https://doi.org/10.5194/se-2021-79
Preprint. Discussion started: 11 June 2021

Solid Earth
Discussions
EGU

Exploration of the data space via trans-dimensional sampling: the case study of seismic double difference data

Nicola Piana Agostinetti[1,2] and Giulia Sgattoni[3]

[1]Department of Earth and Environmental Sciences, Universitá di Milano Bicocca, Milano, Italy
[2]Department of Geology, Universitat Wien, Althanstrasse 14, 1090, Wien, Austria
[3]Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Bologna, Bologna, Italia

Correspondence: Nicola Piana Agostinetti (nicola.pianaagostinetti@unimib.it)

odd phrasing, no hyphen

pairs

Abstract.

Double difference~~s~~ (DD) seismic data are widely used to define elasticity distribution in the Earth's interior, and its variation in time. DD data are often pre-processed from earthquake~~s~~ recordings through expert opinion, where couples of earthquakes are selected based on some user-defined criteria, and DD data are computed from the selected couples. We develop a novel methodology for preparing DD seismic data based on a trans-dimensional algorithm, without imposing pre-defined criteria on the selection of couples of events. We apply it to a seismic database recorded on the flank of Katla volcano (Iceland), where elasticity variations in time ~~has~~ have been indicated. Our approach quantitatively defines the presence of changepoints that separate the seismic events in time-windows. Within each time-window, the DD data are consistent with the hypothesis of time-invariant elasticity in the subsurface, and DD data can be safely used in subsequent analysis. Due to the parsimonious behavior of the trans-dimensional algorithm, only changepoints supported by the data are retrieved. Our results indicate that: (a) retrieved changepoints are consistent with first-order variations in the data (i.e. most striking changes in the DD data are correctly reproduced in the changepoint distribution in time); (b) changepoint locations in time do not correlate neither with changes in seismicity rate, nor with changes in waveforms similarity (measured through the cross-correlation coefficients); and (c) noteworthy, the changepoint distribution in time seems to be insensitive to variations in the seismic network geometry during the experiment. Our results ~~proofs~~ that trans-dimensional algorithms can be ~~positively~~ applied to pre-processing of geophysical data before the application of standard routines (~~i.e.~~ e.g. before using them to solve standard geophysical inverse problems) ~~in the~~ so ~~called exploration of the data space~~.

what kind of changes?

[superfluous phrase]   demonstrate   effectively

1 Introduction

Data preparation is a daily routine in the worklife of geoscientists. Before using data to get insights into the Earth system, geoscientists try to deeply understand their datasets, to avoid introducing, e.g. instrumental issues, redundant data, un-wanted structures like data density anomalies, and many others (Yin and Pillet, 2006; Berardino et al., 2002; Lohman and Simons, 2005). All the activities for preliminary data analysis can be considered as exploration of the "data space" (Tarantola, 2005) and are mainly based on *expert opinion*. Previous experience drives scientists in selecting the most trustable portion of their experiments, cleaning data-sets before using them for getting new knowledge about Earth model parameters. There are two

such as

main reasons for moving a step forward from expert opinion. First, the huge amount of (often multidisciplinary) data, accumulating in geosciences in the last decade, requires more and more data screening and preparation, sometimes involving multidisciplinary expertise. Research activities could greatly benefit from a more automated exploration of the data space able to ease preparatory tasks. Second, expert opinion is a human activity and is mainly based on dual categories, e.g. good/bad data, and can not easily handle a continuous probability distribution over the data (i.e. expert opinion can not easily associate a continuous "confidence" measure to each data-point).

In recent years, in the framework of Bayesian inference, exploration of the data space has been introduced in a few cases to "explore" unknown features of the datasets. For example, the so called *Hierarchical Bayes* approach has been introduced to estimate data uncertainties from the data themselves (Malinverno and Briggs, 2004). More complex Hierarchical Bayes approaches have been developed to measure the data correlation as well (e.g. Bodin et al., 2012a; Galetti et al., 2016) or to evaluate an error model (e.g. Dettmer and Dosso, 2012). The exploration of the data space, in all these studies, implies ~~to~~ considering some additional unknowns (e.g. data uncertainties or error correlation length), so called *hyper-parameters* or *nuisance parameters*, and to estimate them directly from the data. A step forward in exploration of the data space has been ~~re~~presented by Steininger et al. (2013) and Xiang et al. (2018), where the authors used a data space exploration approach to evaluate the performance of two different error models directly from the data. In such studies, the number of hyper-parameters considered is not fixed, but can assume two different values (1 or 2), depending on the error model considered. Another interesting, recent case of exploration of the data space is represented by the work of Tilmann et al. (2020), ~~where~~ in which the authors used Bayesian inference to separate the data in two sets: "outliers" and "regular". In this case, the data themselves are probabilistically evaluated to understand their contribution to the final solution as "regular" data or "outlier", i.e. the data are classified in two different families, according to their coherence with the hypothesis of being "regular" data or not. uncertainties? errors are not known

In this study, we push the exploration of data space in a new direction. We develop an algorithm for computing Bayesian inference specifically for the exploration of the data space. Exploration of the data space is performed through a trans-dimensional algorithm (e.g. Malinverno, 2002; Sambridge et al., 2006) so that the number of hyper-parameters is neither fixed nor limited to 1 or 2. We represent *data structure* as partitions of the covariance matrix of errors, i.e. changepoints that create sub-matrices of the covariance matrix with homogeneous characteristics, where the number of partitions is not dictated by the user, but it is derived by the data themselves, in a Bayesian sense (i.e. we obtain a posterior probability distribution, PPD, of the number of partitions). In this way, similar to Tilmann et al. (2020), portions of data can be classified and used differently in the subsequent steps of the analysis.

We apply our algorithm to prepare a widely exploited type of seismic dataset, the seismic double-difference (DD) dataset, that ~~have~~ has been used as input in seismic tomography for defining subsurface elasticity (e.g. Zhang and Thurber, 2003) and its variation in time (e.g. so called "time-lapse tomography", Caló et al., 2011; Zhang and Zhang, 2015). DD data need to be re-constructed from specific partitions of the original data (i.e. seismic events). Subjective choices have a great impact on the definition of DD data. In particular, such choices can be used to limit the number of DD data itself and the selection, in turn, could introduce biases in the subsequent definition of the elastic model and its variations in time. We apply our algorithm to statistically define,

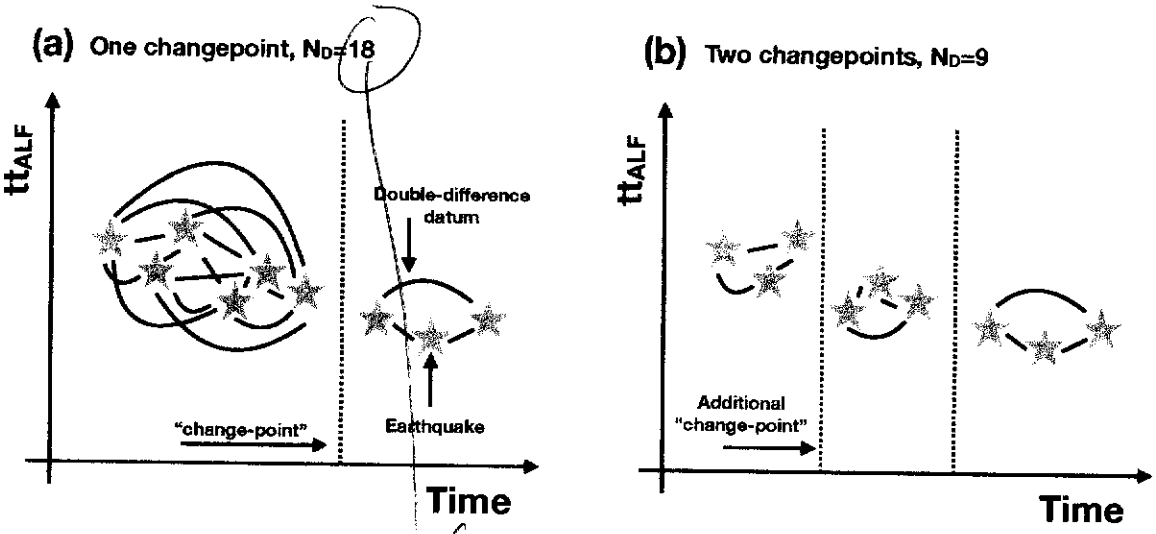
https://doi.org/10.5194/se-2021-79
Preprint. Discussion started: 11 June 2021

