GA is that large and complex solutions can be represented by simple binary strings. These strings are analogous to chromosomes in natural genetics and their components are called genes. The strings can then be manipulated using genetic operators, such as crossover and mutation, to generate new strings (i.e. new solutions). For GA to improve the solutions, it is necessary to reject the worst and to keep the best. This is again an analogy to the 'survival of the fittest' law, which only allows an organism that adapts best to a natural environment to survive. The environment in GA is played by the objective function to be optimised and the fitness is measured by an objective function value.

As summarised by Davis (1987), the GA requires five components: (1) a GA representation of the problem; (2) a method for creating an initial population; (3) a function (playing the role of the environment) verifying the fitness of each individual in the population; (4) GA operators for changing a gene's contents in a chromosome; (5) some constants, such as the size of population, probabilities of crossover and mutation. The next sections will follow this line to formulate a GA for electrical resistivity tomography.

5.4.1 Representation—image coding

For image vector $\mathbf{p} = [p_1, p_2, \ldots, p_M]^T$ where p_i is the resistivity ρ_i or $\ln \rho_i$, if the ρ_i is bounded, i.e. $p_{\min} \leq p_i \leq p_{\max}$, it can be represented as a binary string with d bits

$$p_i^s = b_1 b_2 b_3 \ldots b_d \qquad (b_j = 0 \text{ or } 1, \quad j = 1,2,\ldots,d) \tag{5.42}$$

and the mapping relation from p_i^s to p_i is

$$p_i = p_{\min} + \frac{p_{\max} - p_{\min}}{2^d - 1} \sum_{j=1}^{d} b_j 2^{d-j} \tag{5.43}$$

The resolution (or precision) of the mapped coding is $(p_{\max} - p_{\min})/(2^d - 1)$. For example, if p_i=8 and is known within the range of [2, 17], its representation as a 4 bits binary string would be 0110 and the resolution is 1. Clearly higher resolution requires a longer binary string. Each such string can be taken as an individual 'chromosome', or alternatively in our work, all such strings for the image are concatenated into a single long string. The latter approach is more easy to program.

The initial population of images can be generated by randomly assigning 0 or 1 to each bit of the string. If the size of population is Q, the population is

$$\mathrm{P}(t) = \left\{ \mathrm{p}_1^s(t), \mathrm{p}_2^s(t), \ldots, \mathrm{p}_Q^s(t) \right\} \tag{5.44}$$

where t denotes the generation (i.e. t=0) and the binary string $\mathrm{p}_k^s(t) = b_1(t) b_2(t) \ldots b_{M \times d}(t)$ represents the k th image at the t th generation. The total length of the string is $M \times d$ where M is the number of cells and d is the length of string coded for the image parameter of each cell.

5.4.2 Objective function

The objective function, which plays the role of environment, can be any measure of the goodness of model fitting stated above. In fact because no derivative information is needed in GA the choice of objective function is more flexible compared with the algorithms based on local gradient

information. However, GA is designed to find the chromosome with maximum fitness and hence the minimisation problem has to be transformed into the maximisation. This can be accomplished by the reciprocal transform if the misfit is known to be non-zero or a simple linear relation such as

$$fitness_k = E_{\max} - \left(\|\mathbf{e}_k\| + \lambda \|\mathbb{D}\mathbf{p}_k\|\right) \tag{5.45}$$

where $E_{\max}$ is the maximum possible fitting error, $\mathbf{e}_k$ is the fitting error between observed data and predicted data (through the model associated with the image string), λ is the regularisation parameter and $\mathbb{D}\mathbf{p}_k$ is the spatial difference of the kth image within the population of the current generation. The norm $\|...\|$ can be L_1 or L_2.

In each generation, one can apply equation (5.45) to evaluate each image within the current population and obtain a fitness data set $[fitness_1, fitness_2, ..., fitness_Q]$ where Q is the size of population.

5.4.3 Reproduction

Reproduction is a process of selecting the images within current population and using crossover and mutation operators to generate a new population. It includes three actions: selection, crossover and mutations.

The selection is a stochastic sampling based on the relative performance (fitness) of the individual image. The probability of selection for the k th individual at the t th generation $P_s^k(t)$ is defined as

$$P_s^k(t) = \frac{fitness_k(t)}{\sum_{l=1}^{Q} fitness_l(t)} \tag{5.46}$$

Then the pseudo-procedure for the selection can be

```
procedure selection
  begin
    P_r=random[0,1]          {uniform random number generator over 0 to 1}
    from k=1 to Q
      if P_s^k(t) ≥ P_r then select image p_k^s(t)
    until an image has been selected
  end
```

The crossover is an operation whereby a pair of strings (chromosomes) exchange part of their components with some probability P_c. After the images have been paired via the selection, a random test is carried out to see whether the crossover should occur. If it is true, first the crossover position in the string is randomly selected, say the kth bit from left, and then all characters after this position are swapped between the two (parent) strings, producing two new (child) strings. For example, if there are a pair of 7 bits binary strings A, B and a random test indicates the crossover position is 4 from left, the crossover will generate two new strings C and D as

$$\left.\begin{aligned} A &= a_1a_2a_3a_4\underline{a_5a_6a_7} \\ B &= b_1b_2b_3b_4\underline{b_5b_6b_7} \end{aligned}\right\} \xrightarrow{crossover} \begin{cases} C = a_1a_2a_3a_4\underline{b_5b_6b_7} \\ D = b_1b_2b_3b_4\underline{a_5a_6a_7} \end{cases}$$

where a_i and b_i is 0 or 1. A simple pseudo-procedure for crossover is

```
procedure crossover
  begin
    select parent1           {use selection procedure to pick up two strings
    select parent2            from current population}
    P_r=random[0,1]          {uniform random number generator over 0 to 1}
    if P_c ≥ P_r then        {crossover}
      begin
        k=random[1, l-1]           {l= the length of string}
        child1=first k characters of parent1+last (l-k) character of parent2
        child2=first k characters of parent2+last (l-k) character of parent1
```

```
    end
  else                          {no change }
    begin
      child1=parent1
      child2=parent2
    end
  end
end
```

Crossover is the driving force behind GA. It takes the 'genetic materials' from currently high performance individuals as the building blocks to produce future generations and hence ensures that the population evolves towards a better version. Clearly such a swapping process is simply an exchange of 'genetic materials' within the current population but cannot create the new 'genetic materials' which the population does not possess. For instance, if the population contains only two individuals A=100101 and B=000110, the simple crossover stated above will never be able to produce any string with the second and third positions (from left) being *1*, i.e. any string like *11*** where * represents 0 or 1. Mutation is such an operation to add diversity to the population. It is simply the alteration of a randomly selected bit in the string subject to a specified probability P_m. In the above example if the randomly selected bit happens to be the second, the mutation operation will create structures like *10***. A simple pseudo-procedure for mutation can be

```
procedure mutation
  begin
    select parent               {use selection procedure to pick up one string}
    P_r=random[0,1]             {uniform random number generator over 0 to 1}
    if P_m ≥ P_r then           {mutation occurs}
      begin
        k=random[1, l]                  {l= the length of string}
        child⇐ change the k th bit of parent (i.e.1 to 0, 0 to 1)
      end
    else                        {no change }
      child=parent
    end
  end
```

If crossover and mutation are taken as a recombination of images from current population to produce next generation population, the main pseudo-procedure for GA can be

```
procedure GA
  begin
    t=0
    initial P(t)
    evaluate P(t)                    {compute the fitness for each individual}
    while (termination is not true)
      begin
        select P(t)
        t=t+1
        recombine P(t)
        evaluate P(t)
      end
  end
```

The termination condition can be a pre-specified misfit value, maximum number of generation (iteration) or standard deviation of fitness of the population. The population size Q, crossover probability P_c and mutation probability P_m are important constants which normally have to be pre-specified. Large populations have more diverse images and hence GA is more likely to succeed but the computing cost for each generation (iteration) will be high. Higher mutation probability will also produce more diversity in the population but GA may never be able to converge if it is too high. Goldberg (1989) suggest that $Q \geq 50$, $P_c \approx 0.6$ and $P_m \approx 0.001$ to 0.01. These constants should be tuned for each problem.

There are several points that should be noticed. Firstly GA manipulates the coding string of image parameters, not (directly) the parameters themselves. Secondly GA produces a group (population) of possible (optimal) solution images, not a single one. Thirdly there is no guarantee that GA will find the global minimum although it frequently outperforms other more direct search methods such as gradient descent on simulated difficult problems and has shown good efficiency and robustness (e.g. Goldberg, 1989). Furthermore GA is much slower compared with local search methods such as least squares methods. One pertinent aspect is that GA is an implicitly parallel algorithm where individuals can be evaluated simultaneously and hence it may be tailored to run fast on parallel computers.

5.5 NEURAL NETWORK

Neural networks, which aim to imitate some of the functions of the brain using its nerve structure, comprise one of the fastest developing fields of artificial intelligence (AI). The basic idea is that the brain can be modelled by inter-connected units which serve as model *neurons* (i.e. synthetic neurons). Each unit converts the incoming signals that it receives into a single outgoing activation that it broadcasts to other units. It performs this conversion in two steps. First, it multiplies each incoming activation by the weight on the connection and adds together all these weighted inputs to get a total input (or total activation). This process can be modelled by an overall input function. Second, the unit uses an activation function to transform the total input into the outgoing activity (e.g. Figure 5.1). With such a network, a 'knowledge database' can be learned (taught) and stored, and later when it is stimulated with (similar) input it can use the knowledge stored as weights to give the desired output, or even be able to be re-trained to adapt new knowledge. Recently progress in this field has been summarised by Rumelhart & McClelland (1986) and McClelland & Rumelhart (1988).

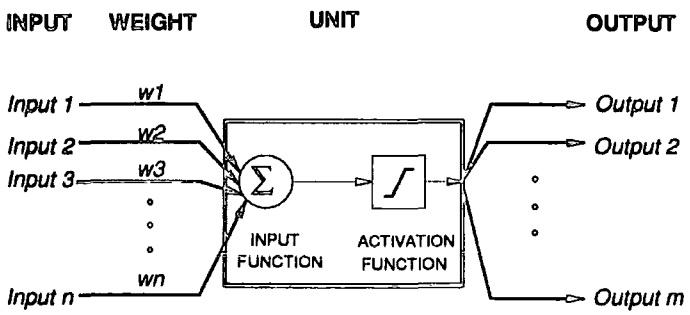


Figure 5.1 Synthetic neuron model

The neuron inputs may be real or Boolean numbers. The total input function usually has the form

$$I_i = \sum_j w_{ij} o_j + bias_i \tag{5.47}$$

where I_i is the overall input, w_{ij} is the connection weight between the ith unit the jth, and o_j is the output from the jth unit to the ith unit. The $bias_i$ is the bias term of the unit which incorporates some a *priori* information such as that the unit has a preference to be turned on or off. The

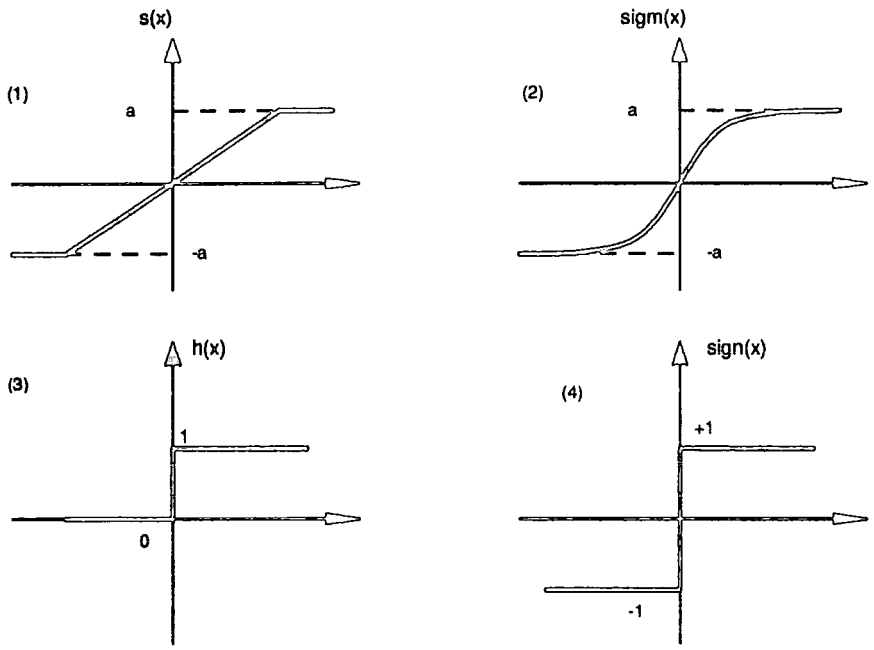


Figure 5.2 Examples of activation function for neural networks: (1) linear threshold function; (2) sigmoid function; (3) and (4) binary functions. (after Rumelhart & McClelland 1986)

activation function is generally chosen to be monotonic and odd (with thresholds), for example, the linear threshold function s(x), the sigmoid function sigm(x) and the binary functions h(x) or sign(x) as shown in Figure 5.2. The commonly applied activation function for layered neural networks is the sigmoid function with the form:

$$f(x)=\frac{1}{1+e^{-x}} \tag{5.48}$$

5.5.1 Multilayer Model

Although neurons may be connected to each other in many different ways, only two types of architecture are frequently used; the multilayer model and the Hopfield model (e.g. McClelland &

Rumelhart, 1988; Hopfield & Tank, 1985). The Hopfield model, in which the neurons are fully interconnected, will not be considered here because it is still under development. Instead, the multilayer model is examined. In the simple multilayer model the network is arranged into successive layers of neurons and the connections are only allowed to be between consecutive layers (see Figure 5.3). The information transmission is in one direction only, i.e. each layer receives

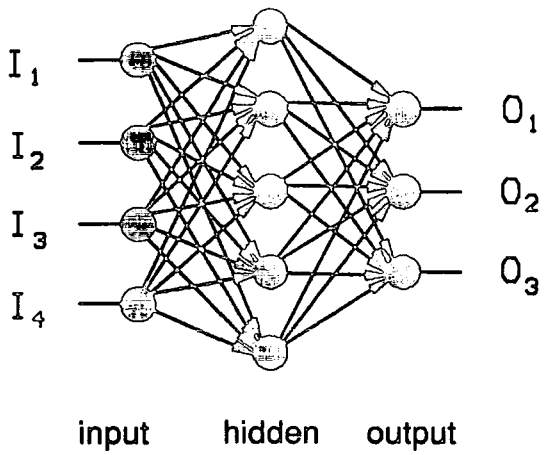


Figure 5.3 An example of multilayer neural network model.

signals only from the previous layer and transmits to the following one. More recent models permit the connections within each layer and the feedback from outputs. The two outermost layers are called the input layer, which receives inputs from the external world, and the output layer which outputs the results of processing, respectively. The intermediate layers are called hidden layers.

When inputs are presented, the network will propagate the signals from input through hidden layers to the output layer. The input and output activations for each neuron can be calculated from equations (5.47) and (5.48). This process can be operated in parallel mode or serial model. Suppose that time is discrete and each time step is marked by a clock tick. In parallel mode, at

each clock tick each neuron calculates its new input activation (when it is available) and its output, and transmits it to the next neurons which it connected. In serial model, each time only one neuron performs such actions. The intrinsic parallelism of neural networks implies that the parallel model can be easily implemented on parallel computers or hardware neural networks.

Association is the main power of such multilayer neural models, i.e. it can associate the inputs to the outputs. For this reason, it is usually used in classification and pattern recognition. However, the layered neural network may also be suitable for resistivity image reconstruction. That is, it may be possible to design such a multilayer neural network such that after training (see next section) its output is the reconstructed image when the boundary potential measurements are presented to its input layer. The nonlinear features between measured data and resistivity structures are then reflected on the hidden layers. There are at least two advantages with such an approach. First, the reconstruction process is fast because it involves only a one-step forward propagation of signals through the network. Second, the training data may be taken from not only resistivity forward modelling but also from field measurements, or even from laboratory experiments. Therefore deficiencies due to the use of a simplified physical model for resistivity structures used in forward modelling may be avoided.

5.5.2 Learning Algorithm: Backpropagation

The learning algorithm generally used for a multilayer model is the error *back-propagation* method which is the extension of the Widrow-Hoff rule (or delta rule) for a single layer model (e.g. McClelland & Rumelhart, 1988). Before training, the network weights are usually initiated with random numbers. The training process has two stages. First, a training example is presented to the input layer of the network. The actual outputs are computed from current network weights, proceeding in order from the input layer, through intermediate layers, to the output layer. This is referred to as 'forward propagation'. Next, the total error between desired outputs and actual outputs is calculated. The error then is backpropagated from the output layer, through

intermediate layers in turn, to the input layer, leading to the modification of each connection weight. This process is repeated, with the same example or a set of examples in turn, until the minimum error is reached.

Mathematically, the training (or learning) is a function minimisation (or optimisation) problem with the objective function similar to equation (5.10), i.e.

$$E(w) = \sum_{k=1}^{m} E^k(w) \tag{5.49a}$$

where

$$E^k(w) = \sum_{i=1}^{n} (y_i^{cal} - y_i^{desir})^2 \tag{5.49b}$$

and y_i^{desir} and y_i^{cal} are the desired outputs and the calculated outputs of the network with the kth example presented, respectively. The error function $E^k(w)$ can be minimised by a gradient descent procedure

$$w_{ij}(k) = w_{ij}(k-1) - \omega \frac{\partial E^k}{\partial w_{ij}} \tag{5.50}$$

where ω is the step size and w_{ij} is the connection weight between the ith neuron to the jth neuron. Rumelhart & McClelland (1986) have deduced that equation (5.50) is equivalent to

$$w_{ij}(k) = w_{ij}(k-1) - \omega \delta_i o_j \tag{5.51}$$

where o_j is the output from the jth neuron to the ith neuron and

$$\delta_i = 2f'(I_i)(y_i^{cal} - y_i^{desir}) \qquad \text{(output layer)} \tag{5.52}$$

$$\delta_i = f'(I_i)\sum_h \delta_h w_{hi} \qquad \text{(hidden layer)} \tag{5.53}$$

The $f'(I_i)$ is the derivation of the activation function of the ith neuron and I_i is the total input of neuron i. The variable h ranges over the number of neurons to which neuron i sends signals (see Figure 5.4). As indicated above, for carrying out the gradient descent procedure it is necessary to start from the output layer, recursively computing the δ_i error term in a successive and backward way. For this reason, the gradient descent procedure here is referred as an error *backpropagation* method.

