Chapter 2

Methodology

In this chapter, a summary of the basic theory of travel time tomography is given based on the derivations and descriptions of Spakman and Nolet (1988), Spakman (1993), Bijwaard (1999) and Spakman and Bijwaard (2001).

2.1 Linearized seismic tomography

In travel time tomography, observed arrival times of seismic waves are compared to theoretical times predicted by a reference velocity model to obtain seismic velocity variations within the Earth with respect to the reference model. The basic observation thereby consists of the arrival time affected by a reading error ($= t_{\text{arrival}} + \epsilon$). The observed arrival time t_{arrival} is composed of the real travel time $T_x(s)$ of the wavefront, travel time effects Δt_s due to station elevation and instrument response and the real origin time $t_{x(\text{origin})}$

$$t_{\text{arrival}} = T_x(s) + \Delta t_s + t_{x(\text{origin})} \tag{2.1}$$

where $T_x(s)$ depends on the Earth's slowness field (the reciprocal of the seismic wave speed), using the true earthquake location x. The slowness field s is unknown in equation 2.1. However, reference models inferred from seismological observations exist as the standard 1-D Earth models of Jeffreys and Bullen (1940), PREM (Dziewonski and Anderson, 1981), IASP91 (Kennett and Engdahl, 1991) or ak135 (Kennett *et al.*, 1995) which can be used to compute the travel time for these reference models. To obtain a better prediction of the travel times, the spherically symmetric reference models can be replaced by more realistic velocity models which vary in three dimensions. Independent of the exact type of reference model used, the computation of the predicted arrival times is given as

$$t^0_{\text{arrival}} = T^0_{x_0}(s_0) + t^0_{x_0(\text{origin})} \tag{2.2}$$

where the index "0" denotes reference model quantities, $T^0_{x_0}$ is the predicted travel time starting from the reference source location x_0 along the linearized ray path in the reference model, s_0 is the slowness of the reference model and $t^0_{x_0(\text{origin})}$ is the origin time computed in the reference model. These reference model quantities are obtained from an earthquake location

procedure where the arrival time observations are used as constraints after correction for the Earth's ellipticity (Kennett and Gudmundsson, 1996). Arrival times of some special phases (e.g. pP, pwP) require an additional bounce point topography/bathymetry correction. Also, generally an earthquake will be mislocated with respect to its true position as a result of the observational errors, of possibly insufficient data constraints, and of the slowness difference between the true Earth and the reference Earth, the slowness anomaly.
The delay time, which will be used to obtain velocity variations in a tomographic inversion, is defined as the difference of observed and predicted arrival time

$$d = t_{\text{arrival}} - t^0_{\text{arrival}} + \epsilon \tag{2.3}$$

Replacing the arrival times by the expressions derived in equation 2.1 and 2.2 gives the delay d as

$$d = T_x(s) - T^0_{x_0}(s_0) + \Delta t_s + \underbrace{t_x - t_{x_0}}_{\Delta t_x} + \epsilon \tag{2.4}$$

where Δt_x contains the timing error due to source mislocation as a result of the slowness anomaly field $s - s_0$ and ϵ describes the observational errors (e.g. picking errors, phase misidentifications, remaining location errors).
In the high-frequency approximation seismic wave propagation theory reduces to seismic ray theory. In this approximation travel times are computed by integration of the slowness along the ray path:

$$T_x(s) = \int_L s \, dl \qquad \text{with} \qquad L = L(s) \tag{2.5}$$

where L is the ray path and dl the ray segment.

$$T^0_{x_0}(s_0) = \int_{L_0} s_0 \, dl_0 \qquad \text{with} \qquad L_0 = L_0(s_0) \tag{2.6}$$

L_0 is the ray path in the reference model and dl_0 is the ray segment.
Substitution of the travel time integral in the delay time equation gives:

$$d = \int_L s \, dl - \int_{L_0} s_0 \, dl_0 + \Delta t_x + \Delta t_s + \epsilon \tag{2.7}$$

