A novel approach for spatial clustering and averaging 2-D earthquake ray paths using a machine learning algorithm

Himanshu Agrawal[1], Supriyo Mitra[1,2]

[1]Department of Earth Sciences, [2]Centre for Climate and Environmental Studies, Indian Institute of Science Education and Research-Kolkata, Mohanpur, Nadia, 741246, West Bengal, India.
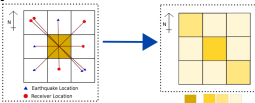
EGU2019-104

ABSTRACT

Spatial data heterogeneity often introduces bias in processing of observed data. Zonation of data sources or receivers in the concerned space is one of the major, yet unavoidable reason for this problem. For earthquake-receiver raypath based studies such biases can arise due to heterogeneous raypath density, which in turn can introduce artifacts in the results obtained from inversion of such dataset, eg. performing a linear tomography. In such problems, clustering of similar raypaths can be used to unformly sample the data space. Our approach to deal with this problem down-weights the denser, more concentrated datasets by a density based clustering model and then averaging out the nodes spatially among the clusters. We introduce the 2-D CLUST algorithm to minimize raypath heterogeneity spread over an area. Our results strongly suggest that 2D-CLUST is potential tool for earthquake-receiver raypath cluster analysis. We tested our algorithm on attenuation tomography computation by Xie and Mitchell (1990) for the dense distribution of earthquake-receiver raypaths and observed a significant improvement in the spatial resolution in 2-D point spreading function recovery maps. Significant improvement is also observed in the checker board tests of surface wave group velocity tomography.

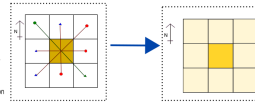## PROBLEM IDENTIFICATION AND SOLVING APPROACH

Problem in general tomography routines using earthqauke ray paths

Heterogeneous raypaths in the gridded area having parameter represented by grid colour is to be recovered by using the tomography technique . The yellow colour at the centre grid and white colour in the remaining grids represents the original parameter structure of the gridded area.

- Highest parameter value — Lowest parameter value

Darker shade represents closeness to orginal highest parameter value (in the left)

Illustrative tomography results representing the recovered parameter results after following the general idea of tomography routines using the ray paths illustrated in the left. The smearing or false values are obtained more along the dense raypaths.

The smearing or false values along the clustered raypaths generally arises because more raypaths give more weight to false values in the initial iterations and which doesn't allow them to go down and then to converge to right values of the particular grids. The subsequent section illustrates idea to resolve this problem.

Solution envisioned for the same

The illustration representing the clustering idea of raypaths. Here closeby raypaths are grouped and the earthquake and receiver locations are averaged out. The said processing attains relative homegenity in the data as compared to that of above in the figure 1.

▲ Earthquake location ● Averaged out earthquake location
● Receiver location ▲ Averaged out receiver location

Figure 4: Illustration representing the observed results of the routine raypath tomography resultsresults after performing it on the clustered raypaths (Colour scheme being same as that of figure 2)

The general idea of solving the problem associated with the heterogeneous data is clustering the closeby raypaths and thus obtaining a relatively homogenous dataset out of the same. which should result in more accurate results. At the same time it should also ensure that the resulting clusters should not highly undersample the region as compared to unclustered ones.

METHODOLOGY
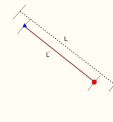
Step 1: Grouping according the slopes of raypaths
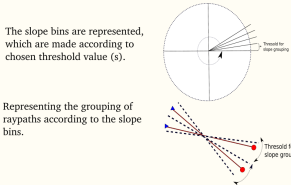
Step2: Clustering midpoints of raypaths using DBSCAN Algorithm

Step 3: Grouping according to the length

The slope bins are represented, which are made according to chosen threshold value (s).

Representing the grouping of raypaths according to the slope bins.

Threold for slope grouping

The sliding window is shown around the midpoint of a raypath. The sliding window has a radius of € units and minimum no. of points is, m
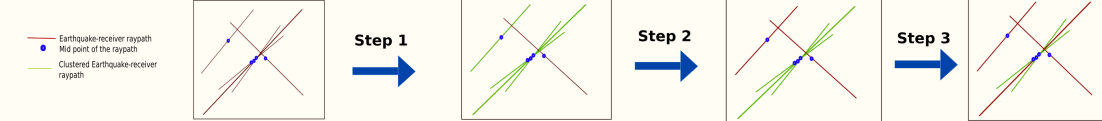
Clustering of a differentially dense set of points is shown, The points around which the sliding window has MinPts available will be outlier

Illustration showing the threshold value = l (L-L') in which raypaths falling under will be clustered and averaged out.