In resistivity image reconstruction, both the measured data and the image parameters can vary over a large range and hence it is necessary to map them into a regular range, say [0, 1], for training and reconstruction.

5.5.3 Difficulties and Limitations

Neural network theory is a relatively young field. Although it has been proved to be useful in a number of problems, it still has some limitations and difficulties which are not yet resolved. First, there is no theoretical basis nor even a satisfactory empirical rule suggesting how a network should be constructed to solve a particular problem. Second, the backpropagation, deduced from the gradient descent method, is a local optimisation method. It therefore may converge to a local minimum or even not converge at all. Third, in spite of its intrinsic parallelism, neural networks implemented on conventional computers are still working in serial mode and the backpropagation, operating on a large scale network or a large training data set, requires large computing resources.

Perhaps the most difficult problem is that there are so many resistivity structures, with different geometries and resistivities that it may prove impossible to train the network with adequate examples. Experience suggests that the multilayer neural networks only work best for structures within the scope of the training set and otherwise they may fail.

CHAPTER 6 IMAGE PRESENTATION AND PROCESSING

In the past, 2D resistivity structures were shown as a tabulated list of resistivity values or a contour map (e.g. Oristaglio & Worthington, 1980). More recently they are represented as grey-scale or pseudo-colour images. By comparison, the method of tabulation is the most difficult to visualise when the number of resistivity values is large although it maintains all reconstructed information, such as the parameter values and their corresponding coordinates. The contoured map can at best be viewed as a crude image with a relatively low spatial resolution and it is still difficult to visualise although it has the advantage that the image values can easily be estimated from the values on the contouring lines. Recent development of using colours or grey scales to shade the areas between contour lines makes it more visual. However, for higher spatial resolution, a grey-scale or pseudo-colour image may be the best for representation of 2D structures.

This chapter will discusses some problems associated with the image representation of 2D resistivity structures. First, the method of the image representation of resistivity structures using grey-scale and pseudo-colour is presented. Then the image enhancement techniques are discussed, such as the histogram equalisation and logarithmic transformation methods.

6.1 IMAGE REPRESENTATION

6.1.1 Grey scale image

In this method, resistivity structures are shown as a grey-scale image (i.e. graded black to white image). As shown in Figure 4.1, the subsurface section is divided into small cells. If the resistivity value of each cell is transformed into a grey value (i.e. luminance intensity), it is then possible to

present the subsurface resistivity structure as a grey-scale image on a display device such as a computer monitor or printer.

Suppose the resistivity values cover a range (ρ_{min}, ρ_{max}) and the display device can show grey levels from m to M (from the lightest to the darkest grey). The transformation is then

$$n_g(x,z) = \frac{(\rho(x,z) - \rho_{min})}{(\rho_{max} - \rho_{min})}(M - m) + m \qquad (6.1)$$

where $n_g(x,z)$ is the grey level corresponding to resistivity value $\rho(x,z)$ of the cell at (x,z). A modern display device, such as a VGA computer monitor or a laser printer, may be able to display up to 256 grey scales or more but human eyes may only be able to distinguish less than one hundred grey levels.

For a VGA computer monitor, a Pascal program has been written to display the reconstructed resistivity structure as a 16 grey-scale image. A postscript printer driver has also been developed to download the image to a postscript laser printer in this thesis.

6.1.2 Pseudo-colour image

The reconstructed resistivity distribution can also be shown as a 2D pseudo-colour image in a similar way as the grey-scale image. Instead of mapping the resistivity value of each cell into a single grey level, it is necessary to transform this resistivity value into a set of three numbers (r, g, b) where r, g and b represent the intensity of red, green and blue respectively (i.e. the colour palette). It is therefore difficult to define an optimum mapping, not only because of the non-unique choices of one dimension (resistivity value) to three dimensions (r, g, b intensity) mapping but also

because of the effects of colour perception. In general, the use of colour in representation of 2D resistivity structures is limited only by artistic imagination.

However, in practice it may be argued that some constraints should be imposed on the mapping. First it is desirable that the target in the image is represented by a colour (and intensity) which is different from its background and hence stands out more clearly to a human viewer. Therefore the mapping may need to be adjusted according to the problem at hand. Second, the change of colour and its intensity from one to another should be smooth. Otherwise it will lead an image with artefacts of colour boundaries. The mapping of resistivity values into different r, g, b values can be accomplished by two steps. Initially, the resistivity values are transformed into grey levels by equation (6.1). Then these grey levels n_g are mapped into different colour intensities, for example, by a *sine* function:

$$r = N_r \times \left| Sin\left(\frac{n_g \pi}{n_{g\max}} \right) \right| \tag{6.2a}$$

$$g = N_g \times \left| Sin\left(\frac{n_g \pi}{2n_{g\max}} + \frac{\pi}{7} \right) \right| \tag{6.2b}$$

$$b = N_b \times \left| Sin\left(\frac{n_g \pi}{3n_{g\max}} + \frac{\pi}{15} \right) \right| \tag{6.2c}$$

where r, g and b are the intensities of red, green and blue respectively, the N_r, N_g and N_b are maximum values of r, g, b respectively. The $n_{g\max}$ is the maximum value of grey levels. The above mapping produces a colour palette that varies smoothly from blue to dark red. Another method is to use a pre-specified colour palette, for example that in Table 6.1.

Table 6.1 Grey scale to colour palette mapping. (after Hall, 1979, p286)

Grey Level	Red	Green	Blue
1	1	255	255
2	1	238	255
3	1	221	255
4	1	204	255
5	1	187	255
6	1	170	255
7	1	153	255
8	1	136	255
9	1	119	255
10	1	102	255
11	1	85	255
12	1	68	255
13	1	51	255
14	1	34	255
15	1	17	255
16	1	1	255
17	1	17	238
18	1	34	238
19	1	51	221
20	1	68	204
21	1	85	187
22	1	102	170
23	1	119	153
24	1	136	136
25	1	153	119
26	1	170	102
27	1	187	85
28	1	204	68
29	1	221	51
30	1	238	34
31	1	255	17

Grey Level	Red	Green	Blue
32	1	255	1
33	17	255	1
34	34	255	1
35	51	255	1
36	68	255	1
37	85	255	1
38	102	255	1
39	119	255	1
40	136	255	1
41	153	255	1
42	170	255	1
43	187	255	1
44	204	255	1
45	221	255	1
46	238	255	1
47	255	255	1
48	255	238	1
49	255	221	1
50	255	204	1
51	255	187	1
52	255	170	1
53	255	153	1
54	255	136	1
55	255	119	1
56	255	102	1
57	255	85	1
58	255	68	1
59	255	51	1
60	255	34	1
61	255	17	1
62	255	1	1
63	255	1	1

The great strength of pseudo-colouring is that the subtle variations within an image can be more easily distinguished because the human visual system is highly sensitive to colour. In addition, colour images are generally much more pleasant to view than black-white images. It is through this aesthetic aspect of colour that the pseudo-colouring method is often referred to as an 'image

enhancement' technique. Alternatively, this strength can also be a disadvantage of using colours because if the colour palette is not arranged properly, it may emphasis some features of an image which may be undesirable, for example, the artefacts of the reconstruction, at the expense of others. Therefore the pseudo colouring should be used cautiously with a carefully arranged colour palette.

A Pascal program has been developed to display a reconstructed resistivity image on a colour VGA monitor using a colour palette of 14 mapped by equations (6.2a-c). The pseudo-colour image can also be download to a colour printer using the mapping relationship of equations (6.2a-c) or Table 6.1. A comparison of a grey scale image and pseudo-colour images for the same image data are shown in Figure 6.1.

6.2 IMAGE ENHANCEMENT

Image enhancement is the processing of images to improve their appearance to human viewers. Methods and objectives of image enhancement vary with applications. For example, low pass filtering may remove some noise in an image and hence improve its visual quality. If image intensities (grey values) in some region are sufficiently high that features with low image intensities cannot be displayed on a given device, then a logarithmic transformation of contrast may be desirable. The criteria for enhancement are often subjective because they rarely can be qualified by a useful objective measure. Various image enhancement methods have been developed (e.g. Hall 1979, p158-214) but only the contrast transformation, histogram equalisation and image smoothing methods are considered here.

6.2.1 Logarithmic transformation

The dynamic range of a grey-scale image is very limited because the human visual system can probably only distinguish tens of grey levels, as mentioned above. However, the subsurface

resistivity values may change from a few ohms to ~10^3 ohm·m or much more. If such high contrasts occur even in a small region of an image, the structures with low grey values will be overshadowed, making them invisible on such an image. In such cases, it may be better to show the resistivity image on a logarithmic scale, i.e. the resistivity values are transformed into a logarithmic scale before being mapped into the grey scale by equation (6.1):

$$p(x,y) = \log \rho(x,z) \tag{6.3}$$

where $\rho(x,z)$ and $p(x,y)$ are the original and the transformed resistivities respectively. This is usually referred as *logarithmic contrast transformation* in standard image enhancement methods.

Logarithmic transformation is nonlinear which expands the contrast of the small grey values and compresses that of large values. It improves an image appearance by making the low-contrast detail more visible. In addition, it transforms multiplicative noise of an image into additive noise. Figure 6.2 compares an image with high resistivity contrast to its logarithmic transformation. The structures with small grey values in the original image are hardly visible but they are clearly indicated on the transformed image.

6.2.2 Histogram equalisation

A more general grey scale transformation is the histogram modification. The *histogram* of an image, denoted by $p(f)$, represents the number of cells that have a specific intensity f (grey level) as a function of f (see Figure 6.3). The histogram modification is to transform grey scales of an image to produce a desired histogram. If the desired histogram is uniformly distributed (i.e. constant, see Figure 6.3), the histogram modification is called *histogram equalisation.* Such a modification can be carried out in two steps. First the cumulative histograms, $S(f)$ for the original image and $S_d(g)$ for the desired image, are computed

$$S(f) = \int_0^f p(t)dt = \sum_{k=0}^{f} p(f) \tag{6.4}$$

$$S_d(g) = \int_0^g p_d(t)dt = \sum_{k=0}^{g} p_d(g) \tag{6.5}$$

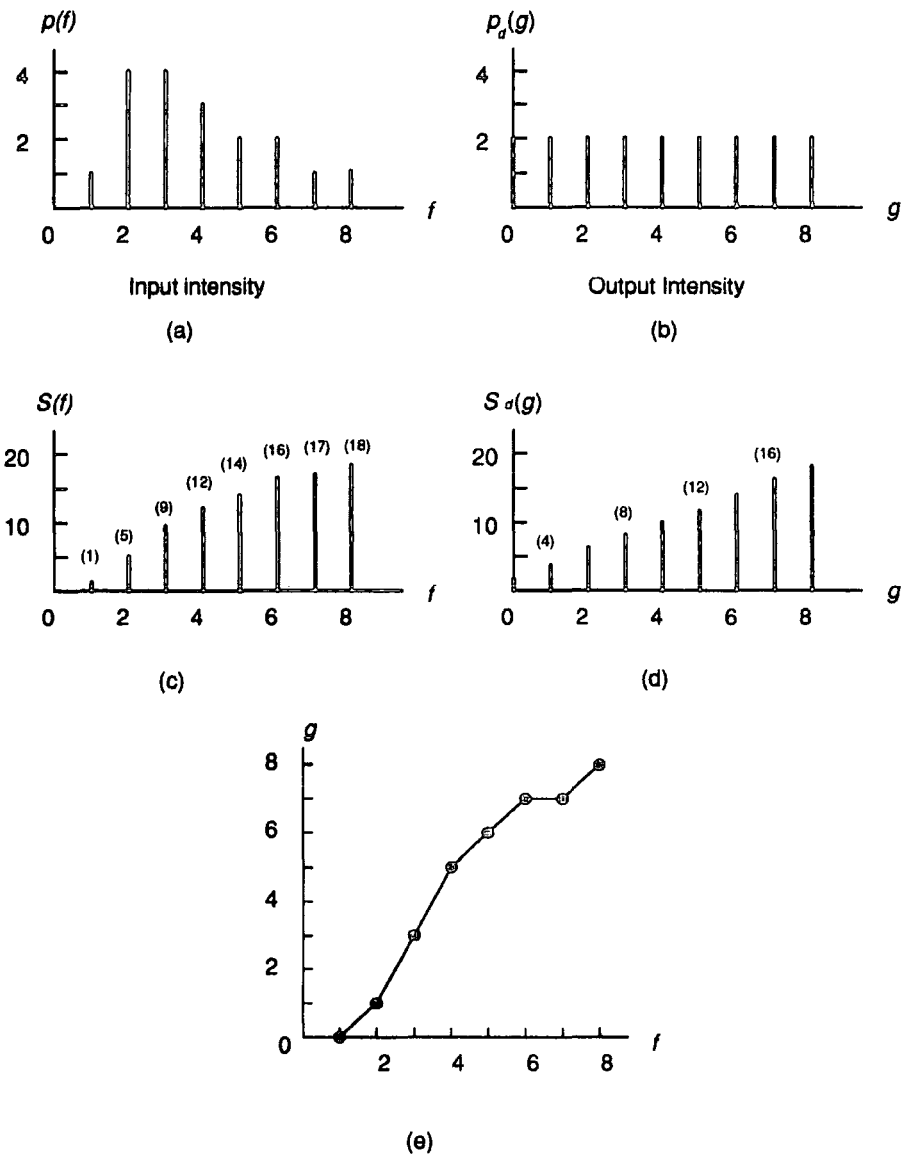


Figure 6.3 Illustration of histogram equalisation of an image: (a) the original histogram and (c) its cumulation; (b) the desired histogram and (d) its cumulation; (e) grey scale transformation function.

where the $p(f)$ and $p_d(g)$ are the original and desired histogram respectively. Then the grey scale transformation $g = T[f]$ is

$$g = S_d^{-1}[S(f)] \tag{6.6}$$

where S_d^{-1} is the inverse function of $S_d(g)$. The above transformation is exact only if the histogram functions are continuous. In practice, histogram functions are usually discrete and the transformation $g = T[f]$ can be obtained by choosing g for each f such that $S_d(g)$ is close to $S(f)$ (see Figure 6.3).

Images processed by histogram equalisation typically have more contrast than unprocessed images. However, such transformation is neither physically based nor mathematically well defined, for example, compared to the logarithmic transformation mentioned above. The processed images may appear somewhat unnatural. It, therefore, may be argued that the logarithmic transformation may be better in processing resistivity images.

6.2.3 Image smoothing

In addition to enhancement of images by modifying their contrast and dynamic range, images can also be enhanced by reducing random noise through image smoothing. Instead of incorporating the spatial smoothing into the reconstruction process as mentioned in Chapter 5, the spatial smoothing methods can be applied after the reconstruction simply as an image processing technique (see Chapter 5).

CHAPTER 7 NUMERICAL EXPERIMENTS

In determining the behaviour of the electrical resistivity reconstruction algorithms stated above, it is important to apply the algorithms to synthetic data corresponding to models with known parameters. In this way, the data noise level and complexity of model structures are controllable and hence it is possible to investigate whether a particular algorithm converges. In this chapter the algorithms proposed in previous chapters will be tested against the synthetic data from several resistivity model structures. The aim is to evaluate the advantages and the weakness of the different algorithms and to investigate how the changes of some algorithm parameters will affect the reconstruction. The tests will focus on the Born inversion and the regularised non-linear least squares algorithm with different choice of regularisation functions and spatial smoothers. Firstly, the problem of how to choose a suitable regularisation parameter is investigated. The influence of different choice of regularisation functions is then demonstrated. The effects of data and model parameter scaling and the influence of initial model choices are also illustrated. As it has been mentioned in Chapter 2, one data set may perform better than others in the image reconstruction and this will also be investigated. Finally, a comparison is made between images reconstructed from the one-step Born inversion and the non-linear iterative algorithms.

A detailed evaluation of the use of genetic algorithms and neural networks in solving the problem of electrical resistivity tomography requires large computing resources and hence a detailed evaluation will not be made.

As it has been mentioned in previous chapters, for this research the subsurface is divided into rectangular (cylinder) cells in both the Born procedure and the method of regularised non-linear least squares. In the follow experiments, the mesh for Born inversion is 52×11 (572 cells) while the forward modeling and inverse meshes for the method of regularised non-linear least squares are

56×13 and 48×9 respectively. All horizontal dimensions of cells are equal except in the first and the last several columns of the mesh where the cells are gradually enlarged to simulate the infinity. The vertical dimensions of the cells are enlarged downwards in an approximately logarithmic scale (e.g. Figure 3.6).

7.1 CONVERGENCE CRITERIA

One of the convergence criteria is the data residual or fitting error. In the following experiments, the data residual or fitting error is measured as

$$E_d = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\frac{u_i^{obs} - u_i^{pre}}{u_i^{obs}}\right)^2} \tag{7.1}$$

where M is the number of data and u_i^{obs} and u_i^{pre} are the observed data and the calculated data, respectively. The E_d is also known as the root-mean-square error (rms).

In the regularised iterative algorithm, the data residual or data fitting error is an important criterion of convergence but it is not the only one since a small data residual does not always imply that the resulting image is close to the true solution. The smoothness of the reconstructed image is another criterion and hence the iteration is convergent if both the data residual and image roughness are *relatively small*. While the criterion of a small data residual may be specified uniformly without considering the image structures, for instance, 5% or 10%, the criterion for small roughness is context dependent and therefore cannot be specified without *a priori* information. However, if the algorithm is known to have the characteristics of strong smoothness constraints and it is unlikely to produce an image with a high level roughness, one may then argue that a small data residual can alone be taken as the only measure of convergence.

The iteration will be terminated if one of following events happen:

(1) data residual E_d is less than a desired level, for example 5% or 10%;

(2) relative change in data residual E_d is less than 1% for three successive iterations;

(3) number of iterations exceeds a maximum number allowed.

It is found that sometimes the final solution at the termination of iteration is not necessarily the 'best' result during the iteration according to *a priori* knowledge and hence in the study all intermediate images during the iteration were recorded for later examination.

7.2 CHOICE OF REGULARISATION PARAMETER

The regularisation parameter is normally referred to as the Lagrange multiplier in Tikhonov (1963) regularised minimisation problems. The discussion here will include the threshold of singular values for Born inversion given in Chapter 4 because it has similar effects as the regularisation parameter in reducing the roughness in the recovered image. For the convenience of comparison, synthetic data from the model shown in Figure 7.1 will be used throughout the following sections where a linear array of 20 electrodes equally spaced at 1m is spread from the left to the right. All data are assumed to be gathered by the circulating dipole-dipole (*c*CCPP) electrode configuration unless otherwise specified.