in a more objective way, the distribution of partitions in the DD data. We show how a more data-driven approach can obviate ~~to~~ expert-driven data selection and can be used as a preliminary tool for, e.g., time-lapse seismic tomography.

1.1 Double-difference data in seismology

Double-difference seismic data are widely used for relocating seismic events and imaging the subsurface (e.g. Waldhauser and Ellsworth, 2000). DD data rely on the assumption of (nearly) co-located events for which seismic recordings have been obtained from the same station (Zhang and Thurber, 2003) or for the same pair of stations (e.g. Guo and Zhang, 2016). The concept of co-located events relies on expert opinion and relates to the average spatial dimension of the local heterogeneity in the seismic velocity field. unclear It is generally assumed a-priori as a maximum distance between hypocenters in order to consider a couple of events to be included in the DD data, together with a high value of cross-correlation for their waveforms. pair A DD datum is the differential travel-time for the selected ~~couple~~ pair of events. ~~even if~~ The same scheme has been applied to more complex analyses, Since ~~like~~ full waveform inversion (Lin and Huang, 2015). Based on the assumption of ~~almost~~ nearly co-located events, the information contained in the DD datum can be used to refine event locations (e.g. small events referred to a master event, Waldhauser and Ellsworth, 2000) or the seismic properties of the rocks in the area where events are clustered (e.g. Zhang and Thurber, 2003). In recent years, seismic monitoring of subsurface processes has also been realized through seismic tomography (e.g. Chiarabba et al., 2020) and in particular with the analysis of DD data: rock weakening due to mining activities (Qian et al., 2018; Ma et al., 2020; Luxbacher et al., 2008), granite fracturing during geothermal well stimulation (Caló et al., 2011; Caló and Dorbath, 2013) and oil & gas operations (Zhang et al., 2006). For monitoring purposes, an additional assumption is considered during DD data preparation: elastic properties of the media traversed by the seismic waves should not change between the occurrence of the selected pairs of events. This fact implies the computation of the so called *time-lapse* analysis, where pre-defined time-windows are considered and static images of the subsurface (Caló et al., 2011), or differential elastic models (Qian et al., 2018), are reconstructed for each time window. In any case, the most relevant issue in time-lapse tomography remains how to define the time windows, which artificially separate events and prevent their coupling to obtain DD data. How many time windows are meaningful to construct DD data? And which should be their time lengths? This issue is critical due to the dependence of the number of DD data ~~from~~ on the number of events coupled and, thus, from the number of time windows, as schematically shown in Figure 1.

The definition of the set of time windows, on which the sequence of 3D time-lapse ~~tomographies~~ inversions should be computed, ~~is~~ demands ~~to the~~ expert opinion. There are three main possibilities in time-lapse tomography: (a) imposing time windows based on known seismic history (before and after a known, relevant seismic event; Young and Maxwell, 1992; Chiarabba et al., 2020); (b) keeping the same length for all time-windows (e.g. one day, Qian et al., 2018), or (c) trying to have the same amount of data in all the time windows (e.g. Patanè et al., 2006; Kerr, 2011; Zhang and Zhang, 2015). In other cases, the lengths of the time windows vary based on research needs (e.g. Caló et al., 2011). A human-defined set of time windows might mask the real variations of the physical properties, ~~and~~ in which case the time-evolution of the elastic model found ~~can~~ could be not associated ~~to~~ with the investigated geophysical process.


(a) One changepoint, N_D=18
tt_{ALF}
Double-difference datum
"change-point"
Earthquake
Time
(b) Two changepoints, N_D=9
tt_{ALF}
Additional "change-point"
Time

fix

Figure 1. Schematic example of standard preparation of DD data in different time-windows. Time-windows are defined by changepoints (also called "hard-partitions"). Here, for sake of simplicity, we represented the travel-time to station ALF for each seismic events (yellow stars) as a function of origin time. A DD datum (curved black line) is prepared for each pair of events not separated by a changepoint. (a) Here, only one changepoint is present, so $N_D = 19$ DD data can be prepared. (b) In case of two changepoints, only $N_D = 9$ DD data can be prepared.

by — does not make sense

Here, we tackle the issue of defining the number and time-length of the time-windows in DD data preparation though a novel approach. To simplify the experiment, we focus on closely associated events recorded on a volcanic edifice in Iceland. Such a cluster of events, which spans no more than 100 meters in diameter, is considered as a punctual source of repeating events, that ~~which~~ are recorded from a seismic station 6 km away, for more than two years continuously. ~~In~~ this way, we assume perfectly co-located events and we can focus on time-variations of DD data. More generally, the novel approach can be applied to both temporal and spatial associations (i.e. to define both time-windows and spatial-length for associating events, and composing DD data).

1.2 Background on Bayesian inference, Markov chain Monte Carlo sampling and trans-dimensional algorithms

Geophysical inverse problems have been solved for a long time following direct search or linearized inversion schemes, due to the limited number of computations needed to obtain a solution. Such solutions have been given in the form of a single "final" model, presented as representative of the Earth's physical properties. Thanks to the computational resources now available, such approaches are outdated for more sophisticated and cpu-time consuming workflows, ~~where~~ multiple models are evaluated and compared, to obtain a wider view of the Earth's physical properties. Algorithms based on Bayesian inference belong to this second category, ~~where~~ the "solution" is no ~~more~~ a single model, but a distribution of probability on the possible value of

for which longer CPU in which

https://doi.org/10.5194/se-2021-79
Preprint. Discussion started: 11 June 2021

the investigated parameters, following Bayes theorem (Bayes, 1763):

$$p(\mathbf{m} \mid \mathbf{d}) = \frac{p(\mathbf{m})p(\mathbf{d} \mid \mathbf{m})}{p(\mathbf{d})} \tag{1}$$

where $p(\mathbf{m} \mid \mathbf{d})$ represent s the information obtained on the model parameters $\mathbf{m}$ through the data $\mathbf{d}$, the so called "posterior probability distribution (PPD)", or simply "posterior". Such information is obtained combining the prior knowledge on the model: $p(\mathbf{m})$, with the likelihood o~~f~~ the model ~~to~~ given the data: $p(\mathbf{d} \mid \mathbf{m})$. The denominator of the right -hand term is called "evidence" and represents the the probability of the data in the model space:

$$p(\mathbf{d}) = \int p(\mathbf{m})p(\mathbf{d} \mid \mathbf{m})d\mathbf{m}\,. \tag{2}$$

The evidence is a high-dimensional integral that normalizes the PPD. It is generally difficult to compute and, thus, methods ~~which~~ that do not require its computation (~~like~~ such as Markov chain Monte Carlo, McMC, see below) are widely used in Bayesian inference.

