Weighing geophysical data with trans-dimensional algorithms: An earthquake location case study

Nicola Piana Agostinetti,[1,2], Alberto Malinverno,[3], Thomas Bodin[4], Christina Dahner[5], Savka Dineva[5,6]and Eduard Kissling[7]

[1]Department of Earth and Environmental Sciences, Universitadi Milano Bicocca, Piazza della Scienza 1, I-20126 Milano, Italy
[2]Department of Geology, University of Vienna, Vienna, Austria
[3]Lamont-Doherty Earth Observatory, Columbia University, NY, USA
[4]Univ Lyon, Univ Lyon 1, ENSL, CNRS, LGL-TPE, F-69622, Villeurbanne, France
[5]Luossavaara-Kiirunavaara AB, Kiruna, Sweden
[6]Department of Civil, Environmental and Natural Resources Engineering, Lulea, Sweden
[7]Department of Earth Sciences, ETH Zurich, Switzerland

Key Points:

- We develop a novel approach for automatic weighting of data in geophysical inverse problems, based on a trans-dimensional algorithm
- We apply the novel approach to seismic event location in mines
- Our approach outperforms standard approaches, when limited information are known about data uncertainties

Corresponding author: Nicola Piana Agostinetti, `nicola.pianaagostinetti@unimib.it`

Abstract

In geophysical inverse problems, the distribution of physical properties in an Earth model is inferred from a set of measured data. A necessary step is to select data that are best suited to the problem at hand. This step is performed ahead of solving the inverse problem, generally on the basis of expert knowledge. However, expert-opinion can introduce bias based on pre-conceptions. Here we apply a trans-dimensional algorithm to automatically weigh data on the basis of how consistent they are with the fundamental assumptions made to solve the inverse problem. We demonstrate this approach by inverting arrival times for the location of a seismic source in an elastic half space, under the assumptions of a point source and constant velocities. The key advantage is that the data do no longer need to be selected by an expert, but they are assigned varying weights during the inversion procedure.

Plain Language Summary

In the Big data era, automated approaches to data evaluation are needed for two main reasons: to be able to process a large amount of data in a limited time, and to avoid bias introduced by data analysists. In this study we present a novel approach to data analysis, where the data themselves measure their consistency with our hypotheses. The approach is applied to earthquake location in mines, where millions of seismic events occur every year, and automatic processing of seismic data is mandatory. We demonstrate that our approach outperforms standard ones when almost nothing is known about the data and their measurement errors.

1 Introduction

Measured scientific data make possible a quantitative analysis of observations (e.g., a seismometer can record seismic waves, which are only felt by humans as transient phenomena). Scientific data are routinely processed before making inferences on the spatio-temporal distribution of physical quantities and/or physical processes (e.g., arrival times for seismic P-waves are extracted from continuous seismic recordings to infer the position of a seismic source). Processing steps can be necessary to remove spurious data (e.g., arrival times from seismic sensors that are not synchronized), but also to enhance data to better represent the most relevant signal for the problem being investigated (e.g., seismic waveforms may be filtered in the frequency domain before picking relative arrival times by cross-correlation (VanDecar & Crosson, 1990), for a clear identification of phases and for removing noise-site-effect interferences with targeted signal wavelet).

Geo-scientific data are especially challenging, because they are generally used to make inferences on physical quantities which are not directly measurable, but need to be estimated by solving an inverse problem (Tarantola, 2005), where processed measurements (e.g., P-wave arrival times or maximum wavelet amplitudes) are combined with hypotheses about the physics of the system (e.g., models of seismic wave propagation in the rock volume or seismic energy released by source). In this case, data processing typically includes selecting a subset of the data that is most relevant for the problem at hand (e.g., by removing arrival times for P-waves that do not travel directly from source to receiver). Additionally, seemingly less accurate data are often excluded or apriori down-weighted to make them less influential in the final solution (e.g., arrival times recorded at distant seismic sensors that are likely to show larger effect of influence by attenuation or scattering along the ray-path). These data processing steps are usually based on expert opinion, but expert decisions made a priori before solving the inverse problem can be somewhat arbitrary and bias the inversion results.

