JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 104, NO. B12, PAGES 28,841–28,857, DECEMBER 10, 1999

An efficient, probabilistic neural network approach to solving inverse problems: Inverting surface wave velocities for Eurasian crustal thickness

R.J.R. Devilee, A. Curtis,[1] and K. Roy-Chowdhury

Geodynamic Research Institute, Department of Geophysics, Utrecht University, the Netherlands

Abstract. Nonlinear inverse problems usually have no analytical solution and may be solved by Monte Carlo methods that create a set of samples, representative of the a posteriori distribution. We show how neural networks can be trained on these samples to give a continuous approximation to the inverse relation in a compact and computationally efficient form. We examine the strengths and weaknesses of this approach and use it to determine the full a posteriori distribution of crustal thickness from surface wave velocities. The solution to this inverse problem shows significant asymmetry and large uncertainties due to trade-off with shear velocity structure around the Moho. We produce maps of maximum likelihood crustal thickness across Eurasia which are in agreement with current knowledge about the crust; thus we provide an independent confirmation of these models. In this application, characterized by repeated inversion of similar data, the neural network algorithm proves to be very efficient.

1. Introduction

Geophysical inverse problems are characterized by the nonlinearity of the physics and by the statistical nature of the solution, the latter being partly due to the physics and partly due to noise in geophysical measurements. The solution to a general inverse problem can be given in the form of a density function (df) $\sigma(\mathbf{m},\mathbf{x})$ [*Tarantola and Valette,* 1982], where $\mathbf{x}$ represents a set of distinct measurements and $\mathbf{m}$ a set of model parameters. A df, when integrated over a range of values of $\mathbf{m}$ and $\mathbf{x}$, defines the probability that $\mathbf{m}$ and $\mathbf{x}$ assume a value in that range. Approximate probabilistic solutions are given by linearizing the problem [*Matsu'ura and Hirata,* 1982]; explicit analytical solutions for nonlinear inverse problems do not usually exist. Instead, approximations are formed by a representative set of samples, obtained by methods such as Monte Carlo [*Wiggins,* 1996; *Mosegaard and Tarantola,* 1995], simulated annealing [*Sen and Stoffa,* 1991; *Zhao et al.,* 1996], and the genetic algorithm [*Stoffa and Sen,* 1991; *Lomax and Snieder,* 1994]. Of these, the Monte Carlo method yields most statistical information, but forming the solution requires many forward calulations. We show that when repeated inversions using similar prior information are required, the cost of forming subsequent solutions can be reduced significantly by using neural networks as an inversion tool.

In the past, feedforward neural networks (henceforth, just networks) have been applied succesfully in a wide range of geophysical situations: to classification problems (e.g., source characterisation [*Pulli and Dysart,* 1990; *Dowla et al.,* 1990]), to first-break picking [*McCormack et al.,* 1993; *Dai and MacBeth,* 1997], and to continuous inverse problems, e.g., yield estimation [*Leach et al.,* 1993], subsurface target location [*Poulton et al.,* 1992], and inversion for seismic velocity models [*Roth and Tarantola,* 1994]. We will show later that such networks yield mean solutions to problems which are probabilistic in nature. We will refer to them as sonnets (single output neural networks), since they use a single output node to represent the mean value of each model parameter.

The use of networks in probabilistic applications is described in a general context by *Bishop* [1995]. We develop specific criteria for constructing networks which provide probabilistic information on geophysical inverse problems by emulating the solution from samples obtained with a Monte Carlo method. The resulting methodology is referred to as neural network inversion (NNI). With proper preprocessing of training data, a sonnet can be used to estimate $\sigma(\mathbf{m},\mathbf{x})$ directly. An alternative approach is to use networks with multiple output nodes to represent the statistical properties of each model parameter. We call these networks monnets (multiple output neural networks). If we split σ as follows:

$$\sigma(\mathbf{m},\mathbf{x}) = P(\mathbf{m}|\mathbf{x})Q(\mathbf{x}), \qquad (1)$$

[1]Now at Schlumberger Research, Cambridge, England, United Kingdom.

Paper number 1999JB900273.
0148-0227/99/1999JB900273$09.00

where $Q(\mathbf{x})$ represents the limitations on the data space imposed by the physics and prior constraints on the model space and $P(\mathbf{m}|\mathbf{x})$ defines the conditional df of models fitting a datum, then a monnet can be used to emulate $P(\mathbf{m}|\mathbf{x})$.

To demonstrate the effectiveness of our methodology, we apply NNI to the nonunique and weakly nonlinear problem of estimating the crustal thickness from velocity dispersion curves of fundamental mode surface waves. The thickness of the crust is an important tectonic parameter, and we investigate how well it is constrained by the data, irrespective of the values of other unknowns such as crustal and mantle velocities. Two independent dispersion data sets are used, consisting of regionalized phase [*Curtis et al.*, 1998] and group velocity [*Wu and Levshin*, 1994] maps of regions of Eurasia. Thus we obtain a unique, consistent view of the patterns of crustal thickness across the continent.

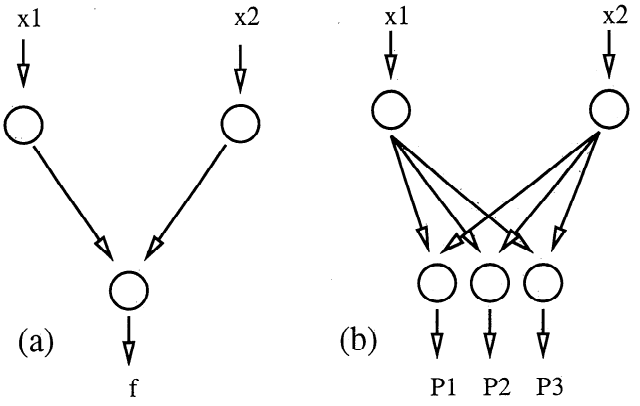

Figure 1. Sketch of the structure of feedforward networks. Arrows indicate that information flows in one direction. (a) Network representing a sonnet, which returns a value f given a set of inputs $\mathbf{x}$. (b) Network representing a monnet, which returns M ($M = 3$) outputs. These networks are trained to give a set of probabilistic estimators $\mathbf{P}$ about a model parameter.

2. Neural Networks

A neural network may be thought of as a black box which filters a given input vector $\mathbf{x}$ through weights $\{w_j\}$, offsets $\{o_j\}$, and transfer functions (for which we use sine functions) to produce an output vector $\mathbf{u}(\mathbf{x}, \{w_j\}, \{o_j\})$. This output should approximate a desired vector $\mathbf{W}(\mathbf{m})$ ($\mathbf{W}$ may have a different number of elements than m), where $\mathbf{x}$ and $\mathbf{m}$ are samples from the df $\sigma(\mathbf{m}, \mathbf{x})$; the function $\mathbf{W}$ will determine the properties of the monnets. A measure of the misfit between the network output $\mathbf{u}(\mathbf{x}, \{w_j\}, \{o_j\})$ and the required output $\mathbf{W}(\mathbf{m})$ over an infinitely large number of samples of $\sigma(\mathbf{m}, \mathbf{x})$ is given by the global error:

$$E(\{w_j\}, \{o_j\}) = \int_{-\infty}^{+\infty} Q(\mathbf{x}) \int_{-\infty}^{+\infty} P(\mathbf{m}|\mathbf{x}) \times \sum_{j=1}^{N} \varepsilon[u_j(\mathbf{x}, \{w_j\}, \{o_j\}), W_j(\mathbf{m})]\, d\mathbf{m} d\mathbf{x}, \quad (2)$$

where we have used the decomposition in equation (1). A measure for the difference between vectors $\mathbf{u}$ and $\mathbf{W}$ for each input $\mathbf{x}$ is given by the local error ε. For the moment, we assume that a suitable training algorithm is used to optimize the weights and offsets of the network such that the global error is minimized. In sections 2.1 and 2.2, we vary the form of ε and $\mathbf{W}$ to train networks to return different types of output to characterize the density function. In these problems the input vector $\mathbf{x}$ will represent some geophysical data, the values of the output vector $\mathbf{u}(\mathbf{x}, \{w_j\}, \{o_j\})$ will be used to approximate a df. Henceforth, we will drop the explicit dependence of $\mathbf{u}$ and E on the weights $\{w_j\}$ and offsets $\{o_j\}$ in our notation.

2.1. Sonnet-Type Networks

The sonnet represents the traditional use of neural networks. Their general structure is sketched in Figure 1a. These networks use a single output to create an estimate of the average value of each parameter. This property can be used to our advantage when the training data are preprocessed to carry probabilistic information.

2.1.1. Averaging property of a sonnet. We consider a sonnet which has N output values $u_j, j = 1, ..., N$ which will be used to approximate the N model parameters of $\mathbf{m}$ directly. Let the local error be given by the L^2 norm:

$$\varepsilon[u_j(\mathbf{x}), m_j] = [u_j(\mathbf{x}) - m_j]^2. \quad (3)$$

Then the misfit function in equation (2) has its minima for values of $u_n(\mathbf{x}), n = 1...N$ given by

$$u_n(\mathbf{x}) = \int_{-\infty}^{+\infty} P(\mathbf{m}|\mathbf{x}) m_n d\mathbf{m}. \quad (4)$$

This result implies that once the network has been configured to minimize the error E, u_n gives the mean value of m_n. This strategy has been used in previous geophysical applications [e.g. *Leach et al.*, 1993; *Poulton et al.*, 1992; *Roth and Tarantola*, 1994]. A simple application of a sonnet is shown in Figure 2a.

2.1.2. Using a sonnet to retrieve the full solution. It is straightforward to use a sonnet to obtain more detailed probabilistic information by proper preprocessing of the training data. Suppose a Monte Carlo method has been used to create a set of samples $\{\mathbf{t}_j\}$ from some parameter space, distributed according to $\sigma(\mathbf{t})$ (for example, we may define $\mathbf{t} \equiv \{\mathbf{m}, \mathbf{x}\}$). We compare each $\mathbf{t}_j$ with K other samples. For any such sample $\mathbf{t}_k$ we assign a weight $w(\mathbf{t}_j, \mathbf{t}_k)$ representative of the distance between the two vectors. For large enough K we can write

$$\frac{1}{K} \sum_k w(\mathbf{t}_j, \mathbf{t}_k) = \int w(\mathbf{t}_j, \mathbf{t}) \sigma(\mathbf{t}) d\mathbf{t}. \quad (5)$$

