# 1

# Introduction

## 1.1 The digital revolution

Recording the output from geophysical instruments has undergone four stages of development during the past century: mechanical, optical, analogue magnetic, and digital. Take the seismometer as a typical example. The principle of the basic sensor remains the same: the swing of a test mass in response to motion of its fixed pivot is monitored and converted to an estimate of the velocity of the pivot.

Inertia and damping determine the response of the sensor to different frequencies of ground motion; different mechanical devices measured different frequency ranges. Ocean waves generate a great deal of noise in the range 0.1–0.5 Hz, the *microseismic noise* band, and it became normal practice to install a short-period instrument to record frequencies above 0.5 Hz and a long-period instrument to record frequencies below 0.1 Hz.

Early mechanical systems used levers to amplify the motion of the mass to drive a pen. The classic short-period, high-gain design used an inverted pendulum to measure the horizontal component of motion. A large mass was required simply to overcome friction in the pen and lever system.

An optical lever reduces the friction dramatically. A light beam is directed onto a mirror, which is twisted by the response of the sensor. The reflected light beam shines onto photographic film. The sensor response deflects the light beam and the motion is recorded on film. The amplification is determined by the distance between the mirror and film. Optical recording is also compact: the film may be enlarged later to a more readable size. Optical recording was in common use in the 1960s and 1970s.

Electromechanical devices allow motion of the mass to be converted to a voltage, which is easy to transmit, amplify and record. Electromagnetic feedback seismometers use a null method, in which an electromagnet maintains the mass in

a constant position. The voltage required to drive the electromagnet is monitored and forms the output of the sensor.

This voltage can be recorded on a simple tape recorder in analogue form. Analogue magnetic records could be converted to paper records simply by playing them through a chart recorder. There is a good analogy with tape recording sound, since seismic waves are physically very similar to low-frequency sound waves. The concept of fidelity of recording carries straight across to seismic recording. A convenient way to search an analogue tape for seismic sources is to simply play it back fast, thus increasing the frequency into the audio range, and listen for bangs.

The digital revolution started in seismology in about 1975, notably when the World Wide Standardized Seismograph Network (WWSSN) was replaced by Seismological Research Observatories (SROs). These were very expensive installations requiring a computer in a building on site. The voltage is sampled in time and converted to a number for input to the computer. The tape recording systems were not able to record the incoming data all the time so the instrument was triggered and a short record retained for each event. Two channels (sometimes three) were output: a finely sampled short-period record for the high-frequency arrivals and a coarsely sampled channel (usually one sample each second) for the longer-period surface waves. Limitations of the recording system meant that SROs did not herald the great revolution in seismology some had hoped for: that had to wait for better mass storage devices.

The great advantage of digital recording is that it allows replotting and processing of the data after recording. If a feature is too small to be seen on the plot, you simply plot it on a larger scale. More sophisticated methods of processing allow us to remove all the energy in the microseismic noise band, obviating the need for separate short- and long-period instruments. It is even possible to simulate an older seismometer simply by processing, provided the sensor records all the information that would have been captured by the simulated instrument. This is sometimes useful when comparing seismograms from different instruments used to record similar earthquakes. Current practice is therefore to record as much of the signal as possible and process it after recording. This has one drawback: storage of an enormous volume of data.

The storage problem was essentially solved in about 1990 by the advent of cheap hard disks and tapes with capacities of several gigabytes. Portable broadband seismometers were developed at about the same time, creating a revolution in digital seismology: prior to 1990, high-quality digital data were available only from a few permanent, manned observatories. After 1990 it was possible to deploy arrays of instruments in temporary sites to study specific problems, with only infrequent visits to change disks or tapes.

## 1.2 Digital recording

The *sensor* is the electromechanical device that converts physical input (e.g. ground motion) into voltage; the *recorder* converts the voltage into numbers and stores them. The ideal sensor would produce a voltage that is proportional to the ground motion but such a device is impossible to make (the instrument response would have to be constant for all frequencies, which requires the instrument to respond instantaneously to any input, see Appendix 2). The next best thing is a *linear response*, in which the output is a convolution of the ground motion with the *transfer function* of the instrument.

