

AI-based unmixing of medium and source signatures from seismograms: ground freezing patterns

René Steinmann¹, Léonard Seydoux^{1,2} and Michel Campillo¹

¹ISTerre, équipe Ondes et Structures, Université Grenoble-Alpes, UMR CNRS 5375, 1381 Rue de la
Piscine, 38610, Gières, France

²Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology,
Cambridge, MA, USA

Key Points:

- With methods of unsupervised learning, we identify source and medium processes in seismograms.
- A data-driven product of the seismogram tracks a continuous medium change due to freezing and thawing of the surface.
- The data-driven product can act as a filter and reveal the hidden signature of the medium change.

Corresponding author: René Steinmann, rene.steinmann@univ-grenoble-alpes.fr

Abstract

Seismograms always result from mixing many sources and medium changes that are complex to disentangle, witnessing many physical phenomena within the Earth. With artificial intelligence (AI), we isolate the signature of surface freezing and thawing in continuous seismograms recorded in a noisy urban environment. We perform a hierarchical clustering of the seismograms and identify a pattern that correlates with ground frost periods. We further investigate the fingerprint of this pattern and use it to track the continuous medium change with high accuracy and resolution in time. Our method isolates the effect of the ground frost and describes how it affects the horizontal wavefield. Our findings show how AI-based strategies can help to identify and understand hidden patterns within seismic data caused either by medium or source changes.

Plain Language Summary

Seismic waves, emitted by a seismic source and then travelling through the Earth, contain crucial information about the sources and the medium. However, often multiple sources emit simultaneously, while the elastic properties of the medium can change over time. Unmixing and identifying the different processes in the seismograms is a complex task, which we try to solve with methods of artificial intelligence (AI). In a completely data-driven fashion, we are able to mute the variation in the seismograms due to anthropogenic seismic sources and reveal a continuous medium change due to freezing and thawing. This approach could reveal hidden information in complex environments such as volcanoes, where many different source and medium processes occur.

1 Introduction

Continuous seismograms are time series of the ground motion recorded at a single location and provide a vast amount of information about processes occurring at the Earth's surface and interior. The recorded ground motion at a given location results from the convolution of the medium's impulse response — expressed as the Green's function — and the seismic waves emitted by various sources, often simultaneously. Thus, continuous seismograms are goldmines to study the medium's properties or sources in time. However, unmixing source or medium changes is often not easy, especially if source and medium changes coincide. For instance, seismic recordings in the vicinity of volcanoes, where many different source and medium effects occur, are challenging and complex datasets to analyze.

To better explore continuous seismic data, seismologists developed many data processing tools to extract valuable information for the task at hand. For example, the Short-Term-Average to Long-Term-Average energy ratio (STA/LTA) scans the continuous recordings for impulsive signals (Allen, 1978). On the other hand, passive image interferometry can interrogate the medium regularly by exploiting the ambient seismic signals of a dataset (Sens-Schönfelder & Wegler, 2006). Undoubtedly, these tools delivered many new insights into the processes happening at and inside the Earth. However, it is important to note that the design of the tools and the related preprocessing favors certain processes in the seismic data. This can be a problem if the source or medium processes encoded in the seismic data are poorly understood. For example, non-volcanic tremors were detected about twenty years ago (Obara, 2002), and still today, the physical mechanism and signal properties of such events are not well apprehended. Therefore, it remains unclear if these signals do not exist in specific environments or if the detection tools are not adapted to the task (Pfohl et al., 2015; Bocchini et al., 2021).

Artificial intelligence (AI) can help overcome those blind spots and discover new signals or hidden patterns within the data. Recently, clustering gained attention as a method to identify families of signals in the continuous seismograms (Köhler et al., 2010; Holtzman et al., 2018; Mousavi et al., 2019; Seydoux et al., 2020; C. W. Johnson et al., 2020; Snover et

90 The freezing and thawing process on a centimeter scale was well tracked with seismic
 91 velocity variations retrieved from passive image interferometry applied to the data from
 92 the three broadband stations WM01, WM02 and WM03 (Steinmann et al., 2021). Freez-
 93 ing periods caused a velocity increase and thawing periods caused a velocity decrease. The
 94 local seismic wavefield comprises many non-stationary seismic sources related to the anthro-
 95 pogenic activity, such as commuter and freight trains in the south, a highway passing in the
 96 southeast (labeled A1 on Figure 1a), a close gravel pit (marked by the two nearby lakes on
 97 Figure 1a) and an industrial neighborhood in the northwest. The combination of the contin-
 98 uously changing medium due to the freezing and thawing and many non-stationary seismic
 99 sources makes it an interesting study case for our approach to disentangle the medium from
 100 the source effects blindly.

