1

Introduction

Early in the 1970s, seismologists realized that a full three-dimensional (3D) interpretation was needed to satisfy variations in the observed seismic travel times. The starting point of modern seismic tomography (from $\sigma\epsilon\iota\sigma\mu\acute{o}\varsigma$ = quake and $\tau\acute{o}\mu o\varsigma$ = slice) is probably the 1974 AGU presentation by MIT's Keiti Aki (Aki et al. [3]) in which arrival times of P-waves were for the first time formally interpreted in terms of an 'image' as opposed to a simple one-dimensional graph of seismic velocity versus depth. That Aki's co-authors came from NORSAR – the Norwegian array to monitor nuclear test ban treaties – was caused by a quirk of history: Aki had originally planned a sabbatical in Chile, but when a military coup d'état brought the Allende government down in 1973 he changed plans and accepted an invitation from his former MIT student Eystein Husebye for a short sabbatical at NORSAR, which was equipped with a state-of-the-art digital seismic network and computing facilities. Even so, in the twenty-first century it is easy to underestimate the difficulties faced by early tomographers, who had to invert matrices of size 256×256 using a CPU with 512 Kbyte of memory. The collaboration between Aki, Husebye and Christoffersson was continued in 1975 at Lincoln Labs in Massachussets (Aki et al. [2, 4]).

The name that was later given to the new imaging technique is more than an accidental reference to medical tomography, because the earliest radiologic tomograms also attempted to get a scan of the body that focuses on a plane of interest, albeit using X-rays rather than seismic waves. This was obtained by moving the X-ray source and the photographic plate in opposite directions, such that objects outside the target plane would be blurred, but those in the plane would always illuminate the photographic plate at the same spot, a technique known as 'backprojection'. This enabled radiologists to reconstruct the tissue or bone density on that plane from observed X-ray intensities. In the 1970s the photographic plate was replaced by sensors, which feed a computer that reconstructs the density from discrete observations (computerized tomography). In seismic travel time tomography, we attempt to reconstruct the local seismic velocity in the Earth from arrival times of body waves observed with seismic sensors. Though very similar to the case of computerized medical tomography, the challenges are far greater. Added complications are that seismic waves do not follow paths that are straight, and that we have little or no control over the experiment – we have to use the earthquakes that are dealt to us, and seismic networks are virtually confined to the continents and a few islands in the oceans.

1.1 Early efforts at seismic tomography

The field of seismic tomography blossomed quickly, being an outgrowth of several developments that matured in the 1970s: Backus and Gilbert [14, 12, 13], Jackson [144] and Wiggins [395] developed geophysical inversion theory and provided the necessary tools to deal with the inevitable underdetermined nature of geophysical inverse problems. Dense digital seismic networks, the largest of which were the LASA and NORSAR arrays in Montana and Norway, had been installed to monitor the testing of nuclear weapons, and supplemented the analogue World Wide Standardized Seismograph Network (WWSSN) which served the same purpose. The monitoring arrays also provided densely spaced data to global seismologists. Sengupta and Toksöz [305] read arrival times from WWSSN seismograms and adapted Aki's method in a first attempt at global tomography. But the International Seismological Centre (ISC, then in Edinburgh, UK) had been assembling many more arrival time data from thousands of station operators around the world since 1964, and, even though some of the early ISC data were archived on 7-track tapes that had become difficult to read, the painstakingly careful efforts by Dziewonski [93] to recover them led to his first try at global tomography and would soon give Harvard a dominant position among the institutions doing global imaging studies.

In the meantime, the MIT efforts resulted in a number of local tomography experiments using Aki's technique: Ellsworth and Koyagani [97] imaged the structure beneath the Kilauea volcano on Hawaii in 1977, and in the same year Mitchell et al. [208] published a tomographic study of the New Madrid seismic zone. This was soon followed by larger experiments on a regional scale: Menke [206], Romanowicz [286, 287] and Taylor and Toksöz [354] pioneered continental tomography with early studies of the upper mantle under the Himalayas, and under the North American and European continents. With time, such local and regional studies became more precise: in 1984, Thurber [360] imaged velocities at Kilauea that were low enough to be interpreted as the underlying magma complex.