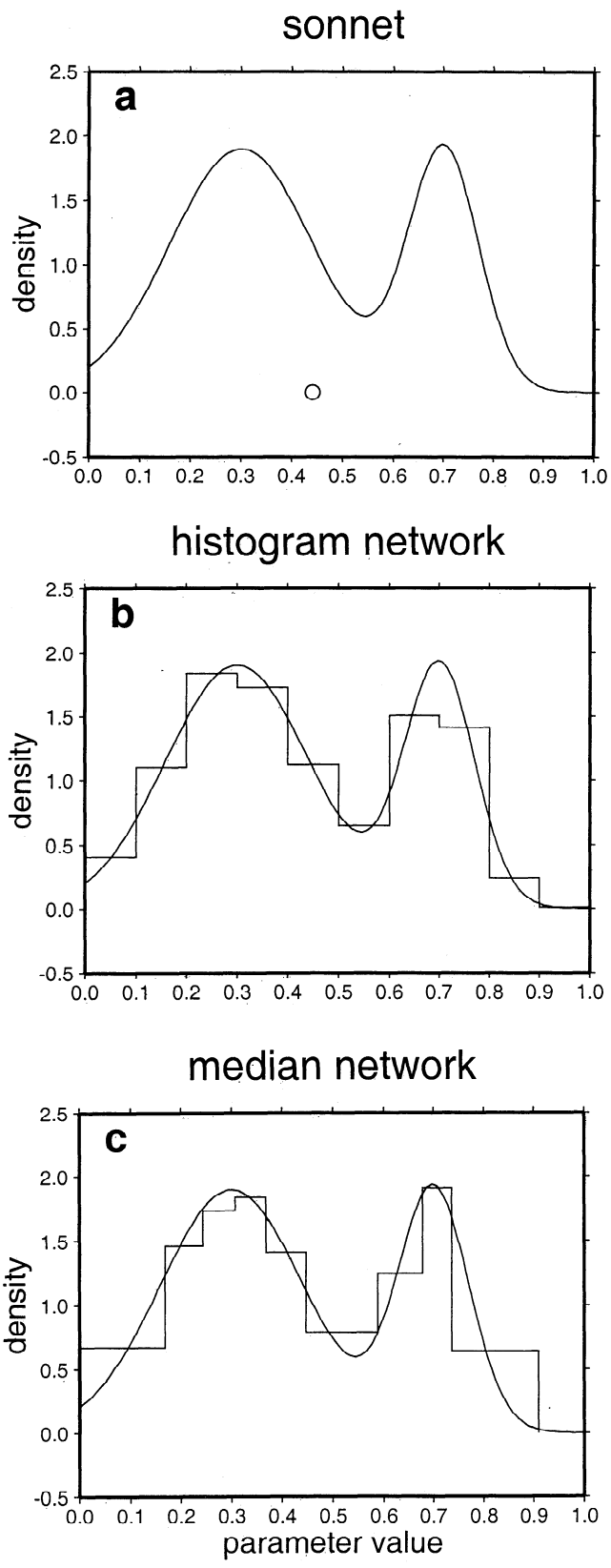

Figure 2. Outputs of (a) a sonnet, (b) a histogram network, and (c) a median network. The networks were trained for fixed input 1.0 to recognize a single parameter which takes values between 0 and 1 according to the smooth distribution plotted in each figure. The sonnet returns a single value, which is indicated by a circle at offset 0. The outputs of the monnets were postprocessed and plotted as discretized density functions.

In particular, if w defines a cubic region around point $\mathbf{t}_j$ (i.e., returns a value 1 if point $\mathbf{t}_k$ falls inside the region and 0 if outside), then the left-hand side of equation (5) equals K'/K, where K' is the number of occurences that vector $\mathbf{t}_k$ lies in the cubic region. If the region V is small enough such that the value $\sigma(\mathbf{t})$ is representative of values within V, we can make the following approximation:

$$\frac{K'}{K} = V\sigma(\mathbf{t}). \tag{6}$$

These values can be used to train a network to approximate the df. During training, the network is averaging over examples, so that effectively $K \to \infty$, and hence the network outputs will approximate $V\sigma(\mathbf{t})$ closely. In practice, K should be chosen large enough to stabilize the training of the network.

If measurements and model parameters are available, i.e., $\mathbf{t} \equiv \{\mathbf{m}, \mathbf{x}\}$, and are preprocessed as follows,

$$P_j(\mathbf{m}_j, \mathbf{x}_j) = \frac{1}{K}\sum_k w(\{\mathbf{m}, \mathbf{x}\}_j, \{\mathbf{m}, \mathbf{x}\}_k), \tag{7}$$

we could train the network on a set of samples $\{\mathbf{m}_j, \mathbf{x}_j, P_j(\mathbf{m}_j, \mathbf{x}_j)\}$, where $\{\mathbf{m}_j, \mathbf{x}_j\}$ forms the input vector and P_j is the output value for each sample j. Thus the network will return an approximation to $V\sigma(\mathbf{m}, \mathbf{x})$. However, a sample set usually does not include examples in regions with zero probability when Monte Carlo methods are used. Indeed, even a random sampling of the model space does not guarantee a random sampling of $(\mathbf{m}, \mathbf{x})$ space, and hence the network remains untrained on a large volume of inputs. Applying the network to data it has not been trained on leads to unpredictable outputs. In section 2.2 we describe monnets which can be used in more general cases.

2.2. Monnet-Type Networks

We now introduce the histogram and median networks which emulate the conditional df $P(m|\mathbf{x})$; their general structure is sketched in Figure 1b. These networks use multiple output nodes to generate equidistant and generalized median estimators of the solution, respectively, and thus provide a finite discretization of the solution distributions. Generally, this is sufficient since the solution is often represented by a finite number of samples from a Monte Carlo analysis. We show how these networks emulate the distribution of a single parameter and conclude with a discussion of how this can be generalized to multiple parameters.

2.2.1. Histogram network. First, we discuss the histogram network, which returns an equidistantly sampled approximation to the df $P(m|\mathbf{x})$. We consider the case of a scalar m and apply the following operator to discretize its values using M segments with lengths Δm:

$$W_j(m) = \begin{cases} 1 & j\Delta m < m < (j+1)\Delta m \\ 0 & \text{otherwise} \end{cases}. \tag{8}$$

Then define the local error for training sample $(m, \mathbf{x})$

to be

$$\varepsilon[u_j(\mathbf{x}), W_j(m)] = [u_j(\mathbf{x}) - W_j(m)]^2. \quad (9)$$

The M outputs $u_n(\mathbf{x})$ of the optimally trained network which minimize this error for a given input x can be inferrred from the relation $\partial E/\partial u_n(x) = 0, n = 1, ..., M$. With E defined as in equation (2) this gives:

$$u_n(\mathbf{x}) = \int_{m_0+n\Delta m}^{m_0+(n+1)\Delta m} P(m|\mathbf{x})dm \simeq p_n(\mathbf{x}). \quad (10)$$

For each set of inputs $\mathbf{x}$, the trained network has M outputs $u_j(\mathbf{x})$ which return the probabilities (not densities) $p_j(\mathbf{x})$ that m takes a value in the jth window of width Δm. With $M = 2$, we obtain a classical application of networks, namely, the classification of an input into one of two states [*Dowla et al.,* 1990; *McCormack et al.,* 1993]. Note that the solution satisfies the constraint

$$\sum p_j = 1; \quad (11)$$

hence one output node is redundant and one output would suffice in this particular example. However, this relation is better used to check consistency of the solution. A simple application of a histogram network is shown in Figure 2b.

2.2.2. Median network. Next, we consider the median network, which contours the topography of the cumulative distribution of $P(m|\mathbf{x})$. Again we use M output nodes to obtain information about a scalar m given an input $\mathbf{x}$. The error measure is now defined to be

$$\varepsilon(u_j(\mathbf{x}), W(m)) = \begin{cases} u_j(\mathbf{x}) - m & u_j(\mathbf{x}) < m \\ 0 & u_j(\mathbf{x}) = m \\ -c_j(u_j(\mathbf{x}) - m) & u_j(\mathbf{x}) > m \end{cases} \quad (12)$$

for some constants c_j. Using this local error, the optimally trained set of $u_j(\mathbf{x})$ for which $\partial E/\partial u_n(x) = 0$ is given by:

$$\int_{-\infty}^{u_m(x)} P(m|\mathbf{x})dm = \frac{c_m}{1 + c_m}. \quad (13)$$

Hence if we set $c_j = j/(M + 1 - j)$, the output values $u_j(\mathbf{x})$ subdivide the df $P(m|\mathbf{x})$ into $M + 1$ equal areas (the two outer areas are only bounded on one side). For $M = 1$, the value of $u_1(\mathbf{x})$ equals the median of the distribution. Areas of high probability are subdivided more densely by $u_j(\mathbf{x})$ than areas of low probability. Fixed outer bounds can be defined separately for the two outermost areas, but it is preferable to let the network approximate these by setting $c_1 = 0.001$ and $c_M = 999.0$ which approximately give a lower bound and an upper bound that separate regions of nonzero and near-zero probabilities. The median network is preferable to the histogram network if the distribution has sharp peaks which need to be mapped in more detail. Note that to emulate a distribution, the histogram and median networks do not require smoothing over the inputs, unlike the smoothing applied in preprocessing when using a sonnet, and that by default the whole model space is sampled. A simple application of a histogram network is shown in Figure 2c.

2.2.3. Multidimensional distributions. For a more general problem with N model parameters $m_1 ... m_N$, a histogram network requires M^N output nodes to describe the solution; however, this is not possible for the median network. This approach leads to an explosive growth of the number of required outputs; the output vector becomes extremely sparse, which makes training quite difficult. Now consider the following decomposition:

$$P(m_1, ..., m_N|\mathbf{x}) = P_1(m_1|\mathbf{x})P_2(m_2|\mathbf{x}, m_1) \ ... \ P_N(m_N|\mathbf{x}, m_1, ..., m_{N-1}). \quad (14)$$

This can be approximated using N separate histogram or median networks, each with M outputs dedicated to one of the distributions P_n. Equation (14) introduces a continuous mapping similar to the one in the sonnet approach which may be poorly approximated by the networks in regions with insufficient training samples.

2.3. Training Algorithm

The optimization of the weights and offsets of the network itself represents an inverse problem. For this purpose we use the backpropagation rule, which is an iterative learning algorithm which uses values of the local error to perform approximately gradient descent of the global error in equation (2) [*Hecht-Nielsen,* 1991]. However, the global error is a highly nonlinear function of the weights and offsets, and in its pure form the backpropagation algorithm is almost sure to end up in a local minimum. Furthermore, the local error can never become zero if conflicting samples are presented to the network during training, which means that the values of the weights do not converge. Since we train on distributions of samples, this problem occurs in all our applications unless we make the step from the local to the global error.

A momentum applied to the weight updates during training may be used to smooth the error function and hence to avoid local minima [*Hecht-Nielsen,* 1991]. If we increase the momentum during training, we create a long running average over the local error and we could achieve convergence in the presence of conflicting data. Training can also be stopped when a state of smallest rms error over some independent data set is reached [*Jarvis and Stuart,* 1996] which is a costly procedure, since the independent data set needs to be comparable in size to the training data set. We use a momentum term and additionally force convergence by scaling the local error by a factor α, which has a value of 1 at the

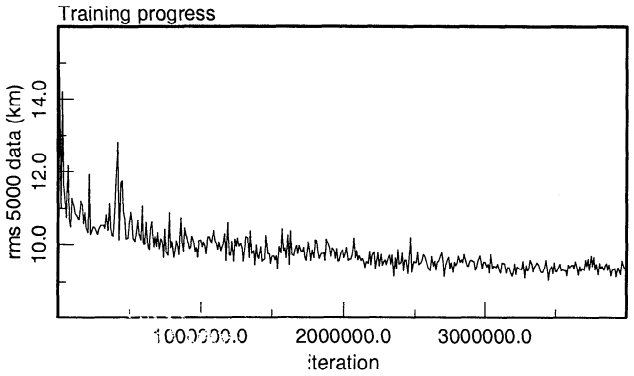

Figure 3. Development of the rms error of a sonnet during training. Each rms error is calculated over 5000 samples randomly selected from the training data set.

start of training and 0 after $i_{\max}$ iterations:

$$\alpha = \left(\frac{i_{\max} - i}{i_{\max}}\right)^2. \quad (15)$$

The value for $i_{\max}$ is determined by trial and error. Although this empirical rule cannot guarantee that a state of smallest error is reached, the chaotic development of the error (Figure 3) during training indicates that the network jumps out of local minima, probably due to conflicting updates in subsequent iterations. As the value of α is gradually reduced, the oscillations become smaller and the overall error decreases.