7.2.1 Choice of threshold of singular values in Born inversion

As it has been mentioned in Chapter 4, the choice of threshold of singular values for Born inversion is a balance of image resolution and distortion. Lower threshold results in higher resolution but it may also lead to higher frequency oscillation in the reconstructed image. Notice that the threshold of singular values λ_t is normally expressed as a fraction of the largest singular values, for example, $\lambda_t = 1.0 \times 10^{-3} \lambda_{max}$, where λ_{max} is the largest singular value of the sensitivity matrix.

The term λ_{max} will be omitted in the following discussion. Because the Born inversion is fast if the weighting function is pre-calculated, it seems best to select the threshold interactively, i.e. a threshold is chosen, say between 1.0×10^{-3} and 1.0×10^{-4}, and a reconstruction is carried out and the resulting image is then examined; this process is repeated until a satisfactory threshold is obtained. As an illustration, Figure 7.2. shows four images reconstructed from the same *c*CCPP data set with thresholds $\lambda_t = 4.5\times10^{-3}$, 4.5×10^{-4}, 4.5×10^{-5} and 4.5×10^{-6} respectively. All images have been smoothed by a 3-by-3 spatial median filter. When $\lambda_t = 4.5\times10^{-3}$, the image is smooth and the two targets are clearly visible but their vertical positions are not well resolved since the apparent targets are displayed upwords by about 0.5m (Figure 7.2(a)). As λ_t decreases, the positions of the targets are better resolved at the cost of increasing oscillation in the image (Figure 7.2(b), (c)). When $\lambda_t = 4.5\times10^{-6}$, the oscillation is so strong that the targets are hardly visible (Figure 7.2 (d)). Therefore $\lambda_t = 4.5\times10^{-4}$ seems to be the best choice in this case. It is also noticed that as the threshold λ_t decreases, the resistivity values of the whole image increase because the effect of the variable offset term b (see equation 4.14). This emphasises the fact that images from a Born inversion can only show the relative variation of resistivities, not their absolute values.

7.2.2 Choice of regularisation parameter

The regularisation parameter λ regulates the trade-off between the data residual and roughness penalty in regularised non-linear least squares methods (see equation (5.27)). Experience shows that a good choice of the regularisation parameter has a crucial importance for the success of the algorithm. In the following discussion, several methods for choosing the parameter λ mentioned in Chapter 5 will be evaluated. Unless it is specifically indicated, the voltage data from a *c*CCPP electrode configuration are assumed and the iterative procedure of the zero-order regularisation method plus median filter is applied as the reconstruction procedure (see equation (5.39)).

1 Marquardt's strategy and its modified version.

As it has been mentioned in Chapter 5, the success of the Marquardt's method depends on the initial choice of regularisation parameter λ as well as its reduction and expanding factors α and β. Two images reconstructed by this method are shown in Figure 7.3 (a1) and (a2). The initial value for λ is selected as 10^5 and the reduction and expanding factors are fixed as α=0.2 and β=5 respectively. Figure 7.3 (a1) is the image after 3 iterations and has a data residual of 9.8% and Figure 7.3 (a2) is the image after 11 iterations with a data residual of 5.6%. The targets are visible in the images but the distortion is high, creating some undesired artefacts. The reason is that the conventional Marquardt's strategy is designed to chose λ as small as possible. This is equivalent to setting the roughness penalty as small as possible and it therefore produces images with strong oscillation structures. Figure 7.4 shows the changes of λ during the iteration and it indicates another disadvantage of such strategy: when λ is reduced to its lower boundary, the choice of λ will wander about the boundary and about half of iterations will fail.

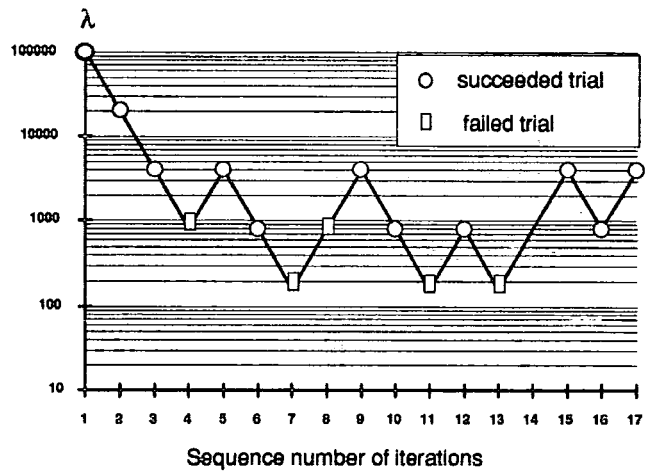


Figure 7.4 The variation of regularisation parameter λ during iterations in Marquardt's method. About half of iterations failed at later stage of the iterative process because the λ wandered about its lower boundary.

Various modified versions of the above strategy are possible. For example, if a higher roughness penalty is desirable, as in the case of 2D resistivity image reconstruction, one can start with a high initial λ value and allow it to decrease only after several successive reductions have occurred in data residual but increase it whenever the residual stops decreasing. The images reconstructed by such a strategy are shown in Figure 7.3 (b1) and (b2) where the initial λ, α and β are the same as the above example but the λ is reduced only after 4 successive decreases in data residual. It is evident that the images (b1) and (b2) are much better than image (a1) and (a2) in terms of

distortion although the fitting errors from the two strategies are nearly identical: images (a1) and (b1) have a fitting error of 9.8% and 10% respectively and images (a2) and (b2) have a fitting error of 5.6% and 5.7% respectively. It is noticed from above example that the fitting error as the only measure of fitting goodness is inadequate here because of under-determination and ill-posedness of the problem.

In general, Marquardt-like methods of choosing λ tends to put smoothing constraints as small as possible and therefore the reconstructed image may have large fluctuations.

2 *Choosing λ as a constant value*

If *a priori* knowledge of model smoothness is available, the regularisation parameter λ can be fixed as a constant value for all iterations (e.g. Sasaki, 1992). Such *a priori* knowledge may be gained by testing the algorithm with several different λ values ranging from small to very large before actually carrying out the reconstruction. Experience shows that a good result can be obtained if a suitable λ value is selected. The computation cost is relatively low because no additional function trials are needed for selecting λ. Figure 7.5 shows such an example in which λ is fixed as 10^5. A total of 20 iterations were carried out and all of them are successful. However, numerical experiments indicate that a good choice of λ is context dependent and there is no uniformly good λ for all problems. Although the algorithm is fast if a good choice of λ has been made, in some cases the overall cost of computing may be not much lower than the method mentioned above (or below) if the computing cost of selecting λ is taken into account.

3 *Choosing λ by 1D search methods*

An 'optimal' λ may be selected at each iteration if it is allowed to sweep from near zero to infinity. The automatic implementation of such a procedure requires adding a 1D optimisation search at each iteration. As it has been mentioned in Chapter 5, the criterion for a good λ may be the minimizer of either data residual or GCV function (see equations (5.40) and (5.41)). The 1D

search method used here is the golden section search described by Press et al. (1989, p277). As a comparison, Figure 7.6 shows two images (a) and (b) reconstructed by selecting λ through minimising the data residual and the GCV function respectively. It is shown that the image (b) has a relatively poor quality in terms of contrast and geometric resolution, for example, although the

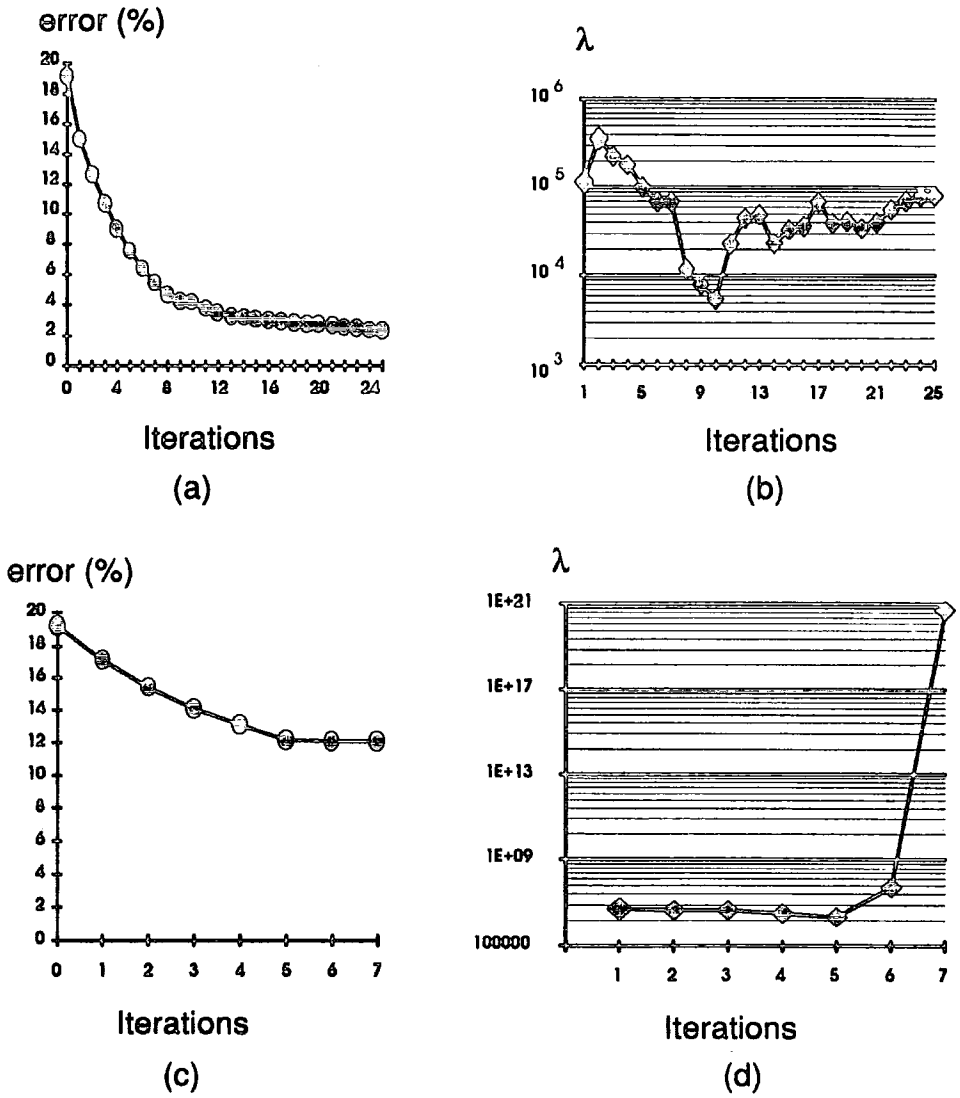


Figure 7.7 The changes of data residual and regularisation parameter λ during iterations: (a) is the data residual and (b) is its corresponding λ which is obtained by minimising the data residual in each iteration; (c) and (d) have the same relationship as (a) and (b) but the λ has been selected by minimising the GCV function.

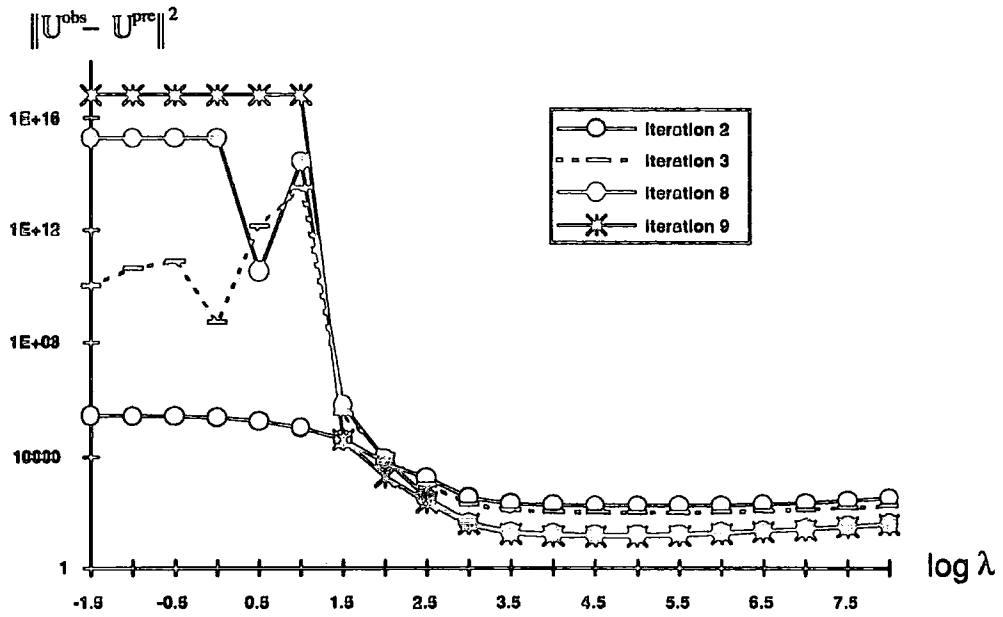
horizontal geometric positions of the targets are well recovered, the vertical positions of the targets seems to move upward by one cell. The reason is that GCV method tends to choose a very large λ after just several iterations even when the data residual is still high (see Figure 7.7). Consequently the iterative process is terminated because the stepsize of model parameter correction, which is roughly proportion to $1/\lambda$, is too small. Other experiments with different regularisation functions also indicate that the GCV method of choosing λ tends to produce an over-smoothed image. In this test, the method of choosing λ as the minimizer of data residual at each iteration performs better than the GCV method.

One may suspect that the λ chosen by minimising the data residual is not necessarily appropriate at all because it only emphasises the fit to the data rather than minimising both the data residual and the roughness of the image, as required by the underlying principle of regularisation methods. However, experience indicates that such approach does produce sufficiently smooth images in most cases although it fails occasionally. Several numerical experiments were carried out to investigate how the data residual and the image roughness vary as a function of λ during iterations. Figure 7.8 shows a typical example of the data residual and the image roughness as a function of regularisation parameter λ. From Figure 7.8 following observations can be made:

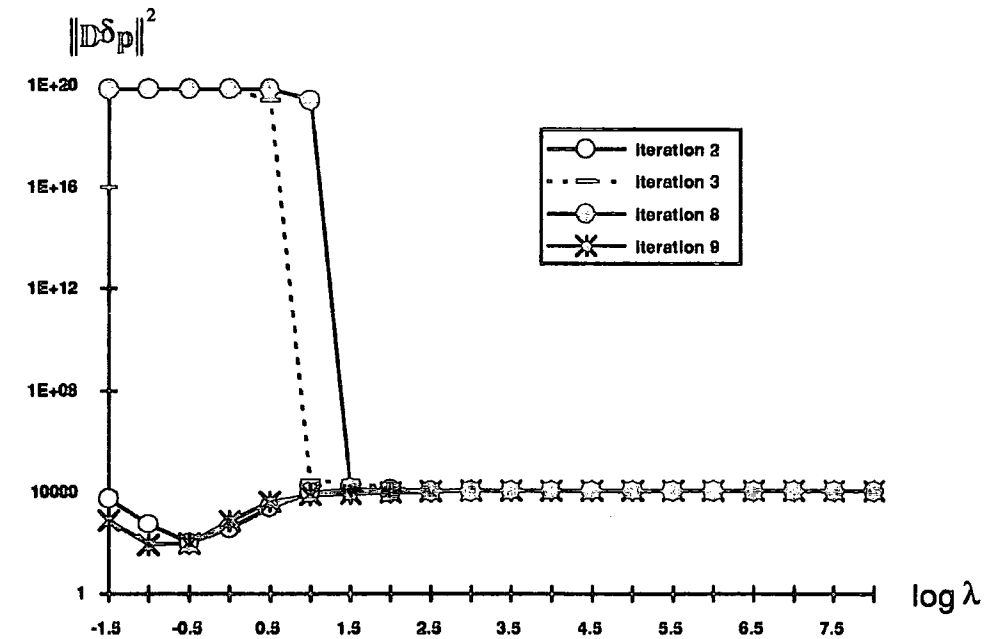
(1) A small λ leads to a larger data residual and sometimes also a larger image roughness than a large λ.

(2) When the λ is too large, the data residual will increase slowly.

(3) The choice of λ as a minimizer of data residual normally produces an image with relatively small roughness.



(a)



(b)

Figure 7.8 The data residual and the image roughness as a function regularisation parameter λ during iterations: (a) data residual; (b) image roughness where D is an identity matrix (zero-order smoothing constraints)

The slow change of data residual and image roughness at large λ also implies that the regularisation parameter may be chosen as a constant value as stated above. In practice, the λ can be chosen as a constant value between 10^5 to 10^6 if the resolution is not the major concern. Choosing λ by a 1D search method may produce a high resolution image but the iteration may not be so stable.

7.3 EFFECTS OF REGULARISATION FUNCTION

There are several regularisation functions which can be chosen and therefore it is necessary to investigate their effects on the reconstructed image. As it has been stated in Chapter 5, there are basically two choices of image functions to which the smoothness constraints can be imposed during successive iteration, namely, the image vector $\mathbf{p}$ and the iterative correction stepsize $\delta\mathbf{p}$, while the choice of the smoothing operator $\mathbf{D}$ can be a zero-, first- or second-order of spatial difference. In addition, a median or spatial average filtering operation can be applied to the image vector during each iteration. The following experiments will examine the effects of different smoothness constraints on the reconstructed image.

7.3.1 Smoothing image $\mathbf{p}$ or stepsize $\delta\mathbf{p}$

A choice to impose smoothness constraints on the image vector $\mathbf{p}$ or the iterative stepsize $\delta\mathbf{p}$ will lead to two different iterative procedures (5.30) and (5.32) respectively. Numerical experiments show that the algorithms which require a smoothing $\mathbf{p}$ appear to converge slightly faster but that higher quality images seem to result from algorithms which impose smoothness on $\delta\mathbf{p}$, as indicated in Figure 7.9. In addition, if $\mathbf{D}$ is a zero-order differential operator, experiments show that the iterative procedure which imposes smoothness on $\mathbf{p}$ normally does not converge while the algorithm which constrains the roughness of $\delta\mathbf{p}$ does. The algorithm which puts penalty on the roughness of $\delta\mathbf{p}$ appears to be more stable.