101 3 Seismic pattern detection with hierarchical waveform clustering

102 We search for the imprint of the ground frost within the continuous three-component
 103 seismograms recorded by a single station with the hierarchical waveform clustering approach
 104 introduced in (Steinmann et al., 2022). Hierarchical clustering observes how a dataset
 105 merges into clusters based on some similarity criterion (Estivill-Castro, 2002). In our case,
 106 we calculate the similarity between waveforms from a set of features derived from a deep
 107 scattering spectrogram, as depicted in Figure 2. Firstly, we calculate the deep scattering
 108 spectrogram of the continuous three-component seismograms with a deep scattering net-
 109 work, as introduced in Andén and Mallat (2014) and adapted to seismology in Seydoux et
 110 al. (2020). A deep scattering network is a deep convolutional neural network, where the
 111 convolutional filters are restricted to wavelets and the activations to modulus operation.
 112 The output of such a network at each layer allows building the deep scattering spectrogram
 113 representation of a continuous multichannel seismogram. This representation of time series
 114 is relevant for classification purposes since it preserves signal phenomena such as attack
 115 and amplitude modulation. Moreover, a deep scattering spectrogram is locally translation
 116 invariant and stable towards small-amplitude time warping deformations (Andén & Mal-
 117 lat, 2014). We depict a two-layer scattering network in Figure 2, where we apply a sliding
 118 window on a single-component seismogram and calculate the first-order scalogram with
 119 the wavelet transform. A second wavelet transform is applied to the first-order scalogram
 120 creating the second-order scalogram. A pooling operation collapses the time axis of the
 121 scalograms and recovers the first- and second-order scattering coefficients. For each compo-
 122 nent of the ground motion record, we calculate the scattering coefficients and concatenate
 123 them. We repeat this for each window and retrieve the deep scattering spectrogram. The
 124 design of the scattering network (number of wavelets, type of pooling, et.c) can be adapted
 125 to the task at hand and is explained more in detail in Text S1 of the supplementary material.

126 Deep scattering spectrograms are redundant and high-dimensional representations, not
 127 directly suited for clustering due to the curse of dimensionality (Bellman, 1966). Therefore,
 128 we extract the most relevant characteristics — or features — and reduce the number of di-
 129 mensions with an ICA, a linear operator for feature extraction, and blind source separation
 130 (Comon, 1994). Before applying the ICA, we whiten the deep scattering spectrogram by
 131 equalizing its covariance matrix eigenvalues, allowing us to disregard patterns’ relative am-
 132 plitudes as much as possible. Finally, the number of most relevant features (or independent
 133 components) is often unknown and should be inferred, which is explained more in detail in
 134 Text S2 of the supplementary material.

135 Lastly, we perform hierarchical clustering in the low-dimensional feature space built by
 136 the unmixed sources. Clustering aims at grouping objects — here defined as data points
 137 in a given feature space — based on a similarity or dissimilarity measurement. With a
 138 bottom-up approach of hierarchical clustering, also called agglomerative clustering, all ob-
 139 jects start in a singleton cluster and merge to larger clusters until all objects unify in a
 140 single cluster (S. C. Johnson, 1967). A dendrogram depicts this process, representing the
 141 inter-cluster similarity in a cluster-distance diagram. The similarity measurement, which