Networks trained on similar data are different but comparable in performance. They represent different samples from the space spanned by the weights and offsets that yield identical fit to the training data.

3. Bayesian Approach to Solving for Crustal Thickness.

We aim to invert for crustal thickness from dispersion data of surface waves. The inverse solution will be constructed in a probabilistic framework, using density functions (df's) [*Tarantola and Valette,* 1982].

3.1. Data and Modeling

We will constrain Earth structure by using dispersive surface wave data. These data are parameterized by a set of N_D distinct velocities $\mathbf{s} \equiv \{s_j\}, j = 1,..,N_D$ at different periods. Phase velocities are used at periods of 30, 40,..., 90, 100 s [*Curtis et al.,* 1998] and group velocities at periods of 10, 15, 20, 25, 30, 40, ..., 90, 100 s [*Wu and Levshin,* 1994]. An upper limit of 100 s is chosen since dispersion at greater periods is hardly sensitive to Earth structure around the Moho. These data were obtained at each period by a linearized inversion for discretized regional dispersion curves using measurements over many paths with different source-receiver geometries across the region under study. We assume that they represent "noiseless" data s_j, which have uncertainties due to noise in the raw data used to create the maps, and that these uncertainties can be described by a normal distribution with average value r_j (given by the values of the velocity maps) and a standard deviation Δ, which we take to be constant over all periods:

$$\rho(\mathbf{s}|\mathbf{r},\Delta) = c\exp\left(-\sum_{j=1}^{N_D}(s_j - r_j)^2/2\Delta^2\right). \quad (16)$$

We represent uncertainties in all velocities with a fairly high standard deviation of $\Delta = 0.1$ km/s, estimated by *Curtis et al.* [1998]. The data may be biased by noise with nonzero mean and may contain additional errors due to limitations in resolution in some areas. However, these contributions cannot be quantified and will introduce an unknown error in the solution.

It is worth noting that in principle, group and phase velocities carry the same information, although group velocities are more sensitive to the shallow structure. Since a larger part of the signal is affected by the crustal structure, the latter type of data will constrain Moho depth better in the presence of noise. The two are related by

$$U(T) = \frac{c(T)}{1 + \frac{T}{c(T)}\frac{dc(T)}{dT}}, \quad (17)$$

where T is period, $U(T)$ is group velocity, and $c(T)$ is phase velocity. This explains why the sensitivity kernels (Figure 4) of group and phase velocity of a particular mode are fairly similar. However, Love waves are sensitive to horizontally polarized SH velocity, which is fundamentally different from the Rayleigh waves, which are sensitive to both P and vertically polarized SV velocities. The sensitivity kernels show that Rayleigh phase velocities are more affected by the velocity structure at Moho depth than Love velocities. Hence the Rayleigh wave data may be more robust for the identification of the crustal thickness if noise is present. Nevertheless, Love waves do yield information, especially about the upper crustal velocity structure, and therefore a combination of the two data types is expected to constrain Moho depth most effectively (assuming $V_{SH} \simeq V_{SV}$ to a first isotropic approximation).

Noiseless synthetic data $\mathbf{s}$ are calculated for a given model $\mathbf{m}$ using normal mode theory $\mathbf{G}$ [*Mendiguren,* 1977], which is represented by the conditional df $\theta(\mathbf{s}|\mathbf{m})$:

$$\theta(\mathbf{s}|\mathbf{m}) = \delta[\mathbf{s} - \mathbf{G}(\mathbf{m})]. \quad (18)$$

The physics of the problem (modeled using $\mathbf{G}(\mathbf{m})$) are such, that crustal thickness can not be uniquely constrained by dispersion data, since each part of each dispersion curve contains information about an average of the velocity structure of the Earth (Figure 4). In principle, if data are available over a wider range of periods, this nonuniqueness can be reduced.

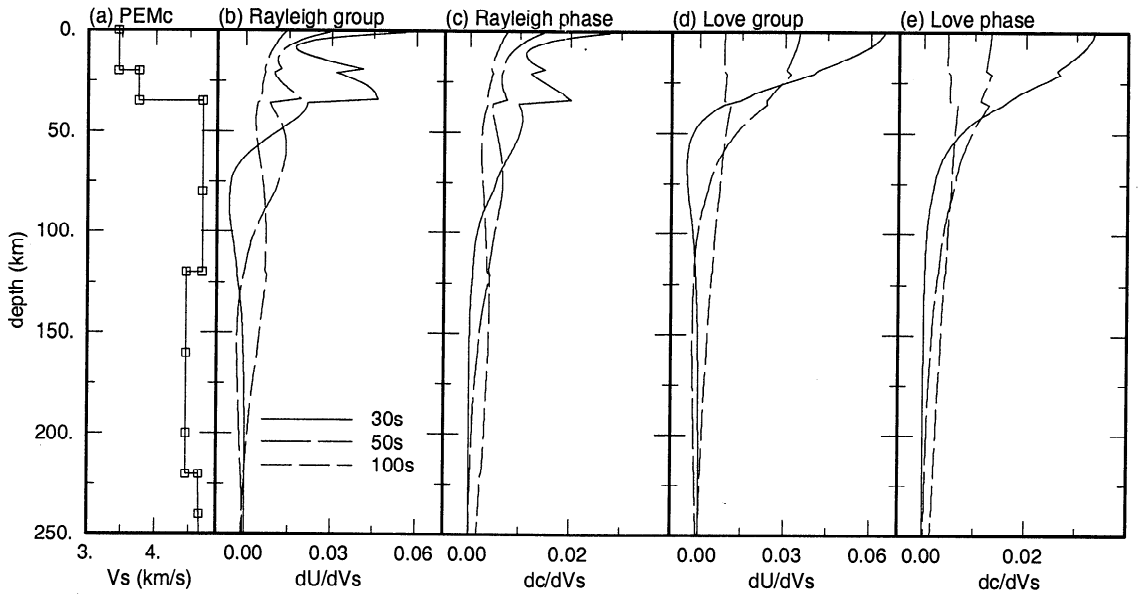

Figure 4. (a) The shear velocity V_s of reference model PEMc. (b)-(e) Sensitivity kernels of group (U) and phase (c) velocities of Rayleigh and Love waves for several periods. V_s was perturbed around each depth by 0.2 km/s in a layer 1 km thick. Compressional velocity V_p was perturbed such that the ratio V_p/V_s of the PEMc model is preserved. The resulting fractional changes dc/dV_s and dU/dV_s in phase and group velocities are shown.

We parameterize Earth structure by a set of parameters $\{m_j\}$ (Table 1). Models are constrained by data and theory but also by prior information, which is represented by the df $\rho_M(\mathbf{m})$. We define prior information of each parameter by a uniform probability density centered around the PEMc model [*Dziewonsky et al.*, 1975] with values given in Table 1. We impose the additional constraint that the Moho is defined by a single interface, with a shear velocity jump of at least 0.3 km/s. In the true Earth, however, the transition between crust and mantle may involve a transition zone with a gradual velocity variation. Since our data will barely allow us to resolve these geometrical differences due to the width of the averaging kernels in Figure 4, inversions will locate the Moho at some average depth within this zone and, indeed, nearly at halfdepth in case of a linear gradient. Acceptable Moho depths are between 10 km, above which the data have identical sensitivity kernels, and 70 km, which we take to be the maximum depth in the Tibetan region. Furthermore, to keep the problem tractable, we assume a constant V_p/V_s ratio. In areas where this ratio is significantly perturbed, Rayleigh inversion (which depends on both shear and compressional velocities) may yield biased results. In addition, inversion of a combination of Love and Rayleigh dispersion curves may be biased by the effect of transverse anisotropy [*Muyzert*, 1998] which is not included in the models of our training set. We expect errors due to these assumptions to be small compared to uncertainties arising from trade-offs in the physics of the problem.

Table 1. Model Parameterization Consisting of Seven Velocities at Different Depths and Variable Thicknesses for the Crust (H_{crust}) and the Sediment Layer (H_{sed})

Interface	H Midvalue, km	Half Width, km
H_{sed}	2.5	2.5
H_{crust}	40.0	30.0
Interface, km	V_s Midvalue, km/s	Half Width, km/s
0	1.5	0.50
H_{sed}	3.45	0.75
H_{crust}	3.75	0.75
H_{crust}	4.69	0.75
100	4.69	0.75
150	4.46	0.75
250	4.66	0.75

Prior information used for depth and shear velocities consists of independent boxcar-like distributions, with central values equal to the PEMc model [*Dziewonsky et al.*, 1975] and halfwidths as given. The V_p/V_s ratio, density, and Q factors of the PEMc model are preserved.

3.2. Forming the Solution

The dependency between model $\mathbf{m}$ and the noiseless data $\mathbf{s}$ is formed by using Bayes theorem (which applies to conditional df's). To assess the combined information carried by two independent unconditional df's, two measures are often used, the entropy [*Rietsch*, 1977] and the conjunction [*Tarantola and Valette*, 1982; *Tarantola*, 1987]. In a geophysical context the latter is normally used. We integrate out $\mathbf{s}$, which forms a link between $\mathbf{r}$ and $\mathbf{m}$ but provides no independent information. The solution is then defined as