1.2 Ocean acoustic tomography

In 1981 a first ocean acoustic tomography experiment was performed near Bermuda, and tomography is now an accepted tool to study sound speed (a proxy for temperature) and flow in the upper layers of the oceans. Because ocean acoustic tomography has already been extensively documented in the monograph by Munk, Worcester and Wunsch [223], this book will not often venture into the domain of oceanography. Though the experimental methods used by oceanographers differ greatly from those used by seismologists or astronomers, the theory behind tomographic interpretation is very much the same. Though high frequencies used in ocean acoustics allow for a broader validity of ray theory, the Fresnel zones in ocean acoustic tomography are often large, and a finite-frequency approach is called for (Skarsoulis and Cornuelle [321]). This has the added effect of reducing the effects of wave chaos in range-dependent ocean models.

It is noteworthy that Simons et al. [317] succeeded in recording a teleseismic P-wave using a hydrophone mounted on a freely floating diver submerged at 700 m depth – a development that may eventually bring ocean acoustics together with global seismic tomography and open up the Southern oceans for dense seismometry and acoustic monitoring (see Chapter 17).

1.3 Global tomography

In solid Earth sciences, seismic tomography soon extended its reach to the inversion of the Earth's eigenfrequencies and dispersion properties of long-period surface waves, using perturbation theory on normal modes (see Chapter 9). In 1982, Masters et al. [199] at UCSD discovered the strong degree-2 pattern in the Earth's heterogeneity. This anomaly was later correlated by Cazenave et al. [43] to the surface topography of the Earth, topography being the dynamic response of the surface to mantle convection processes, confirming an earlier prediction by Hager et al. [128]. It became evident that the 3D seismic structure of the Earth was strongly linked to the deep dynamics of the planet, and that it was important to resolve that structure.

By 1984, Dziewonski's work on ISC delay times resulted in the first reliable global modal of long-wavelength P-wave velocity variations [90], and Woodhouse and Dziewonski [400] began to use long-period data to image the upper mantle using the lowest order spherical harmonics. This led to an immensely fruitful era in which the large scale heterogeneous structure of the Earth was mapped in increasingly finer detail.

In early global tomography it was customary to parametrize the velocity model in terms of a small number of spherical harmonics. This was not just a preference of the early pioneers; a parametrization in terms of a small number of coefficients was also mandated by the limited computer capacity available at the time, and spherical harmonics provide an excellent means of doing so while forcing velocity anomalies to vary smoothly. By parametrizing the Earth in terms of just a few hundred coefficients the least squares matrix – which is $N \times N$ in size for N model parameters – could be diagonalized and inverted using singular value decomposition or other regularization techniques. However, the spherical harmonic expansion made little sense for the investigation of more localized structures, and for this purpose existing iterative matrix solvers were soon introduced into seismic tomography by Clayton and Comer [61], Nolet [233, 234], and Neumann-Denzau and Behrens [229]. These iterative solvers are very flexible in the type of parametrization and impose virtually no limit to the size of the set of model coefficients or data. With iterative solvers, the lack of a formal inverse hampers formal estimation of error or resolution, but sensitivity tests try to make up for that shortcoming (Spakman and Nolet [338]). In a separate development, Jackson [145], Tarantola and Valette [353], and Tarantola and Nercessian [352] circumvented the problem of model parametrization by imposing a-priori correlation properties on the model and introduced the tools of Bayesian inference – which allows for the assignment of *probabilities* to scientific truth – into the tomographic inversion.

Waveform inversions remained restricted to modelling the low frequency wavefield, though, independent from global tomography, 'waveform tomography' was formulated for exploration purposes by Tarantola [350], who used first order scattering theory, finite difference modelling, and an adjoint approach to attack the strongly nonlinear inverse problem for high frequency waveforms. This approach – which is related to seismic migration as used in the oil industry – stands at the basis of some of the more recent developments in seismic tomography. For example, in regional waveform tomography, Bostock and coworkers [31] succeeded in imaging the subducting lithosphere under the Pacific Northwest using converted wave energy to map discontinuities.