Let the voltage output be $a(t)$. The recorder *samples* this function regularly in time, at a *sampling interval* $\Delta t$, and creates a sequence of numbers:

$$\{a_k = a(k\Delta t); \quad k = 0, 1, 2, \ldots, N\}. \tag{1.1}$$

The recorder stores the number as a string of bits in the same way as any computer.

Three quantities describe the limitations of the sensor: the *sensitivity* is the smallest signal that produces non-zero output; the *resolution* is the smallest change in the signal that produces non-zero output; and the *linearity* determines the extent to which the signal can be recovered from the output. For example, the ground motion may be so large that the signal exceeds the maximum level; the record is said to be 'clipped'. The recorded motion is not linearly related to the actual ground motion, which is lost.

The same three quantities can be defined for the recorder. A pen recorder's linearity is between the voltage and movement of the pen, which depends on the electronic circuits and mechanical linkages; its resolution and accuracy are limited by the thickness of the line the pen draws. For a digital recorder, linearity requires faithful conversion of the analogue voltage to a digital count, while resolution is set by the voltage corresponding to one digital count.

The recorder suffers two further limitations: the *dynamic range*, $g$, the ratio of maximum possible to minimum possible recorded signal, usually expressed in deciBels: $20\log_{10}(g)$ dB; and the maximum frequency that can be recorded. For a pen recorder the dynamic range is set by the height of the paper, while the maximum frequency is set by the drum speed and width of the pen. For a digital recorder the dynamic range is set by the number of bits available to store each member of the time sequence, while the maximum frequency is set by the sampling interval $\Delta t$ or *sampling frequency* $\nu_s = 1/\Delta t$ (we shall see later that the maximum meaningful frequency is in fact only half the sampling frequency).

A recorder employed for some of the examples in this book used 16 bits to record the signal as an integer, making the maximum number it can record $2^{16} - 1$ and the minimum is 1. The dynamic range is therefore $20\log_{10}(65\,535) = 96$ dB.

Another popular recorder uses 16 bits but in a slightly more sophisticated way. One bit is used for the sign and 15 are used for the integer if the signal is in the range $\pm 2^{15} - 1$; outside this range it steals one bit from the integer, records an integer in the range $\pm(2^{14} - 1)$, and multiplies by 20, giving a complete range $\pm 327\,680$, or 116 dB. The stolen 'gain' bit is used to indicate the change in gain by a factor of 20; the increase in dynamic range has been achieved at the expense of accuracy, but this is usually unimportant because it occurs only when the signal is large. A typical 'seismic word' for recorders in exploration geophysics consists of 16 bits for the integer ('mantissa'), one bit for the sign, and four bits for successive levels of gain. The technology is changing rapidly and word lengths are increasing; most recorders now being sold have 24 bits.

Storage capacity sets a limit to the dynamic range and sampling frequency. A larger dynamic range requires a larger number and more bits to be stored; a higher sampling frequency requires more numbers per second and therefore more numbers to be stored. The following calculation gives an idea of the logistical considerations involved in running an array of seismometers in the field. In 1990, Leeds University Earth Sciences Department deployed nine three-component broadband instruments in the Tararua Mountain region of North Island, New Zealand, to study body waves travelling through the subducted Pacific Plate. The waves were known to be high frequency, demanding a 50 Hz sampling rate. Array processing (see Section 11.1) is only possible if the signal is coherent across the array, requiring an interval of 10 km between stations. The Reftek recorders used a 16 bit word and had 360 Mb disks that were downloaded onto tape when full.

A single field seismologist was available for routine servicing of the array, which meant he or she had to drive around all nine instruments at regular intervals to download the disks before they filled. How often would they need to be downloaded? Each recorder was storing 16 bits for each of three components 50 times each second, or 2400 bit $s^{-1}$. Allowing 20% overhead for things like the time channel and state-of-health messages gives 2880 bit $s^{-1}$. One byte is eight bits and, dividing 2880 into the storage capacity of the disk, a 360 Mb disk can store $8 \times (3.6\ 10^8)$ bits, which would last about $10^6$ s, or 11.5 days. It would be prudent to visit each instrument at least every 10 days, which is possible for an array stretching up to 150 km from base over good roads. A later deployment used Mars recorders with optical disks that could be changed by a local unskilled operator, requiring only infrequent visits to collect data and replenish the stock of blank disks. In this case, the limiting factor was the time required to back up the optical disks onto tape.