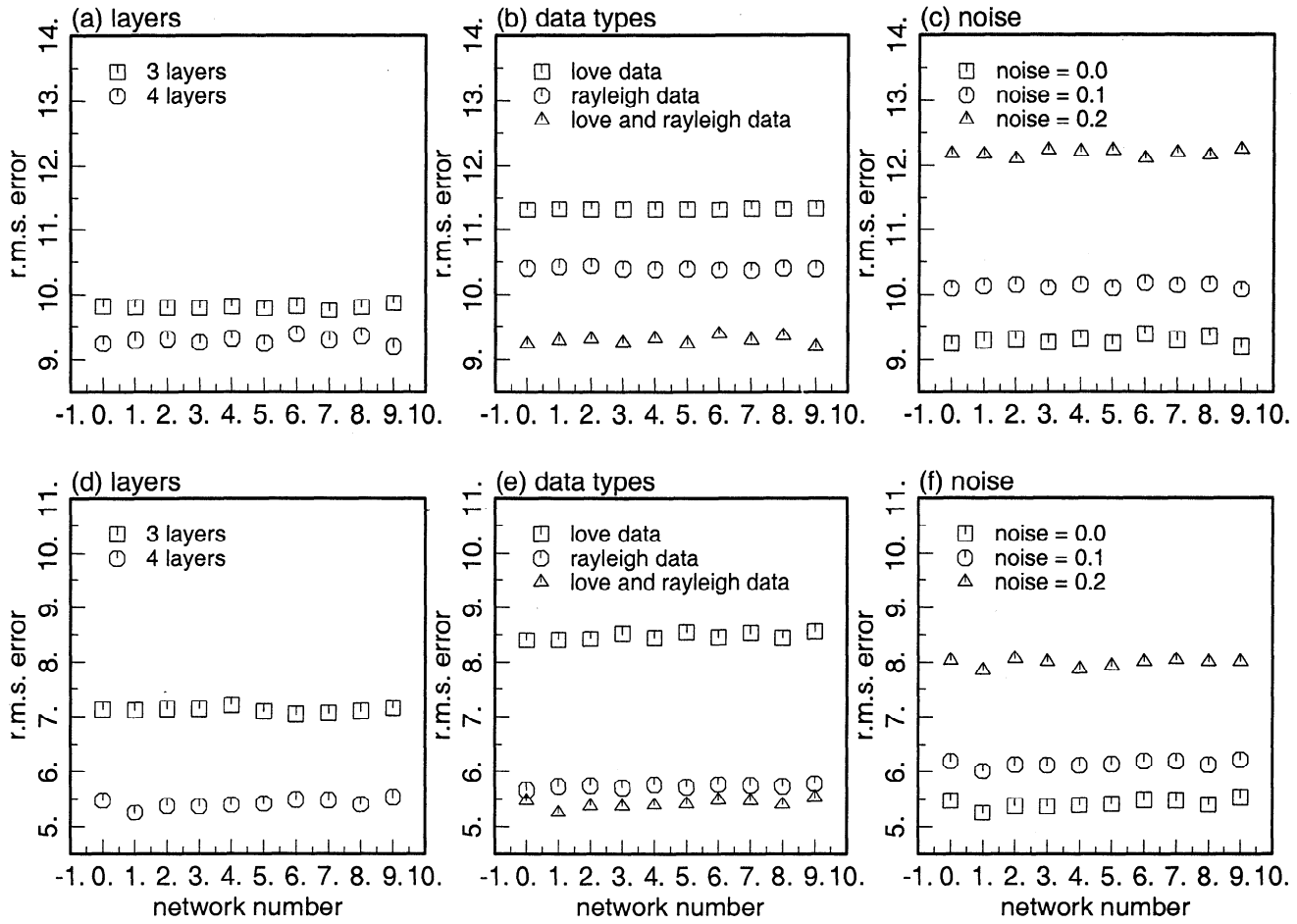

Figure 5. Sonnets were applied (a)-(c) to synthetic phase velocities for periods in the range 30-100 s and (d)-(f) to group velocities for periods in the range 10-100 s. We varied the complexity of the network (Figures 5a and 5d), the types of input data (Figures 5b and 5e), and the amount of white noise added to the input data (Figures 5c and 5f). The rms errors are calculated with respect to the independent test data (see text).

$$\sigma(\mathbf{m}|\mathbf{r}) = \int \frac{\rho(\mathbf{s}|\mathbf{r})\theta(\mathbf{s}|\mathbf{m})\rho_M(\mathbf{m})}{\mu_s(\mathbf{s})} d\mathbf{s}. \quad (19)$$

Using relations (16) and (18) this yields

$$\sigma(\mathbf{m}|\mathbf{r}) = c \exp\left(-\sum_{j=1}^{N_D} \frac{[G_j(\mathbf{m}) - r_j]^2}{2\Delta^2}\right) \rho_M(\mathbf{m}), \quad (20)$$

where the constant c provides some scaling factor, which is not important for the construction of the df. The nulldistribution has shape $\mu(v) = 1/v$, where v represents a velocity [*Tarantola and Valette,* 1982], and has been assumed to be approximately constant for our range of velocities. We may form a solution $\sigma(H|\mathbf{r})$ for crustal thickness H by integrating out all other model parameters other than H in equation (20).

Owing to the complex form of $G(\mathbf{m})$, an analytical solution to equation (20) is not available. We proceed by randomly sampling the space of possible model/data pairs by (1) drawing a model $\mathbf{m}$ from the prior distribution, (2) calculating the corresponding synthetic datum $G(\mathbf{m})$ using the assumed physics, (3) reproducing the average $\mathbf{r}$ using the normal distribution in equation (20), and (4) removing all parameters except H from the sample. These samples can then be used to train neural networks to provide statistics about, or to approximate, $\sigma(\mathbf{m}|\mathbf{r})$. For this purpose, we create a data set of 400K (400,000) models and their corresponding dispersed velocities. We will only invert for crustal thickness and do not require perfect resolution of the solution; hence this number of samples is satisfactory.

4. Neural Network Approach to Solving for Crustal Thickness.

We will use the networks introduced in section 2 to emulate function $P(H|\mathbf{c}_L, \mathbf{c}_R)$ or $P(H|\mathbf{u}_L, \mathbf{u}_R)$. These return the df of depth H of the Moho discontinuity given a set of Love or Rayleigh phase velocities $\mathbf{c}_L, \mathbf{c}_R$ or group velocities $\mathbf{u}_L, \mathbf{u}_R$. In this section we investigate the performance of network inversion using only synthetic velocities.

4.1. Choosing a Network for Inversion

We train sets of sonnets with three and four layers (a layer refers to a group of neurons) for 4M (4×10^6) iterations to return the average value of crustal thickness. Afterward, we check the approximation error by applying the networks to an independent test data set of 50K models. These latter models are varied down to a depth of 520 km to include possible effects of variations in deep earth structure in the data. They contain more complexity than the training set models in order to check that the networks interpolate correctly between the training models and that the simpler training models contain sufficient complexity to represent all possible trade-offs between crustal thickness and velocity.

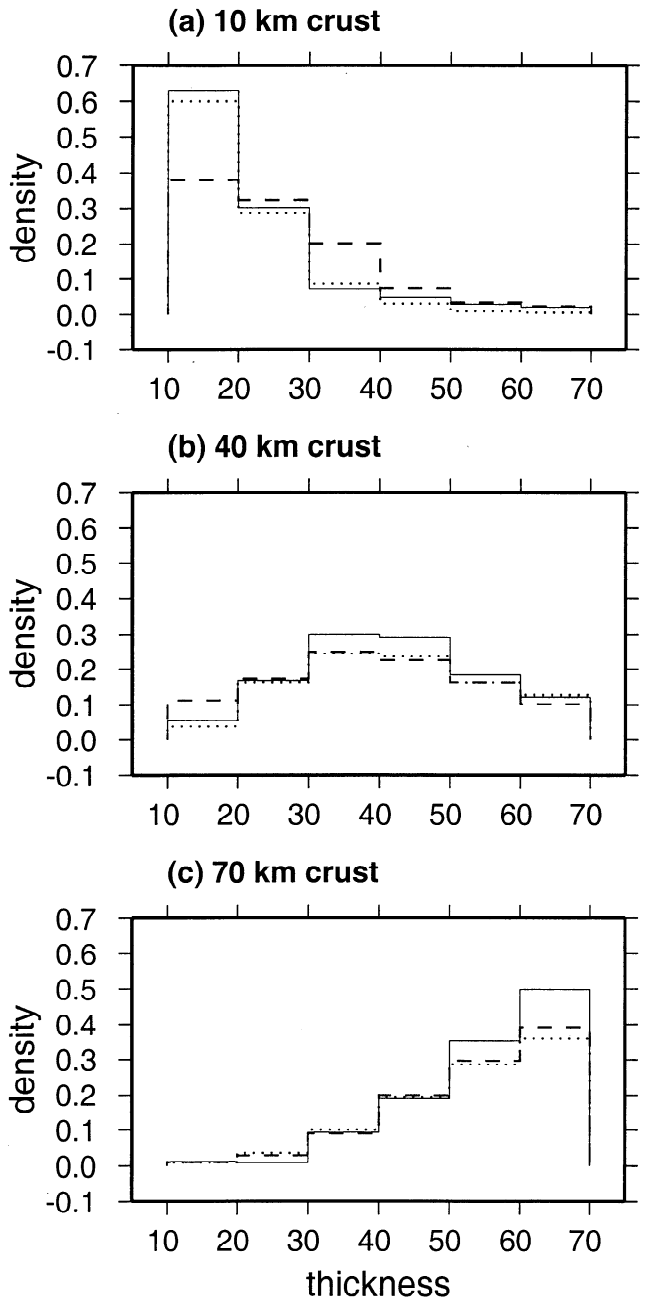

Figure 6. Histogram networks were applied to synthetic phase velocities in the period range of 30-100 s, calculated for PEMc models with adapted crusts of (a) 10 km, (b) 40 km, and (c) 70 km. The complete solutions of crustal thickness are shown for inversion of combined Love and Rayleigh data (solid), Rayleigh data (dashed), and Love data (dotted).

Figure 5 shows that sonnets with four layers, applied to combined Love and Rayleigh phase velocities in the period range of 30 s to 100 s, perform better than ones with only three layers. This indicates that the inverse relation is quite nonlinear. If the data contain white noise, the errors made by the networks increase and the advantage of increased complexity decreases, since noise obscures rapid variations in the function to be mapped. We also see that the training process is stable, since each independently trained network of similar architecture yields approximately the same rms error; that is, the complexity of the neural network is the only parameter influencing the accuracy of the mapping. Networks with four layers will be used for all subsequent inversions.

For phase and group velocity data, input curves have 8 and 12 samples respectively, and either Love, Rayleigh, or both sets of velocities are input to the network. Let us denote the number of input nodes by N. We use M outputs to provide statistics about the df of crustal thickness, which we then use to approximate the df. For the sonnets which produce only average model values we use relatively simple networks with two sets of 15 hidden nodes, which we write as $(N, 15, 15, M)$; the nodes in subsequent layers are fully connected. The histogram networks have to perform a more complicated mapping and thus receive a more complex structure, namely $(N, 60, 120, M)$, with one modification: to reduce the (huge) number of weights, we allow each of the 120 nodes in the last hidden layer to connect only to three nodes in the output layer.

Sonnets produce a single parameter of the distribution and can, in principle, be trained in a relatively small number of 1M iterations. To train a histogram network which approximates the distribution of crustal thickness with a resolution of 10 km (i.e., using six output nodes), we need about 4M iterations. If the resolution is increased to about 5 km, we find that we need many as 12M training iterations. This increase is due to the lower density of sampling of each thickness interval; that is, we need a larger number of iterations to find an equal number of samples per interval.

4.2. Representing the Solution

To investigate the asymmetry of the posterior df of crustal thickness, we invert synthetic phase velocity curves which are derived from PEMc velocity structure but with three different crustal thicknesses (Figure 6). For this purpose we use a histogram network and a discretization interval of 10 km to approximate the solution. The solutions are asymmetric for the 10 and 70 km crusts, which is partly caused by prior constraints

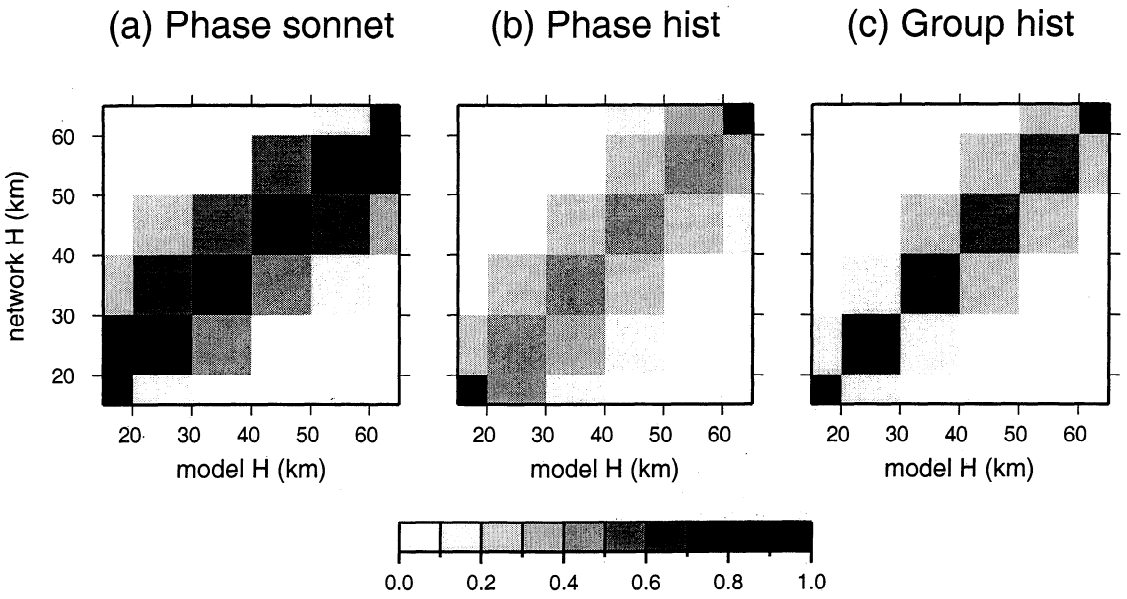

Figure 7. Estimates of crustal thickness from synthetic dispersion data from the training set are shown against the crustal thicknesses of the underlying models. The number of occurrences of pairs of values in 10 x 10 km ranges for 25K comparisons is contoured. (a) Estimations of a sonnet for phase velocities in the period range of 30-100 s. (b) Maximum likelihood crustal thicknesses obtained from the outputs of a histogram network. (c) Last experiment, repeated for group velocities in the range of 10-100 s.