If the difference between the actual Earth and the reference model is sufficiently small, the first travel time integral is linearized using Fermat's Principle. Applying Fermat's Principle effectively results in replacing L by L_0 and requires accounting for the spatial event mislocation $x - x_0$. The effect of the mislocation on the travel time is approximated by a Taylor expansion of the travel time around the reference source x_0. This leads to the delay time equation

$$d = \int_{L_0} (s - s_0) \, dl_0 + (\mathbf{x} - \mathbf{x}_0) \cdot \nabla_0 T^0 + \Delta t_x + \Delta t_s + \epsilon \tag{2.8}$$

which is the forward equation of travel time tomography based on ray theory. In a tomographic analysis, a large set of delay times equation 2.8 derived from many source-station combinations is inverted for estimates of the slowness anomaly field $s - s_0$, the event mislocations $\mathbf{x} - \mathbf{x}_0$, origin time errors Δt_x and the station static terms Δt_s.

2.2 Parameterization and forward (observation) equation

To convert a set of equations 2.8 into a matrix-vector equation for inversion, the piecewise continuous slowness field $\Delta s = s - s_0$ has to be parameterized. Among various methods, which exist for the parameterization, here the irregular cell representation of Spakman and Bijwaard (2001) is chosen. The volume of interest is initially divided into a regular grid of non-overlapping conical cells of a basic (=smallest) cell size. To make the cells independent of their geographical position on the globe, they are constructed in such a way that they have equal surface areas at given depth. Laterally, the size of the basic cells is set dependent on the maximum expected horizontal resolution (smallest detail). The investigated volume is discretized by layers with cells of constant thickness, where the layer thickness varies with depth and is defined to fit cell layers between the first-order seismic discontinuities contained in the reference model and to accommodate the expected resolution with depth. Subsequently, further regular grids are defined with cells which have a bigger lateral extension and are multiples of the basic cells. As a constraint for the cell size on the final irregular grid, the hitcount, i.e. the number of rays crossing a cell, is used and a hitcount threshold is defined. The hitcount is computed on each of the regular grids and the smallest cell size is determined, for which the hitcount is still above the threshold. Thus, the final irregular grid cells are all sampled by a comparable number of rays in regions of sufficient ray coverage. Effectively, the irregular grid results from an optimization procedure which adapts cell volume to obtain more or less equal hitcount in cells. Figure 2.1 shows an example of an equal surface area grid and an irregular grid.
The basic cells are represented by orthonormal cell functions b_j (normalized by $V_j^{-1/2}$ with V_j as volume of cell b_j) which are used to construct the non-overlapping, irregular grid cells c_k:

$$c_k = \sum_{j=1}^{N_b} P_{kj} b_j \tag{2.9}$$

where N_b is the number of basic cells and P_{kj} is the projection coefficient which is zero except when b_j is part of c_k.
The slowness anomaly field Δs ($= s - s_0$) can then be parameterized as follows:

$$\Delta s = \sum_{k=1}^{N} \Delta s_k c_k + \Delta s_p \tag{2.10}$$

where c_k represents the orthonormal cell function, Δs_k the projection coefficient, N the number of irregular cells in the volume and Δs_p the projection error. The projection error reflects the accuracy of the parameterization and will be small/negligible for a sufficiently

detailed parameterization (cf. Spakman, 1991).
The arc length l^c_{ik} in the irregular cells c_k is given by the projection of the arc length in the basic cells l^b_{ij}:

$$l^c_{ik} = \sum_{j=1}^{N_b} P'_{kj} l^b_{ij} \tag{2.11}$$

with P'_{kj} as projection coefficient which is one if the basic cell b_j is part of the irregular cell c_k and zero otherwise.
Using this discretization, the travel time difference for the i-th ray is calculated to

$$d_i \approx \sum_{k=1}^{N} \Delta s_k l^c_{ik} + \sum_{l=1}^{4} \Delta x_l g_{il} + \Delta\sigma_m l_{im} + \epsilon_i \tag{2.12}$$