Here we propose a novel approach to incorporate the choice of weights for the data in the inversion process (or, more precisely, the variance of data noise). Our approach is based on trans-dimensional Markov chain Monte Carlo (McMC) sampling (Piana Agostinetti et al., 2021; Piana Agostinetti & Sgattoni, 2021) and works by proposing and accepting/rejecting data weighing schemes following the Metropolis algorithm (Sambridge & Mosegaard, 2002) where the data weighing schemes have a variable number of parameters (Malinverno, 2002; Sambridge et al., 2006). The complexity of the weighing scheme is dictated by the data themselves, rather than by user-defined choices made during preprocessing. The assigned weights depend on how closely different data match the fundamental assumptions made in solving the inverse problem.

We test our approach in the geophysical inverse problem of locating a seismic point source using P- and S-wave arrival times recorded by sensors in a seismic network. In this inverse problem, data are generally downweighted with distance of the sensor from the seismic source or are removed in a pre-processing step if the sensors are farther than a chosen distance from the source. In our novel approach, we define a set of spherical shells centered on the source (Figure 1a). All sensors within a shell are assigned the same weight (Figure 1b), but more complex weight assignments can be made (e.g., weights that vary linearly with distance from the source within each shell; see Figure 1c). The number of shells, their radii and weights are unknown, and will be defined by the McMC sampling. The stations that receive the largest weights will be those that measure arrival times consistent with the fundamental assumptions made in the inverse problem (namely, a point-wise seismic source and constant P- and S-wave velocities in the rock volume).

Our natural laboratory is Kiirunavaara mine (Sweden), a 6km-long active mine with more than 200 seismic sensors in a 3D configuration that spans along the exploited rock volume (Dineva et al., 2022). Given such an extensive seismic network, events can be well located in three dimensions. We selected two seismic events. The first is a man-made blast, used to calibrate the seismic network (Figure 1d). The actual location of this seismic source is known within < 1 meter and can be immediately used to evaluate our results. The second is a natural M_w 4.2 multi-phase seismic event that occurred on May 18th 2020 (Dineva et al., 2022) and it was recorded on all working sensors in the mine (Figure 1e). Our experiment is structured as follows. We first compute a reference solution for the calibration blast by applying a standard McMC algorithm (see "Materials and Methods" and (Riva & Piana Agostinetti, 2023)). In this reference solution, we do not use our novel approach, but we solve for the source location by removing data from sensors at a range of distances from a preliminary location of the seismic source, as done in standard seismological workflows in mines. This is intended to simulate a range of possible expert opinions on the distance threshold for data selection (here we assume that the hypocentral distance is of such utmost importance that observational quality differences may be neglected, which is not the case in crustal studies). We then apply our novel approach to the complete data set for the calibration blast and compare the results with those in the reference solution. Finally, we apply our methodology to the natural seismic event. All the necessary details of our novel approach are in the Supporting On-line Materials.

2 Results

The reference solution results are in Figure 2. Starting with all the available data (all 57 seismic sensors that recorded the blast to a maximum distance of 800 meters from the source), we get a posterior mean event location which is about 12 meters away from the blast, with estimated uncertainties as large as 7 meters. We then start removing data from sensors farther than 700 meters, 600 meters, etc., in steps of 100 meters (see Figure 2 and "Materials and Methods"). The event location uncertainties and the differences with the actual blast position reach a minimum for a maximum sensor distance

of 300 meters (19 sensors). Considering sensors closer to the source (200 meters, 5 sensors) results in an increase in uncertainties and location error.