The likelihood of the data for a given model is necessary to evaluate and compare different set s of model parameters. It is generally expressed as~~:~~

$$L(\mathbf{m}) = p(\mathbf{d} \mid \mathbf{m}) = \frac{1}{(2\pi|C_e|)^{1/2}} e^{-\frac{1}{2}\phi} \tag{3}$$

where ϕ represents the fit between model prediction p_i and the actual value of the i-th observation o_i, i.e. $e_i = (o_i - p_i)$, through ~~the covariance~~ matrix of the error C_e:

implies this is error but it is misfit

this needs rewriting

C is the data covariance matrix, not "error"

$$\phi = \mathbf{e}^T C_\mathbf{e} \mathbf{e}\,. \tag{4}$$

Due to the difficulties in computing the evidence and the analytic solution of Equation 1, and thanks to the improved computational resources, in the last two decades the emerging trend in Bayesian inference is represented by "sampling methods", where Which the direct computation of Equation 1 is substituted replaced by the sampling of the model space according to the PPD (Sambridge and Mosegaard, 2002). One of the most famous sampling methods is called Markov chain Monte Carlo, where for which the chain samples the model space according to probability rules, like such as Gibbs sampler or Metropolis rule (Metropolis et al., 1953; Gelman et al., 1996). Briefly, starting from a given point in the model space, called the *current model*, a new point of the model space, called the *candidate model* is proposed and visited evaluated? according to some rules based on the PPD. In particular, the Metropolis rule coupled to the approach developed ~~in~~ Mosegaard and Tarantola (1995), which is the workflow adopted in this study, accepts or rejects to move from ~~a~~ current model to a candidate model according to the ratio of their likelihoods, i.e.:

$$\alpha = L(\mathbf{m}_{cand})/L(\mathbf{m}_{cur}). \tag{5}$$

This is a simplified version of a more general formulation of the acceptance probability in Metropolis-based McMC (Gallagher et al., 2009). It is worth noticing noting that our workflow does not directly specify the dimensionality of the model space. In fact, following the recent advancements in the solution of geophysical inverse problems, we do ~~no more~~ not consider models with a fixed number of parameters, but we make use of the so called trans-dimensional (trans-D) algorithm~~s~~ and propose candidate models with a different number of dimensions with respect to the current models. This approach is called trans-~~dimensional~~ D sampling and it has been widely used for the solution of geophysical inverse problems in the last decades~~,~~ (Malinverno, 2002; Sambridge et al., 2006; Bodin et al., 2012a; Dettmer et al., 2014; Mandolesi et al., 2018; Poggiali et al., 2019). Trans-D algorithms have been proven to be intrinsically "parsimonious" (Malinverno, 2002) and, thus, they preferably sample simpler models with respect to complex ones. This is one of the most important characteristics of trans-D algorithms, enabling a fully data-driven~~,~~ solution for the model parameters.

2 Data

initiated

We use data from a cluster of repeating earthquakes located on the southern flank of Katla volcano (Iceland; Figure 2a). This seismic activity originated in July 2011 following an unrest episode of the volcano (Sgattoni et al., 2017) and continued for several years with remarkably similar waveform features over time. The cluster is located at very shallow depth (< 1 km) and consists of small magnitude events (~ −0.5 - 1.2 ML)~~,~~ characterized by emergent P waves and unclear S waves, a narrow-band~~ed~~ frequency content around 3 Hz at most stations, and correlation coefficients well above 0.9 at the nearest stations during the entire sequence. The temporal behavior is also peculiar, with a regular event rate of about 6 events per day during warm seasons gradually decreasing to one event every 1-2 days during cold seasons (Sgattoni et al., 2016b). Sgattoni et al. (2016a) obtained relative locations of 1141 events recorded between July 2011 and July 2013 by designing a method optimized for very small clusters that includes the effects of 3D heterogeneities and tracks uncertainties throughout the calculation. The number of relocated events depends on a selection of the best events among a total of > 1800, based on thresholds on correlation coefficient and amount of detected P and S phases. The resulting size of the cluster is on the order of $25 \times 50 \times 100$ m^3 (easting, northing, depth), with estimated uncertainties on the order of few tens of meters (Figure 2b). Changes in the station network configuration around the cluster occurred due to technical problems, with the greater loss of data in the second part of the sequence, ~~from~~ after January 2012. This coincides with a clear increase in relative location uncertainties, which correlates also with a decrease in correlation coefficients, mainly for S phases. Other temporal changes in waveform correlation were identified by Sgattoni et al. (2016a) in August 2012 and January 2013. In this study we focus on P-wave data recorded at station ALF (part of the Icelandic Meteorological Office seismic network), which is located about 6 km away from the cluster (Figure 2a) and is the only nearby station that has been continuously operating during the entire time. The similarity of the waveforms recorded at ALF is remarkable, with correlation coefficients of the biggest events above 0.99 throughout the entire period of study (Figure 2c). To compute the DD dataset, we use the origin times (OT_i) of $N_e = 1119$ relocated events from Sgattoni et al. (2016a). We remark that the increased location uncertainties due to the network geometry change in January 2012 may affect the quality of the locations the events and, consequently, the determination of their OT, which is relevant for computing uncertainties in the DD data (see Section 2.1).

of

largest

origin times

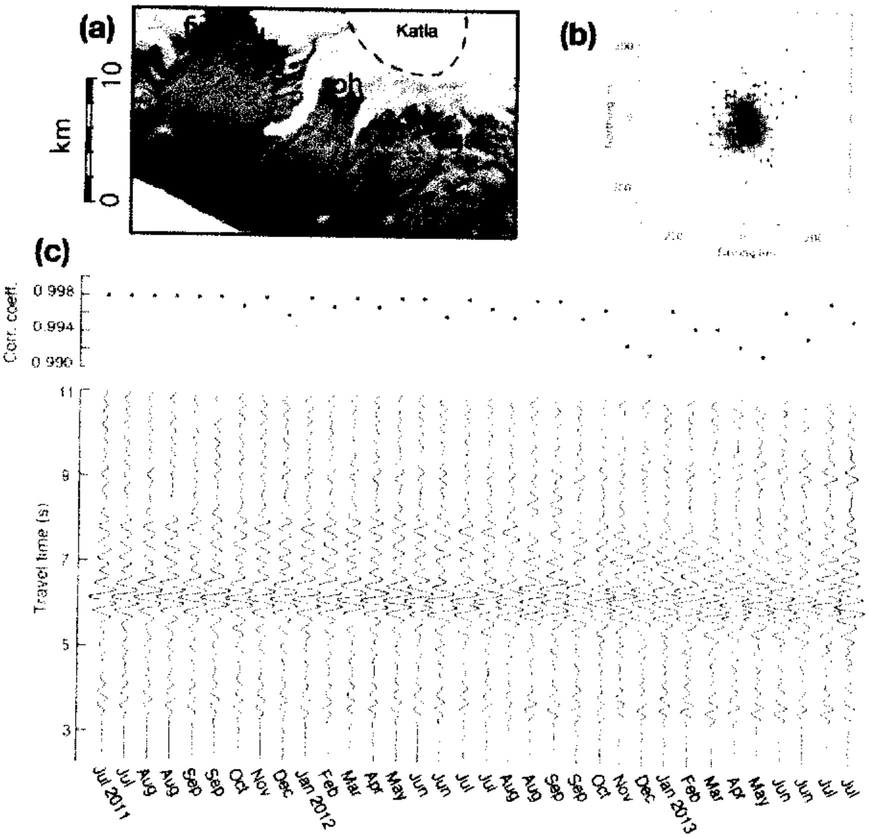


Figure 2. (a) Map of the southern flank of Katla volcano (Iceland; topography information from the National Land Survey of Iceland). The caldera rim is outlined by the black dashed line. White areas are glaciers. The star marks the location of the seismic cluster. Dark brown triangles: permanent Icelandic Meteorological Office (IMO) seismic stations. Orange triangles: temporary Uppsala University seismic stations operating between May 2011 and August 2013. (b) Relative locations (blue points) and uncertainties (± std; grey lines) from Sgattoni et al. (2016a) (c) Example waveforms of the Z component recorded at station ALF throughout the entire period investigated and correlation coefficients of the P waves with respect to the master event used for the relative locations shown in (b). Panels (a) and (b) have been modified from Sgattoni et al. (2016a). Panel (c) has been modified from Sgattoni et al. (2016b).

uncertainties ?