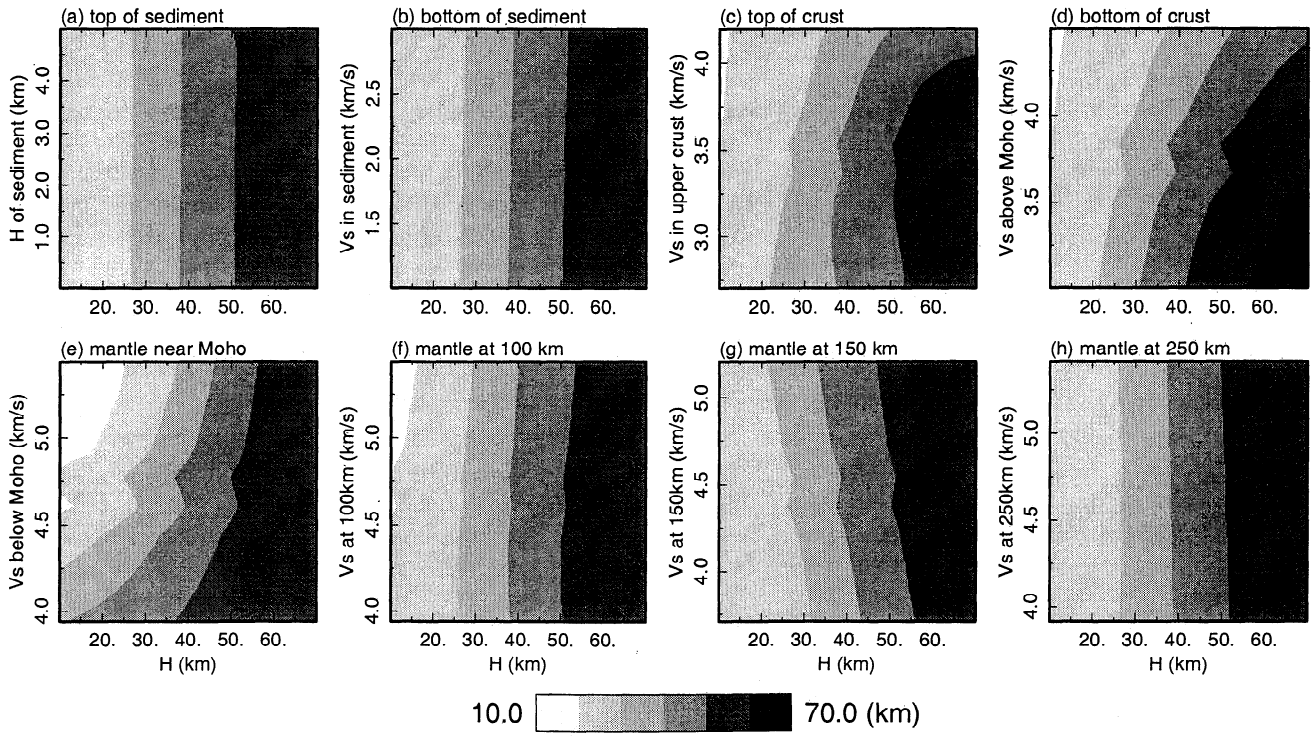

Figure 8. Phase velocities in the period range 30 s to 100 s calculated for 10 x 10 combinations of the pairs of model parameters indicated on the horizontal and vertical axes, while keeping all other parameters fixed at values given in Table 1. The thickness estimates from an averaging sonnet for these data are contoured.

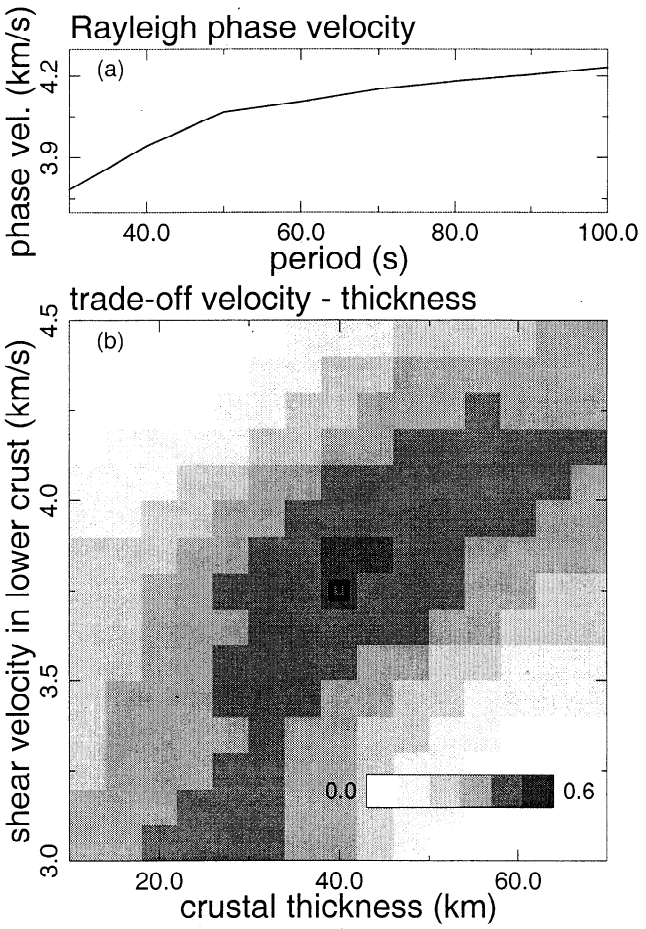

Figure 9. (a) Synthetic Rayleigh phase velocities calculated for the model of Table 1. We trained two histogram networks on the probability of finding crustal thickness given (1) the Rayleigh phase dispersion curve **d**, and (2) both **d** and the shear velocity in the lower crust. (b) Multiplication of the two distributions, giving a contour plot of the trade-off between crustal thickness and the shear velocity in the lower crust. The values of the model used are indicated by the box.

on the models (the model space is cut at the 10 and 70 km bounds) and partly by the physics (distributions decrease sharply to 0 for the 10 km thick case).

We invert Rayleigh phase velocities (30-100 s) for crustal thickness and investigate uncertainties in the crustal thickness estimations based on two definitions of the solution. First, we define the best model by the average crustal thickness, given all possibilities. Second, we define the best value by the crustal thickness that is most likely (the maximum likelihood value). Figures 7a and 7b show diagnostic plots created by inverting all training data for these two definitions and showing them as a function of model crustal thickness. The estimated crustal thicknesses are clearly correlated to the model thicknesses, although the average thickness shows a bias at the smallest and largest values due to asymmetry of the solution in these cases. The experimental standard deviations are of the order of 10 km and are reduced to 5 km when group velocities are used over a wider period range (Figure 7c).

Owing to the asymmetry, the values of average crustal thickness predicted by a sonnet may be very different from the maximum likelihood estimates obtained from a histogram network. We prefer the maximum likelihood models since these are related to the best observed datum; however, this is a matter of taste; the only unbiased way of viewing the result is by looking at the whole distribution for each datum.

4.3. Trade-offs between Model Parameters

We examine the trade-off between crustal thickness and other model parameters using a sonnet. This network inverts synthetic Love and Rayleigh phase velocities for average crustal thickness. The data are in the period range of 30 s to 100 s and are calculated from various models having 10 x 10 combinations of two free parameters while keeping all other parameters fixed at values of Table 1. The crustal thicknesses obtained from inverting these data are contoured in Figure 8. In the absence of trade-off the estimates would be perfectly correlated to crustal thickness, giving vertical contours as is approximately the case for the sediment layer and the deep mantle. The estimated thicknesses themselves are biased since the estimates of the averaging sonnet are sensitive to asymmetry of the distribution. As expected, crustal thickness estimates trade off most strongly with velocity variations just above and just below the Moho. At 150 km depth we see trade-offs in the opposite sense and at 250 km depth we see no correlation with crustal thickness estimates, since the sensitivity kernels (Figure 4) that constrain the crust hardly sample this depth range.

Next, we examine the trade-off in the inversion between the two parameters H, crustal thickness, and v, the shear velocity just above the Moho (Figure 9), given a single set of synthetic Love and Rayleigh phase velocities $\mathbf{x}_{\mathrm{PEMc}}$ calculated from the PEMc reference model with 40 km crust. This is done using two histogram networks which approximate

$$P(H, v|\mathbf{x}_{\mathrm{PEMc}}) = P_1(H|\mathbf{x}_{\mathrm{PEMc}})P_2(v|\mathbf{x}_{\mathrm{PEMc}}, H) \quad (21)$$

similarly to equation (14), using 10-km intervals. The first network is trained to emulate $P_1(H|\mathbf{x}_{\mathrm{PEMc}})$. The second network emulates $P_2(v|\mathbf{x}_{\mathrm{PEMc}}, H)$, i.e., H is provided as an additional input. The contours show the likelihoods for different combinations of H and v. Figure 9 shows that crustal thickness trades off almost linearly with velocity structure. Note that the average solution of crustal thickness from this experiment produces a single estimation in the set of tests in Figure 8. For this particular datum, the maximum likelihood point coincides with the parameter values of the input model (shown as a box).

Thus we have shown that neural networks can be used to describe trade-off between crustal thickness and velocity structure. This trade-off causes the uncertainties in crustal thickness estimates.