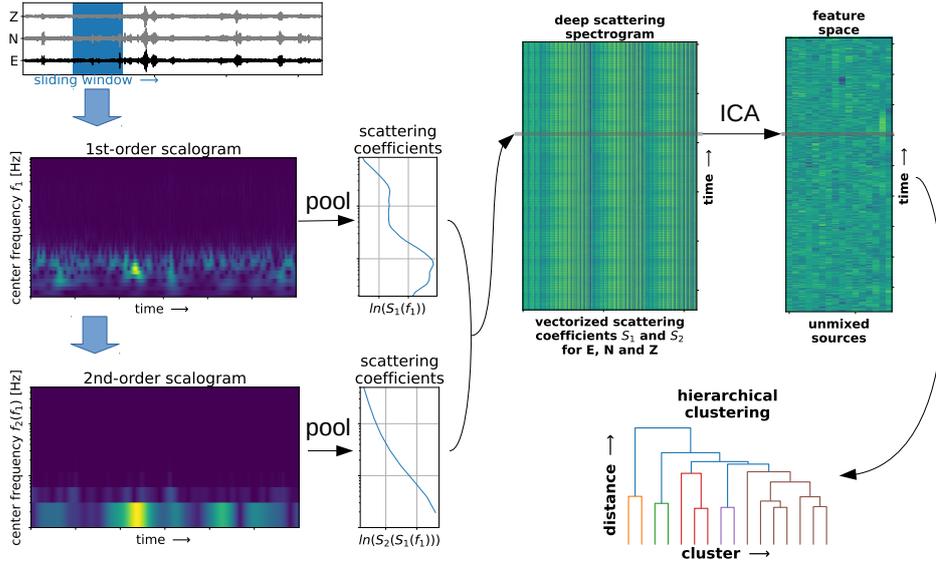


Figure 2. Sketch of the hierarchical waveform clustering approach. A two-layer scattering network with wavelet transforms, modulus and pooling operations calculates the deep scattering spectrogram. An independent component analysis (ICA) extracts the most relevant features, which are used for hierarchical clustering.

142 drives the cluster merging, is often a distance in the feature space between the objects.
 143 Thus, the type of distance is the only choice to be made here and determines the structure
 144 of the dendrogram. We use Ward’s method as a criterion to merge clusters in hierarchical
 145 clustering and produce the dendrogram. Clusters are merged with the objective to keep the
 146 increase of the total within-cluster variance minimal (Ward Jr, 1963). This allows to find
 147 cluster of various size, which fits the nature of seismic data, where ambient seismic activity
 148 often outweighs transient signals. Finally, depending on the truncation distance explored in
 149 the dendrogram, one can obtain a different number of clusters. This allows exploring the
 150 dataset’s structure and searching for a cluster of seismic signals related to the ground frost.

151 4 Cluster of signals occurs during ground frost

152 We show a truncated dendrogram of the continuous three-component seismogram recorded
 153 at station WM01 from January to April 2018 in Figure 3a, using a truncation distance to
 154 end up with 16 clusters in this case. A data point in the feature space represents 10 min
 155 of continuous waveform data without overlap. Moreover, the feature space contains 16 un-
 156 mixed sources, as a trade-off between keeping enough information and low dimensionality
 157 (see Text S2 and Figure S1 in the supplementary material). Note that finding a cluster
 158 related to ground frost effects is an exploratory task where we do not know where such a
 159 cluster would appear in the dendrogram nor if it even exists. As suggested in Steinmann
 160 et al. (2022), we extract a few large clusters at a high distance threshold to overview the
 161 whole dataset. We can then focus on certain branches in the dendrogram and extract sub-
 162 clusters hierarchically to get a more detailed cluster analysis if needed. In our case, we
 163 extract five clusters (hereafter denoted A, B, C, D, and E) at a distance threshold of 0.9
 164 (Figure 3a). In the following lines, we will interpret the clusters and assign meaningful la-
 165 bels with certain inherent clusters properties such as the normalized cumulative detections
 166 in time (Figure 3b–f), the number of detections per hour during the day (Figure 3g–k), the
 167 number of detections per weekday (Figure 3l–p), and the first-order scattering coefficients

168 averaged for each input channel (Figure 3q–u). In particular, the normalized cumulative
 169 detections in time can help identify a cluster related to the presence of ground frost since
 170 the temperature time series indicate the periods of freezing air temperature.

171 Cluster A seems to detect in a linear-piecewise way, with no relation to the temperature
 172 time series or occurrence of ground frost (Figure 3b). This cluster detects only between 05:00
 173 and 18:00 local time from Monday to Friday (Figure 3g and i). Note that around 09:00 and
 174 12:00, the detections reach a minimum, coinciding with the typical breakfast and lunch
 175 break during workdays. Compared to the other clusters, the averaged first-order scattering
 176 coefficients show larger values for frequencies above 1 Hz with a local maximum around 8 Hz
 177 on the vertical component (Figure 3q). The analysis of these parameters indicates that
 178 this cluster contains seismic signals related to anthropogenic sources, mainly active during
 179 classical labor hours. The gravel pit with trucks in the direct neighborhood of this measuring
 180 site could be a possible source (Figure 1a).