1.4 Some major discoveries

I refer the reader to the excellent reviews by Romanowicz [288, 291] for the later history of global tomography but note several of the major discoveries: In 1988, Spakman et al. [340] showed that the Hellenic subduction zone extends far deeper than the cut-off depth for seismicity, thereby opening up a vista that up to then seemed a mere hypothesis: could slabs sink deeper than the depth of the deepest located earthquakes near 700 km? Such penetration of slabs into the lower mantle had earlier been inferred by Creager and Jordan [69] from the statistics of travel time anomalies without the benefit of tomography. I still remember the day that

Suzan van der Lee, then an undergraduate at the University of Utrecht working on her senior thesis with Wim Spakman, produced a stunning image of the Aegean slab subducting into the lower mantle. This was in 1989 and her thesis was met with some scepticism – including from me – but by the time it found its way into the refereed literature [339] there were other, very strong indications from regions with deep seismicity for lower mantle subduction. A first image in which both the Farallon plate and the Tonga subduction zone enter the lower mantle came from global tomography by Inoue et al. [141]. In this early study, Inoue also imaged the major low velocity regions near the core–mantle boundary that we now call 'superplumes'. But it was van der Hilst et al. [372, 373], who used P and surface-reflected P-waves to construct the first tomographic images of the subducting slabs under the Northwest Pacific that contained sufficient detail to convince the geophysical community that several slabs indeed penetrate into the lower mantle. When Grand's S-wave models [123, 125] showed very similar cross-sections of slabs sinking into the lower mantle, any remaining doubts about the reliability of seismic tomography quickly melted away and the hypothesis of two-layered convection in the Earth was dealt a serious blow.

The first global model of topography on the 660 km discontinuity was constructed by Shearer and Masters [310] in 1992. The first global attenuation model was derived by Romanowicz [289], but to date little agreement exists among different attenuation models for the Earth, and this remains one of the major challenges of tomographic research. In 1997, Su and Dziewonski [345] discovered a negative correlation between anomalies in bulk and shear modulus in the lowermost mantle where the Pacific and African superplumes are located. Ishii and Tromp [142, 143] determined that the Pacific superplume has a higher density than the surrounding mantle, indicating that these are chemically distinct entities, though this finding is somewhat controversial.

The hypothesized – much thinner – thermal mantle plumes remained elusive in tomographic images, apart from a few contested interpretations. In this context, the debate over the origin of the Iceland plume is very illustrative. Following two field campaigns with portable seismometers in Iceland, seismic tomography by Wolfe et al. [396] and Allen et al. [7] left little doubt that Iceland caps a strong upper mantle plume. At the same time, Bijwaard and Spakman [22] presented tomographic evidence for a lower mantle extension of this plume and argued that the plume originates from a broad upwelling at the base of the mantle. This interpretation was strongly contested by Foulger et al. [104]. In 2004 Montelli et al. [215] used the enhanced resolving power of finite-frequency tomography to image more than a dozen plume-like anomalies in the lower mantle in detail, but argued that the Iceland plume disappears at mid-mantle depth, an interpretation that was supported by ray-theoretical tomographic imaging by Zhao [414]. More

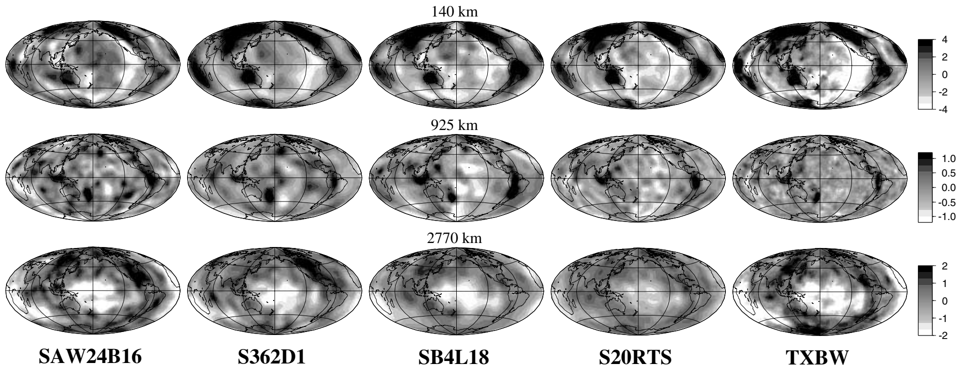

Fig. 1.1. A comparison of perturbations in S-wave velocity in five recent global tomographic models permits judgement of the degree of agreement at three depth levels (top to bottom): upper, mid- and lower mantle. The greyscale reflects perturbations in per cent. Figure courtesy Barbara Romanowicz, reprinted, with permission, from [291] © 2003 by Annual Reviews www.annualreviews.org.

recent finite-frequency tomography using S-waves by Montelli et al. [214] seems to re-open the question of the depth of the Iceland plume since the S-wave images suggest a weak connection between low velocity anomalies in the upper- and lowermost mantle.