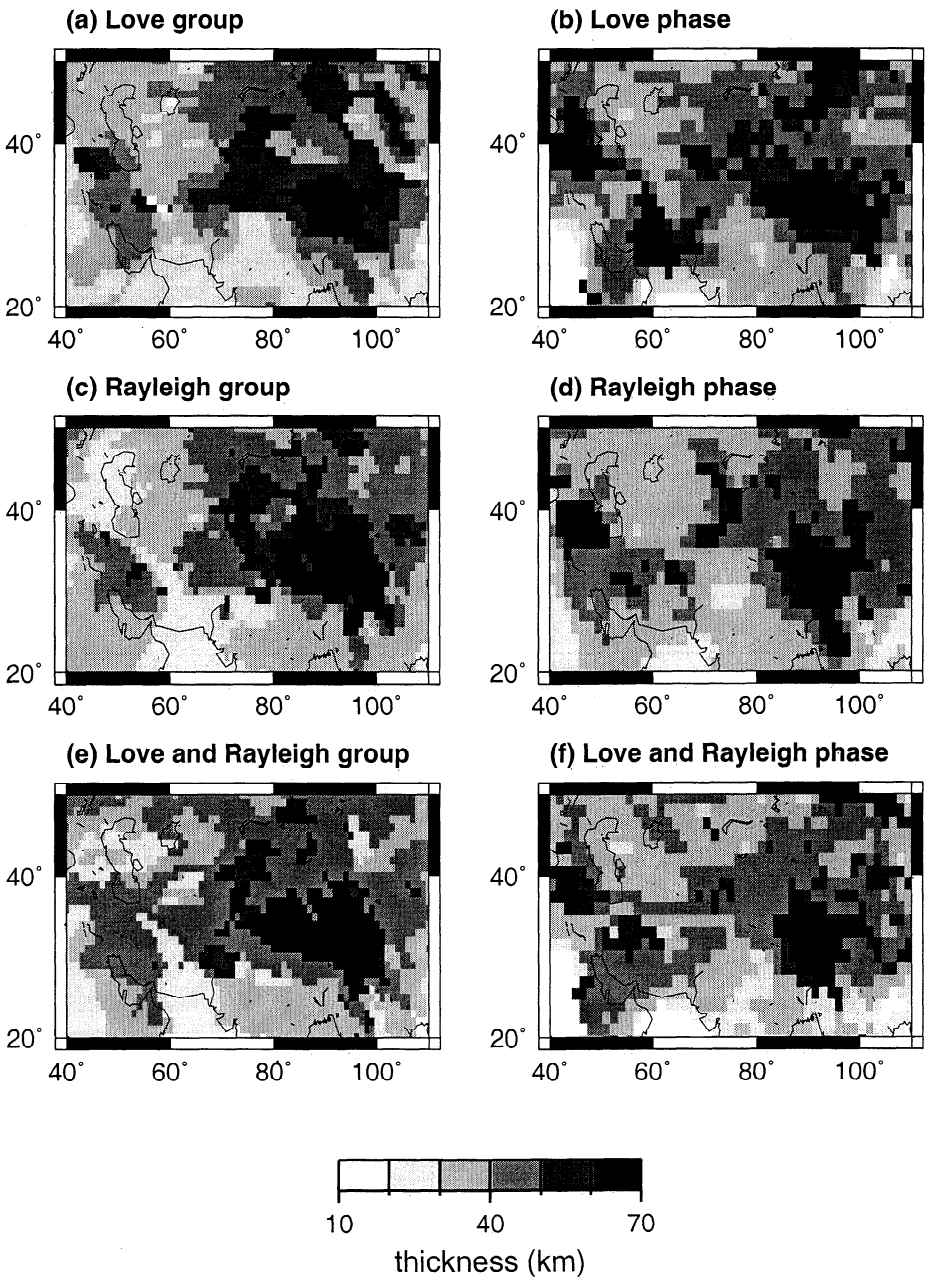

Figure 10. Maximum likelihood crustal thickness maps obtained using a histogram network. Results (a), (c), and (e) obtained from inversion of the group velocities of data set I (see text). Results (b), (d), and (f) obtained from network inversion of the phase velocities of data set II (see text).

4.4. Efficiency of NNI

Creating a training set of 400K samples of the posterior distribution took 500 hours on a Sparc 20. However, this is an aspect of the inversion process which needs to be done only once and is equivalent to the time of a single run of a Monte Carlo inversion using a similar number of examples. Training of a sonnet on these data (plus added noise) during 1M iterations took 1 hour on a Pentium II 350 Mhz; training the most complex histogram network during 12M iterations took 20 hours. Again, this part of the procedure needs to be done only once for each network and may be improved by using a more efficient training algorithm (which falls outside the scope of this paper). The network then emulates the distribution constrained by the samples of the Monte Carlo run.

Applying a network to phase velocities to invert for average crustal thickness or a crustal thickness distribution takes a fraction of a second. This illustrates that a trained network is an extremely fast inversion tool, which can be reapplied almost instantaneously to any new data set and which in principle provides arbitrarily detailed information about the inverse problem solution. In contrast, an adaptive Monte Carlo method would require a full inversion for every da-

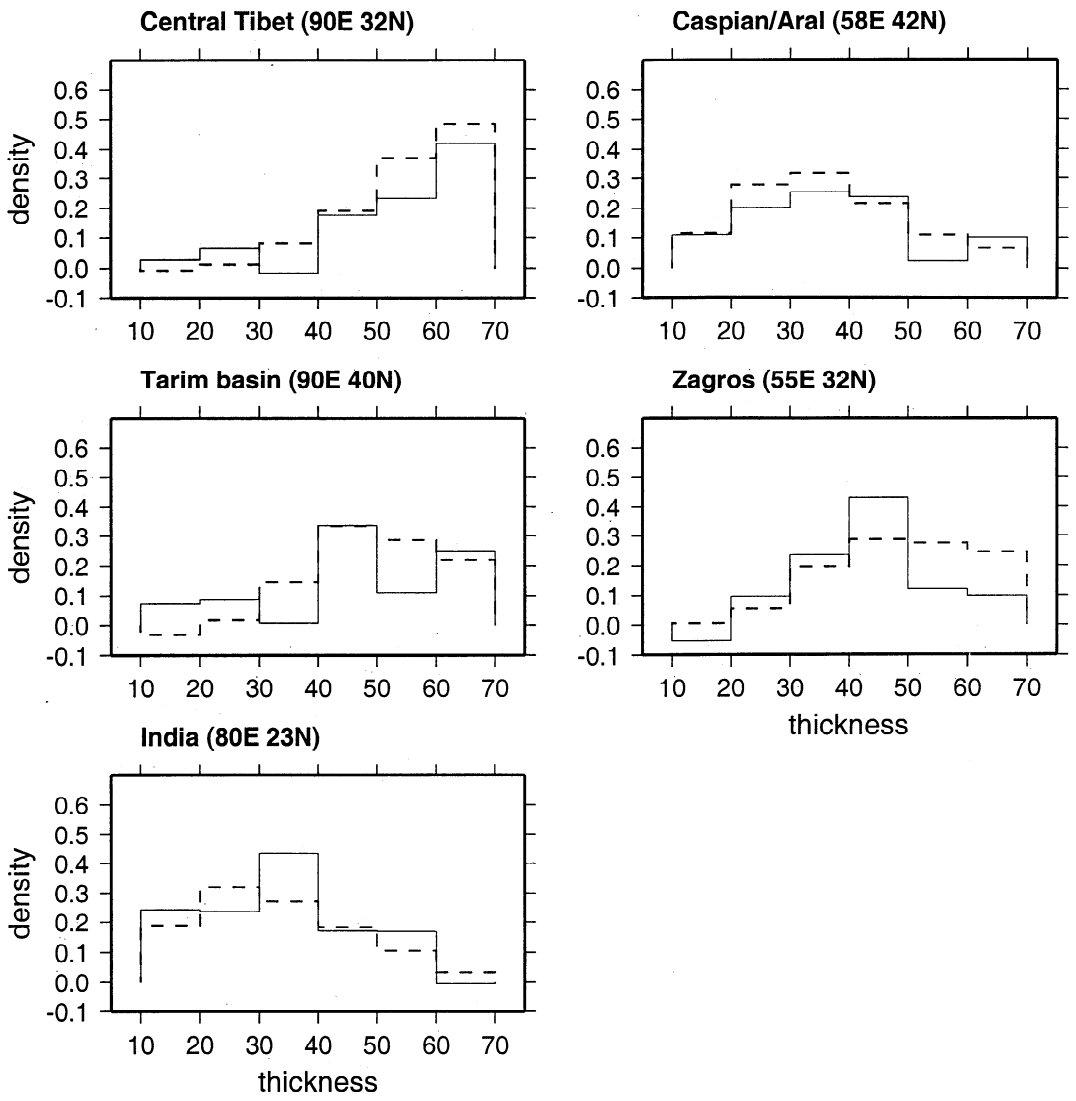

Figure 11. The full crustal thickness distributions for several coordinates in the Tibetan area, given combined Love and Rayleigh data from sets I (solid) and II (dashed). A set of inversions was performed; only representative solutions are shown.

tum, which is extremely expensive. Nonadaptive Monte Carlo methods would require millions of samples (thousands of hours) to obtain the same level of precision as the networks, since the latter interpolate between training samples. Although this calculation would also need to be done only once, comparing a single newly measured datum to this massive database of samples to obtain a solution would itself take tens of minutes on the same machine. NNI (with trained networks) is also about a factor of 10K faster than linearized inverse procedures and has the additional advantage that the solution is independent of a reference model (although not independent of other types of prior information).

5. Application of NNI to the Continent of Eurasia

Now we invert real phase velocity data for crustal thickness using two data sets: (1) the set of group velocity data having a period range of 10-100 s and covering the area around Tibet [*Wu and Levshin,* 1994] (now referred to as data set I), and (2) the set of phase velocity data in the period range 30-100 s across Eurasia [*Curtis et al.,* 1998] (now referred to as data set II). Thus the effectiveness of different data types to constrain crustal thickness can be assessed. A crustal thickness map of Eurasia based on data set II is presented.

5.1. Inversion of Group and Phase Velocities for a Small Region of Eurasia.

We use a histogram network to invert the data from set I for the df of crustal thickness, using a discretization interval of 10 km. The maximum likelihood crustal thicknesses are extracted from the full solution obtained from inverting the Love data (Figure 10a), the Rayleigh data (Figure 10c), and paired Love and Rayleigh data (Figure 10e). The inversion based on the Love wave data shows a band of thick crust outlining the Pamir, Tien Shan, and the Altai mountains.

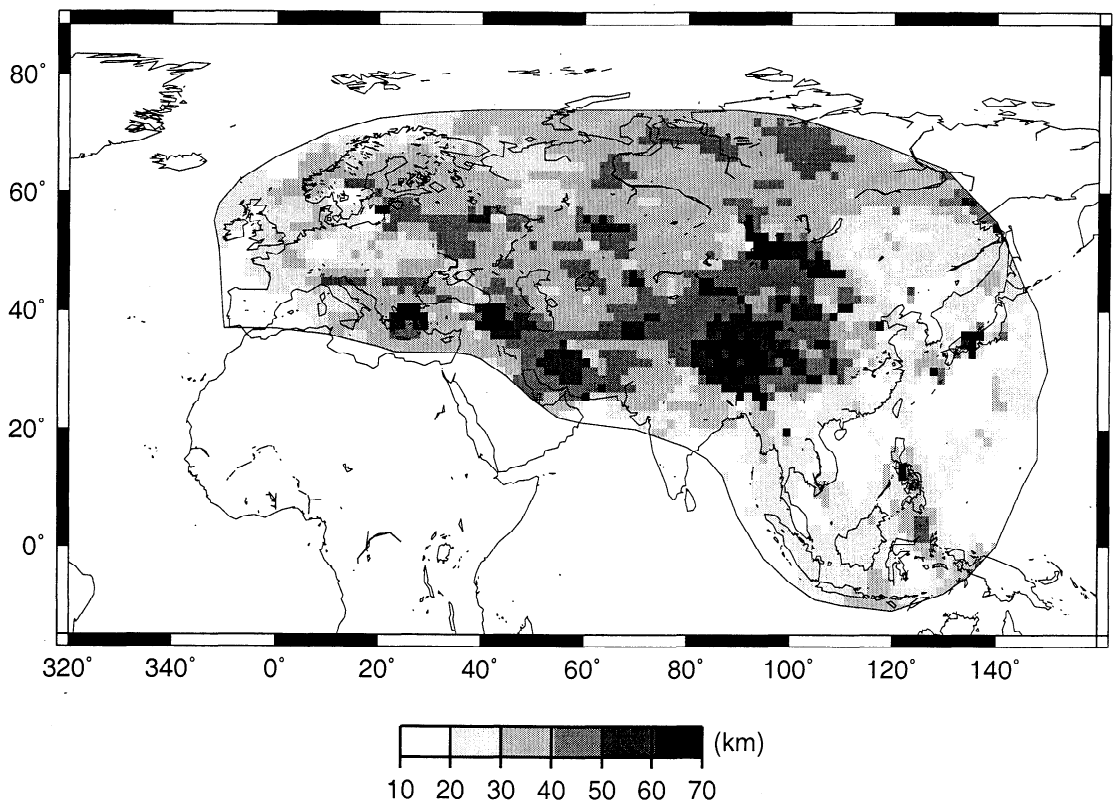

Figure 12. Maximum likelihood crustal thickness map across Eurasia obtained from the inversion of combined Love and Rayleigh wave data of data set II with a histogram network. The discretization interval of the solution is 10 km. Only areas in which phase velocities are resolved at length scales of < 1000 km are shown.