It is well worth making the effort to capture all the information available in a broadband seismogram because the information content is so extraordinarily rich. The seismogram shown in Figures 1.1 and 1.2 provides a good example. *P*, *S*, and

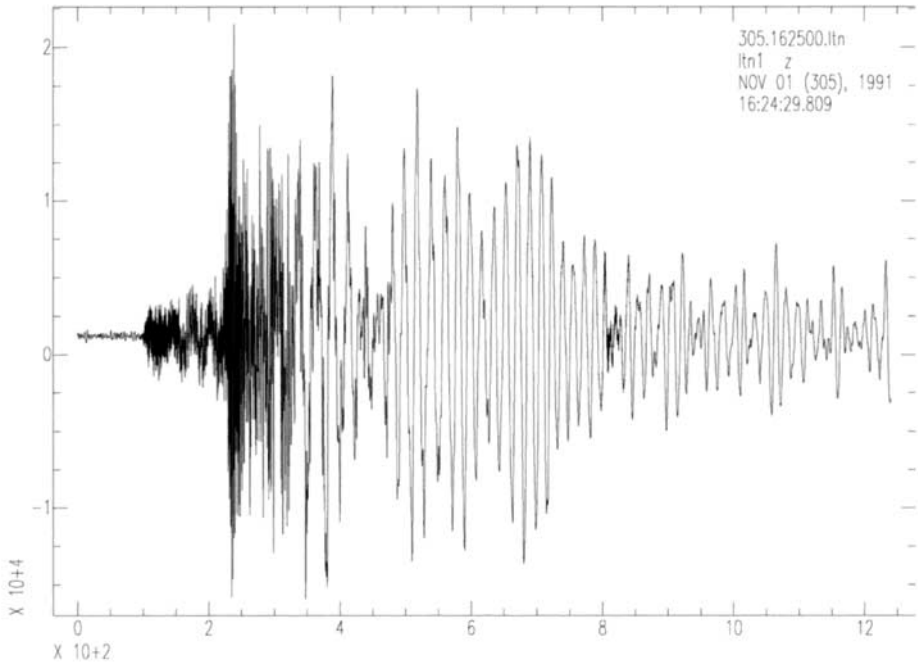


Fig. 1.1. Seismogram of an event in the Kermadec Islands on 1 November 1991 recorded in North Island, New Zealand. Note the *P*, *S*, and longer-period surface waves. The dynamic range meant maximum signals of ±16 383 digital counts, and the scale shows that this was nearly exceeded. The time scale is in minutes, and the surface waves have a dominant period of about 20 s.

surface waves are clearly seen. The surface waves are almost clipped (Figure 1.1), yet the onset of the *P* wave has an amplitude of just one digital count (Figure 1.2). The frequency of the surface waves is about 20 s, yet frequencies of 10 Hz and above may be seen in the body waves. The full dynamic range and frequency bandwidth were therefore needed to record all the ground motion.

## 1.3 Processing

Suppose now that we have collected some important data. What are we going to do with them? Every geophysicist should know that the basic raw data, plus that field note-book, constitute the maximum factual information he or she will ever have. This is what we learn on field courses. Data processing is about extracting a few nuggets from this dataset; it involves changing the original numbers, which *always* means losing information. So we always keep the original data.

*Processing* involves operating on data in order to isolate a *signal*, the message we are interested in, and to separate it from 'noise' , which nobody is interested in,