2.1 Data uncertainties from full-waveform investigation

To apply our novel Bayesian approach, we need to estimate a covariance matrix of the errors in the DD data. Having an origin time for each event (given by the location obtained in Sgattoni et al. (2016a) using the full seismic network), we derive the DD data and their uncertainties directly comparing the raw waveforms and finding the absolute delays between each couple of events. From the absolute delays of the P arrivals, the subtraction of the time differences in the OTs of two events gives the DD datum for each pair ~~such couple~~. We estimate the absolute time delay between two events following the Bayesian approach described in Piana Agostinetti and Martini (2019). Briefly, we collected a 20-s record of each event, centred on the approximate P-wave arrival time. We compute a stacked version of the events, so called "wavelet" (Figure 3a). From the wavelet, we compute ~~compute~~ the residuals for each event (Figure 3b). Event residuals define a standard deviation function $\sigma(t)$ for the 20s record. Event residuals are also auto-correlated to obtain an averaged auto-correlation function $r(t)$. The standard deviation and the auto-correlation function are used to define a covariance matrix $\mathbf{C_{e,w}}$ (the same for all the waveforms, Piana Agostinetti and Malinverno, 2018) using the equation:

pair

not a normal use of the term wavelet

this needs a clearer and more complete explanation

$$\mathbf{C_{e,w}} = \mathbf{S}\,\mathbf{R}\,\mathbf{S}, \qquad (6)$$

uncertainty ?

uncertainties ?

where $\mathbf{C_{e,w}}$ is the covariance matrix of the waveform errors, $\mathbf{S}$ is a diagonal matrix containing the standard deviation $\sigma(t)$ computed from the residuals, and $\mathbf{R}$ is a symmetric Toeplitz matrix whose rows and columns contain the auto-correlation function $r(t)$ with $t = 0$ on the diagonal. In this way, we reconstruct a full covariance matrix, which can be used to obtain realistic error estimates for our DD data. Noteworthy, the use of a diagonal covariance matrix instead of a full covariance matrix would risk to underestimate the errors, biasing the subsequent analysis for defining the DD time windows. Having the error model for the waveforms, for each couple of waveforms, we perform a Markov chain Monte Carlo sampling (Mosegaard and Tarantola, 1995) to reconstruct the PPD of the time-shift between the two waveforms. Following Sgattoni et al. (2016a), we use a 1s-long time window to compute the likelihood of the waveforms, centred on the approximate P-wave arrival time. Starting with $N_e = 1119$ events, we obtain $N_e \times (N_e - 1)/2$ DD data. The total number of DD data is $N_D = 625521$. DD data value d_{ij} and uncertainties σ_{ij}, associated to events i and j, are reported in Figure 4. Striking changes in DD values suggest the presence of clustering of data in time, but the exact number and positions of such clusters are not easy to define by visual inspection. Moreover, some of those changes could be given by modifications of the seismic network, in principle mapped in our DD data. In fact, our DD data depend on the Origin Time computed exploiting all the recordings from the seismic network. Thus, a change in the seismic network configuration could influence the quality of seismic network detections and locations, which in turn could introduce a bias in our data, as a shift in the DD value or an increase in DD error. Given the independent processes used to define each single DD datum, as a first approximation we consider our final covariance matrix of the errors in the DD data $\mathbf{C}_e^*$ as a 625521×625521 diagonal matrix, with the square of the uncertainties presented in Figure 4b along the diagonal, omitting the correlation between errors given by, e.g., biases in OT determination.

how do you really justify ignoring off-diagonal values ?

with

OT

due to

uncertainties ?

https://doi.org/10.5194/se-2021-79
Preprint. Discussion started: 11 June 2021

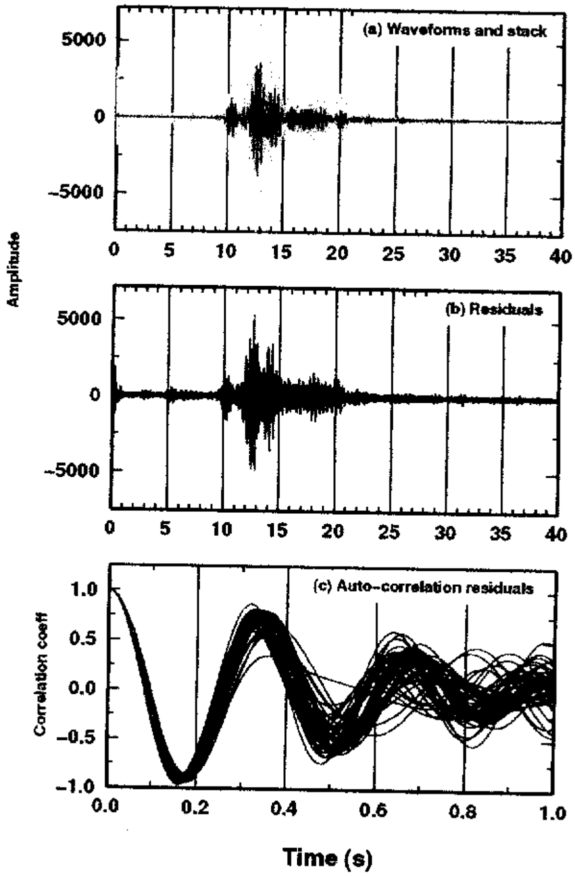


Figure 3. (a) Original waveforms (grey lines) and their stack, called "wavelet" (orange line). (b) Residual of each single waveform (grey lines) with respect to the "wavelet". As a reference, the residual for the first (last) trace is shown as a blue (red) line. (c) Auto-correlation of the single residuals (grey line) and averaged value of the autocorrelation (orange line).

3 An algorithm for exploration of double-difference data space

What happens to the DD data-set if we create a hard partition in time, i.e., if we artificially separate some events from the others? As clearly illustrated in Figure 1, the number of data N_D in the DD data set varies, decreasing for increasing number 200 of hard partitions. From a Bayesian point of view, this is not admissible, because Equations 3 and 5 need to consider the same number of data points in two models to allow their comparison (see also Tilmann et al., 2020).

Our novel approach to solve this issue relies on the introduction of a family of "hyper-parameters", which represent the partitions of the events, and such hyper-parameters are used for scaling the different entries in the data covariance matrix $\mathbf{C}_e^*$. In our approach, the number of "hyper-parameters" in the family is not fixed, but it is directly derived from the data 205 themselves. Following a Bayesian inference approach, we reconstruct the statistical distribution of the hyper-parameters (i.e.