7.3.2 Effects of different smoothing operators

The choice of different smoothing operators in iterative procedures (5.38) and (5.39) may have different effects on the recovered image. The following experiments address this problem. For the reason stated above, only the iterative procedure (5.39) which imposes smoothness constraints on $\delta\mathbf{p}$ will be considered here. Different choice of spatial smoother $\mathbb{S}$ and difference operator $\mathbb{D}$ in the procedure (5.39) would result in more than 6 different iterative algorithms but only five of them will be tested, namely,

(1) the zero-order regularisation or damped least squares method (DLS);

(2) DLS plus median image filter;

(3) DLS plus spatial average image filter;

(4) the first-order regularisation method;

(5) the second-order regularisation method.

Figure 7.10 shows images recovered by the above five algorithms from noise-free data respectively. The images recovered from noisy data (5% Gaussian noise) are shown in Figure 7.11. The maximum number of iterations allowed was 20 and the initial model was a homogenous half space having a resistivity of 100 Ωm. The test results can be summarised as:

1 Zero-order regularisation or damped least squares method (DLS)

The zero-order regularisation or DLS method alone seems to have the highest resolution in terms of image value and geometrical information but it is less stable. It tends to overfit the data by generating some isolated spikes in the image, especially when noisy data are present. Figure 7.11 (a) shows the result after 11 iterations. In fact, all 20 iterations were successful. However, whilst the data residual still shows a decrease during further iteration, pixel values for the high resistivity target approach infinity. Therefore the image after 11 iteration is not realistic because it is known

that the highest resistivity in the model used to produce the voltage data is 500 ohm-m (see Figure 7.1).

2 DLS plus median filter (DLSM)

The DLS plus median filter method appears to produce the best images from both noise-free and noisy data, compared with the original model, as shown in Figures 7.10 (b) and 7.11(b). The isolated spikes from DLS are removed effectively while the boundaries of the targets are preserved. The images are relatively smooth and the positions of the targets are well resolved in comparison with the other methods. The pixel values of the targets are under estimated but this is reasonable because such an algorithm is unlikely to be capable of recovering all information due to the smoothness constraints and the underdetermined and ill-posed nature of the problem.

3 DLS plus spatial average filter (DLSA)

The method of DLS plus the spatial average filter produces the smoothest images and therefore it is the most stable algorithm. In fact the two images in Figures 7.10(c) and 7.11(c) are nearly identical, despite the fact that one was recovered from the data with 5% Gaussian noise while the other was noise-free. The price paid for the higher stability is a lower image resolution, i.e. the images appear to be over-smoothed. However, as will be seen later, the algorithm is remarkably stable when different initial models are applied.

4 First-order regularisation

The first-order regularisation performs well with noise-free data as it is shown in Figure 7.10(d). The image of the low resistivity target appears to be over-smoothed but the high resistivity target is clearly indicated. However, it is unstable when 5% Gaussian noise is added to the data and it performs just a little better than DLS method as it is indicated in Figure 7.11(d). The problem is similar to that of the DLS algorithm where the data residual shows a continuous decrease during 20 function trials but the pixel values (resistivities) for the high resistivity target rise to

unrealistically high levels after 13 iterations. In addition a distinct 'salt and pepper' like noise is present in the image.

5 Second-order regularisation

The second-order regularisation method yields moderate resolution and stability (see Figures 7.10(e) and 7.11(e)). There are some artefacts in both images recovered from noise-free and noisy data but they are not very significant. For noisy data, the image reconstructed is better than those recovered by DLS, DLS-plus-spatial-average and first-order-regularisation methods

The test results indicate that the method of DLS plus median filter performs best while the DLS alone and first-order regularisation methods produce relatively poor images if the data are noisy. However, the choice of smoothing operators may be best understood to be context dependent. For example, if the data noise level is known to be low, then the DLS method may be chosen to take advantage of high achievable resolution. The method of DLS plus spatial average filter may work well if the resistivity variation of the subsurface is low whereas when targets have clear boundaries the method of DLS plus median filter may be optimal.

The above experiments clearly demonstrate the necessity for a trade-off between stability and resolution in electrical resistivity reconstruction. They also illustrate that although the data residual (or misfit) is an important measure of an algorithm's performance, it is not the only measure, for instance, in the DLS method the data residual can be very small but the corresponding image may have unrealistic fluctuations. It is also noticed that an overfit of the data and an over-smoothness of the image may result from excessive number of iterations and therefore it is recommended that all intermediate images be recorded for later evaluation.

7.4 EFFECTS OF SCALING

The data residual can be scaled or weighted if some data are considered to be more reliable than others and hence worthy of greater influence on the solution. An obvious measure of reliability is the standard deviation of the data but it is rarely available in practice. Alternatively, the data residual may be scaled by the measurements themselves as has been stated in equation (5.19b). The argument behind such scaling is that a given percent difference between the measured and calculated data for each data point should have the same influence on the data residual regardless of its signal level (Hohmann & Raiche, 1988). For similar reasons, the model parameters can also be scaled or weighted. A widely applied scaling method of model parameters in electrical resistivity inversion is the logarithm transformation (see equation (5.19b)). Several experiments have been carried out to test the effects of such scaling and Figure 7.12 shows one of the resulting images.

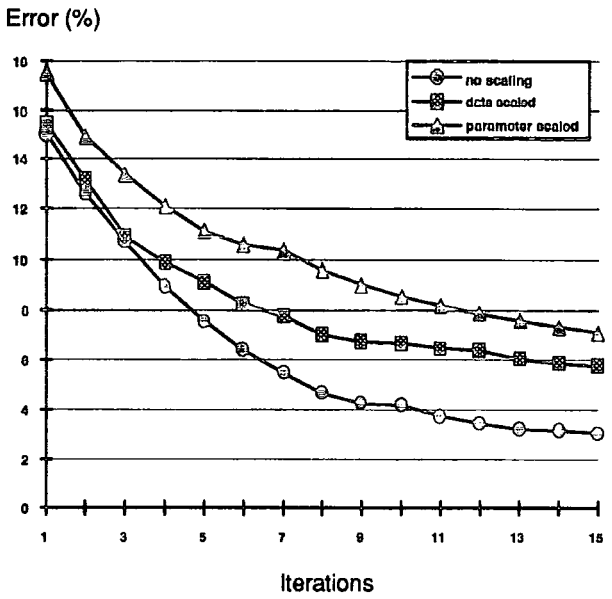


Figure 7.13 A comparison of data residual of the algorithm with and without scaling during the first 15 iterations. The algorithm used is the method of damped least squares plus a 2D median filter.

The experiments indicate that although the algorithms with such scaling methods do converge, they tend to produce some undesirable artefacts in the images (see Figure 7.12) and that they converge more slowly than the algorithms without scaling as shown in Figure 7.13.

7.5 EFFECTS OF INITIAL MODEL AND NON-UNIQUENESS

It is known that the final solution in non-linear iterative algorithms may depend on the initial model from which the iteration starts and thus the solution may not be unique (e.g. Menke, 1989, p153-156). The underlying reason is that the data residual function may have more than one minimum point and therefore the algorithm may be trapped in any one of them. Such an example is shown in Figures 7.14 and 7.15 where the initial model R_0 is assumed to be a homogeneous half space with resistivity 20, 50, 100, 200 and 500 Ωm respectively. Figure 7.14 shows that all iterations may be regarded as having converged in terms of the data residual except when

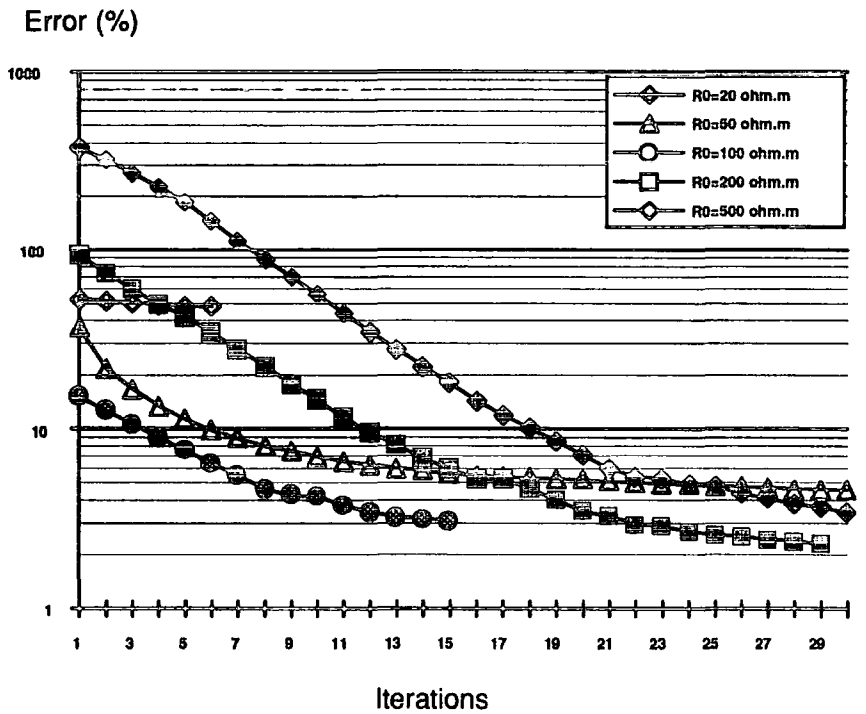


Figure 7.14 The variation of data residual during iterations when the iterative process started from different initial models. The initial model is a homogenous half space with a resistivity R0.

$R_0 = 20\Omega m$ in which the iteration was trapped in a local minimum with about 50% data residual. The initial model with $R_0 = 100\Omega m$ can be taken as the closest to the true solution among the five

models since the initial data residual is small and the algorithm converges rapidly. The initial models with $R_0 = 200\Omega m$ and $500\Omega m$ are 'far' away from the true solution in terms of initial data residuals but the algorithm did finally converge, indicating that the algorithm has a certain degree of robustness towards different initial models. In the case of $R_0 = 50\Omega m$, the algorithm behaved in a more complex manner: the data residual decreased very quickly within the first several iterations and then decayed very slowly. An examination of the resulting images in Figure 7.15 indicates that the algorithm in this case actually converged to the 'wrong' image, i.e. it converged to a local minimum with small data residual (about 4%). In summary, if the data residual is the only convergence criterion, the non-linear iterative algorithm may follow three patterns when initial model is 'far' away from the true solution:

(a) It does converge towards the true solution but requires more iterations to reach a desirable data residual, as in the case of $R_0 = 200\Omega m$ or $500\Omega m$.

(b) It does not converge at all or converges to a local minimum with a high data residual (e.g. $R_0 = 20\Omega m$). This problem normally can be easily identified and a new initial model can be tried.

(c) It converges to a local minimum with a small data residual. This problem is difficult to identify, especially when field data with noise are applied. A small data residual would lead one to believe that the true solution has been obtained but in reality the result may be totally irrelevant because a small difference in data residual may lead to a big difference in the solution due to the ill-posedness and under determination of the inversion.

The above example clearly demonstrates the importance of the initial model and the associated problem of non-uniqueness for the non-linear iterative algorithm. If *a priori* information is available, the initial model may be appropriately designed. In the absence of such knowledge, the apparent resistivity pseudosection or images from Born inversion may be used to estimate the initial model. In practical applications, it is recommended that several initial models should be

tried. However, one must bear in mind that there are no simple methods for determining whether the iterative algorithm does converge globally or not without *a priori* information.

It was found that in some cases the iterative algorithm failed to produce a desired image because it generated some highly oscillatory image structures at an early stage and could not remove them in successive iterations, i.e. the reconstruction was under-smoothed. In such cases, a stronger smoothing constraint may improve the results. For instance, in the above example where the initial model is 50Ωm, the true image structures are shadowed by a strong artefact at the bottom of the image (see Figure 7.15(a)) This problem can be solved by imposing stronger smoothness constraints, for example, using the method of DLS plus spatial averaging filter. Figure 7.16 indicates that such an approach does remove the artefact although some of the image resolution is lost.

The influence of initial models on the images reconstructed from Born inversion seems not so significant. As shown in Figure 7.17, the images from initial models of 20, 50, 100, 200Ωm are nearly identical. The image recovered with the initial model of 500Ωm is also close to the others if one excludes the pixel rows near the surface which show a regular fluctuation, probably due to the inaccuracy in the approximation of the weighting function in this region.

7.6 COMPARISON OF IMAGES RECONSTRUCTED FROM DIFFERENT DATA SETS

As stated in Chapter 2, there are various data sets which may be used as the inputs to the ERT inversion and hence a question may arise as to their relative merits. To answer this question, an experiment was carried out to invert various data sets with and without adding noise to the data. Specifically, the following data sets were evaluated:

(1) five complete data sets, namely, pole-pole (ℂℙ), circulating pole-dipole (*c*ℂℙℙ), circulating dipole-dipole (*c*ℂℂℙℙ), circulating PCPC (*c*ℙℂℙℂ) and circulating CPPC (*c*ℂℙℙℂ) data sets;

(2) two conventional data sets, namely, the Wenner data set (i.e. data in Wenner pseudosection) and the dipole-dipole data set (ℂℂℙℙ).

In the first case, the data were generated from the model shown in Figure 7.1. The experiment indicates that all images from complete data sets are nearly identical if the data are noise free (see Figure 7.18). However, if a 5% Gaussian noise is added into each data, the images show considerable variations as indicated in Figure 7.19: the image produced from *c*ℂℂℙℙ data set is most close to the true image while the image produced from *c*ℙℂℙℂ seems to be the second best, with the objects apparently moving upward nearly two cells. All other images have strong distortions which overshadow the true structures. Therefore the *c*ℂℂℙℙ data set performs best in this test.

The images reconstructed from Wenner and dipole-dipole data sets are shown in Figure 7.20 which indicates that with noise free data, both data sets produce images resembling to the true structure but that the image from the Wenner data set shows a lower resolution in terms of geometric position and resistivity contrast. When a 5% Gaussian noise is added, the ℂℂℙℙ data set still produced a reasonable image while the Wenner data set failed because of extreme image distortion. One of the reasons may be that the number of data in the Wenner data set is too small and hence the inversion is severely under-determined: in this case the number of data in the dipole-dipole data set is only 10% less than that of the complete data set of a four electrode configuration whilst the number of data in the Wenner data set is 66% less for a linear array of 20 electrodes.

However, the number of data is not the only factor which influences the final result, although it may be an important one. As it has already been demonstrated in Figure 7.19, a larger data set

does not always produce a better image, for instance, the ℂℙ and *c*ℂℙℙ data sets have about 10% more data than the *c*ℂℂℙℙ data set while they still show a poor performance when noise is present in the data. One of the underlying reasons may be that the data sets from different electrode configurations are not equally sensitive for the same resistivity structures. In the above example, it is found that the ℂℙ data set shows about 4% difference on average when compared with that of 100Ωm homogenous half space whilst the *c*ℂℂℙℙ data set presents about 12%. Therefore, in this case a 5% Gaussian noise may overshadow most of information contained in a ℂℙ data set but not that in a *c*ℂℂℙℙ data set.

Clearly, the sensitivity distribution of a data set is associated with the electrode configuration and varies with the geometric position of the image cells (see equation (4.8) in Chapter 4). A large current electrode spacing may enable more current to be injected to greater depths and therefore is more sensitive to deep targets when compared with smaller electrode spacings. Thus one may argue that the choice of data sets is also context dependent. Figure 7.21 shows four images of a vertical boundary reconstructed from four data sets with 5% noise added, namely pole-pole, ***c*CCPP**, dipole-dipole and Wenner data sets, respectively. The vertical boundary is 10m from the left and the resistivity is 100Ωm left of the boundary and 500Ωm to the right. It is shown that although the *c*ℂℂℙℙ data set still gives the best overall performance, the lower part of boundary seems to be better resolved in the image reconstructed from a pole-pole data set if the image oscillations due to the noise are excluded. The dipole-dipole data set also performs well except that the reconstruction at the two lower corners of the image is inadequate. This seems to suggest that the additional data in a ***c*CCPP** data set contains some information which is lost in the dipole-dipole data set. The Wenner data set performs reasonably well, considering the relatively small number of data used, although the area at the lower right hand corner of the image is clearly under-estimated.

Various numerical experiments indicate that the *c*CCPP data set usually produces the best image if the targets are within a depth of 2 or 3 inter-electrode spacings for a linear array of 20 electrodes. Other data sets may contain more information about deeper targets but they appear to be not so suitable for image reconstruction if noise is present. One of the problems in using the *c*CCPP data set is that some of the data may be too small to be measured accurately because of the rapid decay of voltage with the increase of electrode spacing between current and potential electrodes. For a homogeneous half space having a resistivity of ρ, the smallest voltage V_{min} in the *c*CCPP data set for a linear array of 20 electrodes is

$$V_{min} = \frac{I\rho}{\pi \times a \times 17 \times 18 \times 19} = 5.48 \times 10^{-5} \frac{I\rho}{a} \tag{7.2}$$

where I=current strength, a= inter-electrode spacing. Therefore, a liner array of 20 electrodes is about the maximum size (in term of number of electrodes) applicable if I=20mA, ρ=100Ωm and a=1~5m, giving V_{min}=0.1~0.02mV, since the resistivity meters normally only have a precision of 0.01~0.05mV.

7.7 COMPARISON OF THE IMAGE RECONSTRUCTED BY BORN INVERSION AND NON-LINEAR REGULARISATION METHODS

7.7.1 Effects of noise

It is anticipated that the data noise will have a strong influence on the reconstructed image because of the ill-posedness of the problem. Imposing stronger smoothness constraints on the image would reduce such effects but then there will be a trade-off in the resolution. In general, the one-step Born inversion (or 'backprojection') is more susceptible to data noise than regularised iterative algorithms. For instance, if the *c*CCPP data set from the model shown in Figure 7.1 is applied, a 10% Gaussian noise in the data would destroy the image reconstructed by Born inversion while the

targets are still visible on the image recovered by the regularised iterative algorithm (see Figure 7.22). However, if the image structure is simple, the Born inversion may also produce a result comparable to that of regularised iterative algorithms even when the data are noisy. Figure 7.23 shows such an example where the resistivity model can be regarded as an oil pollution problem. The Born inversion produces a clean image with 1% data noise (Figure 7.23 (b)). When 10% Gaussian noise is added into the data, the image recovered by the Born inversion is blurred but it is nearly as clear as the image reconstructed by the regularised iterative algorithm (see Figure 7.23 (c) and (d)).