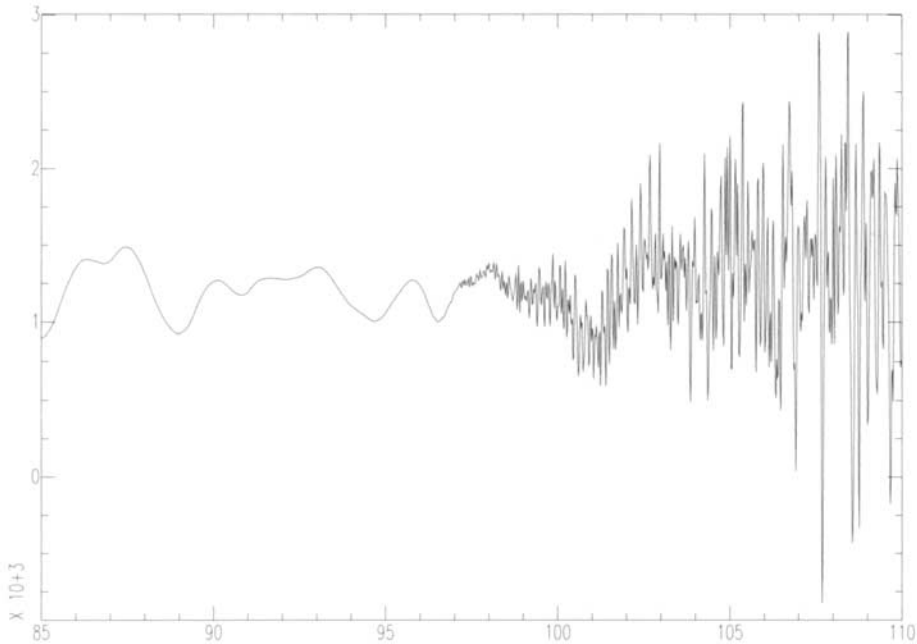


Fig. 1.2. Expanded plot of the first part of the seismogram in Figure 1.1. Note the very small amplitude of the *P* arrival, which rides on microseisms of longer period. The arrival is easy to see because its frequency is different from that of the microseisms.

and unwanted signals, which do not interest us at this particular time. For example, we may wish to examine the *P* wave on a seismogram but not the surface waves, or we may wish to remove the Earth's main magnetic field from a magnetic measurement in order to determine the magnetisation of local rocks because it will help us understand the regional geology. On another day we may wish to remove the local magnetic anomalies in order to determine the main magnetic field because we want to understand what is going on in the Earth's core.

Often we do not have a firm idea of what we ought to find, or exactly where we should find it. Under these circumstances it is desirable to keep the methods as flexible as possible, and a very important part of modern processing is the interaction between the interpreter and the computer. The graphical display is a vital aid in interpretation, and often the first thing we do is plot the data in some way. For example, we may process a seismogram to enhance a particular arrival by filtering (Section 4.1). We know the arrival time roughly, but not its exact time. In fact, our main aim is to measure the arrival time as precisely as possible; in practice, the arrival will be visible on some unprocessed seismograms, on others it will be visible only after processing, while on yet more it may never be visible at all.

The first half of this book deals with the analysis of time series and sequences, the commonest technique for processing data interactively without strict prior prejudice about the detailed cause of the signals or the noise. This book adopts Claerbout's 1992 usage and restricts the word *processing* to distinguish this activity from the more formal process of *inversion* explained below. Processing is interactive and flexible; we are looking for something interesting in the data without being too sure of what might be there.

Suppose again that we wish to separate $P$ waves from surface waves. We plot the seismogram and find the $P$ waves arriving earlier because they travel faster than surface waves, so we just cut out the later part of the seismogram. This does not work when the $P$ waves have been delayed (by reflecting from some distant interface, for example) and arrive at the same time as the surface waves. The surface waves are much bigger than body waves and the $P$ wave is probably completely lost in the raw seismogram.

$P$ waves have higher frequencies than surface waves (period 1 s against 20 s for earthquake seismograms, or 0.1 s against 1 s for a typical seismic exploration experiment). This gives us an alternative means of separation. The Fourier transform, in its various forms (Appendix 2), decomposes a time series into its component frequencies. We calculate and plot the transform, identify the low-frequency contributions from the surface waves, zero them and transform back to leave the higher-frequency waves. This process is called *filtering* (Section 4.1). In these examples we need only look at the seismogram or its Fourier transform to see two separate signals; having identified them visually it is an easy matter to separate them. This is processing.