The long-wavelength structure of the Earth is very similar across recent tomographic models, as can be judged from Figure 1.1, even though there is disagreement on the amplitudes of even the largest structures. But for smaller length scales the correlation between the models becomes less. Disagreements such as for Iceland abound for many weakly resolved features. Yet, barring undiscovered errors in the analyses, we must assume that all tomographic models have an acceptable fit to the data. The lack of agreement comes only partly from the fact that different investigators use different data sets, because in fact all global data sets rely heavily on the network of broadband, digital seismometers that has been growing over the past 20 years and there is a significant amount of overlap between the information used to construct each one of these models. A major factor that creates differences between tomographic solutions is that the inverse problem is underdetermined and requires regularization, which leaves the choice of 'optimum' model among all models that satisfy the data to the investigator. Apart from that (or perhaps because this aspect of arbitrariness has not forced us to face the issue) there is at present not much of a tradition among tomographers to use the same statistical criteria for goodness of fit. We console our scientific conscience with the argument that 'resolution tests' will convey how much wriggle room there is for other solutions. But such resolution tests are almost always limited in scope, depend on the misfit criterion used, are subject to the same theoretical approximations used in the inversion and usually give no quantitative information on the covariance of the solution.

1.5 Helioseismology

Helioseismology began in 1961, when Leighton et al. [178] reported the discovery of 'a striking repetitive correlation with a period $T = 296 \pm 3$ sec' in Doppler shifts of solar spectral lines. An explanation of such 'five minute oscillations' as standing waves trapped in the solar interior was provided in 1970 by Ulrich [370], who also predicted that discrete spectral lines in a frequency–wavenumber diagram would be observed if the spectral resolution could be improved. Deubner [85] was the first to observe such dispersion lines, using observations lasting several hours from the Anacapri observatory on Capri, Italy.

The results from helioseismology have been spectacular. Since 1995, the Global Oscillations Network Group (GONG) makes it possible to obtain 24-hour coverage of the solar observations through a network of six telescopes around the globe, leading to an important increase in the precision of the Doppler measurements. Even better, after December 1995 the Michelson Doppler Imager (MDI) began sending back data from the Solar and Heliospheric Observatory (SOHO), circling the Sun in a nearly stationary position with respect to Earth, around the L_1 Lagrangian point. These observations are crucial in the testing of models of energy generation in the Sun. Early determinations of the sound speed as a function of depth in the Sun seemed to indicate broad agreement with the standard model (Bahcall et al. [16]) – and thus provided no relief for the discrepancy in neutrino flux exhibited at the time by that model: the observed flux was about a third of that predicted (the discrepancy was later resolved when particle physicists realized that some of the neutrinos change their nature as they flow out of the Sun and were unobservable with the available detectors).

Recent reviews on helioseismology are given by Christensen-Dalsgaard [58] and Gizon and Birch [118]. The first 'sunquake' – oscillations in response to a flare of gas above the solar surface – was observed on July 9, 1996 by Kosovichev and Zharkova [167], but normally the Sun's interior is imaged using random waves. Helioseismic observations have also led to the discovery of 'weather patterns' in the upper layers of the convective zone, which are due to convective motions and flows associated with concentrations of magnetic activity (e.g. Kosovichev et al. [166]).

Originally, helioseismology relied on the inversion of dispersion relationships of surface waves. But in 1993 time–distance helioseismology was developed by Duvall et al. [89] using cross-correlation techniques to retrieve the Green's function for a fixed distance on the solar surface (Claerbout's conjecture [278]). This helped greatly the development of true 3D imaging of the Sun's convective zone. The helioseismic inversion methods are inspired by the work in terrestrial seismic tomography and there are many similarities.

1.6 Finite-frequency tomography

This book is for a large part motivated by the most recent development in the theory of body wave and surface wave tomography: the recognition that progress in sharpening the images requires that we step away from the shortcomings of ray theory. Ray theory is an infinite-frequency (or zero wavelength) approximation, but in reality seismic waves have wavelengths that range from 10 to 1000 km or even more, and wavelengths and Fresnel zones are often much larger than heterogeneities of considerable interest. As the models became more and more detailed, scepticism about the use of ray theory grew. To be valid, ray theory requires that wavelengths are short and Fresnel zones narrow. We now have simple but powerful diagnostics – Equation (7.2) for example – for the correctness of ray theory. Though much of the tomographic work so far made acceptable use of the ray-theoretical approximation, such tests also tell us that more sophisticated analytical techniques are called for if we want to move beyond the current limits on resolution.