It is also noticed that in the Born inversion the pixel values of the image will increase as a function of the data noise because they are summations of weighted data deviations between the true and the reference models and hence the data noise in such cases is directly coupled into the pixel values (see equation 4.13). Such effects are not significant in the regularised iterative algorithms for there is no direct coupling between data noise and pixel values.

7.7.2 Resolution

The resolution of the electrical resistivity image can be described in two separate contexts: (1) the spatial resolution which indicates the smallest region (pixel) in which the resistivity can be determined independently, or more loosely, how well the positions of the targets are resolved; (2) the resistivity value resolution which can be measured as the percentage error of resistivity in each cell relative to its true resistivity value. Because of the complex non-linear relationship between the data and the reconstructed image, it is difficult to describe the resolution in terms of closed analytic functions. It may only be inferred from numerical experiments.

Figures 7.24 presents the images reconstructed from the *c*CCPP data set of a fault model by the Born inversion and the regularised nonlinear least squares algorithm. A model of two dykes intruded into the lower of two horizontal layers is shown in Figure 7.25 (a) and its reconstructed images

are shown in Figure 7.25 (b) and (c). All data are contaminated with 5% Gaussian noise. The results appear to indicate:

(1) Regularised nonlinear algorithm has better overall resolution than the Born inversion.

In the fault model, the Born inversion can only indicate the position of the fault, not the relative variation of resistivity values from the left to the right hand side and in the near surface layer of the model, while the regularised nonlinear method does resolve such information although it is not fully recovered (Figure 7.24). In the dyke model (Figure 7.25), the Born inversion clearly resolves the spatial information of two layer structure but fails to show the two dykes embedded within the lower layer. The regularised nonlinear algorithm resolves both the layer structure and the dykes with a certain degree of smoothness.

(2) Born inversion normally can only provide the spatial information of the targets or relative variations of resistivity distribution, not the absolute resistivity values.

As is apparent in Figure 7.24 (b), although there are indications of low resistivity in the image area of the fault, the resistivity values in these areas are far from the true resistivities of the fault. As already mentioned above, the overall pixel values of an image will increase as a function of data noise level and hence the images reconstructed by such methods can at best be regarded as semi-quantitative.

(3) Resolution decreases with increase of smoothness constraints.

The resolution in regularised nonlinear algorithm usually has to be trade-off of smoothness to achieve a reasonable stability. It may be argued that the images reconstructed by the method of least squares plus spatial average smoother in Figures 7.24 and 7.25 are over-smoothed. In fact, a higher resolution can be achieved if the method of regularised non-linear least squares plus median smoother is applied. The problem is that the resistivities at the high resistivity area of the image

will rise to unrealistic high values at the later stage of the iteration while the data residual still decreases.

(4) Resolution decreases with the increasing depth.

This is clearly indicated in Figure 7.24 (c) where lower part of the fault appear much wider than its upper part.

7.7.3 Effects of inaccuracy of forward modelling

The ill-posedness of ERT implies that the FDM forward modelling in the iterative algorithm has to be highly accurate because a small difference in the data may imply a large difference in the model parameters. Barber (1989) suggests that voltage data have to be accurate to 1% or 2%. Experience indicates that such an accuracy is difficult to achieve without high cost of computation. In the experiments, the FDM forward modelling is implemented on a mesh of 56×13 and only has about 3% accuracy when tested against analytic resistivity models of a three layered earth with contrasts ranging from 0 to ± 0.82. This proved to be a problem when the signal from resistivity structures is small, for example, when their dimensions or contrasts are small compared with the inter-electrode spacing or their burial depth is large. As an illustration, Figure 7.26 shows two images reconstructed from a ***c*CCPP** data set in which the data only have about 1% difference on average to that of homogenous half space of 100Ωm. The regularised iterative algorithm failed to resolve the target because the forward modelling error (about 3%) is large in comparison to the signal (Figure 7.26 (a)). As a comparison, the image reconstructed by the Born inversion is shown in Figure 7.26 (b) which clearly shows the target. The resistivity model is a 2m×1m prism of 200 Ωm embedded in 100Ωm medium. The prism has a burial depth of 1m and 160m long in the strike direction. The linear array bisects the prism at a right angle to its strike. The ***c*CCPP** data set is generated by the finite boundary method with a symmetric error of 0.15% (i.e. the error results from the inter change of the positions of current and potential electrodes).

However, if the signal from the targets is strong and only a smooth image is required, the accuracy of forward modelling is not so significant. Figure 7.27 shows two images reconstructed from the synthetic data (*c*CCPP data set) of a three layer model. For the image 7.27 (b), the data can be regarded as highly accurate for the inversion because they are computed by the same forward code used in the iterative algorithm, although they have about 3% error on average when compared to the data in the image 7.27 (c) which are computed by the method proposed by Mooney & Orellana (1966) and are accurate to 0.1%. It is shown that there is no significant difference between images (b) and (c) in Figure 7.27: the main feature of the three layer structures is shown in both images although the boundaries and the resistivity values of layers are not well resolved, partly due to the smoothness constraints.

7.7.4 Imaging 3D structures

Subsurface geologic and archaeological structures are commonly 3D. In some cases a 3D structure may well be approximated as a 2D structure while in other cases such an approximation may not be justified It is therefore interesting to know how the 2D algorithm developed so far will react when the data from 3D structures are to be reconstructed.

The data from a 3D prism of 2m×1m with three different lengths in Y (strike) direction, namely 40m, 20m and 2m, were used to evaluate the abilities of the algorithm to resolve the resistivity distribution which is not strictly two-dimensional. In each case the prism (200Ωm) is embedded in a 100Ωm half space and is buried at 1m depth beneath the middle of the section. The linear array bisects the cylinder resistivity with a right angle to its strike and the boundary data are collected by the ***c*CCPP** electrode configuration and computed by the finite boundary method with a symmetric error less than 1%. The reconstructed images are shown in Figure 7.28 in which images (a), (b) and (c) are recovered by Born inversion and image (d) by the method of regularised non-linear LSQ plus spatial average smoother with an initial model of 200Ωm homogeneous half space. The results indicate that for the Born algorithm the image contrast decreases and artefacts increase,

both slightly, with the reduction of the length of the cylinder and hence it may be applied to image 2D as well as 3D structures. The iterative algorithm in such cases has the tendency to produce smooth and blurred images with low contrast. For instance, when the length of cylinder L=2m, the reconstructed image nearly reaches the homogeneous background with resistivities ranging only from 94 to 108 (see Figure 7.28 (d)). It seems that the regularised iterative algorithm is not so sensitive to 3D structures in comparison to the one-step Born inversion.

7.7.5 Speed of inversion

If the weighting coefficients are pre-computed, the Born inversion can be carried out on PC computers with each inversion taking about two minutes. The non-linear iteration algorithms currently can only be carried out on the SUN or HP workstations. The Jacobian matrix is both computation and memory demanding and hence the algorithms cannot yet be implemented on PC computers. In the iterative algorithm, the mesh size (cells) of finite difference forward modelling is 56×13 and the number of unknowns in the inversion is 432 (i.e.48×9). Tests on the HP-9000/710 workstation shows that the CPU time for each iteration is about 553s in which more than half of the time (387s, i.e. 70% of total) is spent on computation of the Jacobian matrix and only about 2s (0.4% of total) on forward calculation of voltage data. It is clear that the major computing cost goes to the Jacobian matrix.

7.8 GENETIC ALGORITHM

The genetic algorithm (GA) is not efficient in terms of computation in comparison to the algorithms base on gradient information such as the LSQ method. For example, for a three layer structure, the damped least squares method may invert the data with 5 to 20 function trials whilst the GA may require 100 to 1000 function trials to reach a similar result. Essentially, the GA is a Monte Carlo like procedure and therefore the number of function trials will increase rapidly with the increasing number of unknowns. For this reason, in the test the 432 cells mentioned above are

grouped into 45 large cells (9×5) to reduce the number of unknowns while the mesh for forward modelling is still 56×13. The GA parameters used in the test are

population size n = 100

crossover probability $p_c = 0.6$

mutation probability $p_m = 0.001$

length of binary code string for each unknown=10

Figure 7.29 shows the resistivity model and its image reconstructed by GA after 500 generations. The whole process takes about 60 hours on the HP9000/730 workstation. The target is visible on the image but there are strong artefacts beneath and around it. Further iteration may improve the result but the computing cost is already too high. In general, although the GA may be robust and may eventually produce a better result than the regularised non-linear least squares method, the computing cost is too high to be practicable.

7.9 NEURAL NETWORK

It is difficult to construct a neural network for imaging a general 2D resistivity distribution in a half space because of the large scale of the required network and the difficulties in training. For this reason, the primary test of a neural network was limited to three layered earth models. The neural network tested is shown in Figure 7.30. It has four layers and a total of 72 neurones, of which 17 input units receive the

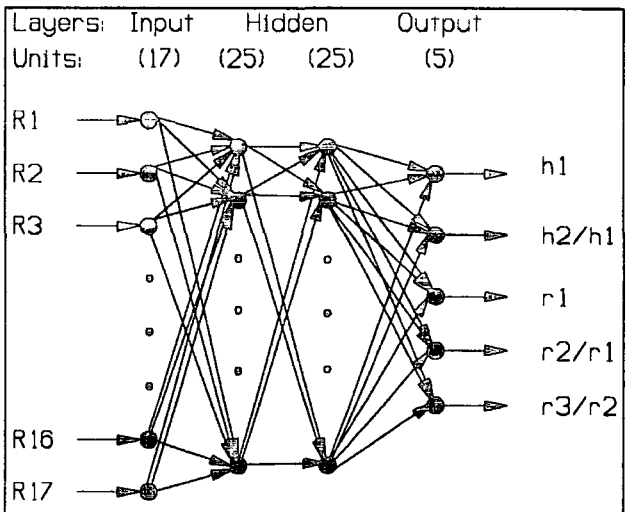


Figure 7.30 Neural network used in processing Wenner and Offset Wenner sounding data over three layered earth models. R1, R2, ..., R17 = apparent resistivities; h1 and h2= thickness of first and second layer respectively; r1, r2 and r3 = the resistivities of the first, second and third layer respectively.

resistivity sounding data and 5 output units give the predicted model parameters. The rest of the units are equally distributed within two hidden layers. The sounding data consist of 9-point Wenner and 8-point Offset Wenner soundings corresponding to inter-electrode spacing of a = 1, 1.5, 2, 3, 4, 6, 8, 12, 16, 24, 32, 48, 64, 96, 128, 192, 256 (i.e. the spacing is expanded at a ln2 interval on a logarithmic scale). The 5 outputs are :

h_1 = thickness of the first layer;

h_2 / h_1 = thickness ratio of the second layer to the first layer;

r_1 = resistivity of the first layer;

r_2 / r_1 = resistivity ratio of the second layer to the first layer;

r_3 / r_2 = resistivity ratio of the third layer to the second layer.

Table 7.1 Resistivities of three layer models used to generate sounding data for the training of the neural network. The thickness of the first layer h_1=10m and that of the second layer h_2=100m.

r_1 (Ωm)	r_2 / r_1	r_3 / r_2		
		a	b	c
10	10	5		1/3
10	5	5	1.5	1/3
10	3	5	1.5	1/3
10	2	5	1.5	
10	1.5	5	1.5	1/3
10	1/3	5	1.5	1/3
5	10	5	1.5	

Note: r_1, r_2 and r_3 are the resistivities of the first, second and third layers respectively.

An interactive back propagation program implemented by McClelland and Rumelhart (1988) was used in the experiment. A total of 18 resistivity models listed Table 7.1 were presented to the network in sequence or selected at random, during the training. The learning converged after 2500

iterations with the total sum of square errors of 0.145. An attempt to train the network over 500 different models failed to converge, possibly due to the small size of the network.

Several tests were carried out to see whether the network can predict the source models when it was presented with the sounding data from resistivity models for which it was not taught. The first two models tested are close to what have be taught where it was found that the network predicted the models fairly accurately (see Table 7.2 (a) and (b)). In the second two test models ((c) and (d)

Table 7.2 The three layer resistivity models and their predictions when the sounding data are presented at the input units of the neural network.

(a)

	Model	Prediction	Error(%)
h_1	10.0	10.0	0.0
h_2 / h_1	10.0	10.0	0.0
r_1	10.0	9.99	1.1
r_2 / r_1	10.0	9.96	0.45
r_3 / r_2	0.667	0.682	-2.3

(b)

	Model	Prediction	Error(%)
h_1	10.0	10.0	0.0
h_2 / h_1	10.0	10.0	0.0
r_1	10.0	10.0	0.0
r_2 / r_1	2.0	2.07	-3.6
r_3 / r_2	0.667	0.683	-2.5

(c)

	Model	Prediction	Error(%)
h_1	5.0	10.0	-100
h_2 / h_1	5.0	10.0	-100
r_1	10.0	10.0	0.0
r_2 / r_1	0.667	0.698	-4.8
r_3 / r_2	0.667	0.673	-0.9

(d)

	Model	Prediction	Error(%)
h_1	3.0	10.0	-233
h_2 / h_1	3.0	10.0	-233
r_1	5.0	9.92	-98
r_2 / r_1	10.0	10.0	0.0
r_3 / r_2	5.0	0.67	87

in Table 7.2), the thickness of first layer and its ratio to that of second layer are out of the range of what have been taught and hence the network fails to predict them. However, the resistivities in

case (c) of Table 7.2 are predicted within 5% error, indicating that the neural network is robust when dealing with new data for which it has not been taught.

The experiments seem to suggest that the network works well if the model to be predicted is close to, or within the range of, the models used in training and otherwise it may fail. This produces a major difficulty for its use in electrical resistivity tomography (ERT) because a geological resistivity distribution is a complex problem which can be hardly described by a few models even for simple three layer structures. The neural network analogy is still at its primary stage of development and there are some problems remaining to be solved as has mentioned in Chapter 5. Its application to ERT problems may rely on its development in the future.

CHAPTER 8 FIELD EXPERIMENTS

This chapter presents results of several field tests obtained by using the techniques discussed in the previous chapters. An attempt has been made to cover applications ranging from relatively large scale problems such as geological investigations to small scale archaeological applications and from typical 2D structures such as dykes to the nearly 3D or dynamic structures encountered in environmental problems. In the experiments, some of the target structures are known and can thus be used to evaluate the algorithm while others are less well understood and the results have to be interpreted in the light of any other information available.

In all experiments, the field instrumentation system used is the ABEM Terrameter SAS300B resistivity meter which multiplexes to an equally spaced linear array of 20 electrodes through a mechanical switching box or computer controlled multi-cable system provided by Geoscan Research. During the data collection, some of the measurements were randomly repeated to check the data quality and reproducibility.

The meshes used here are the same as these have been stated in Chapter 7: in the method of regularised non-linear least squares, the forward modeling and inverse meshes (cells) are again 56× 13 and 48×9 respectively. In the Born inversion, the mesh is 52×11.

8.1 GEOLOGICAL INVESTIGATIONS

8.1.1 Cave detection

The electrical potential data over surveyed subsurface structures would be very valuable for evaluating the algorithms developed so far. Ideally the structure should be simple 2D with known geometry and resistivity contrast. Its depth should be moderate so that it is detectable by the surface linear array. Its geographical position is known and hence the electrode array can be accurately located with respect to the target. It was decided that a segment of the Long Churn

Cave System, near Ingleborough, Yorkshire would fulfil these requirements because it was explored and surveyed (Milner & Milner, 1977).

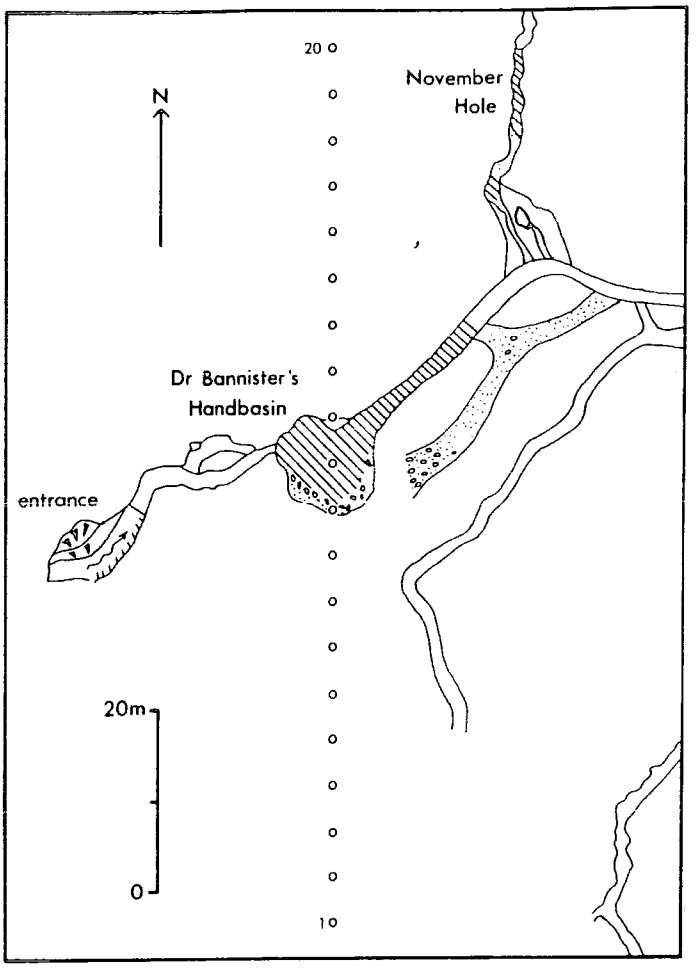


Figure 8.1 The position of a linear array of 20 electrodes in relation to the large chamber (Dr Bannister's Hand Basin) in the Upper Long Churn Cave system. The inter-electrode spacing is 5m and the pole-pole electrode configuration was applied, with two additional electrodes placed 750m away from the array on opposite directions.

The Long Churn Cave System is developed within the limestones of Carboniferous age which in this area form a unit dipping gently north-east. The limestones, known as the Great Scar Limestone, are about 200m thick and made up of limestones of great vertical variation but little lateral variation (Sweeting, 1974). The vertical variation in the beds of the Great Scar Limestone in the north-west Yorkshire is a fundamental feature in the development of both the relief and caves of the region. The ground surface at the test site is flat except that the limestones are partially

exposed, forming clints and grykes with soil filled clints. The location of the test was made by reference to the survey compiled by Milner & Milner (1977).