181 Cluster B seems to detect more continuously than cluster A (Figure 3c). It is active
 182 during the daytime, with a few detections during the nighttime (Figure 3h). Interestingly,
 183 this cluster peaks at 09:00 and 12:00 when cluster A reaches a minimum of detections.
 184 The weekdays show clearly more detections than the weekends, with a peak of detection
 185 on Fridays when cluster A shows a minimum of detection during the week (Figure 3l and
 186 m). The averaged first-order scattering coefficients show similar frequency characteristics
 187 as cluster A. However, cluster B indicates no bumps around 8 Hz (Figure 3r). The analysis
 188 of cluster B suggests that this cluster also relates to anthropogenic activity. Since it shows
 189 elevated activity when cluster A reduces its activity (Fridays and 09:00 and 12:00 local
 190 time), it is probably related to a different anthropogenic seismic source. Because cluster
 191 B also contains some detections during the nighttime and weekends, it possibly contains
 192 seismic signals related to nearby road traffic.

193 Cluster C is the second-largest cluster of the whole dataset (Figure 3a). It detects
 194 irregularly at all hours and all days (Figure 3d, i and n). During the morning and afternoon
 195 its detection rate decreases (Figure 3i). Moreover, the averaged first-order scattering coef-
 196 ficients show no particular pattern (Figure 3s). It is unclear what type of seismic signals
 197 cluster C contains. We can only note that it is not related to ground frost since its detections
 198 rate does not correlate with freezing temperatures.

199 Cluster D activates mainly during two periods (Figure 3e). At the beginning of Febru-
 200 ary, it accumulates 25 % of its size followed by a slight pause. Then, at the end of February
 201 and beginning of March it detects the remaining 75 % of its total size. The detection periods
 202 occur during the coldest temperatures recorded at 5 cm depth. Therefore, cluster D most
 203 likely groups seismic signals related to ground frost. Cluster D detects during all hours
 204 and all days. However, slightly more detections appear during the weekend and nighttime
 205 (Figure 3j and i). There are probably two effects that explain this behavior. Firstly, due to
 206 colder temperatures, ground frost occurs predominantly at night and so do the associated
 207 seismic signals (Figure 1b). Secondly, due to anthropogenic activity, the seismic wavefield in
 208 an urban environment changes significantly between day and night and weekdays and week-
 209 ends. Thus, the changing wavefield modulates the signature of the ground frost recorded
 210 by continuous seismograms. For instance, a seismogram containing seismic signals gener-
 211 ated by road traffic during ground frost could be found in cluster B or D. Indeed, inside
 212 cluster B, we can identify subcluster B.1 as anthropogenic seismic signals effected by the
 213 ground frost (see Figure 3a and Figure S2 in the supplementary materials). This points out
 214 a limitation of clustering: a seismogram containing multiple types of signals is assigned to
 215 a single cluster, which oversimplifies the nature of the data and has been already noted by
 216 Steinmann et al. (2022). The averaged first-order scattering coefficients show no clear and
 217 distinct pattern (Figure 3t). Cluster D seems different from Cluster A and B due to lower
 218 scattering coefficients for higher frequencies. However, it is unclear how cluster D differs
 219 from clusters C and E. We can note that the averaged first-order scattering coefficients do
 220 not deliver a unique signature related to these signals.

Cluster E is the largest cluster of the whole dataset (Figure 3a). It detects continuously with a decreased detection rate during February when ground frost occurs, with more detections during night and weekends (Figure 3f, k, and p). Moreover, the cluster shows lower averaged first-order scattering coefficients at higher frequencies (Figure 3u), distinguishing them from clusters A and B but D. The analysis of cluster E indicates that it groups ambient seismic noise without particular transients and ground frost. In fact, it appears that cluster D and E summarize the stationary ambient wave field separated only due to the occurrence of ground frost. Indeed, the combined clusters seems to detect almost continuously during weekends and nights (see Figure S2 in the supplementary materials).