Our novel approach applied to the blast data gives results that are consistent with those obtained in the reference solution (Figures 3 and **??**). The variation of weights with distance for both P- and S-wave arrival times follows a simple pattern, with a single step decrease at about 380±30 meters from the source (Figure 3b,c). The weights for S-wave arrival times decrease much more sharply than those for P-waves. This main step is well defined, as seen from the histogram of the sampled shell radii (Figure 3d), although the histogram of the number of shells has a maximum between 5 and 7 (Figure 3a). The sampled weights result in a cloud of event locations that closely reproduces what was found in the reference solution for a maximum distance of 400 meters (red vs. black dots in Figure 3e).

In crustal studies, it has been observed that event location uncertainties depend on the azimuthal coverage (Husen et al., n.d.). Here we computed the azimuthal coverage of the 3D distribution of seismic sensors (see "Materials and Methods"). Azimuthal coverage reaches a nearly stable value at a distance of ca. 300 meters from the source, and it does not change substantially at greater distances (Figure 3d). The best reference solution was found when selecting stations only within 300 m from the source, which is also close to the distance where the weights obtained in our new method decrease substantially.

We apply our data-space exploration algorithm to the arrival times of the natural event (Figure 4). This event has a magnitude M_w 4.2, it is composed of several subsequent processes, where the extent of the very first sub-event S1 is likely ca. 100-200 meters (Dineva et al., 2022)). The final posterior distribution of the source location is close to that initially estimated (Figure 4a, b). The pattern of weights with distance is more complex compared to that obtained for the blast. There is a main step at about 1230±70 m, but also three other maxima in the histogram of shell radii (marked with colored arrows in Figure 4c). The weights for the P-wave arrival times slightly increase from the origin to 150±50 meters (grey arrow) and remain near a maximum value between 150±50 and 500±60 meters (red dashed arrow). At greater distances, the weight decrease slightly to a nearly constant value out to 1230±70 meters (red arrow), where there is a sharp decrease of almost one order of magnitude. The weights increase again at about 1900 ± 60 meters (blue arrow).

We also conducted a test to check whether the overall pattern of weights with distance is significantly affected by the simple parameterization of constant weights in each spherical shell. To this end, we implemented an alternative parameterization where weights are defined at the shell boundaries and vary linearly within each shell (Figure 1c). The pattern of weights with distance obtained with linearly varying weights is very similar to that obtained with constant weights (see yellow contours in Figure 4c and Supporting On-Line Materials). The choice of parameterization does not seem to strongly control the variation of weights with distance.

3 Discussion

In our first test with a controlled blast the reference solution seems to outperform our novel approach, as the best event location is slightly closer to the blast position for sensors at a maximum distance of 300 meters from the source (whereas the weights in our approach decrease at distances > 380 meters). This difference is small, however (less than 2 meters), and the pattern of the weights closely mimics a step function. We conclude that in this simple case the performance of the two methods is similar (i.e the classical approach outperforms our approach only in the case where the maximum distance is correctly chosen, 300 m, which is rarely the case).

Comparing the results obtained in the two tests carried out with our novel approach, we note that the pattern of weights with distance seems to be event-dependent and is not a constant in a particular sensor network. While further research would be necessary to determine which event parameters (e.g., magnitude, location) affect the weight pattern, the results indicate that a static workflow for all events would probably introduce artifacts and underestimate the actual uncertainties. In contrast, our approach is adapted to each single event, giving a solution that is statistically consistent and parsimonious (in terms of complexity of the weight pattern parameterization).

The relationship between azimuthal coverage of the event and our results is not straightforward. In the controlled blast, the main decrease in the weights we obtains is near the distance where the azimuthal coverage increases substantially (Figures 3d and 4c). On the other hand, there is no clear correspondence between weight patterns and azimuthal coverage in the test with a natural event. This suggests that azimuthal coverage is only one of the factors affecting the reliability of the inverted source location. A workflow based on this parameter (e.g., where distant seismic sensors are removed once the azimuthal gap decreases below a certain threshold) may not give optimal results. In fact, if the gap is larger than 180 degrees with stations in the epicenter near vicinity, a moderately distant station closing this gap may be very useful if the real subsurface velocities are not perfectly well known (which is almost never the case). On the other hand, closing a gap to significantly less than 180 degrees with a single very distant station is at least questionable (if not useless) when considering the uncertainties of phase identification and frequency difference in first arriving/visible wavelets.