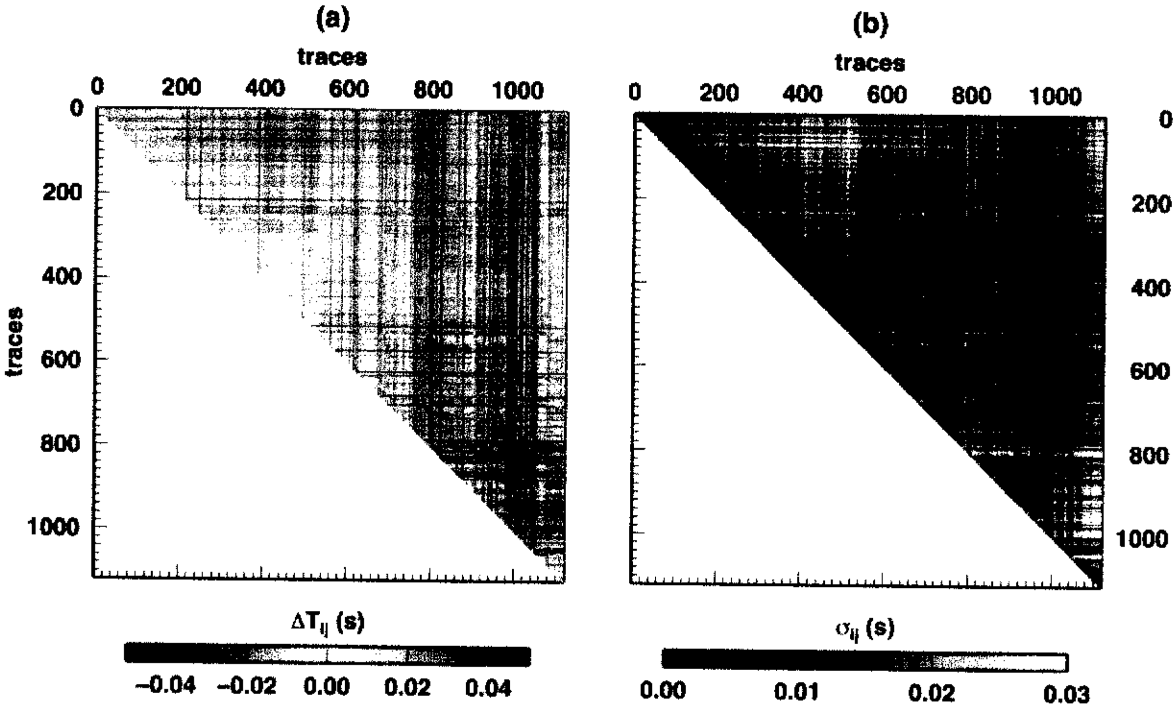


Figure 4. DD data presented as a matrix of ij-couples: (a) values; and (b) uncertainties.

such as

uncertainty ?

event partitions) in time through ~~a~~ trans-dimensional McMC sampling. Hyper-parameters have been introduced in geophysical inversions, for estimating the data uncertainties, expressed, for example, as the variance of a Gaussian distribution (Malinverno and Briggs, 2004). Hyper-parameters are generally part of the model vector together with physical parameters. As stated in Bodin et al. (2012b), estimated hyper-parameters do not only account for measurement errors, but include other contributions that build up the uncertainty in the geophysical inversions, like simplification of the physics included in the forward solutions, or simplified model parameterization. Hyper-parameters have been used to estimate error models (Dettmer and Dosso, 2012; Galetti et al., 2016) or to discriminate between two different families of error models (Steininger et al., 2013; Xiang et al., 2018). In this last case, the number of hyper-parameters belonging to a model vector is not constant, but can be one or two, depending on the family. More recently, a nuisance parameter has been introduced to evaluate the probability for each datum to belong to the "regular data" or to the "outliers" (Tilmann et al., 2020).

For the DD case, we introduce a family of hyper-parameters to estimate which portions of the DD data violate our initial assumptions. In fact, in our assumptions, double-difference data are computed from couple of seismic events occurred in the same rock volume, recorded at the same seismic station. For perfectly co-located events, and in absence of any change in the

pairs

that

rock seismic velocity field between the first and the second event, double-difference measurements should have mean 0 and should be distributed following the Gaussian error model defined in Section 2.1, represented by the covariance matrix $\mathbf{C}_\mathbf{e}^*$. In this case, the value of the fit $\tilde{\phi}$, expressed as:

$$\tilde{\phi} = \mathbf{d}^T \mathbf{C}_\mathbf{e}^* \mathbf{d} \tag{7}$$

of

with $\mathbf{d}$ the DD data vector, should be close to N_D.

When the value $\tilde{\phi}$ significantly deviates from N_D, a modified covariance matrix $\mathbf{C}_\mathbf{e}(\mathbf{m})$ should be considered, where the portion of the data ~~un~~ in consistent with the hypotheses are considered differently from the portions of DD data ~~which~~ that do not violate the hypotheses. The new modified covariance matrix $\mathbf{C}_\mathbf{e}(\mathbf{m})$ is obtained as

$$\mathbf{C}_\mathbf{e}(\mathbf{m}) = W(\mathbf{m})\ \mathbf{C}_\mathbf{e}^*\ W(\mathbf{m}) \tag{8}$$

where the matrix $W(\mathbf{m})$ is a diagonal matrix ~~which~~ that contains a weight for each DD datum, based on the hyper-parameters. ? Noteworthy, if we use Equation 8 in Equation 3, we see that in our case the dependence of the Likelihood function on the model does not reside anymore in the residuals, as is generally ~~done~~ the case in geophysical inverse problems, but only in the Covariance matrix. However, for a simple case like ours, we highlight that this dependence could be in principle moved back to the residuals, if we allow the physical assumptions to be variable in time (i.e., if we allow the elastic model to change in time, which in our assumption can not). such as

The fully novel idea in our algorithm resides in the trans-dimensional behaviour of our exploration of the data space. In fact, the number of hyper-parameters in the model (and, thus, the number of partitions of $\mathbf{C}_\mathbf{e}^*$) is not fixed and can change along the McMC sampling ~~between a minimum and a maximum~~. At the end of the McMC sampling, we can compute a PPD of the number of partitions in the problem, an information fully dictated from the data and priors. defined or prescribed

3.1 Model parameterization

In our algorithm, a model is described by a set of k changepoints that define the partitions of $\mathbf{C}_\mathbf{e}^*$, and their associated quantities, that is: $\mathbf{m} = (k, T_k, \pi_k)$. The k-vector T_k represents the time-occurrence of the k changepoints, while the k-vector π_k contains the weights associated to each changepoint. We assume that a DD datum d_{ij}, associated to event i and j, retains its original variance σ_{ij}^2 if no changepoint occurs between OT_i and OT_j. Otherwise, its importance is modified with weight $\mathbf{W}_{\mathbf{ij}}(\mathbf{m})$:

With   events   whereas

$$\mathbf{W}_{\mathbf{ij}}(\mathbf{m}) = 10^{w_{ij}(\mathbf{m})}, \tag{9}$$

where w_{ij} is computed as:

$$w_{ij}(\mathbf{m}) = \sum_{m=1}^{n} \pi_m \qquad if \qquad OT_i < T_m < OT_j \tag{10}$$

This is really confusing!

① m is a vector and an index of a sum?