Summarized, the dendrogram delivers a data-driven overview about the content of the data containing both source and medium effects. We can clearly identify cluster A and B with anthropogenic seismic sources. Inside cluster B we identified a small subcluster containing anthropogenic signals effected by the ground frost. We have reasons to assume that a more detailed cluster solution would reveal a similar subcluster in A. We can not find a meaningful label for cluster C. The largest part of the data is located within cluster E: ambient seismic noise, which is not effected by ground frost. Cluster D seems to be the only cluster related to the freezing of the surface without particular transient signals from anthropogenic activity. The hierarchical clustering approach, together with an interpretation of a cluster solution at a high distance threshold, allowed us to give a detailed analysis of the content of the seismic data. In particular, the cumulative detection curve identifies cluster D as of interest in our study because it relates purely to ground frost. Hence, we do not need to extract a more detailed cluster solution. In the following lines, we analyze how the freezing and thawing process is encoded in the data.

5 Disentagling the ground-frost from the urban imprint

Hierarchical clustering built the dendrogram within the feature space extracted by an ICA from the deep scattering spectrogram (Figure 2). The features likely reveal insights about the signature of cluster D and, thus, about the ground frost signature. Steinmann et al. (2022) already showed that single features retrieved from the scattering coefficients with an ICA could reveal interesting patterns in the seismogram. Therefore, we can likely identify a single feature in our dataset that encodes the seismic signature of the ground frost. We calculate the absolute centroid of cluster D and observe its coordinates in the 16-dimensional feature space (Figure 4a). We note that if all features are equally important in defining a cluster, they should contribute equally to the centroid coordinates. If a few or single features are more important than others, the centroid should have a stronger contribution from them. We observe that the centroid of cluster D shows a substantial value for feature 15 (Figure 4a) regarding the other features. This suggests that cluster D is active when large absolute values on feature 15 occur.

We can also observe how feature 15 evolves in time (Figure 4b). Feature 15 shows a significant amplitude decrease at the end of February and the beginning of March. During that time, it seems to mimic the low-frequent trend of the air temperature with a slight offset in time. The beginning of February and mid-March show smaller amplitude decreases after a few consecutive nights of freezing air temperature. Unfortunately, we have no ground truth about the occurrence of ground frost. However, we know that the occurrence of ground frost depends on the amount of time and the amplitude of freezing air temperature. Moreover, thawing air temperatures during the day counteract the nightly built-up of ground frost. A more extended and continuous period of freezing air temperature (like the one at the end of February) results in a thicker layer of ground frost. A colder air temperature can also decrease the temperature inside the layer of ground frost and, thus, increase its stiffness and shear wave velocity (Miao et al., 2019). These facts, combined with the observation of feature 15 and the air temperature, suggest that this feature tracks the freezing and thawing process of the surface at a high-resolution timescale of 10 min. We emphasize that feature 15 is an entirely data-driven product from a three-component seismogram with minimal