The maximum likelihood crustal thicknesses, obtained from the inversion of combinations of Love and Rayleigh data from data set II with a histogram network, are shown for the same region in Figures 10b, 10d and 10f. This data set has less spatial resolution and a shorter period-range; hence only the major features can be identified reliably.

The two inversions of the two data types agree on the main pattern of crustal thickness. However, results from Love and Rayleigh data show strikingly different thickness estimates in the mountaineous regions. Given that trade-offs such as those in Figure 8e will vary between different data types, this may be indicative of the presence of very low shear velocities in the upper mantle.

The full solutions of crustal thickness obtained by histogram networks are shown in Figure 11 for several locations within the region of Figure 10. Solutions obtained from a set of identically trained networks are only slightly different, and a representative result is shown. The group velocity data, which have a broad period range, constrain shallow crusts better than the phase velocity data but yield similar results for thicker crusts. The distributions are very asymmetric in some cases and show large uncertainties. We emphasize that the maximum likelihood map gives a best solution for Earth structure but shows rather limited information about the complete solution. The results for the Tibetan Plateau, Tarim Basin, and Indian Shield, obtained from data set II, correspond very well to the estimates of *Curtis and Woodhouse* [1997], who also used interevent phase velocity data. The group velocity data do not resolve the crust in the Tarim Basin, which indicates that the data contain significant noise there.

5.2. Inversion of Phase Velocities for the Whole of Eurasia.

A histogram network has been used to invert combined Rayleigh and Love phase velocities from data set II across the whole of Eurasia. The most likely crustal thickness is extracted from the full solution (discretized at 10 km) and plotted at each coordinate in Figure 12. The continental crustal thicknesses lie in the range 41 $\pm$ 6.2 km given by *Christensen and Mooney* [1995] for the global average continental crustal thickness. Some areas, such as Greece, Zagros, and some places in the Pacific, show anomalously large maximum likelihood crustal thicknesses. We show the complete distributions in Figure 13. In Greece, there is no unambiguous preference for a crustal thickness between 20 and 70 km. The same holds for the Pacific, where any value between 10 and say 50 km seems reasonable. In the Zagros region the inversion using Love waves is significantly different from the inversion using Rayleigh waves. This may indicate that our assumption that P and S velocities are coupled is wrong. Other regions such as Tibet and

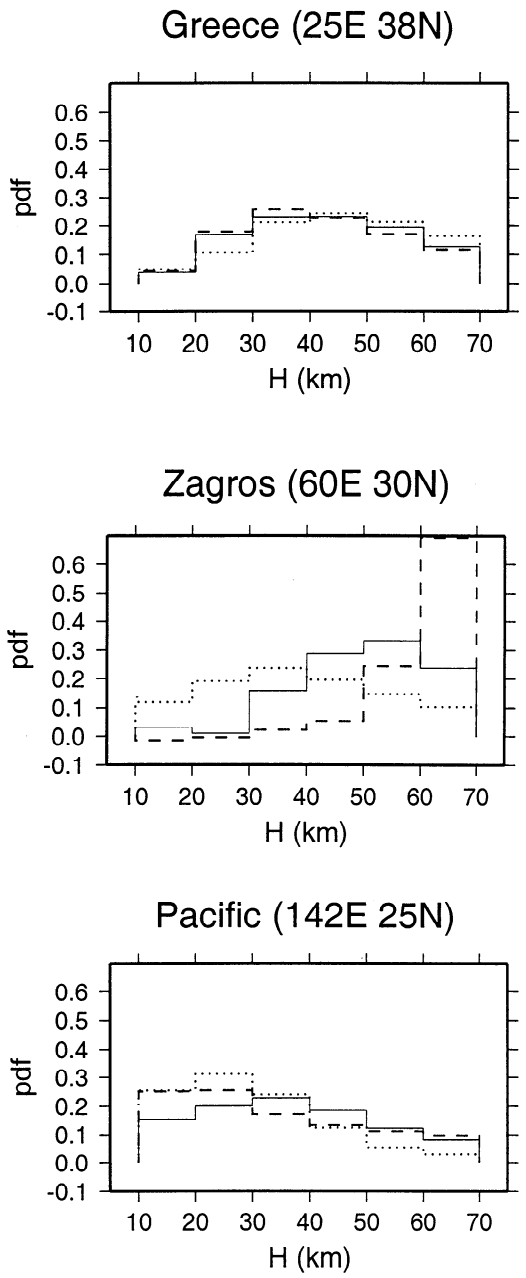

Figure 13. Full distributions from inversions of Love and Rayleigh (solid), Love (dashed) and Rayleigh (dotted) velocities of data set II. Three points were examined, in Greece (25°E,38°N), Zagros (60°E,30°N) and the Pacific (142°E,25°N).

Europe show thickness patterns that are in reasonable agreement with current knowledge about the crust.

6. Discussion

We compare the maximum likelihood thicknesses obtained with NNI of the Love and Rayleigh data from set I (Figure 10e) to the crustal thicknesses of a shear velocity model obtained with a linearized inversion (Figure 14) ([*Wu and Levshin,* 1994]). The patterns are very similar, indeed the correlation coefficient is 0.70; in Figure 15a we show in more detail the scatter of values between these two models. Some clustering of the model of Wu and Levshin (1994) around 40 km could be due to their choice of reference model. Some differences between the two models could be due to the fact that from the histogram network we obtain the model at the maximum of the crustal thickness marginal distribution, whereas linearized inversion in principle obtains the most likely combination of all relevant crustal parameters, of which only Moho depth is plotted here. Also, linearized inversion requires a starting model, whereas our inversion does not (this could be an important consideration in the Tarim Basin area).

Model CRUST5.1 [*Mooney et al.,* 1998] which gives 5°x5° averaged features (Figure 16), represents current knowledge about the crust. In Figure 15b we compare the values of this model with crustal thickness values from our phase velocity inversion (Figure 12). Vertical alignment of points at, e.g., 60 km results from the tectonic zonation of CRUST5.1. The region under study is dominated by continental structure, which is reflected in the large number of observations in the 30-50 km range. The correlation factor is 0.68; hence we can say that the crustal thicknesses from our inversion correlate well with those of CRUST5.1.

The comparison to the model of *Wu and Levshin* [1994] shows that NNI compares favourably to linearized inversion. Since the pattern of crustal thickness we find is also similar to that of CRUST5.1, NNI of these data provides robust results.

7. Conclusions

We have shown that for nonunique inverse problems for which a set of possible models exists for each datum, a classical application of a (neural) network will return the mean model. By properly treating the input data, such a network can map the distribution of output values, but this approach fails if the input and output space are poorly sampled. Alternatively, we have developed networks whose outputs produce histogram-like information about the model distribution, or whose outputs subdivide the distribution into segments with equal areas. Thus we have several approaches to obtain probabilistic estimators of the solution.

We apply these networks to estimate crustal thickness given fundamental mode group or phase velocity curves. Regionalized phase [*Curtis et al.,* 1998] and group [*Wu and Levshin,* 1994] velocities across Eurasia are inverted to produce maximum likelihood crustal thickness maps. Histogram and median networks are used to find a discrete approximation to the complete crustal thickness distribution, showing asymmetry of the solution. This inverse problem is weakly nonlinear, but in principle, networks can also be applied to more strongly nonlinear problems.

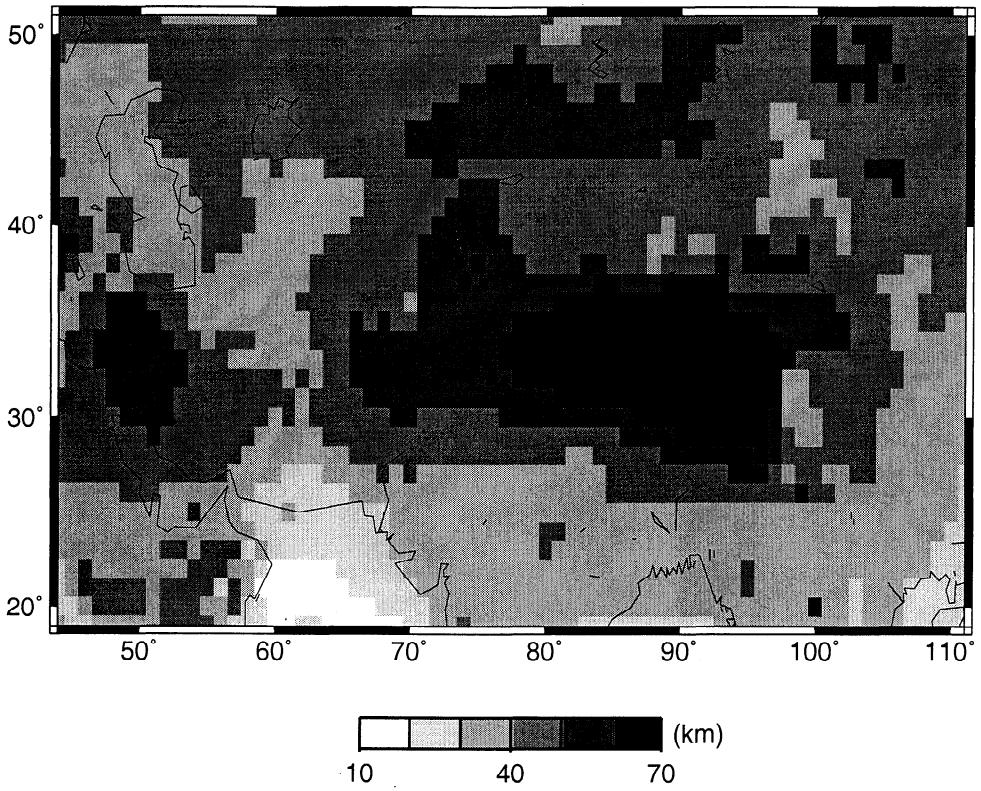

Figure 14. The crustal thickness from a linearized inversion model [*Wu and Levshin,* 1994] using group velocity data with periods between 10 and 200 s.