② Change point present means W_{ij} is large? (sum over many π) paired

meaning?

with

recalling that **W** is a $N_D \times N_D$ diagonal matrix and W_{ij} represents the element along the diagonal associated to DD datum d_{ij}. Following our approach for defining the $\mathbf{W}_{ij}(\mathbf{m})$, specifically the summatory of the weights associated to the relevant changepoints we assume that a couple of "distant" events in time has more probability of being "separated" by one or more changepoints. This assumption reflects the standard process of DD data, where distant (in space and/or time) events are almost never coupled in DD data-sets. However, if no changepoints are present between distant events, our trans-dimensional approach works out. Several synthetic tests, not shown here, demonstrate that non-necessary changepoints are removed from the family due to the parsimoniosity of the trans-dimensional algorithm, and new ones have limited probability of being accepted in the family.

rephrase please

is this a word?

in which

include these!! Even as appendix

3.2 Candidate selection

3.2.1 Recipe

carrying out?

cpu

Having an efficient workflow for moving the McMC sampling is fundamental for keeping the ~~cpu~~-time within acceptable limits. From a theoretical point of view, any recipe can be implemented at the core of the McMC due to the fact that results (i.e. Equation 1) do not depend on the McMC details[1], i.e., the same prior information jointed with the same data will give the same PPD, whatever recipe is selected for the McMC sampling. However, inefficient recipes can take too long to sample adequately the PPD and, thus, from a practical point of view, the users should spend some time in defining a proper recipe. In our case, to perturb the current model and propose a new candidate model, we randomly select one of the following four "moves":

reword

1. (This move is randomly selected with probability $P_1 = 0.4$) The i-th changepoint is moved from its time-position T_i. There are two equally-possible perturbations: the changepoint time-position T_i is randomly selected from the prior, or the changepoint time-position T_i is slightly perturbed from the original value in the current model with a micro-McMC approach (Appendix A2, Piana Agostinetti and Malinverno, 2010)

   Meaning of micro?

2. ($P_2 = 0.4$) The weight π_i of the i-th changepoint is perturbed with a micro-McMC approach (Appendix A2, Piana Agostinetti and Malinverno, 2010)
3. ($P_3 = 0.1$) Birth of a changepoint: a new changepoint is added to the current model
4. ($P_4 = 0.1$) Death of a changepoint: a changepoint is removed from the current model.

How about addition and removal instead?

in which

The last two moves represent the trans-dimensional moves, where the dimensionality of the model is changed from the current model to the candidate. For move (3), we follow the approach described in Mosegaard and Tarantola (1995) and we propose a completely new changepoint with T_{k+1} and π_{k+1} randomly sampled from their prior distributions. For move (4), we simply randomly select one changepoint and remove it from the model.

[1] as long as the recipe follows the necessary probabilistic rules (Sambridge and Mosegaard, 2002; Mosegaard and Sambridge, 2002)

How is the "move" actually selected?
Is model "fitness" used as in Metropolis-Hastings?

3.2.2 Prior information

reword

Uniform prior probability distributions are selected for our inverse problem. Here, the number of changepoints is comprised between 0 and 100. Changepoints can be distributed everywhere in time between 2011.5 and 2013.7. To make the algorithm more efficient, we set a minimum distance between two changepoints as large as 0.5 day (Malinverno, 2002). Changepoint weights π_k follow ~~a~~ uniform prior probability distribution between 0.0 and 1.0.

4 Finding data-driven time-variations of rock elasticity during Katla's seismic swarm

changepoints?

We apply our novel methodology for the definition of the hard-partitions in DD data to the data set recorded on Katla volcano in Iceland, during a two year monitoring experiment. Based on the observations of the limited dimension of the cluster with respect to the events-station distance (100 m versus 6.0 km) and the overall high similarity of the waveforms (correlation coefficient always larger than 0.9), our algorithm is able to map out which portions of the data violate our underlying hypotheses: co-located events and constant elasticity field in time. ~~While~~ separating those two effects with a single station would be challenging, but here we want to illustrate in detail how the time-occurrence of the changepoints is defined and compare them to other potential approaches for the definition of hard-boundaries in DD data, namely, variations in seismicity rate and waveforms cross-correlation.

Stop using different words

As shown in Figure 1, defining hard-partitions for DD data using expert opinion is a dangerous task, due to the limitation in the number of data available for subsequent uses. For example, seismologists could be tempted to test if using less data could give better images in a subsequent seismic tomography, based on pre-defined ideas on the subsurfaces structures. In fact, some changes in DD data are obviously present in the observations (Figure 4, between close to event 550 and 1050 for example), but others are more subtle to be defined.

fewer

tomographic inversion

seismic tomography is a method

please rephrase

about

for which

We compute the data-space exploration running 100 independent McMC samplings, where each chain sampled 2 Millions ~~of~~ of changepoint models. We discarded the first million ~~of~~ models and collected one model every 1000 in the subsequent Million models. Our final pool of models used for reconstructing the PPD is composed by 100 000 models. The full cpu-time for running the algorithm is about 19 hours on a 100-CPU cluster. The value of ~~the~~ chi-square decreases in the first half-million ~~of~~ models, together with the logarithmic value of the normalizing factor in Equation 3 (Figure 5a). The number of changepoints reaches a stable value around 15 to 20 after 1 Million ~~of~~ models confirming the length of the burn-in period used (Figure 5b). The ratio between the number of event pairs not separated from any chagepoint (N_F) and the total number of DD data (N_D) is also stable around 0.2 after 1 Million ~~of~~ models, indicating that no relevant changepoint is added in the second half of the McMC sampling. It is worth noticing that N_F should approximately indicate the number of DD data to be used in any subsequent analysis.

Looking to the full details of the PPD reconstructed from the McMC samplings, we observe the presence of long time-windows completely without any changepoint (e.g. between 2011.7 and 2012.4), demonstrating the parsimoniosity of the trans-D approach: if changepoints are not supported by the data they are removed during the sampling and do not appear in the final PPD (Figure 6e). Moreover, the most probable (relevant) changepoint (changepoint number 3 in Figure 6e) perfectly aligns

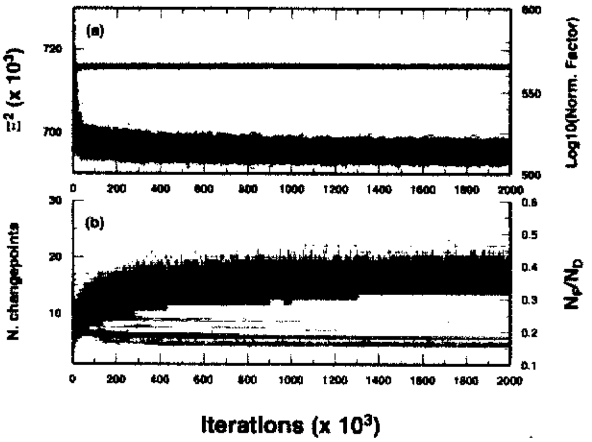


Figure 5. Evolution of some parameters along the McMC sampling. (a) Chi-square value (blue crosses) and logarithmic value of the normalizing factor in the likelihood function (red dots). (b) Number of changepoints in the sampled models (blue crosses) and ratio between the number of un-affected data (i.e. DD data where the two events are not separated by any changepoint) and the total number of DD data, N_F/N_D (red dots).