## 1.4 Inversion

The processed data are rarely the final end product: some further interpretation or calculation is needed. Usually we will need to convert the processed data into other quantities more closely related to the physical properties of the target. We might want to measure the arrival time of a $P$ wave to determine the depth of a reflector, then interpret that reflector in the context of a hydrocarbon reservoir. We might measure spatial variations in the Earth's gravity, but we really want to find the density anomalies that cause those gravity anomalies, then understand the hidden geological structure that caused the gravity variations.

Inversion is a way of transforming the data into more easily interpretable physical quantities: in the example above we want to invert the gravity variations for density. Unlike processing, inversion is a formal, rigid procedure. We have already decided what is causing the gravity variations, and probably have an idea of the depth, extent, and even the shape, of the gravity anomalies. We have a

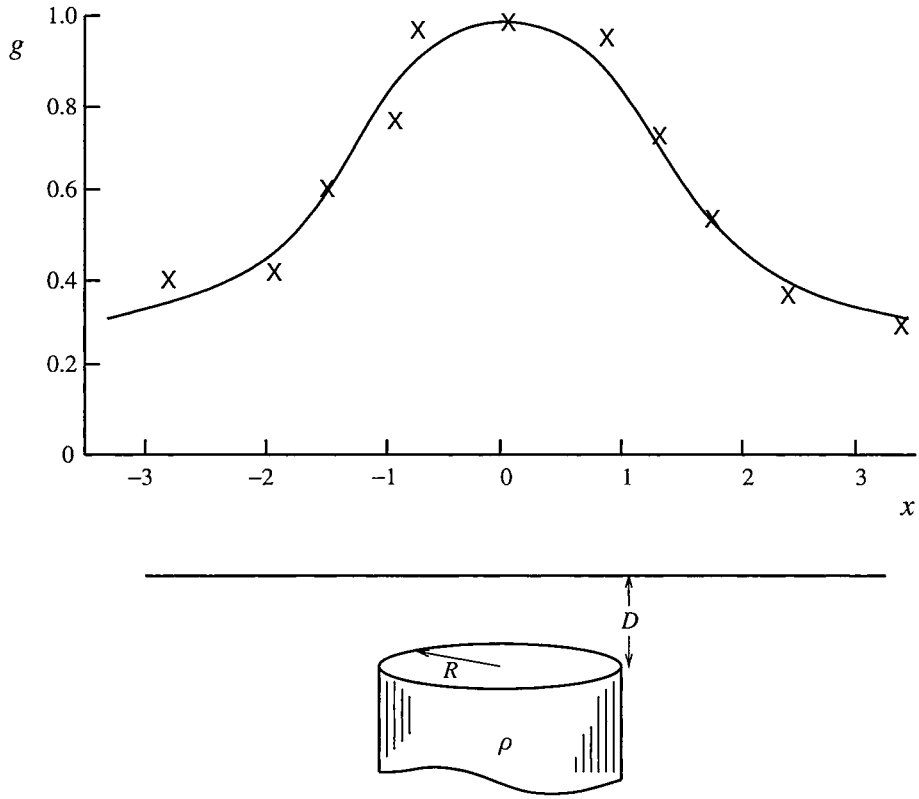


Fig. 1.3. Gravity anomaly caused by a dense, buried cylinder.

mathematical *model* that we wish to test and refine using this new data. The inversion excludes any radically different interpretation. For example, if we invert the seismic arrival times for the depth of a pair of horizontal reflectors we would never discover that they really come from a single, dipping reflector.

Consider the gravity traverse illustrated in Figure1.3, which we intend to invert for density. The traditional method, before widespread use of computers, was to compare the shape of the anomaly with theoretical curves computed from a range of plausible density models simple enough for the gravity to be calculated analytically. This is called *forward modelling*. It involves using the laws of physics to predict the observations from a model. Computers can calculate gravity signals from very complicated density models. They can also be used to search a large set of models to find those that fit the data. A strategy or algorithm is needed to direct the search. Ideally we should start with the most plausible models and refine our ideas of plausibility as the search proceeds. These techniques are sometimes called *Monte Carlo* methods.

There are two separate problems, *existence* and *uniqueness*. The first requirement is to find one model, any model, that fits the data. If none is found, the data are incompatible with the model. This is rare. Either there is something wrong with the model or we have overestimated the accuracy of the data. Having found one model we search for others. If other models are found the solution is nonunique; we must take account of it in any further interpretation. This always happens. The nonuniqueness is described by the subset of models that fits the data.