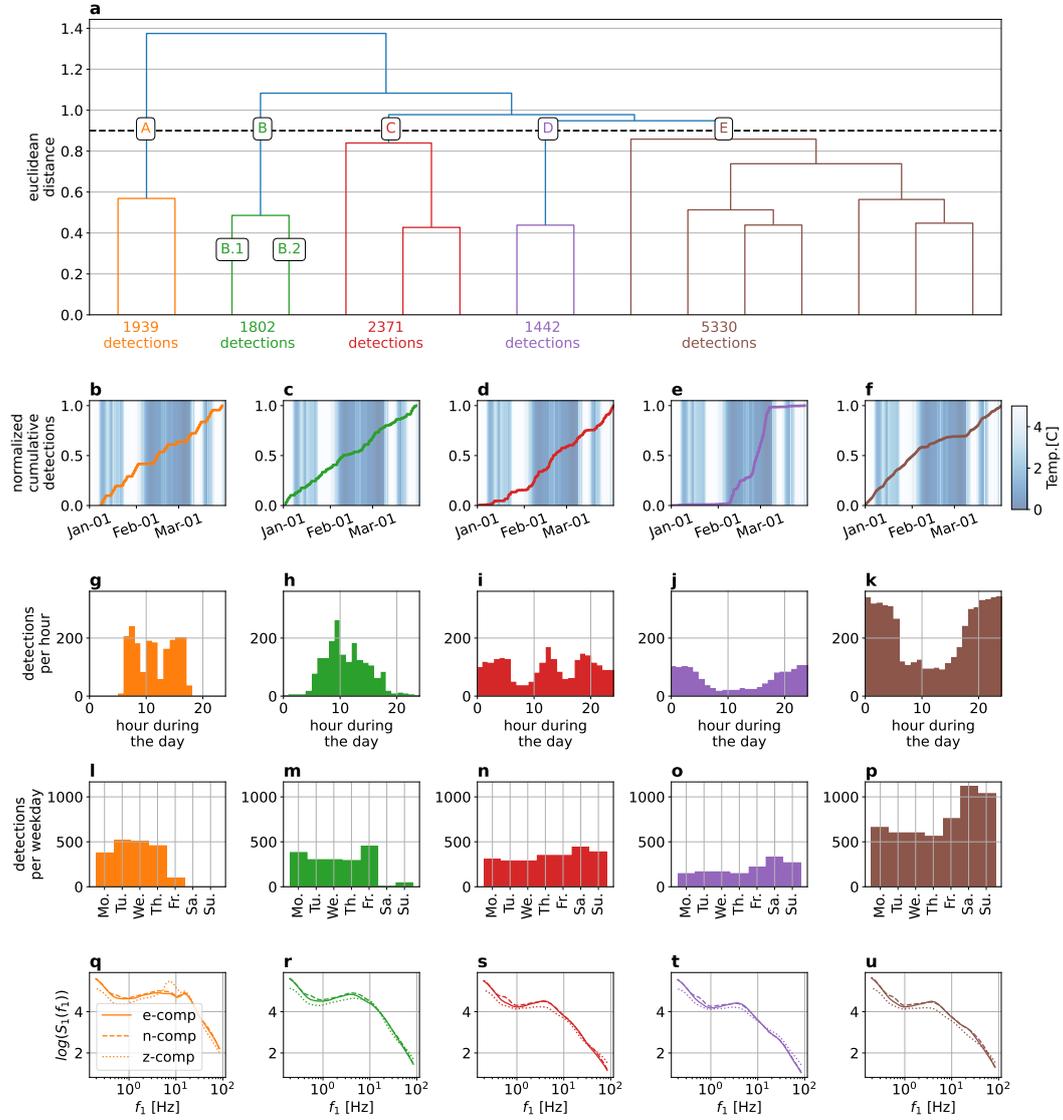


Figure 3. Results of seismic data clustering from the three-component broadband station WM01 between 1 January to 1 April 2018. (a) dendrogram with a truncation distance set to obtain 16 clusters. (b–f) normalized cumulative detection. (g–k) daily occurrence. (l–p) weekly occurrence. (q–u) averaged first-order scattering coefficients.

273 processing. In comparison, Steinmann et al. (2021) tracked the same freezing and thawing
 274 process with data from two seismic stations, heavier preprocessing, and a time resolution of
 275 2 days.

276 Since ICA is a linear operator, we can use only feature 15 to reconstruct the scatter-
 277 ing coefficients out of the mixing matrix, defined as the pseudo-inverse of the unmixing
 278 matrix (Comon, 1994). This procedure acts as a filter process since we zero all features
 279 except feature 15. Due to the large size of first- and second-order scattering coefficients,
 280 Figure 4c–h show only the first-order original and reconstructed scattering coefficients for
 281 all three components. The original coefficients show clearly the urban imprint in the seismic
 282 data: fringes appear during daytime and pause at the weekends (Figure 4c, e and g). No
 283 clear pattern appears during ground frost building periods, such as at the end of February
 284 (Figure 4b). The reconstructed coefficients do not contain the fringes due to urban activity
 285 since these signals were probably encoded in one of the muted features (Figure 4d, f and
 286 h). The filtering effect reveals a slight amplitude decrease for the horizontal components
 287 at frequencies above 1 Hz during the end of February, coinciding with the coldest period
 288 of the dataset. During that time, a faint amplitude decrease can also be observed at the
 289 vertical component. At times with consecutive cold nights such as at the beginning of
 290 February or mid-March, these decreases are also faintly visible. These observations confirm
 291 that the wavefield experiences an energy decrease during ground frost with a discrepancy
 292 between horizontal and vertical components. Indeed, the ratio of horizontal and vertical
 293 scattering coefficients show a clear broadband high-frequent decrease at the beginning and
 294 end of February for both original and reconstructed data (Figure 4i and j). It appears that
 295 the broadband decrease in the ratio becomes stronger with increasing time or amplitude
 296 of the freezing air temperature. The ratio of horizontal and vertical scattering coefficients
 297 resembles the classical Horizontal-to-Vertical-Spectral-Ratio (HVSr) based on the Fourier
 298 transform. Indeed, models based on the diffusive field assumption confirm an HVSr de-
 299 crease due to a thin layer of ground frost (see Text S3 and S4, and Figure S3 and S4) in the
 300 supplementary materials).