For surface waves, the complexity of the finite-frequency sensitivity of waveforms to the heterogeneity in the Earth became apparent in an early study by Woodhouse and Girnius [401] who used a first order perturbation theory. Snieder [325, 327, 334] used first order perturbation theory for high frequency surface waves and was the first to attempt to use these in regional tomography [329]. Because first order perturbation methods were originally developed by the German – later naturalized British – physicist Max Born (1882–1970) within the context of quantum mechanics, first order perturbation theory is often referred to as 'Born theory'. For waves, perturbations lead to scattered waves and yet another term enters the jargon: first order or 'single scattering'. In this book I shall freely mix such terminology.

For body waves, the fact that the sensitivity of travel times and amplitudes extends over the full width of the Fresnel zone and that this extended sensitivity influences seismic tomography was pointed out 20 years ago by Wielandt [394] and Nolet [235, 237]. However, they did not recognize the power of first order perturbation theory to provide a quantitative analysis. Earlier, Hudson [137] had modelled the coda of P-waves using Born scattering theory. Coates and Chapman [63] showed that Born theory adequately models the changes in arrival time and amplitude of body waves using ray theory, opening up the possibility that one could improve on ray theory using ray theory! A formalism based on a first-order perturbation treatment of the wave equation was provided by Woodward [403] and Luo and Schuster [190], who proposed a waveform inversion method and compared this to travel time inversion for acoustic waves in a 2D (cross-well) tomographic experiment. Yomogida [409] was the first to recognize the power of paraxial ray theory in combination with first order perturbation theory. At the same time, Jin et al. [149] developed ray–Born inversion methods that were extended to attenuating media by Ribodetti et al. [276] and applied to invert the results of laboratory experiments. By the mid 1990s, the time seemed also ripe for a change in 3D global tomography, which was still relying heavily on pure ray theory for the inversion of body waves – the type of waves that, at least in theory, provide the best resolution.

Not surprisingly, the first attempts to provide a finite-frequency treatment of teleseismic body waves followed up on the earlier Born theory for normal modes and surface waves. Li and Tanimoto [186] studied the effect of coupling of normal modes on the waveforms of body waves, and Li and Romanowicz [185] used it to develop a tomographic algorithm, that Zhao and Jordan [417] extended to anisotropy. Tanimoto [346] used the slope of the waveform to translate the waveform perturbation directly into a travel time shift. These studies, however, were still restricted to inversions in a two-dimensional setting in which the Earth's properties are assumed independent of the third, horizontal, coordinate. However, Zhao and Dahlen [416] derived Fréchet kernels for body–waveform perturbations in a 3D Earth using asymptotic normal mode theory. At the same time, Snieder and Lomax [333] tried to incorporate finite-frequency effects directly to travel times by applying a smoothing function to the velocity medium.

A full 3D treatment of finite-frequency effects on travel times of elastic waves in the global Earth was given by Marquering et al. [196, 195], who coupled higher modes of surface waves for the case of three-dimensional perturbations and investigated the effect on the cross-correlograms of body waves, summing a large number of surface wave modes to synthesize the S-wave. Zhao et al. [418] subsequently developed three-dimensional kernels using a discrete normal mode formalism.

Because of the surprising result that the travel time of a body wave is insensitive to a perturbation located exactly on the ray itself, Marquering labelled the sensitivity kernels 'banana-doughnut' kernels, in reference to their banana-like shape with a sensitivity 'hole' in the centre. For the computation of three-dimensional kernels, Marquering's surface wave mode summation is considerably more efficient than the summation of discrete normal modes used by Li Zhao et al. [418] in the terrestrial case, and by Birch and Kosovichev [24] in solar tomography, but still not fast enough for routine application in seismic tomography with large data sets. The breakthrough that made finite-frequency tomography a practical possibility was provided by Dahlen et al. [76], who used dynamic ray tracing to make the computation of the sensitivity (or Fréchet) kernels efficient enough for application in large, global tomography inversions, and who provided a comprehensive theory for all types of teleseismic body waves, including the effects of caustics.