where l^c_{ik} represents the length of the i-th ray in the k-th irregular cell, $\Delta\mathbf{x} = \mathbf{x} - \mathbf{x}_0$, g_{il} describes the elements of $\nabla_0 T^0$ (now including the source time error), $\Delta\sigma_m$ represents the unknown slowness error of the m-th station, (l_{im}) is the identity matrix and ϵ_i contains all errors and signal not accounted for by the previous terms.
The described approximations lead to the observation equation

$$\mathbf{d} = \mathbf{A} \cdot \mathbf{m} + \epsilon \tag{2.13}$$

with $\mathbf{A} = (l_{ik}, g_{il}, l_{im})^T$, $\mathbf{m} = (\Delta s_k, \Delta x_l, \Delta\sigma_m)^T$, $\mathbf{d}$ as data vector with the observed travel time residuals and ϵ as the error vector.
Equation 2.13 forms a system of inconsistent linear equations, usually ill-conditioned as a result of insufficient data, that are to be inverted to obtain a model.

2.3 Inversion and regularization

Due to data inconsistency and usual rank deficiency of $\mathbf{A}$, a unique solution of 2.13 does not exist. To deal with data inconsistency a measure of data fit, that determines how well model predictions approximate the real data, is required for selecting a solution (or a range of solutions in case of rank deficiency). In large size inverse problems, the usual strategy is to find the best-fitting model in the least squares sense, i.e. the model that gives the smallest sum of the squared differences between the observed data vector and its prediction computed from the model. The least square criterion is implemented as follows:
In equation 2.13, the data are first weighted by the data covariance $\mathbf{C}_d$ and next the minimization of $\epsilon^T \mathbf{C}_d^{-1} \epsilon$ is performed, or equivalently the minimization of the cost function

$$\Phi(\mathbf{m}) = (\mathbf{d} - \mathbf{A}\mathbf{m})^T \mathbf{C}_d^{-1} (\mathbf{d} - \mathbf{A}\mathbf{m}) \tag{2.14}$$

This results in the weighted least squares solution:

$$\mathbf{m} = (\mathbf{A}^T {\mathbf{C}_d}^{-1} \mathbf{A})^{-1} \mathbf{A}^T {\mathbf{C}_d}^{-1} \mathbf{d} \tag{2.15}$$