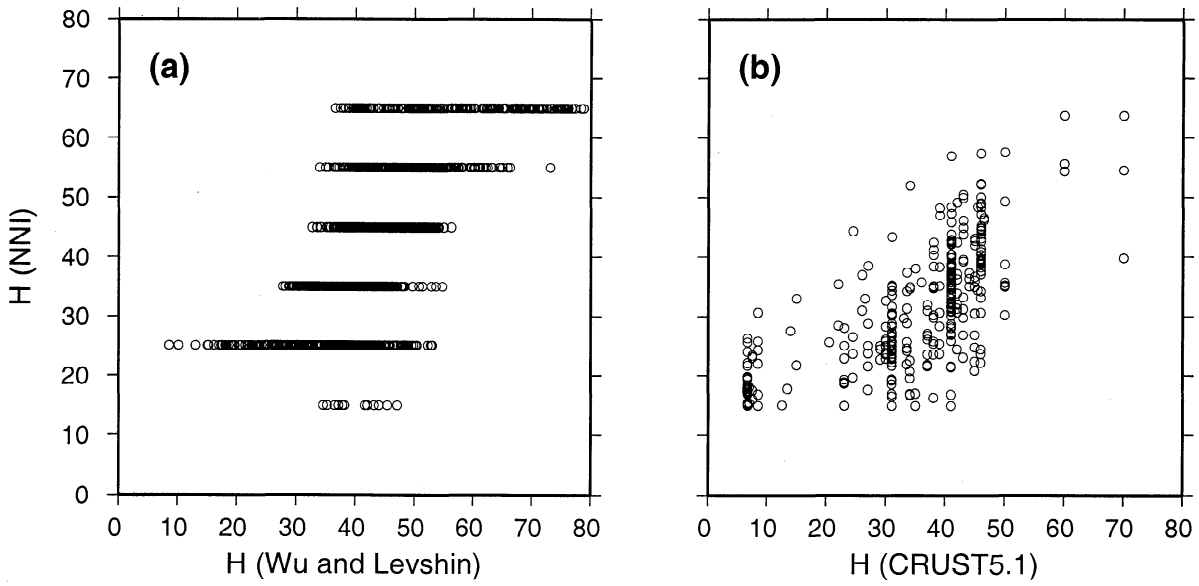

Figure 15. (a) The maximum likelihood crustal thicknesses (obtained in 10-km intervals) obtained from data set I shown against the (continuous) thickness values of the model by *Wu and Levshin* [1994]. (b) The maximum likelihood crustal thicknesses from the inversion of data set II, averaged onto 5 x 5 degree blocks like those of CRUST5.1, pairwise shown against the crustal thicknesses of model CRUST5.1. Values outside the nonresolved region of the data of *Curtis et al.* [1998] are not included.

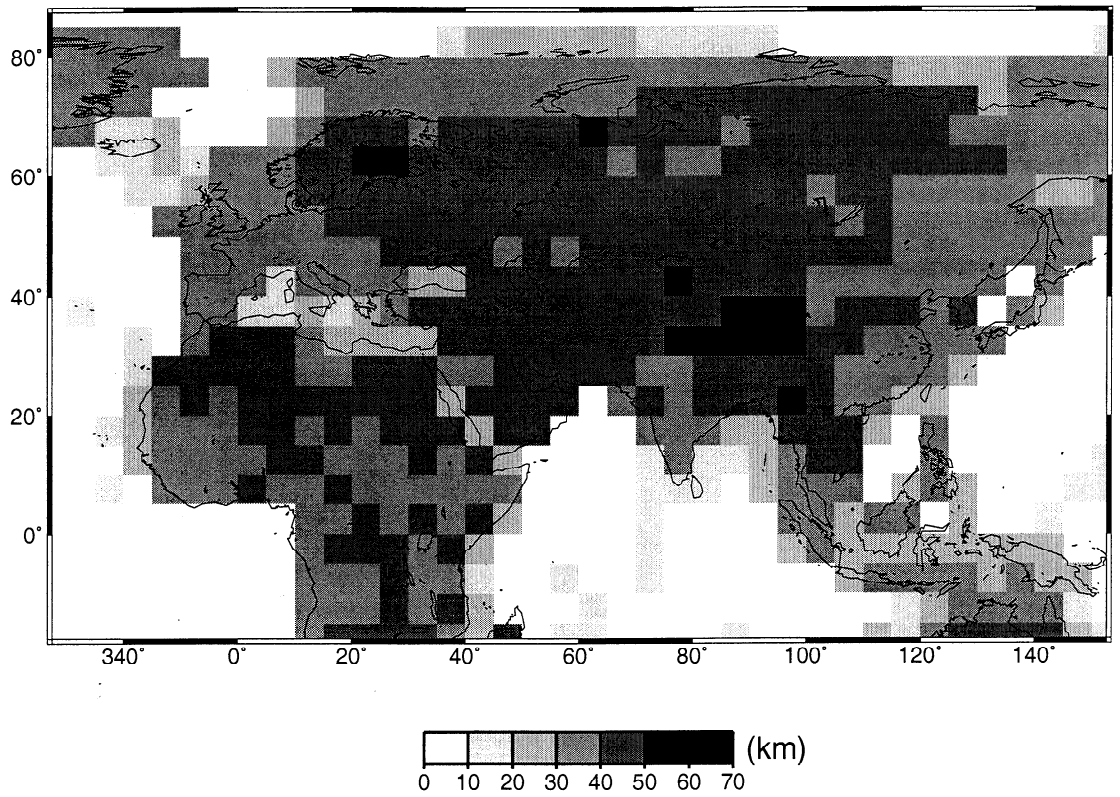

Figure 16. The crustal thickness of model CRUST5.1 [*Mooney et al.,* 1998] which shows 5 x 5 degrees averaged values.

The large-scale patterns of crustal thickness and many of the smaller-scale features which we found are consistent with current knowledge of the crust in Eurasia and largely follow the patterns of topography of the Earth's surface. We found no new crustal thickness results, but we provide further evidence supporting existing models. This is valuable in itself, since we obtained the results from independent data sets. We can also turn this around: the fact that the results are similar means that the data we used are reliable and that the network approach works.

The speed of the inversion and the possibility of creating complete probabilistic information about the solution makes the neural network an important tool for applications that require many similar inversions.

Acknowledgments. We would like to thank A. Levshin for making his group velocity data generally available, and W. Mooney et al. for doing the same regarding their CRUST5.1 model. G. Masters supplied the code for calculating surface modes. We have had helpful discussions with A. Lomax, J. Trampert, R. Snieder, and H. Douma. This research was carried out under the auspices of the Vening Meinesz Research School of Geodynamics in the Netherlands.

References

Bishop, C.M., Neural Networks for Pattern Recognition, Oxford Univ. Press, New York, 1995.

Christensen, N.I., W.D. Mooney. Seismic velocity structure and composition of the continental crust: A global view, *J. Geophys. Res., 100,* 9761-9788, 1995.

Curtis, A. and J.H. Woodhouse. Crust and upper mantle shear velocity structure beneath the Tibetan plateau and surrounding regions from interevent surface wave phase velocity inversion, *J. Geophys. Res., 102,* 11,789-11,813, 1997.

Curtis, A., B. Dost, J. Trampert and R. Snieder. Eurasian fundamental mode surface wave phase velocities and their relationship with tectonic structures, *J. Geophys. Res., 103,* 26,919-26,947, 1998.

Dai, H. and C. MacBeth. The application of back-propagating neural network to automatic picking seismic arrivals from single-component recordings, *J. Geophys. Res., 102,* 15,105-15,113, 1997.

Dowla, F.U., S.R. Taylor and R.W. Anderson. Seismic discrimination with artificial neural networks: preliminary results with regional spectral data, *Bull. Seism. Soc. Am., 80,* 1346-1373, 1990.

Dziewonsky, A.M., A.L. Hales, and E.R. Lapwood. Parametrically simple Earth models consistent with geophysical data, *Phys. Earth Planet. Inter., 10,* 12-48, 1975.

Hecht-Nielsen R., 1991, Neurocomputing, Addison-Wesley Publishing Co., USA.

Jarvis, C.H. and N. Stuart. The sensitivity of a neural network for classifying remotely sensed imagery, *Comput. and Geosci., 22,* 959-967, 1996.

Leach, R.R., F.U. Dowla, E.S. Vergino. Yield estimation using bandpass-filtered seismograms: preliminary results using neural networks with mb(Pn) short-time, long-time and coda energy measurements, *Bull. Seism. Soc. Am., 83,* 488-508, 1993.

Lomax A. and R. Snieder. Finding sets of acceptable solutions with a genetic algorithm with application to surface wave group dispersion in Europe, *Geophys. Res. Lett., 21*, 2617-2620, 1994.

Matsu'ura, M. and N. Hirata. Generalised least-squares solutions to quasi-linear inverse problems with a-priori information, *J. Phys. Earth, 30*, 451-468, 1982.

McCormack, M.D., D.E. Zaucha and D.W. Dushek. First-break refraction event picking and seismic data trace editing using neural networks, *Geophysics, 58*, 67-78, 1993.

Mendiguren, J.A.. Inversion of surface wave data in source mechanism studies, *J. Geophys. Res., 82*, 889-894, 1977.

Mooney, W.D., G. Laski and T.G. Masters. CRUST 5.1: a global crustal model at 5 degrees x 5 degrees, *J. Geophys. Res., 103*, 727-747, 1998.

Mosegaard, K. and A. Tarantola. Monte Carlo sampling of solutions to inverse problems, *J. Geophys. Res., 100*, 12,431-12,447, 1995.

Muyzert, E.J., Monte Carlo waveform inversion and deep continental structure, thesis, Utrecht University, Utrecht, Netherlands, 1998.

Poulton, M.M., B.K. Sternberg and C.E. Glass. Location of subsurface targets in geophysical data using neural networks, *Geophysics, 57*, 1354-1544, 1992.

Pulli, J.J and P.S. Dysart. An experiment in the use of trained neural networks for regional seismic event classification, *Geophys. Res. Lett., 17*, 977-980, 1990.

Rietsch, E.. The maximum entropy approach to inverse problems, *J. Geophys., 42*, 489-506, 1977.

Roth, G. and A. Tarantola. Neural networks and inversion of seismic data, *J. Geophys. Res., 99*, 6753-6768, 1994.

Sen, M.K. and P.L. Stoffa. Nonlinear one-dimensional seismic waveform inversion using simulated annealing, *Geophysics, 56*, 1624-1638, 1991.

Stoffa, P.L. and M.K. Sen. Nonlinear multiparameter optimization using genetic algorithms: Inversion of plane-wave seismograms, *Geophysics, 56*, 1794-1810, 1991.

Tarantola, A., Inverse Problem Theory, Elsevier Sci., New York, 1987.

Tarantola, A. and B. Valette. Inverse Problems: Quest for information, *J. Geophys., 50*, 159-170, 1982.

Wiggins, R.A.. Monte Carlo inversion of body-wave observations, *J. Geophys. Res., 74*, 3171-3181, 1996.

Wu, F.T. and A. Levshin. Surface wave group velocity tomography of East Asia, *Phys. Earth Planet. Inter., 84*, 59-77, 1994.

Zhao, L., M.K. Sen, P.L. Stoffa, C. Frohlich. Application of very fast simulated annealing to the determination of the crustal structure beneath Tibet, *Geophys. J. Int., 125*, 355-370, 1996.

A. Curtis, Schlumberger Cambridge Research, High Cross, Madingley Road, Cambridge CB3 0EL, England, U.K. (curtis@cambridge.scr.slb.com)

R.J.R. Devilee and K. Roy-Chowdhury, Geodynamic Research Institute, Department of Geophysics, Utrecht University, 3508 TA Utrecht, Netherlands. (devilee@geo.uu.nl; kabir@geo.uu.nl)

(Received September 16, 1998; revised June 21, 1999; accepted August 9, 1999.)