A linear array of 20 electrodes with a 5m inter-electrode spacing was placed crossing the position of a large chamber approximately 20m from the upstream entrance to the Upper Long Churn Cave (Figure 8.1). Installing the array at this site presented some problems due to the restricted choice of soil pockets among the limestone pavement and the whole array was carefully located to avoid the electrodes being planted on the grykes. A total of 190 potential data was collected by a pole-pole electrode configuration where the two 'infinite' electrodes were placed 750m away from the linear array in two opposite directions. Field data checks showed that the potential reproduction error was less than 2%, indicating that the electrical noise level at the site was low.

The first image shown in Figure 8.2(a) was reconstructed from the data on a laptop computer immediately after the field work using the Born inversion. The singular value threshold $\lambda_f=4.5\times 10^{-4}\lambda_{max}$ and the initial model was a uniform half space with a resistivity of 3000 ohm-m, estimated from the Wenner apparent resistivities. The data was then inverted on a HP workstation using the method of zero-order regularised least squares plus a spatial averaging smoother. Starting from the same initial model of 3000 ohm-m with an initial data residual of 30%, the convergence was achieved with a *rms* of 8.5% after 9 iterations and the resulting image is shown in Figure 8.2(b). The regularisation parameter λ was automatically selected as the minimizer of data residual function during each iteration.

On the image reconstructed by the Born inversion (Figure 8.2(a)), the passage is clearly visible as a diffuse, high resistivity anomaly directly beneath the midpoint of the multielectrode array with a dimension of about 40m×15m and 10m in depth to the top of the cave chamber. The small, high resistance features which appear to extend almost to the surface may represent open joints in the roof of the chamber, or alternatively, they may be the image artefacts. Both the dimensions and the

depth to the top of the cave are larger than those estimated from published surveys which indicate that the cave has a dimension of 11m×6m and a depth to the roof of 8m. This error almost certainly arises from the assumption made in the Born procedure that the distortion of current paths due to the inhomogeneities is negligible: the air filled chamber has a resistivity of infinity and hence the assumption is severely violated. The success in resolving the cave passage with such a high resistivity contrast indicates the robustness of the method. The resolution can be improved by the regularised non-linear iteration method as shown in Figure 8.2 (b) in which both the horizontal and vertical dimensions of the cave appear to be about 10m. The burial depth and the height of the cave are slightly larger than the published data, possibly due to both the decay of the resolution and the rather coarse model cells used along vertical direction. As a comparison, the data of a Wenner pseudosection over the same array were produced from field data by superposition and the results are shown in Figure 8 (c). In this image the cave is hardly visible.

8.1.2 Dyke detection

The Hett Dyke is one of the major whin dykes of N.E. England and near to Hamsterley village in South Durham it was thought that a more complex structure might occur (Goodwillie, 1988). An electrical resistivity survey was carried out on the Black Hill Top farm land to investigate this matter. In addition, the dyke is a typical 2D geological structure and hence it is an ideal test site for the algorithms, although its exact geometry is unknown because it is beneath the weathered zones at the site. The main aims of the experiments were:

i) To test whether the algorithms developed in previous chapters can resolve the Hett Dyke at the site.

ii) To investigate whether the Hett Dyke at the site has a more complicated structure than a single, vertical 2D structure as it was previously thought.

iii) To reconstruct the electrical resistivity image of the dyke so that the dimensions and the depth to the top of the dyke can be estimated.

iv) To carry out an experiment of the data transform method described in Chapter 2 by collecting two different data sets on the same section and then transforming one into the other to see if the method does work with field data.

The geological map of north-east England indicates that the test area is covered by the Coal Measures of Carboniferous age, which mainly consists of interbedded sandstone and mudstone with seatearths and coal seams. The Hett Dyke is known to be a quartz-dolerite dyke which was intruded along tension fissures during the late Carboniferous earth movements and it traverses across the field on a strike of approximately 070^0. The test site is covered by a weathered zone of several metres deep and slopes away to the North at about 8^0.

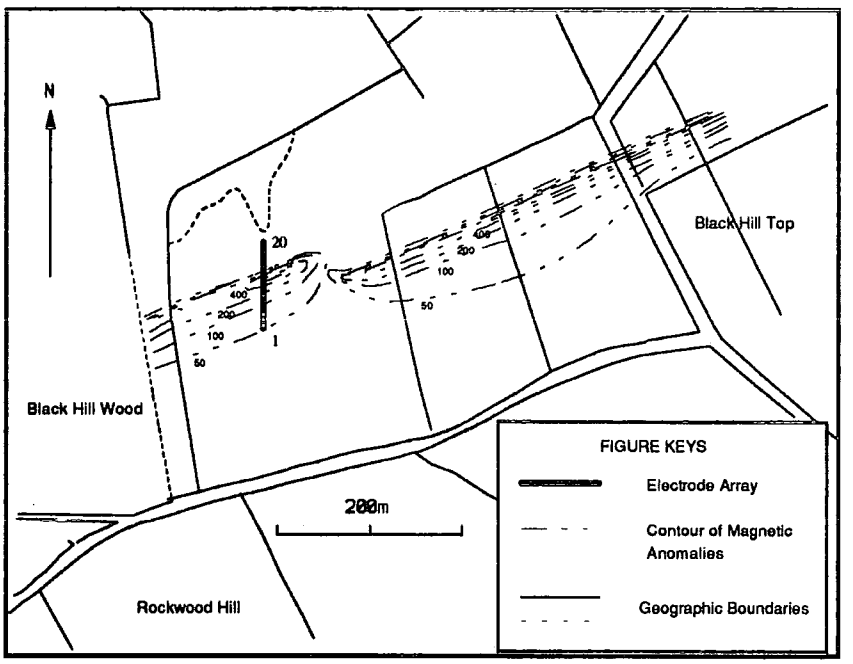


Figure 8.3 The electrode array position with respect to the magnetic anomalies and geographic locations. The first electrode is at the south end of the profile while the last at the north end. The contour map of magnetic anomalies is adapted from Goodwillie (1988). The unit of magnetic anomalies are nano-Tesla.

A linear array of 20 electrodes equally spaced at 5m intervals was planted northward crossing the position of the dyke indicated by a magnetometer survey (see Figure 8.3). The position was located by reference to the pit lay during the magnetic survey. A total of 170 potential data was collected by the *c*CCPP electrode configuration and the results are shown in Figure 8.4 in which images (a) and (b) are reconstructed by Born inversion and the method of zero-order regularised least squares plus a spatial averaging smoother respectively. The initial model for both Born inversion and the iterative algorithm was a homogeneous half space with a resistivity 300 ohm-m. The data residual (or rms) decreased from 25.4% to 12.6% after 4 successive iterations and further iterations showed no improvement. Such a data residual appeared to be relatively high and therefore several other initial models were tried to check if the iteration was trapped at a local minimum of high data residual. It was found that the resulting images resembled image (b) in Figure 8.3 but with even higher data residual. The relatively high data residual may be attributable

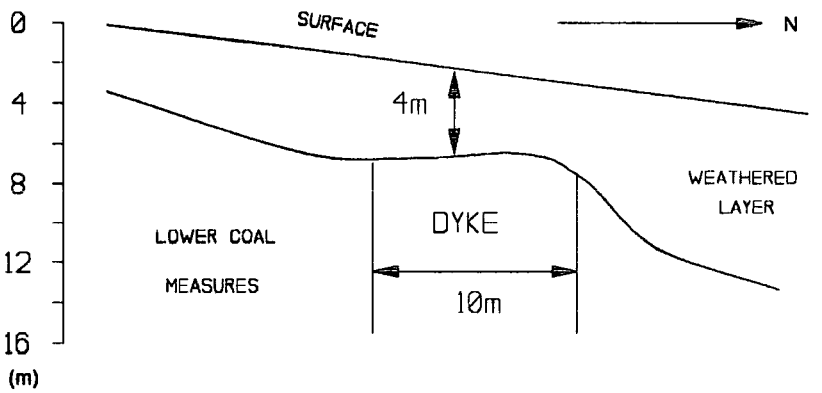


Figure 8.5 Depth to rock head near the Hett Dyke along the electrical resistivity survey line interpreted from seismic refraction surveying (after Todd, 1988).

to the errors in the forward modeling which was known to be relatively high for a **cCCPP** electrode configuration and the inconsistency between the physical model, including the smoothness constraints, and the real earth. Nevertheless, the dyke seems to have been located satisfactorily.

Figure 8.3 shows that the dyke appears as a vertical high resistivity zone with a width of about 8m and a depth of about 5m to its top, suggesting that it is indeed a simple vertical sheet-like dyke. Although the dyke is not detectable beyond a depth of 20m on both images (a) and (b), this is probably due to the rapid decay of resolution with increasing depth, rather than to a genuine change in the structure. This aspect has been mentioned in the previous chapters which showed that data from an equally spaced linear surface array of 20 electrodes may only be able to recover adequately structures to a depth of about three inter-electrode spacings if the resistivity contrasts and the dimensions of the targets are not large. The low resistivity zones on both sides of the dyke are considered to be a weathered layer which has been measured by a seismic refraction survey (see Figure 8.5). Hydrologically, the dyke could be an impermeable barrier to the ground water at the test site because it cuts the north-downwards slope at nearly a right angle and it is therefore possible that the low resistance block towards the left hand (south) side of the dyke is also due to ground water accumulated along the upper slope side of the dyke.

The dyke width (~8m) and depth to its top (~5m) estimated from resistivity images (a) and (b) seems comparable to the results from magnetic and seismic refraction surveys near the ERT profile: the 2D magnetic modelling yielded a dyke width of 13m and a depth to its top of 6m while the seismic refraction gave 10m and 4m respectively (Goodwillie, 1988; Todd, 1988).

In the field survey, after the *c*CCPP data set was collected, the electrodes of the same array were multiplexed to the resistivity meter to collect the data set of a Wenner pseudo-section. The result is shown in Figure 8.4 (c) which also indicates the presence of high resistivity dyke and the low resistivity weathered layer. A comparison of the data collected and those superposed from *c*CCPP data set has been shown in Table 2.7 of Chapter 2 which indicates that the difference on average is about 1.5%, suggesting that the data transform method can be applied in practice.

8.2 ARCHAEOLOGICAL SITE EVALUATION

8.2.1 Lanchester Roman Fort

The Roman fort of *Longovicium* is sited beside the B6296 road, 0.8km west of Lanchester, County Durham. The most extensive excavations were made by Kenneth Steer in 1937 and he concluded that the first occupation of the fort was Hadrianic on the basis of pottery, the style of the gateways, the probable orientation of the interior and structural analogies with Hadrian's Wall (Noel, 1990). Most sources state that the fort was left unoccupied from c.190-240 AD and when under the Emperor Gordian III (238-244) there was a period of rebuilding which included the defences, *principia* and armouries along with the construction of a new bath house with basilica. The soil which covers the fort is an acid sandy clay with a subsoil of yellow clay overlying sandstone and brash within the Upper Coal Measures of Carboniferous age (Noel, 1990).

Noel (1990) mapped the site with a fluxgate magnetic gradiometer at a resolution of 0.5m×0.5m. On his geomagnetic map, most of the individual building structures appear to be recognisable but their vertical extension, burial depth and exactly geometry remain unknown. It was thought that the ERT methods could provide such information and therefore a field experiment was carried out at the site.

A linear array of 20 electrodes with 1m inter electrode spacings was placed at a right angle to the two parallel linear structures which were thought to be two separate walls at the south side of the headquarters building (see Figure 8.6). The array position was located by reference to the pegs which were put down to locate during the magnetic survey. A pole-pole electrode configuration was used where the two remote electrodes were planted about 300m away from the array at opposite directions. A total of 190 potential data was collected. The image was then reconstructed from the data by the method of zero-order regularised least squares plus a spatial averaging smoother. The initial model was 100 ohm-m homogeneous half space and after 11 iterations the data residual reduced from 22% to 7%. The resulting image is shown in Figure 8.7 where the two

blurred, high resistivity objects are buried about 2m deep with approximately 1.5 to 3m in width. The distortions near the surface and the upper left corner of the image may be attributed to the data noise or the localised inhomogeneity of the soil.

8.2.2 Barnard Castle

Barnard Castle is one of the most important historic remains in the North England and was constructed in the 11th century. An electrical resistivity survey was carried out to investigate

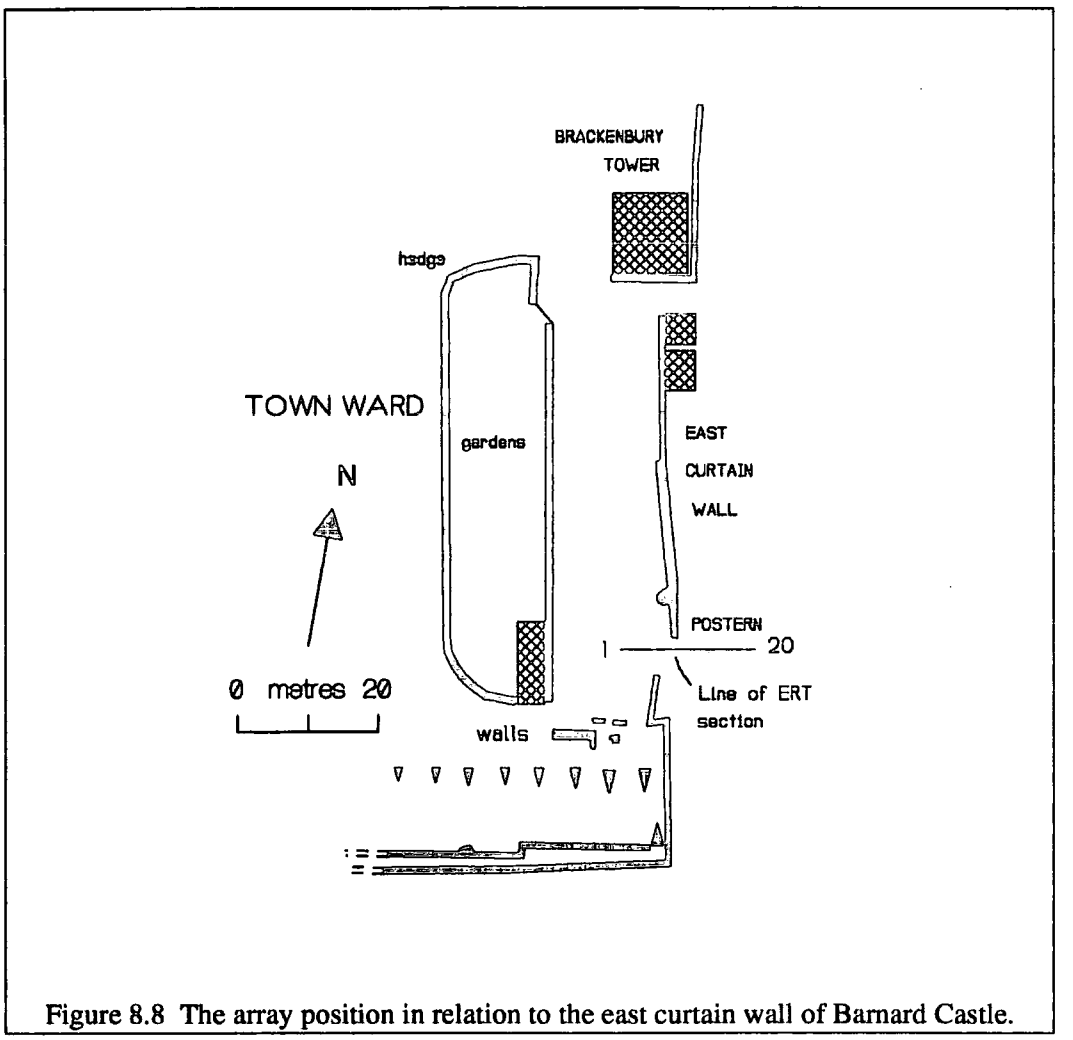


Figure 8.8 The array position in relation to the east curtain wall of Barnard Castle.

whether there is a ditch outside its East curtain wall and whether a tower, thought to exist at the postern, is indeed present (see Figure 8.8). A linear array of 20 electrodes equally spaced at 1m

intervals was placed eastward across the wall as shown in Figure 8.8 and 170 data were collected by the *c*ℙℂℙℂ electrode configuration. The image reconstructed by the Born inversion is shown in Figure 8.9 where a low resistivity zone appeared in the middle, with approximately 4m in width and 1.5m in depth from the surface to the top. This seems to suggest the presence of a defence ditch. The curtain wall, which is still partially visible at the site, is shown on the image as a high resistivity block of about 2m×1m, near to the surface at 7m from left. The other near surface high resistivity features are considered to be the recently backfilled materials such as bricks which are still visible. These localised, near surface high resistive, inhomogeneous materials are considered to be the main cause of the failure of the regularised non-linear least squares method in this test: the iteration was terminated with a very high data residual. No firm evidence was found for the foundation courses of the postulated tower.