The pattern of weights allows us to interpret the results in terms of specific properties of the rock volumes at different distances from the source. We suggest that the seismic sensors closest to the natural event (at distances < 150 meters, first grey circle in Figure 4a), very likely are in the source area, where the assumption of a point-wise seismic source is not realistic for such a large event. Between the grey and the dashed red circle (distances of 150-500 meters) the weights reach their highest values, indicating where the inverse problem assumptions should be valid. Indeed, all sensors within the red dashed circle in Figure 4a are located on the same side of the ore body, where the rock volume is expected to be comparatively homogeneous. Between the dashed and solid red circles in Figure 4a (distances of 500-1200 meters) the weights are still high, but less than in the previous interval. This is likely due to some ray-paths partially crossing the ore body and thus violating the homogeneous rock assumption. Farther than 1200 meters from the source (red circle in Figure 4a), the seismic rays start to densely sample the ore body and the surrounding rocks on both sides of the ore body itself. Here we can expect that the assumption of a homogeneous rock finally breaks down, and the weights decrease significantly. Further investigations are needed to confirm our hypothesis and to check how complex pattern in weights could be related to a less circularity in the data distribution around the seismic source in the case of the natural event than in the case of the blast.

In a more general context, our novel approach can be applied to most of the scientific inference problems, where huge amount of data need to be pre-processed in some way, without introducing bias related to preconceptions of the data-analysts. We mention that our approach only works if data can be ordered or clusterized in some way. Here for example, they are “ordered” with regards to the source-sensor distance. In this case, “ordering” is necessary, but it is not the only way of performing the trans-dimensional data-space exploration. To apply our approach, we need either a metric to be used to ”measure” some kind of data-point distance in the data-space, or, equivalently, some kind of data characterization which enables data clustering, where the trans-dimentional approach is used to define the number of data cluster from the data themselves.

Software and Data Availability Statement

Software and data (i.e., P- and S- arrival times for the blast occurred in the mine) has been archived on Mendeley Data Repository (REF) at https://data.mendeley.com/XXXXX. *For the Editor: Software and data will be made available upon publication*

Acknowledgments

We thanks Mirjana Boskovic for discussion on calibration blasts in Kiruna mine. N.P.A. research is conducted with the financial support of the Ministero dell'Istruzione e del Merito in the framework of the programme "Dipartimenti di eccellenza (2018 - 2022)". The Generic Mapping Tools software has been used for plotting the figures of this manuscript (Wessel & Smith, 1998).

References

Dineva, S., Dahner, C., Malovichko, D., Lund, B., Gospodinov, D., Piana Agostinetti, N., & Rudzinski, L. (2022). Analysis of the magnitude 4.2 seismic event on 18 may 2020 in the Kiirunavaara mine, Sweden. *RaSiM Conference,*.

Husen, S., Kissling, E., & von Deschwanden, A. (n.d.). Induced seismicity during the construction of the Gotthard Base Tunnel, Switzerland: hypocenter locations and source dimensions. *J Seismol*. doi: 10.1007/s10950-011-9261-8

Malinverno, A. (2002). Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem. *Geophys. J. Int.*, *151*(3), 675-688.