bold not italics previously

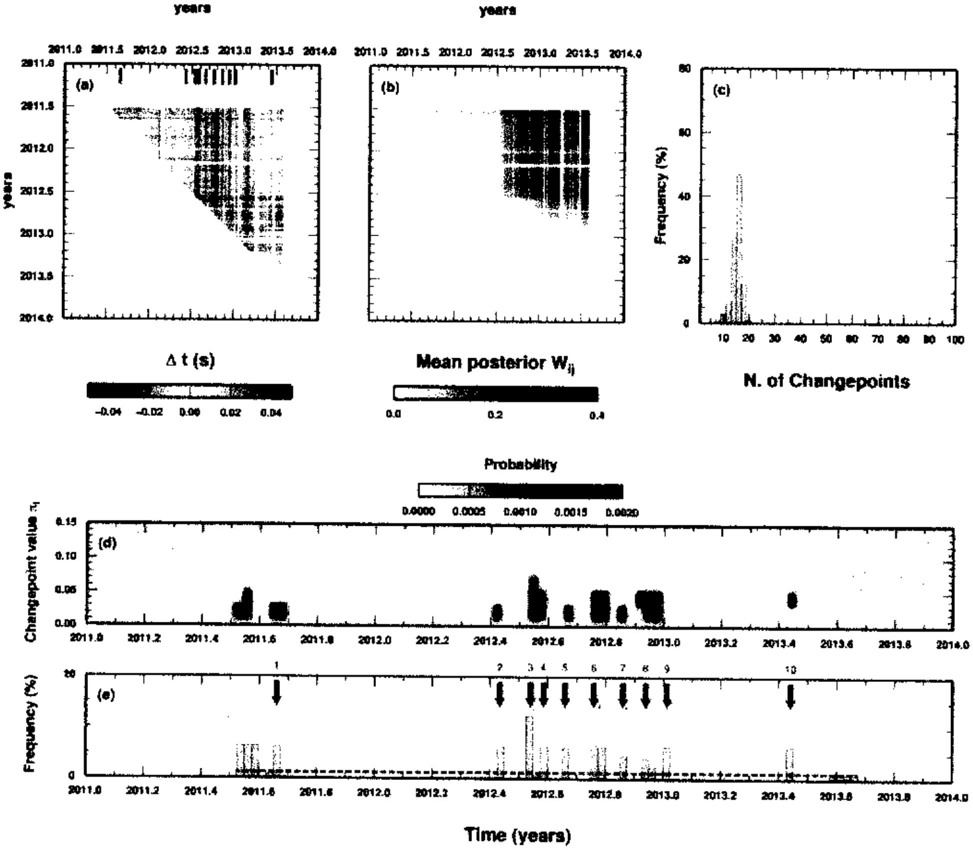
with one of the most striking changes in the DD data as shown in Figure 6a, confirming the goodness of the approach. The distribution of the weights clearly defines the partition of ~~the~~ $C_e(\mathbf{m})$, where initial data (i.e., DD data related to events occurring at the initial stages of the swarm, in 2011) slowly release their "connection" to later events and, thus, indicate that they should not be included in the subsequent analysis. From the histogram of the number of changepoints in each sampled model, we can see how the trans-D algorithm works: no less than 10 and no more than 20 changepoints are generally considered, even if we allow ~~such~~ the number to increase to ~~one hundred~~ 100. Combining this information with the distribution of chagepoints in time given in Figure 6e, we define eleven relevant changepoints (red arrows). We acknowledge that ~~such~~ this number could be, again, a subjective choice, however, looking ~~to~~ at Figure 6d, we see that changepoints can be "ranked" in some sense given their mean PPD weights. For example, changepoints 2, 5, and 9 have clearly associated lower weights ~~with respect~~ compared to the others and, thus, should be considered as less relevant. Our methodology does not solve all issues connected to the preparation of DD data, but, at least, it can be used to quantify the occurrence of changepoints and their importance, and such quantification can be exploited in a later stage depending on the subsequent analysis planned.

I only see 10 in Fig. 6e

We compare the time occurrence of the resulting 11 changepoints with the cross-correlation coefficients between each event in the seismic swarm and the largest one (see Sgattoni et al., 2016a, for details). Both P-wave and S-wave cross-correlation coefficients display some degree of variability and defined patterns in the time window used in this study (Figure 7), even if we ~~should~~ consider that the smaller values are always larger than 0.9. We observe that there is no clear correlation between changepoint position in time and cross-correlation values. In mid 2011, around evetn 300, and early 2012, around evetn 700, we have two changepoints where cross-correlation seems stable for both P and S waves. At the beginning of 2012, when

for which

meaning?



with

Figure 6. Results of the application of the algorithm to the Katla data set. (a) DD data and position of the most probable changepoints, see panel (e). Changepoints occurrence in time is indicated by red arrows on top. (b) Mean posterior values for the weights associated to each DD datum. (c) Histogram of the number of changepoints in the sampled models. (d) 1D marginal PPD for the values of the changepoints in time. (e) Histogram of the distribution of changepoints in time. Red arrows and numbers indicate the most probable time occurrence for a changepoint.

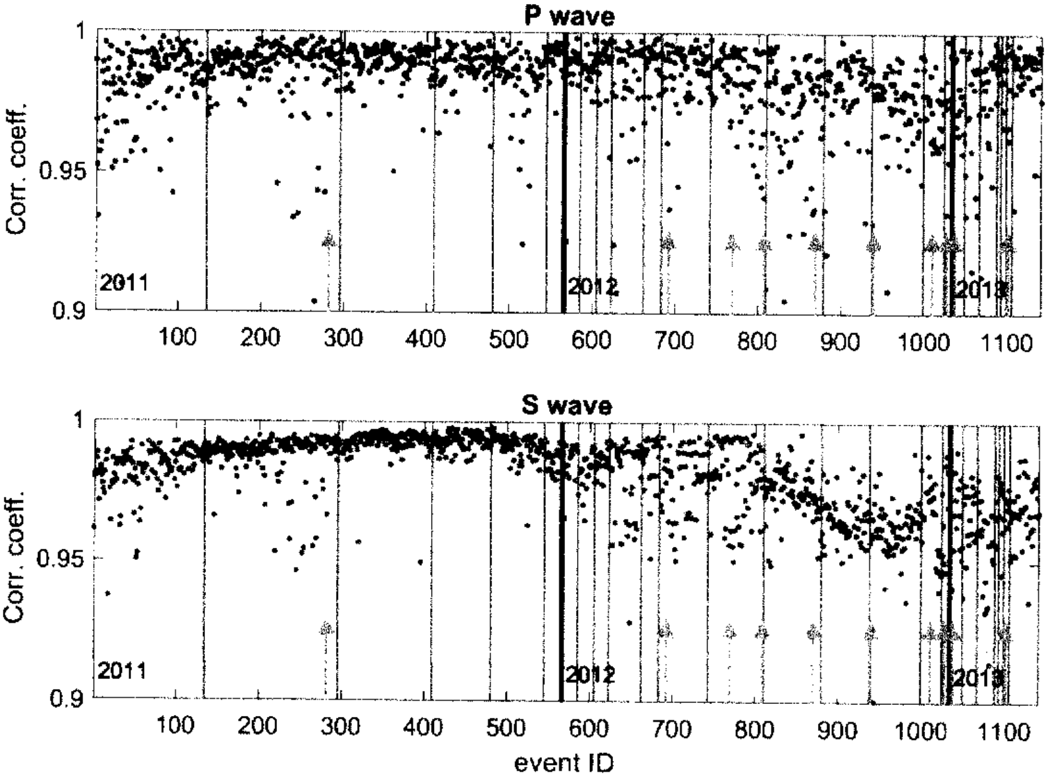


Figure 7. Correlation coefficients for P-waves (a) and S-waves (b) for each event with respect to the largest event (see Sgattoni et al., 2016a, for details). In each plot, the most probable time-occurrences for the 11 relevant changepoints are also reported (yellow arrows).

the seismic network ~~has been~~ was redefined, the cross-correlation for S-waves changes dramatically, ~~while~~ whereas the change in cross-correlation for P-waves is less evident, and no changepoint is found at all. Our results seem to indicate that variations in cross-correlation coefficients (for example, computed for repeating earthquakes) could indicate an unrealistic variations in elasticity and could be a problematic choice for a monitoring system of the subsurface.