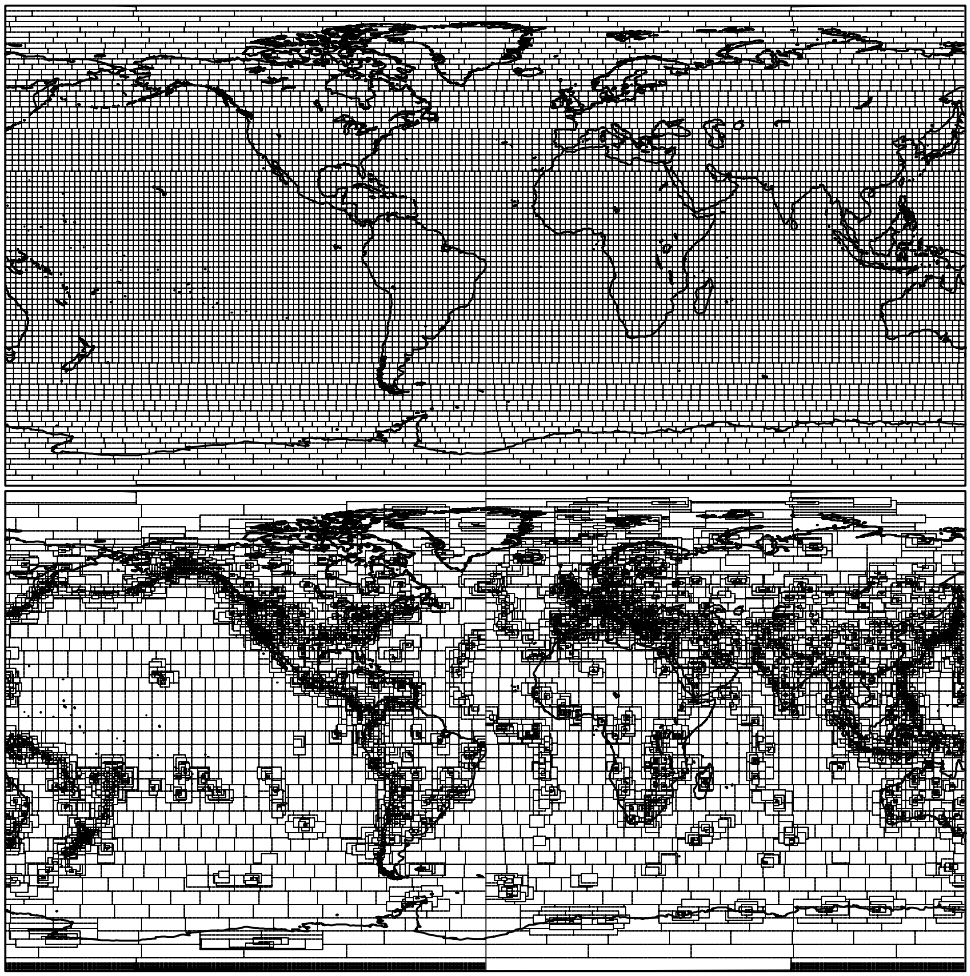

Figure 2.1: Example of two types of grid parameterizations for the topmost layer of the model. On the top, a regular $2.0^\circ \times 2.0^\circ$ grid with cells of equal surface area is displayed where the number of cells per latitude decreases from the equator towards the poles. On the bottom, an irregular grid is shown varying in cell size from $0.5^\circ \times 0.5^\circ$ to $5.0^\circ \times 5.0^\circ$. The cell size is reduced noticeably in areas with a large number of stations and/or earthquakes as it depends on the hitcount (i.e. the number of rays that traverse each cell).

Clearly, the least squares solution can only be computed if $(\mathbf{A}^T {\mathbf{C}_d}^{-1} \mathbf{A})^{-1}$ exists, i.e. for (numerically) overdetermined equations. In tomography this occurs only in exceptional cases. More often $\mathbf{A}$ has zero eigenvalues (there is no unique solution) and/or $\mathbf{A}$ is ill-conditioned. Very small eigenvalues exist due to, for instance, a dominance of nearly parallel ray paths sampling a model region. As a result of ill-conditioning, small errors in the data can lead to large amplitude errors in the model. The usual remedy to deal with rank deficiency and ill-conditioning is to resort to regularized least squares by extending the cost function with a term $\lambda^2 \mathbf{m}^T \mathbf{C}^T \mathbf{C}\mathbf{m}$ that imposes amplitude, roughness and/or smoothness constraints on the model. This results in the following cost function:

$$\Phi(\mathbf{m}) = (\mathbf{d} - \mathbf{A}\mathbf{m})^T \mathbf{C}_d^{-1} (\mathbf{d} - \mathbf{A}\mathbf{m}) + \lambda^2 \mathbf{m}^T \mathbf{C}^T \mathbf{C}\mathbf{m} \tag{2.16}$$

Minimization of Φ gives:

$$\mathbf{m} = \left[\mathbf{A}^T \mathbf{C}_d^{-1} \mathbf{A} + \lambda^2 \mathbf{C}^T \mathbf{C}\right]^{-1} \mathbf{A}^T \mathbf{C}_d^{-1} \mathbf{d} \tag{2.17}$$

where λ is a tuning parameter which controls the trade-off between minimizing the data misfit and finding the minimum of the scaled model norm. $\mathbf{C}$ is a matrix imposing model amplitude damping if $\mathbf{C}$ is the identity matrix and smoothing if $\mathbf{C}$ is a second derivative finite difference operator. A model with reduced roughness is obtained by taking a first derivative operator. As derivative operators have non-trivial null spaces, they are often combined with amplitude damping for a complete regularization.
This formulation of the regularized weighted least squares inversion is equivalent to finding the normal least squares solution of the following equation:

$$\begin{pmatrix} \mathbf{C}_d^{-1/2} \mathbf{A} \\ \lambda \mathbf{C} \end{pmatrix} \mathbf{m} = \begin{pmatrix} \mathbf{C}_d^{-1/2} \mathbf{d} \\ 0 \end{pmatrix} \tag{2.18}$$