Piana Agostinetti, N., Kotsi, M., & Malcolm, A. (2021). Exploration of data space through trans-dimensional sampling: A case study of 4D seismics. *Journal Geophysical Research*, *126*. doi: 10.1029/2021JB022343

Piana Agostinetti, N., & Sgattoni, G. (2021). Changepoint detection in seismic double-difference data: application of a trans-dimensional algorithm to data-space exploration. *Solid Earth*, *12*, 2717–2733. doi: 10.5194/se-2021-79

Riva, F., & Piana Agostinetti, N. (2023). *The micro-seismicity of Co. Donegal (Ireland): defining baseline seismicity in a region of slow lithospheric deformation.* (submitted to *Terra Nova*)

Sambridge, M., Gallagher, K., Jackson, A., & Rickwood, P. (2006). Trans-dimensional inverse problems, model comparison and the evidence. *Geophys. J. Int.*, *167*(2), 528-542. (doi:10.1111/j.1365-246X.2006.03155.x.)

Sambridge, M., & Mosegaard, K. (2002). Monte Carlo methods in geophysical inverse problems. *Rev. Geophys.*, *40*(3), doi:10.1029/2000RG000089.

Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation.* SIAM.

VanDecar, J. C., & Crosson, R. S. (1990). Determination of teleseismic relative phase arrival times using multi-channel cross-correlation and least squares. *Bulletin of the Seismological Society of America*, *80*(1), 150.

Wessel, P., & Smith, W. H. F. (1998). New, improved version of the generic mapping tools released. *EOS Trans. AGU*, *79*, 579.

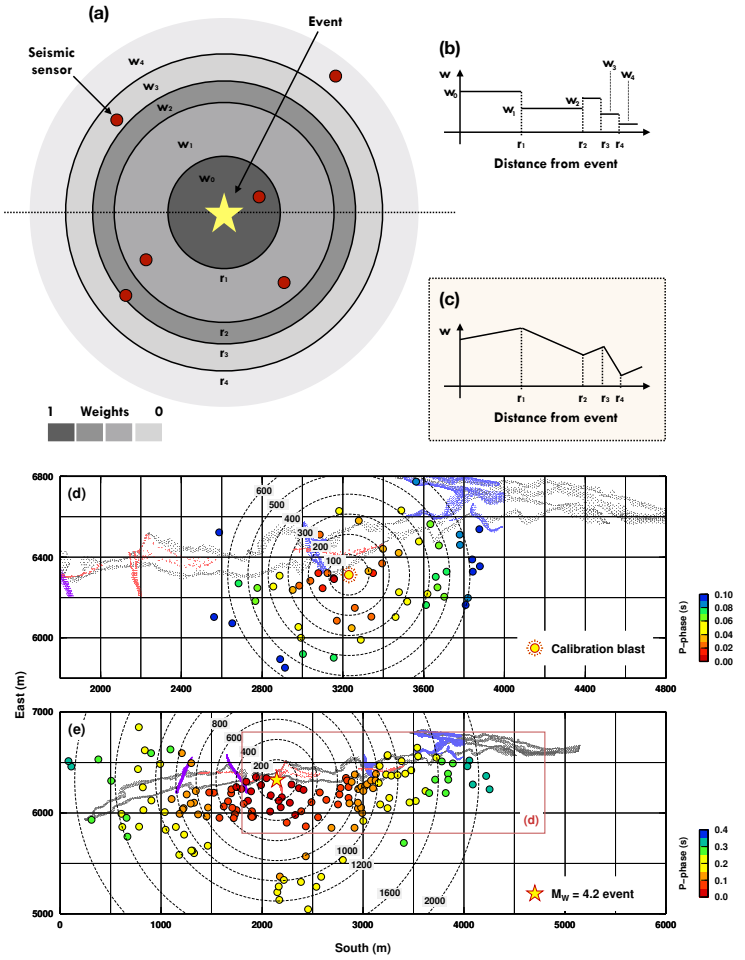

Figure 1. The arrival time weights $\mathbf{w}_k$ are associated with a set of k concentric 3D spherical shells with radii $\mathbf{r}_k$, centered on a preliminary event location. (a) 2D representation of the spherical shells. (b) Constant weights within each shell. (c) Alternative parameterization with linearly varying weights within each shell. (d-e) Seismic data used in this study: sensor locations projected on a horizontal plane (circles) and arrival times for P-waves relative to the earliest recorded arrival time (circle colors). Dashed circles show the distance in meters from the preliminary event location. Important geological features are depicted with coloured dots: grey = ore body; blue = clay zones; red/purple = diapir/diabase. Panel (d) shows seismic arrival times for a calibration blast (yellow Sun symbol), and (e) for a M_w 4.2 event (yellow star). The red box in (e) shows the smaller area plotted in (d).