With

We also compare the position of the retrieved changepoints with the seismicity rate, another parameter usually associated to time-variations of the subsurface properties (e.g. Dou et al., 2018). no hyphen In Figure 8, we report the seismicity rate every two weeks. The rate of events is highly variable along the studied time-window, with values ranging between a few and more than 30 events/week. The seismicity rate decreases in 2012, with some bursts up to 15 events/week in late 2012. As for the cross-correlation coefficients, the position of the retrieved changepoints does not simply correlate with the time-history of the seismicity rate. We have found changepoints where for which the seismicity rate is very high (2011.6) and very low (e.g. 2013.4). The most probable changepoint (changepoint number 3) occurs in a period of sustained seismicity rate ~~which~~ that starts 5-6 weeks

https://doi.org/10.5194/se-2021-79
Preprint. Discussion started: 11 June 2021

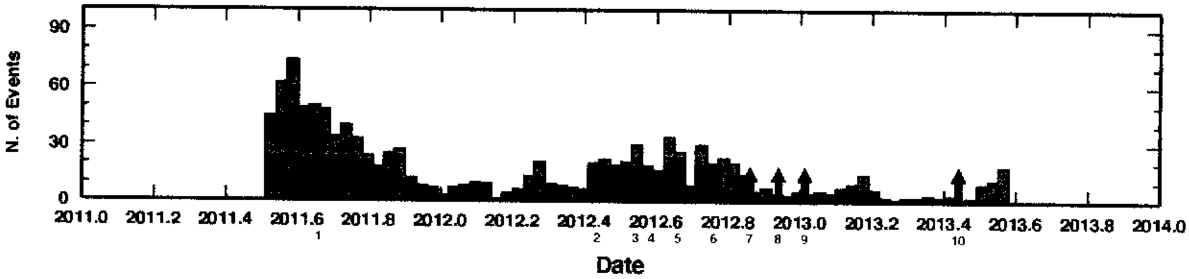


Figure 8. Comparison between the seismicity rate (grey histogram) and the time-occurrences for each changepoint (red arrows and numbers).

before. If our changepoints indicate variations of subsurface elasticity, the time-history of seismicity rate should be carefully evaluated before using it for tracking elasticity changes in time.

It would be more valuable as a paper if you did something tangible with your new method.

5 Discussion

The DD data recorded on Katla volcano and the results presented here clearly indicate that time variations in elastic properties ~~should have~~ occurred between 2011 and 2014 on the south flank of the volcanic edifice. Thus, data-driven time windows can be found using our approach to define where when to apply standard DD analysis for retrieving elasticity variations, with no need of preparing the DD data following subjective choices on the coupled paired events. ~~Being~~ With the algorithm being naturally parsimonious, there is almost no possibility of having "no DD data" (i.e. one partition per event). Data partitions are always ~~in a~~ limited in number, even if, strictly speaking, their number should be given by the user because, as final output, we have the full PPD and not just one set of partitions. Defining the exact number of partitions to use in subsequent analysis depends on the analysis itself. awkward, rewrite

Our approach quantifies the presence and the relevance of the changepoints. Using such information could be straightforward in some cases (e.g. if we look for one most probable changepoint only) or more complex (e.g. if we also wish to appreciate correlation between changepoints, which can be measured using the PPD). It is worth noticing that, in simple cases, our algorithm generally performs as expert opinion (e.g. in the case of the search for one most probable changepoint), and this confirms the overall performance of our methodology. In more complex cases, the weights associated ~~to~~ with the changepoints should be used to classify the changepoints themselves, and this allows selecting the most relevant changepoints using quantified information.

The network of seismic stations deployed around Katla volcano changed in time. This fact has been previously indicated as a potential "bias" in the analysis of the seismic data themselves, as the location uncertainties increased after major network operations (January 2012). Our results point out that the changepoints found do not correlate with such change in the seismic network. Being a statistical analysis, our methodology seems to be insensitive to changes in the acquisition system. Alternatively, the changes in location uncertainties ~~could be not~~ may not be large enough to affect our procedure. In both cases, our approach demon-

correlation in what sense? Unclear

https://doi.org/10.5194/se-2021-79
Preprint. Discussion started: 11 June 2021

strates to be well-suited for handling long-lived databases, in which ~~where~~ changes in the spatial distribution of observational points is likely to occur from time to time.

Finally, we investigate how our changepoints relate to more commonly-used indicators of sub-surface variations of elasticity, Such ~~alike~~ time-series of cross-correlation coefficients and seismicity rate. In both cases, we found poor correlation between our results and the time-series of the two quantities. While this observation is not totally unexpected since the two time-series are based on different seismological observables, it suggests that care should taken when investigating time-variations of elasticity retrieved from methodologies based on cross-correlation, and to reasses approaches based on variations of the seismicity rate as a proxy of "rock instabilities" (Dou et al., 2018).

Although

6 Conclusions

We developed an algorithm for defining data-driven partitions in a seismic database, for a more objective definition of double-difference data. The algorithm is based on the trans-dimensional sampling of data-structures, here represented as partitions of the covariance matrix. The algorithm has been tested in the case of a seismic database acquired in a volcanic setting, where subsurface variations of rock elasticity have probably occurred over a period of two years. Our results indicate that: Ok

1. trans-dimesional algorithms can be efficiently used to map data-structures in the case of double-differences data, namely separating events with a number of changepoints which defines time windows consistent with the underlying hypothesis (here a constant-in-time elasticity field between station and event cluster); that
2. changepoints are quantitatively defined and, thus, can be ranked based on their relevance (i.e. weights) and probability of occurrence at a given time;
3. the results obtained are insensitive to changes in network geometry during the seismic experiment.

Future development and testing will provide additional insights into the use of trans-dimensional algorithms for the exploration of the data-space. For example, in this specific case, our algorithm can be applied to the jont inversion of both P-wave and S-wave databases following the approach described in Piana Agostinetti and Bodin (2018), to reconstruct a set of changepoints based on P-waves data and a set of changepoints based on S-waves data. Comparing the two sets of changepoints, "decoupled changepoints" (i.e. changepoints occurring for one set of waves and not for the other) would properly map out elasticity variations, resolving the trade-off (still existent now) between elasticity changes and changes in event locations. In fact, variations in event location would be indicated by "coupled changepoints", i.e. changepoints occuring in both sets (Piana Agostinetti and Bodin, 2018).

Data availability. Waveform data used in this study come from the Icelandic permanent seismic network run by the Icelandic Meteorological Office (IMO). The data are available upon request to IMO.

Competing interests. The authors declare that they have no conflict of interest.

Author contributions. N.P.A. planned and conduced the data analysis. G.S. provided the seismic data and the relevant data pre-processing. N.P.A. and G.S. equally contributed to the discussion on the results and the implication for monitoring the subsurface. N.P.A. and G.S. wrote the draft of the manuscript and draw the relevant figures.

drew of

Acknowledgements. This study has been published with the financial support of the Austrian Science Fund (FWF), project number: M2218-N29. NPA and GS thank the Icelandic Meteorological Office for providing the original seismic data. The Generic Mapping Tools software has been used for plotting the figures of this manuscript (Wessel and Smith, 1998).