301 6 Conclusion

302 In this study, we made the first attempts towards inferring blindly medium changes
 303 from the wavefield recorded by a single station. For our case study, the medium continu-
 304 ously changes due to surface freezing and thawing, while anthropogenic activity creates a
 305 complex and non-stationary seismic wavefield. An AI-based approach, based on the deep
 306 scattering network, an ICA and hierarchical clustering, helped us explore the seismic data
 307 and search for possible patterns induced by the ground frost without assuming how the
 308 seismic data could be affected. One of the main outcomes of this study is that the AI-
 309 based approach blindly extracts a feature that isolates the seismic response to the medium
 310 change and mutes other non-stationary processes. This opens new possibilities to utilize sin-
 311 gle station data for monitoring purposes, especially in environments with many source and
 312 medium processes such as permafrost (e.g. Köhler & Weidle, 2019) or volcanoes. AI-based
 313 strategies could complement other passive seismic methods used for permafrost monitoring
 314 (e.g. James et al., 2019; Lindner et al., 2021). This could give new insight into the response
 315 of permafrost to climate change given the decade-long availability of single seismic stations
 316 near permafrost areas. Future research could also investigate if other types of medium
 317 changes (e.g., groundwater fluctuations) could be directly extracted from the seismograms
 318 in a data-driven fashion.

319 Moreover, the revealed signature combined with the HVSr model indicates that su-
 320 perfacial freezing might impact the modal energy distribution. To our knowledge, this effect
 321 has not yet been considered in permafrost studies using passive seismic methods. On the
 322 one hand, it could corrupt velocity variation measurements retrieved from surface waves in
 323 cross-correlograms. On the other hand, it would also be an opportunity since more modes
 324 increase the amount of information about the subsurface. Future research is needed to