(a)

Max R (m)	N	X (m)	Y (m)	Z (m)	dX / Xe (m)	dY / Ye (m)	dZ /Ze (m)	D (m)
200.0	5	3227	6313	-1075	-4 / 4	1 / 7	3 / 6	**6**
300.0	19	3227	6312	-1080	-4 / 2	0 / 2	-2 / 2	**4**
400.0	27	3227	6311	-1081	-4 / 2	-1 / 2	-3 / 2	**6**
500.0	40	3225	6312	-1081	-6 / 2	0 / 3	-3 / 3	**7**
600.0	46	3224	6310	-1082	-7 / 2	-2 / 3	-4 / 3	**9**
700.0	56	3224	6305	-1086	-7 / 2	-7 / 4	-8 / 4	**13**
800.0	57	3224	6305	-1085	-7 / 2	-7 / 4	-7 / 4	**13**

(b)

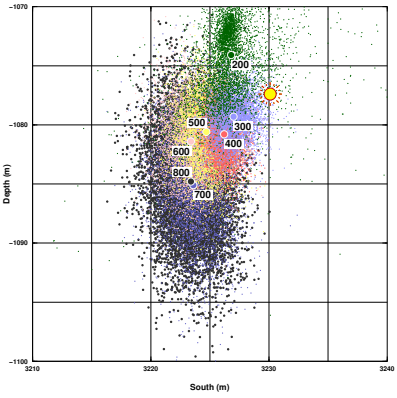

Figure 2. (a) Blast locations obtained in the reference solution with data for sensors that are within different maximum distances from the actual source location. The last column reports the distance D between the posterior mean and true location of the source. The best location (where D is minimum) is obtained with data from sensors up to 300 m from the blast. The X (South), Y (East), and Z (depth) columns list the posterior mean value for the location coordinates. (b) Source locations sampled by the McMC algorithm for different maximum source-sensor distances projected on the X-Z vertical plane (dots). The maximum source-sensor distances are 800 m (black dots), 700 m (dark blue), 600 m (pink), 500 m (yellow), 400 m (red), 300 m (light blue), and 200 m (green). Colored circles are posterior mean locations. The yellow sun indicates the true position of the calibration blast.