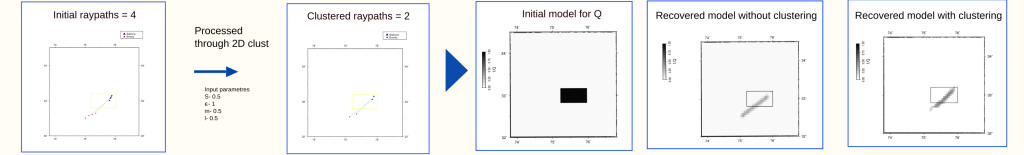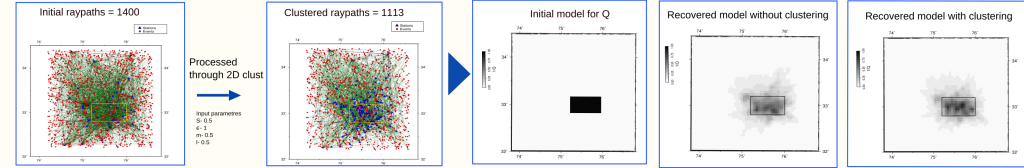**Illustration representing the application of the algorithm on the example dataset**

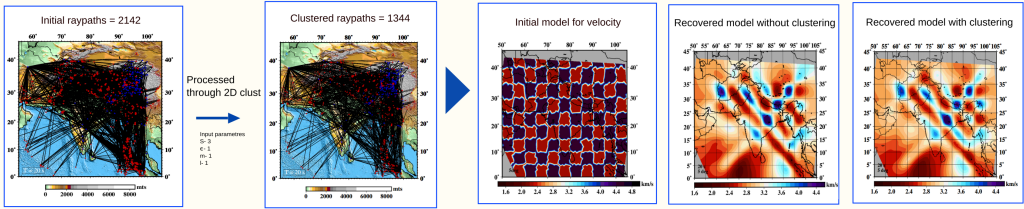— Earthquake-receiver raypath
● Mid point of the raypath
— Clustered Earthquake-receiver raypath

Step 1 → Step 2 → Step 3

RESULTS

1. Point spreading function results using the 4 raypath tomography

TOMOGRAPHY RESOLUTION TEST RESULTS

Initial raypaths = 4 → Processed through 2D clust → Clustered raypaths = 2

Input parametres
S- 0.5
∈ - 1
m- 0.5
l- 0.5

Initial model for Q | Recovered model without clustering | Recovered model with clustering

2. Point spreading function results using the dense raypath coverage

Initial raypaths = 1400 → Processed through 2D clust → Clustered raypaths = 1113

Input parameters
S- 0.5
∈ - 1
m- 0.5
l- 0.5

Initial model for Q | Recovered model without clustering | Recovered model with clustering

3. Surface wave group velocity tomography checkerboard tests for Indo Eurasian region.

Initial raypaths = 2142 → Processed through 2D clust → Clustered raypaths = 1344

Input parametres
S- 1
∈ - 1
m- 1
l- 1

Initial model for velocity | Recovered model without clustering | Recovered model with clustering

TOMOGRAPHY COMPUTATION

Parameterize the region in square grids and assign equal value Q_0 to each grid

Compute the path overlap (S_{gp}) of each grid with the ellipse and calculate weightage average of Q_0

$$\frac{1}{Q_i} = \frac{1}{\sum_{g=1}^{N_i} S_{gp}} \sum_{g=1}^{N_i} \frac{S_{gp}}{Q_g^i} + \varepsilon_{p'}$$

Back-project the residual misfit in the above equation to calculate next set of Q values for each grid

$$\frac{1}{Q_g^{i+1}} = \frac{1}{Q_g^i} + \frac{\sum_{p=1}^{N_g} \frac{S_{gp}^i \Delta_{gp}}{S_{gp}}}{\sum_{p=1}^{N_g} S_{gp}^i}$$

Use above formulation to compute residual misfit between observed (Q_o) and calculated Q

$$\Delta_{pi} = \left| \frac{\sum C_i^{ol}}{\sum C_i} - 1 \right|$$

Stopping criteria: if Misfit function falls below cut-off value or maximum iteration reached

ERROR CALCULATIONS

Accuracy in case of Point spreading function resolution test is estimated using the following scheme-
The weights are given according to the following colour scheme shown in the figure which ARE multipled to calculated Q value and aggregated value for whole space is calculated

1 -1.1 -1.2 -1.3 -1.4 ......
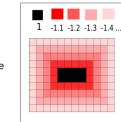
Following results are observed-
O Point spreading function test for 4 raypaths-
Without clustering : -3.46   With clustering: 1.49
O Point spreading function test for dense raypath coverage-
Without clustering : -92.45   With clustering: -72.22

DISCUSSION

● **Considering the improvements in the resolution test results, we can infer that the clustering algorithm (2D clust) may serve a prominant tool for improving tomography results.**

● Since the algorithm uses DBSCAN algorithm hence the chain spreadness of the clustering is taken care of. The Clusteres will have the mean standard deviation within the value given as ∈. if the points are falling beyond this value the clusteres will be broken off.

● It should be noted that the algorithm provides the significant improvement in spatial resolution but at the same time due to decreament in the data points it produces lack of magnitude resolution hence it is advised to superimpose the results alongwith the results obtained by the clustered data for better analysis

References: Xie & Mitchell (1990), GJI, 100, 161-181 Ester et al. (1996), KDD-96, 226-231