Station statics and mislocation parameters are associated with much smaller singular values of $\mathbf{A}$ than the slowness parameters of cells. Also, in our irregular model parameterization small cells, placed in mantle regions where we expect resolution for small detail, are usually associated with systematically smaller singular values than large cells which are placed in regions where only larger scale structure can be imaged. To balance the singular value spectrum of $\mathbf{A}$, (i.e. to relatively raise the singular values associated with small cells, mislocation and station parameters) a parameter scaling matrix $\mathbf{S}$ is implemented depending on cell size and hitcount, or parameter type. This leads to the following equation:

$$\begin{pmatrix} \mathbf{C}_d^{-1/2} \mathbf{A} \\ \lambda \mathbf{C} \end{pmatrix} \mathbf{S}^{-\frac{1}{2}} \mathbf{m}' = \begin{pmatrix} \mathbf{C}_d^{-1/2} \mathbf{d} \\ 0 \end{pmatrix} \tag{2.19}$$

with $(S_{jj}) = h_j V_j$ and h_j and V_j as hitcount and cell volume respectively and $\mathbf{m}' = \mathbf{S}^{\frac{1}{2}} \mathbf{m}$ as scaled model vector. The diagonal elements of $\mathbf{S}$ pertaining to mislocation, origin time error, and station static parameters are taken such that they scale these parameters to an expected slowness anomaly amplitude. The model parameter scaling is beneficial in case eq. 2.18 is to be solved with truncated Singular Value Decomposition (SVD) or with iterative solvers (e.g. conjugate gradients) which are stopped before formal convergence has been achieved.

Many types of solvers exist to compute the inverse matrix. Singular Value Decomposition (applied to eq. 2.19 and delivering as solution eq. 2.17, unless truncated SVD is used) requires full storage of the matrix and data-related matrices and is only applicable to relatively small inverse problems. For larger model sizes than SVD can deal with, the inverse matrix of eq. 2.17 can still be computed, for example, with Cholesky decomposition. Examples of inversions computed with SVD or Cholesky decomposition are studies by Ritsema *et al.* (1999) and Boschi (2003) respectively. Huge inverse problems dealing with a very large number of data (millions) and model parameters (hundreds of thousands), can presently only be solved with iterative solvers (e.g. conjugate gradient methods) which, when stopped before formal convergence, compute an approximate solution of (2.17). Throughout this thesis, the tomographic inversion is performed with the iterative conjugate gradient algorithm LSQR of Paige and Saunders (1982) which is applied to (2.19) and in full convergence yields solution (2.17), or an approximation when stopped before full convergence. LSQR resembles in the early iterations SVD (e.g. van der Sluis and van der Vorst, 1987) where the solution is constructed starting with the components associated with the highest singular values. Stopping the algorithm before formal convergence occurs has a comparable damping effect as truncated SVD.

2.4 Assessment of model quality

The quality of a solution is defined by data fit, spatial resolution and model covariance. A drawback of iterative techniques like LSQR is that only a solution and corresponding data fit are computed. Although approximate methods for finding resolution and covariance in huge inversions have been proposed (Nolet *et al.*, 1999, 2001; Yao *et al.*, 1999, 2001), it is practically not yet possible to obtain these exactly.
Spatial resolution is defined as follows: Let $\mathbf{m}_{true}$ be the true cell slowness anomaly field. The corresponding true data are defined as $\mathbf{d}_{true} = \mathbf{A}\mathbf{m}_{true}$. They are different from the real data $\mathbf{d}$ by an error ϵ which includes observational errors, effects of parameterization, linearization and other theoretical approximations. Writing (2.17) as

$$\mathbf{m} = \mathbf{G}\mathbf{d} \tag{2.20}$$

with $\mathbf{G}$ being the generalized inverse of $\mathbf{A}$ and

$$\mathbf{d} = \mathbf{d}_{true} + \epsilon \tag{2.21}$$
$$= \mathbf{A}\mathbf{m}_{true} + \epsilon \tag{2.22}$$

we find

$$\mathbf{m} = \mathbf{R}\mathbf{m}_{true} + \mathbf{G}\epsilon \tag{2.23}$$

where $\mathbf{R} = \mathbf{G}\mathbf{A}$ is the resolution kernel. $\mathbf{R}$ describes how the true cell model is mapped into the solution $\mathbf{m}$ and the term $\mathbf{G}\epsilon$ shows how errors/bias in the data map into the solution.
When the resolution matrix $\mathbf{R}$ cannot be explicitly computed (as in our inversions), sensitivity tests with synthetic slowness models can be conducted to implicitly obtain knowledge of $\mathbf{R}$