8.3 ENVIRONMENTAL INVESTIGATIONS

8.3.1 Oil pollution

Ground water and soil pollution problems have gained a great deal of attention in recent years and hence it would be very useful to test if the methods developed in previous chapters can be applied to investigate such environmental problems. A test site was made available by British Gas at an old industrial site where lagoons were used to receive separator sludge and spent clay resulting from a lube oil finishing process which ended in 1963. The exact location is confidential. Previous trial pit survey indicated that the contaminates are mainly the localised spent clay severely

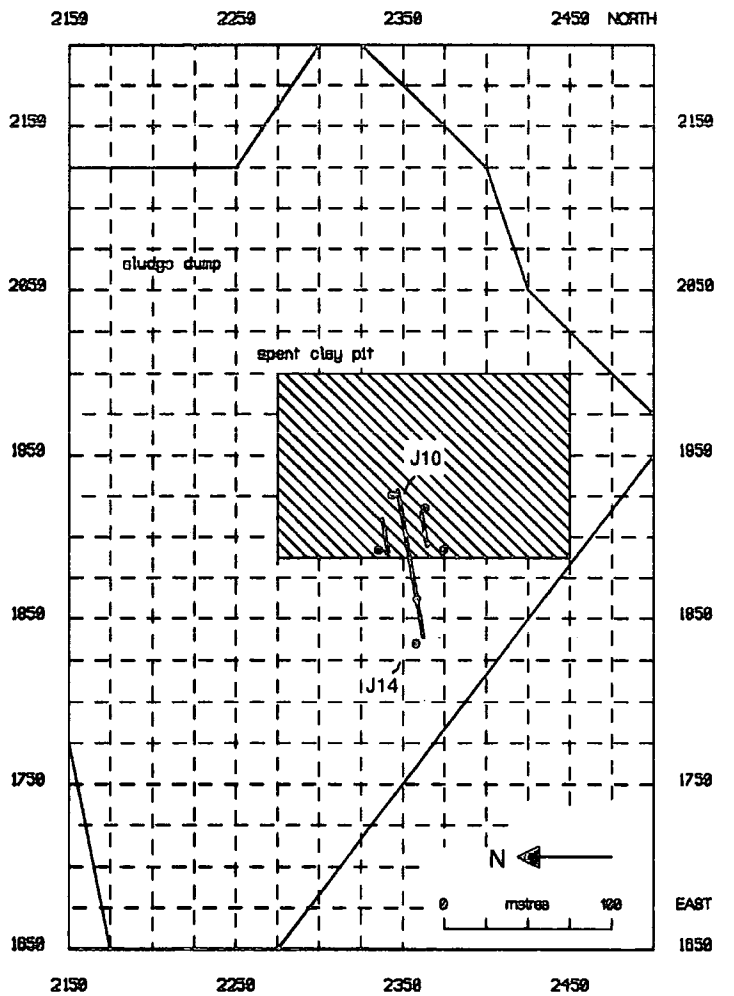


Figure 8.10 The array positions in relation to the spent clay pit.

contaminated with oil and the widely distributed, less oil-contaminated sands and gravels. The study found that the contaminated spent clay is strongly hydrophobic, with oil comprising about 20% by weight, and thus likely to have a higher electrical resistivity than the clean soil or the underlying clay unit. The aims of the electrical resistivity survey were:

i) To investigate the lateral and vertical extension of the contamination within the spent clay pit.

ii) To see if there is any indication that the contamination has been migrated across the west boundary of the spent clay pit into the neighbouring soil.

Figure 8.10 shows the positions of resistivity survey lines with respect to the boundaries of the sludge dump and spent clay pit which was reconstructed from old records and aerial photographs. The *c*CCPP electrode configuration was applied and two separate but related field experiments were carried out:

a) Long traversing line

This is a baseline which was established extending from post J10 to cover the west boundary of the spent clay pit and consists of five short sections (1.1+1.2+1.3+1.4+1.5), as shown in Figure 8.11. Each short section has a length of 19m, with a 1m overlap between successive sections and a 1m inter-electrode spacing. The objective of the experiment was to investigate whether the contamination has migrated westwards across the spent clay pit boundary and if it has, how far it has extended. The 1m overlap between successive sections was designed to check the edge distortion in the reconstructed images which may appear. Data from each of the five sections were inverted by the Born procedure to produce the image shown in Figure 8.12.

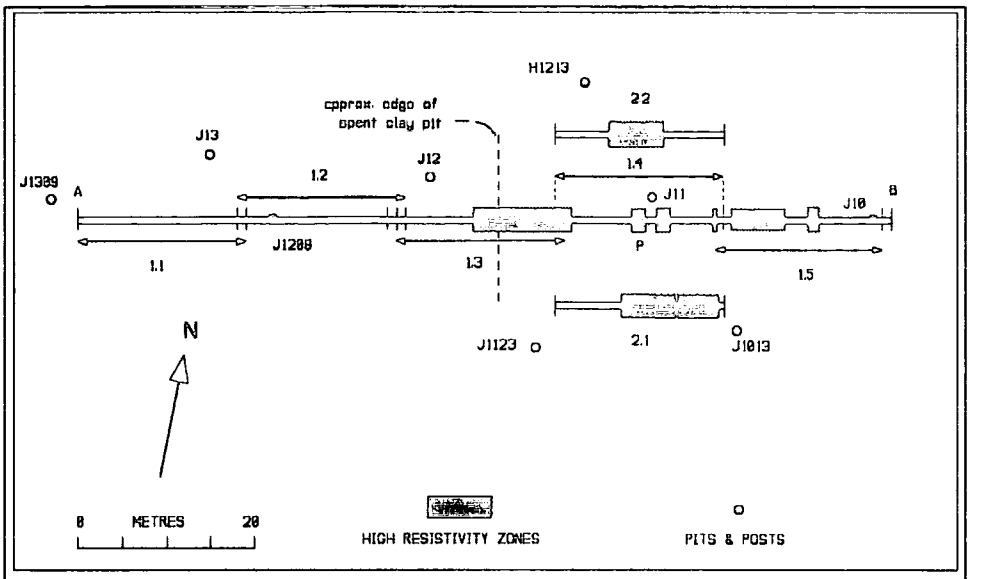


Figure 8.11 The array positions on the long traversing line (baseline) and the short parallel lines.

b) Short parallel lines

Two sets of data were obtained with the array placed parallel to the baseline but displaced, first 10m to the south and then 10m to the north (sections 2.1 and 2.2). The midpoints of these arrays were coaxial with the midpoint of section 1.4. The objective of such section layouts was to explore the changes in subsurface resistivity on the two sides of the baseline. The inter-electrode spacing was again 1m and the images were reconstructed by the Born procedure. The results are shown in Figure 8.13 in which the image of section 1.4 is included again for the convenience of comparison.

Two characteristics appear on all reconstructed images: a relatively lower resistivity layer extends from the surface to 1~2m depth and some artefacts appear on the two upper corners of each image. Experience suggests that such image artefacts usually arise in Born inversion when some of the readings in a *c*CCPP data set are too small to be measured accurately, mainly those readings with large electrode spacings. The low resistivity layer probably reflects the less contaminated materials saturated with water after several weeks of heavy rain prior to the survey (the pools of water were visible on the site). Because of this near surface low resistivity layer, some of the meter

readings approached zero. These very small, inaccurate data are considered to be the main cause of the artefacts as mentioned above.

In spite of the image artefacts, the major contamination still seems to be located and they appear as irregular and diffuse high resistivity blocks extending from near surface down to about 4m. The contamination is mainly within the spent clay pit, as shown on the image sections 1.3, 1.4. 1.5, 2.1 and 2.2. On the long traversing line, the image sections 1.1 and 1.2 seem to comprise a two-layer structure with some image artefacts and thus they appear to suggest that there is no major contamination outside the west boundary of the spent clay pit. The boundary of the spent clay pit is about 11m from west on the image section 1.3 in which the contamination appears mainly on the east side of the boundary, i.e. within the spent clay pit, extending from near surface to about 4m deep. The image section 1.3 also seems to suggest that part of the contamination has diffused westward across the pit boundary. The contamination on the image section 1.3 extends into image section 1.4 horizontally but it is not continuous. On the image section 1.5, near to the upper left hand side there is low resistivity block extending from near the surface to about 2m and immediately beneath this is a high resistivity block. This image structure may be interpreted as less contaminated materials saturated with water which sit on the top of the highly contaminated spent clay with a lower permeability. Alternatively, this feature may simply be the image artefacts which are known to be relatively strong at these positions in the image.

Although there are high resistivity zones on the two short parallel image sections, no obvious correspondence even is seen between these zones beneath the section 2.1, 2.2 and the baseline section 1.3, suggesting that the contamination is localised on a horizontal scale which is smaller than the separation of these sections (10m), i.e. it is three dimensional.

Attempts to reconstruct the images by the methods of regularised non-linear least squares have proved unsuccessful, probably due to the 3D nature of the contamination while the iterative

algorithms developed here assume that the data are collected over 2D structures. The iterations were normally terminated at as high as 60% misfit error.

8.3.2 Watering experiment

Environmental monitoring often requires measurement of dynamic changes of ground water. Recent laboratory experiments on sandstone and clay samples indicated that both the organic and inorganic contamination will cause a change of resistivity (Börner *et al.*, 1993). Therefore electrical resistivity survey methods can be used to monitor the dynamic change or motion of subsurface water in and around waste dump sites in order to provide information about pollution leakage or migration. Several field experiments have been carried out to assess the usefulness of electrical resistivity methods in this area. For example, Bevc and Morrison (1991) conducted an experiment in which saline water was injected into a fresh acquifer while the resistivity was monitored using a multichannel borehole to borehole system; Van et al. (1991) monitored the leaks from storage ponds using resistivity methods. However, near the University of Durham, there is no waste disposals available for an experiment. It was thus decided to simulate the case by a watering experiment, i.e. to create dynamic resistivity change of the subsurface soil by watering the ground. A data set was collected before watering and another after. Then two images corresponding to the two data sets can be reconstructed and their differences may indicate the dynamic resistivity change.

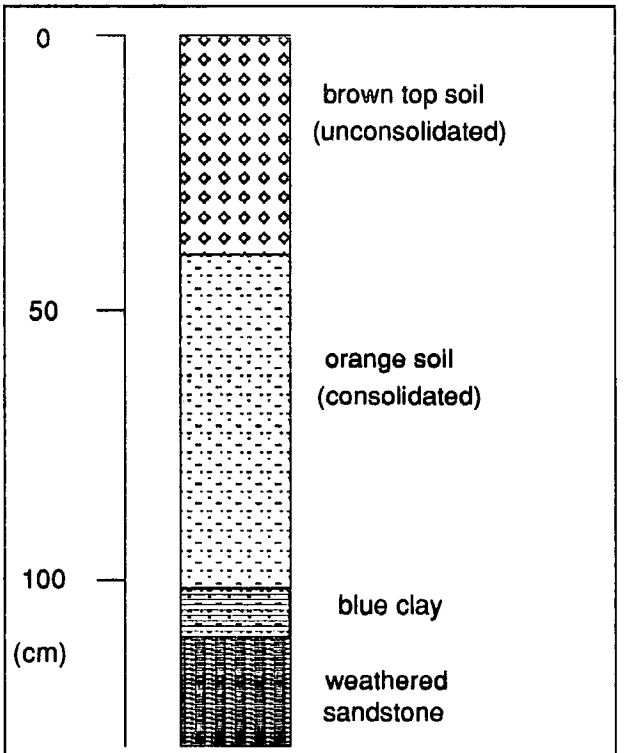


Figure 8.14 The vertical section of soils at the Botanic Garden, University of Durham. The sampling position is at the middle of the electrode array in the watering experiment.

The experiment site was selected at the Botanic Garden, University of Durham, for its relatively low noise level and readily available fresh water (tap) supply. The experiment area is covered by soil and samples from a shallow auger survey indicated that the top soil layer of about 40cm thick is loose and beneath this there is a more condensed layer of soil underlain by a thin clay unit with a total thickness of about 70cm (see Figure 8.14). The depth to the top of the weathered sandstone is about 1.1m.

A linear array of 20 electrodes equally spaced at 1m intervals was setup and 170 data were collected using the *c*CCPP electrode configuration. Then the fresh water of about 0.611m^3 was injected from the ground surface in the middle of the array at a speed of 0.785m^3/h for about 47 minutes. After about 20 minutes, another data set was collected with the same electrode array at the same position. The 20 minutes delay between the end of the watering and the beginning of the second data set collection was considered to allow the fresh water within the soil to reach a relatively stable status. The sequence of the experiment is shown in Figure 8.15.

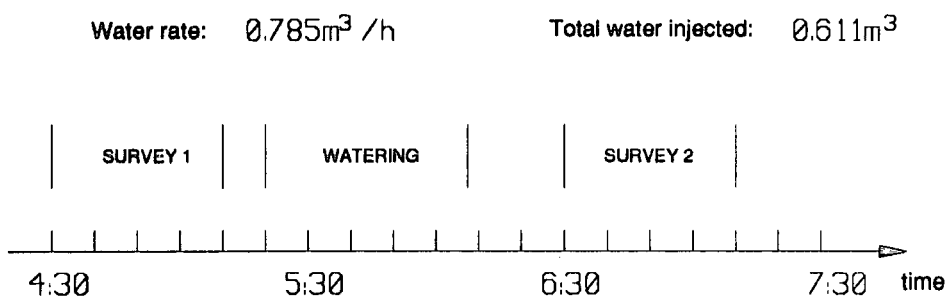


Figure 8.15 The time sequence of the watering experiment: an electrical resistivity survey was carried out before the watering and another afterward.

The images (a) and (b) in Figure 8.16 were reconstructed by Born inversion from the two data sets while the image (c) is a result of image (b) minus image (a). The resistivity changes due to the water injection are hardly visible on image (b) but they are clear indicated as a low resistivity block

at the middle of the array, spreading downward about 1.5m and horizontally about 2m. The 1.5m vertical extension of the low resistivity block suggests that the water flow was stopped when it reached the clay which is less permeable.

CHAPTER 9 CONCLUSIONS AND SUGGESTIONS FOR FUTURE WORK

9.1 CONCLUSIONS

In the previous chapters, the methods of electrical resistivity image reconstruction using linear surface arrays have been developed and tested. The techniques include data collection methods, the linearised and the non-linear image reconstruction algorithms and image presentation methods. From the discussions in previous chapters, the following conclusions can be made:

1 A complete or primary data set which contains all linearly independent measurements of apparent resistivity can be collected using two, three or four electrode configurations on a linear array placed on the ground surface. From this complete data set, it has been shown that any other data sets obtainable on the linear array can be synthesised through a process of superposition. If the array is required to be moved along a traversing line in order to cover a broader area, the minimal redundancy and the completeness of the data set can be achieved by the roll-along data collection method. The number of independent data in a complete data set is normally larger than that of conventional data sets, such as the data set used to compute a Wenner pseudosection.

2 In most cases, the complete data set is likely to produce a better inversion or image than conventional data set for pseudo-section generation. Among the complete data sets, the circulating dipole-dipole (***c*CCPP**) data set usually produces the best results in the presence of data noise although the number of independent data is less than that of pole-pole data set.

3 In comparison to the regularised non-linear least squares algorithms, the Born inversion is fast and not so sensitive to the reference (or initial) models and can achieve modest

resolution if the resistivity structure is not complicated. It can also be applied to image the cross-section of 3D structures or dynamic changes of resistivities. The application of Born inversion is limited by its lack of resolution when a resistivity structure is complicated or resistivity contrast is high. It is also relatively sensitive to noise. Another drawback of Born inversion is that it can only recover relative variation of resistivities, not their absolute values.

4 The algorithms of the regularised non-linear least squares plus various spatial smoothers normally produce better results than Born inversion if the following conditions are satisfied: (i) the initial model is not far away from the true solution; (ii) the balance between the image resolution and the image smoothness is properly regulated. This includes a proper choice of regularisation parameter λ and the smoothing functions, including spatial smoothers, for the problem at hand; (iii) the forward modelling is accurate enough. Experiments indicate

a) The choice of regularising parameter λ has a crucial importance for the success of the iteration. It is context dependent and thus *a priori* knowledge is desirable. The conventional Marquardt's strategy of choosing λ usually failed when substantial noise is present in the data. Better results can be obtained by choosing λ as a minimizer of the data residual or GCV functions although there is no guarantee that these strategies are suitable for all cases. The λ can also be chosen as a constant value by experiments although it should be adjusted according to the data noise level and the roughness of the geologic structures being investigated.

b) Among the various choices of smoothing functions, the zero-order regularisation function plus spatial smoothers seem to provide a better balance of the stability and resolution. The spatial median smoother can effectively filter out the isolated spikes of

the image during the iterations while it still preserves the sharp boundaries which may exist. However, the spatial averaging smoother, weighted by cell size, imposes stronger smoothing constraints on the solution and the corresponding algorithm (the method of regularised least squares plus spatial averaging smoother) has the highest stability among the tested algorithms.

c) Experiments indicate that the stronger smoothing constraints not only produce a stable image in the presence of data noise but also reduce the influence of the initial model on the final results. In the absence of any knowledge of the underlying geologic structures and the data noise level, the image reconstructed under stronger smoothing constraints is thus more reliable.

5 Image resolution decays rapidly with increasing depth in both Born inversion and regularised non-linear least squares algorithms. Although no analytic formulation is developed to characterised such decays, experience suggest that these methods may only be able to image objects within a depth of 2-3 inter-electrode spacings for an equally spaced linear surface array of 20 electrodes. Deeper structures may only be visible on the image if their dimensions are large or resistivity contrasts are high.

6 Initial numerical experiments suggest that although the genetic algorithms or neural network methods may be applied to process the data over 2D resistivity structures in the future, currently their use in practice is severely limited by the high cost of computation.

9.2 SUGGESTIONS FOR FUTURE WORK

Although some encouraging results have been obtained using the algorithms developed so far, the method of ERT is still at its early stage of development and more field experiments should be carried out to test their performance on different geologic and archaeological problems. Further

improvement in the algorithms is also desirable in the future development. First, the non-linear iterative algorithms developed in the previous chapters are still slow and can only be implemented on workstation class computers. It is possible to transfer the program to a personal computer by re-coding the algorithm to reduce computing costs, in both speed and memory. For example, it is possible to increase the speed by eliminating the redundant computations during iterations currently in the program. As the Jacobian matrix is the most intensive task of the computation, special efforts should be made to ensure that the matrix is computed most efficiently. Secondly, the results may be improved by increasing the accuracy of forward modelling in the iterative algorithms. Thirdly, a robust strategy for the selection of regularisation parameter, which can take account the data noise level and image complexity, is still desirable.

The methods developed so far focus on the use of a linear array on the ground surface alone. A natural extension of current work would be to expand the data collection techniques and the image reconstruction algorithms to the cross borehole and borehole to surface geometries. The data collection methods developed in Chapter 2 could also be extended to the situation where the array is distributed as a 2D pattern and the corresponding data processing methods have not been explored yet.

In the development of the image reconstruction algorithm, there are some other possibilities which may be explored in the future. For example, it may be possible to develop a 'ray-tracing algorithm' for electrical resistivity image reconstruction if the voltage between two current transmitters can be measured accurately. If the task is to reconstruct a dynamic resistivity image, a conformal transformation may be applied to transfer the subsurface (half space) into a circular region and then the fast algorithm for medical imaging (Barber & Brown, 1984) may be applied. Finally, the developments of relative fast and yet reliable data collect system is important for the practical application of ERT.

REFERENCES

Alfano, L., 1959, Introduction to the interpretation of resistivity measurements for complicated structural conditions: Geophysical Prospecting, 7, 311-366

Acworth, R.I., and Griffiths, D.H., 1985, Simple data processing of tripotential apparent resistivity measurements as an aid to the interpretation of subsurface structure: Geophysical Prospecting, 33, 861-887.

Backus, G., and Gilbert, F., 1968, The resolving power of gross earth data: Geophys.J.Roy.Astr.Soc., 16, 169-205.