It is more efficient, when possible, to solve directly for the model from the data. This is not just an exercise in solving the equations relating the data to the model for a model solution, we must also characterise the nonuniqueness by finding the complete set of compatible solutions and placing probabilities on the correctness of each individual model.

Like most formal procedures, the mathematics of inverse theory is very attractive: it is easy to become seduced into thinking that the process is more important than it really is. Throughout this book I try to emphasise the importance of the measured data and the desired end goal: these are much more important than the theory, which is just the vehicle that allows you to proceed from the data to a meaningful interpretation. It is like a car that lets you take luggage to a destination: you must pack the right luggage and get to the right place (and, incidentally, the place should be interesting and worthy of a visit!). You need to know enough about the car to drive it safely and get there in reasonable time, but detailed knowledge of its workings can be left to the mechanic.

There are two sources of error that contribute to the final solution: one arising because the original measurements contain errors, and one arising because the measurements failed to sample some part of the model. Suppose we can prove for a particular problem that perfect, error-free data would invert to a single model: then in the real case of imperfect data we need only worry about measurement errors mapping into the model. In this book I call this *parameter estimation* to distinguish it from true inversion.

Modern inversion deals with the more general case when perfect data fail to provide a unique model. True inversion is often confused with parameter estimation in the literature. The distinction is vital in geophysics, because the largest source of error in the model usually comes from failure to obtain enough of the right sort of data, rather than sufficiently accurate data.

The distinction between inversion and parameter estimation should not become blurred. It is tempting to restate the inverse problem we should be solving by restricting the model until the available data are capable of determining a unique solution in the absence of errors, but this is philosophically wrong. The model is predetermined by our knowledge of the problem at the start of the experiment, not by what we are about to do.

Some quotes illustrate the diversity of meaning attached to the term 'inversion'. J. Claerbout, in his book *Earth Soundings Analysis*: *Processing Versus Inversion*, calls it matrix inversion. This is true only if the forward problem is posed as matrix multiplication. In exploration geophysics one sometimes hears inversion called *deconvolution* (Section 4.3), an even more restrictive definition. Deconvolution is treated as an inverse problem in Section 10.2. 'Given the solution of a differential equation, find its coefficients' was the definition used by Gel'fand and Levitan (1955). More specifically 'given the eigenvalues of a differential operator, find the operator', a problem that finds application in the use of normal mode frequencies in determining average Earth structure. A more poetic statement of the same problem is 'can we hear the shape of a drum?', which has received much attention from pure mathematicians. None of these definitions covers the full extent of inverse problems currently being studied in geophysics.

## 1.5 About this book

The book is divided into three parts: processing, inversion, and applications that combine techniques introduced in both of the first parts. Emphasis is placed on discrete, rather than continuous, formulations, and on deterministic, rather than random, signals. This departure from most texts on signal processing and inversion demands some preliminary explanation.

The digital revolution has made it easy to justify treating data as a discrete sequence of numbers. Analogue instruments that produce continuous output in the form of a paper chart record might have the appearance of continuity, but they never had perfect time resolution and are equivalent to discrete recordings with interpolation imposed by the nature of the instrument itself – an interpolation that is all too often beyond the control and sometimes even the knowledge of the operator. The impression of continuity is an illusion. Part I therefore uses the discrete form of the Fourier transform, which does not require prior knowledge of the integral form.

It is much harder to justify discretising the model in an inversion. In this book I try to incorporate the philosophy of continuous inversion within the discrete formulation and clarify the distinction between true inversion, in which we can only ever discover a part of the true solution, and parameter estimation, where the model is adequately described by a few specially chosen numbers. Chapter 9 contains a brief overview of continuous inversion.

In the example of Figure 1.3 the model is defined mathematically by a set of parameters: $R$, the radius of the cylinder, $D$, the depth of burial, and $\rho$, the density or density difference with the surroundings. The corresponding gravity field is computed using Newton's inverse square law for each mass element $\mathrm{d}\rho$ and