(Spakman and Nolet, 1988). As $\mathbf{m} = \mathbf{R}\mathbf{m}_{true}$ is a linear equation, one can invent a synthetic model $\mathbf{m}_{true}$, compute synthetic data

$$\mathbf{d}_{true} = \mathbf{A}\mathbf{m}_{true} \tag{2.24}$$

and solve

$$\mathbf{d}_{true} = \mathbf{A}\mathbf{m}_{s} \tag{2.25}$$

for the tomographic recovery $\mathbf{m}_s$ of the synthetic model to make inferences about $\mathbf{R}$ by comparison of $\mathbf{m}_s$ and $\mathbf{m}_{true}$. However, synthetic slowness anomaly models can only be designed to detect lack of resolution. When the synthetic model $\mathbf{m}_{true}$ is entirely in the null space of $\mathbf{A}$, it will be in the null space of $\mathbf{R}$ and $\mathbf{m}_s$ will only contain zeros. Conversely, when a particular synthetic model happens to be entirely in the range of $\mathbf{A}$, then the model will be recovered completely by tomography and one might be inclined to infer perfect resolution. However, Leveque *et al.* (1993) demonstrated that while a specific synthetic model is recovered, other synthetic models for the same region, but with different dominant structural wavelengths, can still be entirely in the nullspace. To ensure detection of lack of resolution it is therefore necessary to perform sensitivity tests with a large variety of synthetic models, particularly of different structural wavelength content (e.g. Bijwaard *et al.*, 1998). If, for a particular mantle region, no appreciable lack of resolution is detected for a wide variety of synthetic models, i.e. all models are well recovered, then this observation is taken as a measure of high resolution at the wavelengths used, although formally there is still a possibility that resolution is lacking.

The only synthetic model that leads to equivocal interpretations is an overall zero anomaly model except for one cell with a non-zero amplitude – a spike. The tomographic imaging of this model leads effectively to the computation of a column of the resolution kernel. As computing $\mathbf{m}_s$ requires an inversion with LSQR similar to the real data inversion, it is not feasible to reconstruct $\mathbf{R}$ from such synthetic tests when the model consists of a large number of parameters ($\sim 10^5$). A more efficient test, related to a single spike test, was designed by Spakman and Nolet (1988) where the synthetic model contains a 3-D network of spatially well separated spikes. The recovery of this synthetic model shows immediately lack of amplitude recovery of single cells and allows for detection of directional anomaly smearing. In this thesis, the spike tests will be conducted for a variety of spike sizes to assess lack of resolution.

Model covariance, i.e. amplitude errors and their correlations, cannot be formally computed, although the error correlations are implicitly contained in the resolution kernel. The effect of data errors on the solution amplitudes (i.e. the term $\mathbf{G}\epsilon$) is assessed by taking a randomly permuted version $\mathbf{d}_p$ of the data vector $\mathbf{d}$ for ϵ and solving $\mathbf{d}_p = \mathbf{A}\mathbf{m}$ (Spakman, 1991). This random data vector has the same bulk statistics (average, standard deviation, and higher order moments) as the real data vector. In this test, however, all correlation between the real data and their associated ray paths (the matrix $\mathbf{A}$) is destroyed and the permuted data vector $\mathbf{d}_p$ represents an upper limit of data noise. A good outcome of this noise test is a model with low amplitudes and only random amplitude variations. A poor outcome would be if imaged amplitudes and structural wavelengths are comparable to those in the model obtained from inverting the real data which would signal that data and ray paths are not correlated.