Banerjee, B., and Sengupta, B.J., 1987, Transformation of dipolar, Wenner and two-electrode curves to Schlumberger apparent resistivity sounding curves: Geophysical Prospecting, 35, 445-453.

Barber, D.C., and Brown, B.H., 1984, Applied potential tomography: Journal of Physics E: Scientific Instrument, 17,723-733.

Barber, D.C., and Seagar, A.D.,1987, Fast reconstruction of resistance images: Clinical Phys. Physiol. Meas., 8, Supp. A, 47-54.

Barber, D.C., 1989, A review of image reconstruction techniques for electrical impedance tomography: Medical Physics, 16, 162-169.

Barber, D.C., 1990, Image Reconstruction in Applied Potential Tomography—Electrical Impedance Tomography: Internal technical report, INSERM Unite 305, Toulouse, France.

Barker, R.D., 1981, The offset system of electrical resistivity sounding and its use with a multicore cable: Geophysical Prospecting, **29**, 128-143.

Barker, R.D., 1989, Depth investigation of colinear symmetrical four-electrode array: Geophysics, **54**, 1031-1037.

Barker, R.D., 1992, A simple algorithm for electrical imaging of the subsurface: First Break, **10**, 53-62

Bates, R.H.T, McKinnon, G.C., and Seagar, A.D., 1980, A limitation on systems for imaging electrical conductivity distributions: IEEE Trans. Biomed. Eng.: **BME-27**, 418-420.

Bates, R.H.T., Garden, K.L., and Peters, T.M., 1983, Overview of computerized tomography with emphasis on future developments: Proceedings of IEEE, **71**, 356-372.

Bevc, D., and Morrison, H.F., 1991, Borehole-to-surface electrical resistivity monitoring of salt water injection experiment: Geophysics, **56**, 769-777.

Bishop, T.N., Bube, K.P., Culter, R.T, Langan, R.T., Love, P.L, Resnick, J.R., Shuey, R.T., Spindler, D.A., and Wyld, H.W., 1985, Tomographic determination of velocity and depth in laterally varying media: Geophysics, **50**, 903-923

Börner, F., Gruhne, M., and Schön, J., 1993, Contamination indications derived from electrical properties in the low frequency range: Geophysical Prospecting, **41**, 83-98.

Bregman, N.D., Chapman, C.H., and Bailey, R.C., 1989, Crosshole seismic tomography: Geophysics, **54**, 200-215.

Bristow, C.M., 1966, A new graphical resistivity technique for detecting air-filled cavities: Study in Speleology, **1**, 204-227.

Brown, B., Barber, D., and Tarassenko, L., *ed.*, 1987, Electrical Impedance Tomography—Applied Potential Tomography, *in* Clinical Physics and Physiological Measurement, **8**, Suppl. A.

Brown, B., Barber, D., and Jossinet, J., *ed.*, 1988, Electrical Impedance Tomography—Applied Potential Tomography, *in* Clinical Physics and Physiological Measurement, **9**, Suppl. A.

Brown, B.H., and Seagar, A.D., 1987, The Sheffield data collection system: Clinical Phys. Physiol. meas., **8**, Suppl. A, 91-98.

Carpenter, E.W., 1955, Some notes concerning the Wenner configuration: Geophysical Prospecting, **3**, 388-402

Carpenter, E.W., and Habberjam, G.M., 1956, A tri-potential method of resistivity prospecting: Geophysics, **21**, 455-469.

Carpenter, P.J., Calkin, S.F., and Kaufmann, R.S., 1991, Assessing a fractured landfill cover using electrical resistivity and seismic refraction techniques: Geophysics, **56**, 1896-1904.

Coggon, J.H., 1971, Electromagnetic and electrical modelling by the finite element method: Geophysics, **36**, 132-155.

Cohen, J.K. and Bleistein, N., 1979, Velocity inversion for acoustic waves: Geophysics, **44**, 1077-1087.

Constable, S.C., Parker, R.L., and Constable, C.G., 1987, Occam's inversion: a practical algorithm for generating smooth models from EM sounding data: Geophysics, **52**, 289-300.

Cheng, K.S., Simske, S.J., Isaacson, D., Newell, J.C., and Gisser, D.G., 1990, Errors due to measuring voltage on current-carrying electrodes in electric current computed tomography: IEEE Trans. Biomed. Eng., BME-37, 60-65.

Daily, W., 1984, Underground oil-shale retort monitoring using geotomography: Geophysics, 49, 1701-1707.

Daily, W., and Owen, E., 1991, Cross-borehole resistivity tomography: Geophysics, 56, 1228-1235.

Davis, J. L., and Annan, A. P., 1989, Ground-penetrating radar for high-resolution mapping of soil and rock stratigraphy: Geophysical Prospecting, 37, 531-551.

Davis, L. (Ed.), 1987, Genetic Algorithms and Simulated Annealing, Pitman Publishing, London.

Dey, A. and Morrison, H.F., 1979, Resistivity modelling for arbitrarily shaped two-dimensional structures: Geophysical Prospecting, 27, 106-136.

deGroot-Hedlin, C., and Constable, S., 1990, Occam's inversion to generate smooth, two-dimensional models for magnetotelluric data: Geophysics, 55, 1613-1624.

Dines, K.A, and Lytle, R.J., 1979, Computerized geophysical tomography: Proceedings of IEEE, 67, 1065-1073.

Dines, K.A., and Lytle, R.J.,1981, Analysis of electrical conductivity imaging: Geophysics, 46, 1025-1036.

Dyer, B.C., and Worthington, M.H., 1988, Some sources of distortion in tomographic velocity images: Geophysical Prospecting, 36, 209-222.

Edwards, L.S., 1977, A modified pseudosection for resistivity and IP: Geophysics, 42, 1020-1036.

Eskola, L., 1979, Calculation of galvanic effects by means of the method of sub-areas, Geophysical Prospecting, 27, 616-627.

Fisher, E., McMechan, G.A., and Annan, A.P., 1992, Acquisition and processing of wide-aperture ground-penetrating radar data: Geophysics, 57, 495-504.

Fox, R.C., Hohmann, G.W., Killpack, T.J., and Rijo, L., 1980, Topographic effects in resistivity and induced-polarization surveys: Geophysics, 45, 75-93.

Ghosh, D.P., 1971a, The application of linear filter theory to the direct interpretation of geoelectrical resistivity sounding measurements: Geophysical Prospecting, 19, 192-217.

Ghosh, D.P., 1971b, Inverse filter coefficients for the computation of apparent resistivity standard curves for a horizontally stratified earth: Geophysical Prospecting, 19, 769-775.

Goldberg, D.E., 1989, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley Publishing Company, Inc.

Golub, G., Heath, M., & Wahba, 1979, Generalised cross-validation as a method for choosing a good ridge parameter: Technometrics, 21, 215-223.

Goodwillie, A., 1988, An intensive magnetometer survey over the Hett Dyke near Hamsterley, BSc. dissertation, Department of Geological Sciences, University of Durham.

Griffiths, D.H., and Turnbull, J., 1985, A multi-electrode array for resistivity surveying: First Break, 3, 16-20.

Griffiths, D.H., Turnbull, J., and Olayinka, A.I., 1990, Two-dimensional resistivity mapping with a computer-controlled array: First Break, 8, 121-129.

Hall, E.L., 1979, Computer Image Processing and Recognition, Academic Press.

Helgason, S., 1980, The Radon Transform, Birkhauser.

Herman, G.T., 1980, Image Reconstruction From Projections---the fundamentals of computerized tomography, Academic Press.

Hohmann, G.W., and Raiche,A.P., 1988, Inversion of controlled-source electromagnetic data, in Electromagnetic Methods in Applied Geophysics, Volume 1, Edited by Nabighian, M.N.,SEG, Tulsa

Hohmann, G.W., 1988, Numerical modeling for electromagnetic method of geophysics, in Electromagnetic Methods in Applied Geophysics, Volume 1, Edited by Nabighian, M.N.,SEG, Tulsa

Holland, J.H., 1975, Adaptation in Natural and Artificial Systems, University of Michigan Press.

Hopfield, J.J., and Tank, D.W., 1985, Neural computation of decisions in optimisation problems: Biological Cybernetics, 52, 141-152.

Hua, P., Woo, E.J., Webster, J.G., and Tompkins, W.J., 1991, Iterative reconstruction methods using regularization and optimal current patterns in electrical-impedance tomography: IEEE Transactions on Medical Imaging, 10, 621-628.

Hua, P., Woo, E.J., Webster, J.G., and Tompkins, W.J., 1993, Using compound electrodes in electrical impedance tomography: IEEE Transactions on Biomedical Engineering, 40, 29-34.

Ider,Y.Z. Gencer, N.G., Atalar, E., Tosun, H., 1990, Electrical impedance tomography of translationally uniform cylindrical objects with general cross-section boundaries: IEEE Trans. Medical Imaging, 9, 49-59.

Inman, J.R., 1975, Resistivity inversion with ridge regression: Geophysics, 40, 798-817.

Ivansson, S., 1986, Seismic borehole tomography--Theory and computational methods: Proceedings of IEEE, 74, 328-338.

Jackson, D.D., 1972, Interpretation of inaccurate, insufficient and inconsistent data: Geophys.J.Roy.Astr.Soc., 28, 97-109.

Jerry, A.J., 1977, The Shannon sampling theorem—its various extensions and applications: a tutorial review: Proceedings of IEEE, 65, 1565-1596.

Jol, H.M., and Smith, D.G., 1991, Ground penetrating radar of Northern Lacustring deltas: Canadian Journal of Earth Sciences, 28, 1939-

Jupp, D.L.B., and Vozoff, K., 1975, Stable iterative methods for inversion of geophysical data: Geophys.J.Roy.Astr. Soc., 42, 957-976.

Keller, G.V. and Frischknecht, F.C., 1966, Electrical Methods in Geophysical Prospecting. Pergamon Press Inc.

Kim,Y., Webster,J.G., and Tompkins, W.J., 1983, Electrical impedance imaging of the thorax: J. Microwave Power, 18, 245-257.

Koefoed, O., 1977, Comments on paper by Kumar and Das: Geophysical Prospecting, 25, 790.

Kumar, R., and Das, U.C., 1977, Transformation of dipole to Schlumberger sounding curves by means of digital linear filters: Geophysical Prospecting, 25, 780-789.

Li, Y. and Oldenburg, D. W., 1991, Aspects of charge accumulation in d.c. resistivity experiments, Geophysical Prospecting, 39, 803-826.

Madden, T.R., 1972, Transmission systems and network analogies to geophysical forward and inverse problem: Report No. 72-3, Dept. of Geology and Geophysics, M.I.T.

Marquardt, D.W., 1963, An algorithm for least squares estimation of nonlinear parameters: J. Soc. Indust. Appl. Math., 11, 431-441.

McClelland, J.L., and Rumelhart, D.E., 1988, Explorations in Parallel Distributed Processing: a handbook of models, programs and excises, MIT Press.

McMechan, G., 1983, Seismic tomography in boreholes: Geophysical Journal of the Royal Astronomical Society, 74, 601-612.

Menke, W., 1989, Geophysical Data Analysis: Discrete Inverse Theory, Rev. Ed., Academic Press, Inc.

Milner, A. and Milner, B., 1977, The Alum Pot System: Survey published by University of Leeds Speleological Association.

Murat, M. E., and Rudman, A. J., 1992, Automated first arrival picking: a neural network approach: Geophysical Prospecting, 41, 83-98.

Natterer, F., 1986, The Mathematics of Computerized Tomography, John Wiley & Sons Ltd.

Noel, M., and Walker, R., 1990, Development of an electrical resistivity tomography system for imaging archaeological structures: Archaeometry '90, Pernicka, E. and Wagner, G.A.(eds.), Birkhauser, Basel, 767-776.

Noel, M. 1990, Geological and topographic surveys of Lanchester Roman Fort, GeoQest Associates, Lanchester, Durham.

Noel, M. and Xu, B., 1991, Archaeological investigation by electrical resistivity tomography: a preliminary study Geophys.J.Int.,107,95-102.

Okabe, M., 1981, Boundary element method for the arbitrary inhomogenetities problem in electrical prospecting: Geophysical Prospecting, **29**, 39-59.

Olsson, O., Falk, L., Forslund, O., Lundmark, L., and Sandberg, E., 1992, Borehole radar applied to the characterization of hydraulically conductive fracture zones in crystalline rock: Geophysical Prospecting, **40**, 109-142.

Orellana, E, and Mooney, H.M., 1966, Master tables and curves for vertical electrical sounding over layered structures: Interciencia.

Oristaglio, M.L., and Worthington, M.H., 1980, Inversion of surface and borehole electromagnetic data for two-dimensional electrical conductivity models: Geophysical Prospecting, **28**, 633-657.

O'Sullivan, F. & Wahba, G., 1985, A cross validated Bayesian retrieval algorithm for nonlinear remote sensing experiments: Journal of Computational Physics, **59**, 441-455.

Patella, D.,1974, On the transformation of dipole to Schlumberger sounding curves: Geophysical Prospecting, 22, 315-329.

Pelton, W.H., Rijo, L. and Swift, C.M.,Jr., 1978, Inversion of two-dimensional resistivity and induced-polarization data: Geophysics, 43, 788-803.

Phillips, D.L., 1962, A technique for the numerical solution of certain integral equation of the first kind: Journal of Association for Computing Machinery, 9, 84-97.

Powell, H.M., Barber, D.C., and Freeston, I.L., 1987, Impedance imaging using linear electrode arrays: *Clinical Phys. Physiol. meas.*, 8, *Suppl. A*, 109-118.

Pratt, R.G., and Goulty, N.R., 1991, Combining wave-equation imaging with traveltime tomography to form high-resolution images from crosshole data: Geophysics, 56, 208-224.

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T., 1989, Numerical recipes, Cambridge University Press, New York.

Ramirez, A, Daily, W., LaBrecque, D., Owen, E., and Chesnut, D., 1993, Monitoring an underground steam injection process using electrical resistance tomography: Water Resources Research, 29, 73-87.

Rodi, W.L., 1976, A technique for improving the accuracy of finite element solutions for magnetotelluric data: Geophys. J. Roy. Astr. Soc., 44, 483-506.

Roy, A., and Apparao, A., 1971, Depth of investigation in direct current methods: Geophysics, 36, 943-959.

Rumelhart, D.E., and McClelland, J.L., 1986, Parallel Distributed Processing: exploration in the microstructure of cognition v1: Foundations, MIT Press.

Sasaki, Y., 1992, Resolution of resistivity tomography inferred from numerical simulation: Geophysical Prospecting, 40, 453-463.

Schomberg, H., 1980, Non-linear image reconstruction from projections of ultrasonic travel times and electrical current densities, *in* Proceedings of Mathematical Aspects of Computerised Tomography, *in* Lecture Notes in Medical Informatics, ed. by Reichertz Lindberg, Springer, Berlin, 272-290.

Shima, H., 1990, Two-dimensional automatic resistivity inversion technique using alpha centers: Geophysics, 55, 682-694.

Shima, H., 1992, 2-D and 3-D resistivity image reconstruction using crosshole data: Geophysics, **57**, 1270-1281.

Snyder, D.D., 1976, A method for modelling the resistivity and IP responses of two-dimensional bodies: Geophysics, **56**, 192-200.

Stefanescu, S., 1970, Nouvelles applications de la theorie des milleux alpha harmoniques a la prospection electrique en courant continu: Geophysical Prospecting, **18**, 787-799.

Stevenson, A.F., 1934, On the theoretical determination of earth resistance from surface potential measurements: Physics, **5**, 114-124.

Stoffa, P.L., and Sen, M.K., 1991, Nonlinear multiparameter optimization using genetic algorithms: Inversion of plane-wave seismograms: Geophysics, **56**, 1794-1810.

Sweeting, M. M., 1974, Karst geomorphology in north-west England, *in* The Limestones and Caves of North-West England, *edited by* Waltham, A. C., The British Cave Research Association.

Telford, W.M., Geldart, L.P., and Sheriff, R.E., 1990, Applied Geophysics, 2nd ed., Cambridge University Press.

Tikhonov, A.N., 1963, Solution of incorrectly formulated problems and the regularization method: Soviet Mathematics, **4**, 1035-1038

Todd, L. C., 1988, A shallow seismic refraction survey across the Hett Dyke, BSc. dissertation, Department of Geological Sciences, University of Durham.

Tong, L.T. and Yang, C.H., 1990, Incorporation of topography into two-dimensional resistivity inversion: Geophysics, **55**, 354-361

Tripp, A.C., Hohmann, G.W. and Swift, C.M., Jr., 1984, Two-dimensional resistivity inversion: Geophysics, **49**, 1708-1717.

Van, G..P., Park, S.K., and Hamilton, P., 1991, Monitoring leaks from storage ponds using resistivity methods: Geophysics, **56**, 1267-1270.

Van Nostrand, R.G., and Cook, K.L., 1966, Interpretation of resistivity data: Geological Survey Professional Paper **499**, U.S. Government Printing Office.

Van Overmeeren, R.A. and Ritsema, I.L., 1988, Continuous vertical sounding: First Break, **6**, 313-324.

Vozoff, K., 1960, Numerical resistivity interpretation: general inhomogeneity: Geophysics, **25**, 1184-1194

Wexler, A., Fry, B. and Neuman, M.R.,1985, Impedance-computed tomography algorithm and system: Applied Optics, 24, 3985-3992.

Wexler, A., 1988, Electrical impedance imaging in two and three dimensions: Clinical Physics and Physiological Measurement, 9, *Supplement A*, 29-33.

Wu, R.S., and Toksoz, M.N., 1987, Diffraction tomography and multisource holography applied to seismic imaging: Geophysics, 52, 11-25.

Xu, B and Noel, M, 1993, On the completeness of data sets with multielectrode systems for electrical resistivity survey, Geophysical Prospecting, 41, 791-801.

Young, C.T., and Droege, D.R., 1986, Archaeological application of resistivity and magnetic methods at Fort Wilkins State Park, Michigan: Geophysics, 51, 568-575.

Yorkey, T.J., Webster, J.G. and Tompkins, I.L.,1987, Comparing reconstruction algorithms for electrical impedance tomography: IEEE Trans. Biomed. Eng., BME-34, 843-852.

Zohdy, A.A.R., 1989, A new method for the automatic interpretation of Schlumberger and Wenner sounding curves: Geophysics, 54, 245-253.

Zohdy, A.A.R., and Bisdorf, R.J., 1990, Schlumberger soundings near Medicine Lake, California: Geophysics, 55, 956-964.