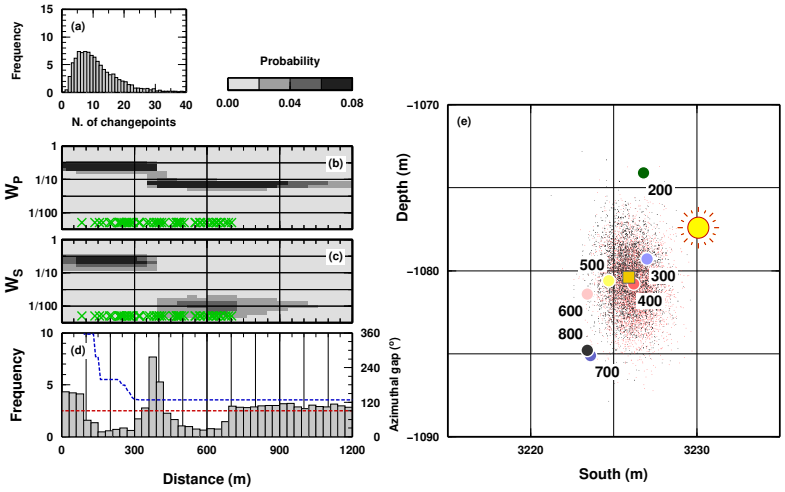

Figure 3. Application of the novel data-weighing method to blast recordings. (a) Posterior probability density function (PDF) of the number of spherical shells, approximated by the histogram obtained by McMC sampling. (b) Posterior PDF of the weights assigned to P-wave arrival times as a function of source-sensor distance. Green crosses indicate the distance of each sensor from the source. (c) As in (b) for S-wave arrival times. (d) Posterior PDF of shell radii. The blue dashed line indicates the azimuthal gap as a function of distance from the source (see "Materials and Methods" for a definition). The red dashed line indicate the prior probability distribution for the shell distance. (e) Sampled source locations projected onto a X-Z vertical plane (black dots) compared to source locations in the reference solution for a maximum source-sensor distance of 400 meters (red dots; see Figure 2b). The yellow square shows the posterior mean source location obtained with the data-weighing method. Colored circles are the posterior mean source locations obtained in the reference solution (same as in Figure 2b). The yellow sun indicates the true position of the calibration blast.

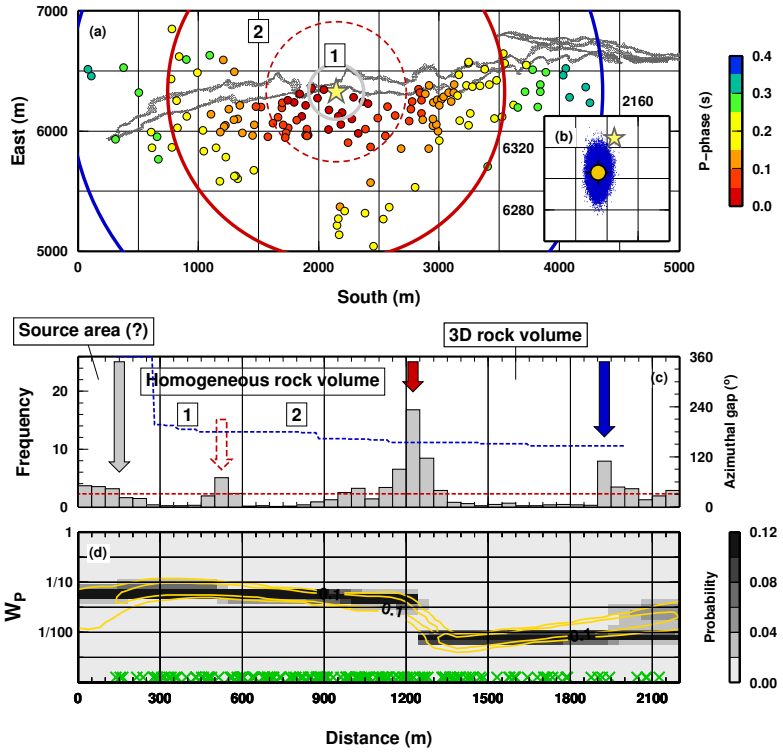

Figure 4. Application of the novel data-weighing method to recordings of the M_w 4.2 natural event. (a) Seismic network geometry (same as in Figure 1e). Colored circles report the position of the main modes in the histogram of sampled shell radii, indicated with colored arrows in panel (c). The inset (b) plots the sampled source locations projected onto a X-Z vertical plane (blue dots) compared to the preliminary location of the event (yellow star). (c) Posterior probability density function (PDF) of shell radii, approximated by the histogram obtained by McMC sampling. The colored arrows indicate the main modes in the posterior PDF, corresponding to the boundaries of the source area (gray arrow), homogeneous rock volume with all sensors on the same side of the ore body (dashed red), homogeneous rock volume (red), heterogeneous rock volume (blue). (d) Posterior PDF of the weights assigned to P-wave arrival times as a function of source-sensor distance. Green crosses indicate the distance of each sensor from the source. The yellow contours display the posterior PDF of the data weights obtained with the alternative parameterization in Figure 1c (see also Figure Supportin On-line Materials).