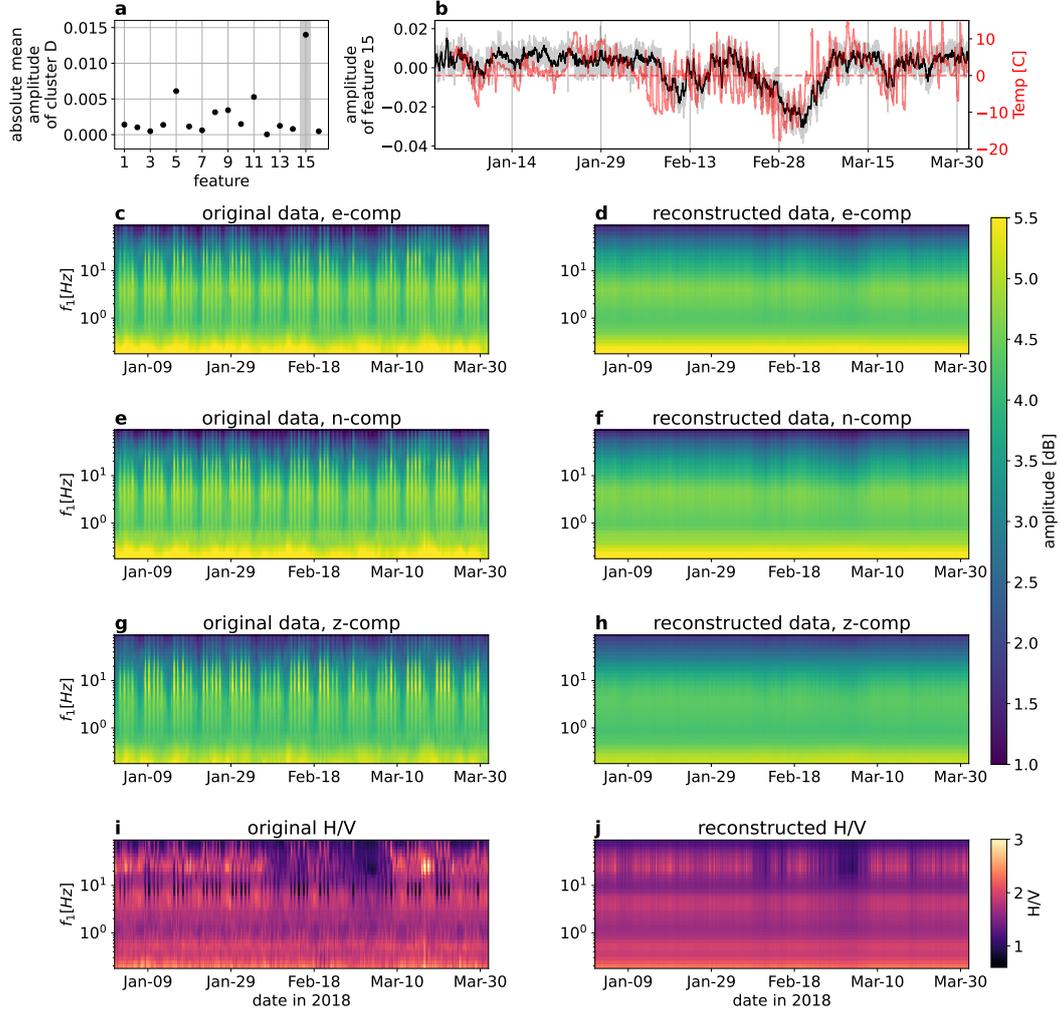


Figure 4. The signature of freezing (a) coordinates of the centroid of cluster D in the eight-dimensional feature space. (b) feature 15 as a smoothed time-series (black) compared to the temperature time-series recorded above ground (red). The original feature without smoothing is represented in grey. (c,e,g) Original first-order scattering coefficients for the east, north and vertical component, respectively. (d,f,h) Reconstructed first-order scattering coefficients based solely on feature 5 for the east, north and vertical component, respectively. (i) Ratio between horizontal and vertical components based on the original first order scattering coefficients. (j) Ratio between horizontal and vertical components based on the reconstructed first order scattering coefficients.

325 understand better the interaction between different surface wave modes in the presence of
 326 frozen surface layers.

327 7 Open Research

328 The seismic data was downloaded from Steinmann et al. (2020) and the temper-
 329 ature data were provided by the Meteorological Institute of Hamburg. The main code
 330 for calculating the scattering coefficients, features and linkage matrix can be found un-
 331 der <https://zenodo.org/badge/latestdoi/460424596>. The work relies heavily on the python
 332 packages ObsPy (Beyreuther et al., 2010), scikit-learn (Pedregosa et al., 2011) and SciPy
 333 (Virtanen et al., 2020). The map was produced with map tiles by Stamen Design, under
 334 CC BY 3.0. Data by OpenStreetMap, under ODbL.

335 Acknowledgments

336 The authors acknowledge support from the European Research Council under the European
 337 Union Horizon 2020 research and innovation program (grant agreement no. 742335, F-
 338 IMAGE). This work has also been supported by MIAI@Grenoble Alpes (ANR-19-P3IA-
 339 0003).

340 References

- 341 Allen, R. V. (1978). Automatic earthquake recognition and timing from single traces.
 342 *Bulletin of the seismological society of America*, 68(5), 1521–1532.
- 343 Andén, J., & Mallat, S. (2014). Deep scattering spectrum. *IEEE Transactions on Signal*
 344 *Processing*, 62(16), 4114–4128.
- 345 Bellman, R. (1966). Dynamic programming. *Science*, 153(3731), 34–37.
- 346 Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., & Wassermann, J. (2010).
 347 Obspy: A python toolbox for seismology. *Seismological Research Letters*, 81(3), 530–
 348 533.
- 349 Bocchini, G., Martínez-Garzón, P., Harrington, R., & Bohnhoff, M. (2021). Does deep tec-
 350 tonic tremor occur in the central-eastern mediterranean basin? *Journal of Geophysical*
 351 *Research: Solid Earth*, 126(1), 2020JB020448.
- 352 Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*,
 353 36(3), 287–314.
- 354 Estivill-Castro, V. (2002). Why so many clustering algorithms: a position paper. *ACM*
 355 *SIGKDD explorations newsletter*, 4(1), 65–75.
- 356 Holtzman, B. K., Paté, A., Paisley, J., Waldhauser, F., & Repetto, D. (2018). Machine
 357 learning reveals cyclic changes in seismic source spectra in geysers geothermal field.
 358 *Science advances*, 4(5), eaao2929.
- 359 James, S., Knox, H., Abbott, R., Panning, M., & Sreaton, E. (2019). Insights into per-
 360 mafrost and seasonal active-layer dynamics from ambient seismic noise monitoring.
 361 *Journal of Geophysical Research: Earth Surface*, 124(7), 1798–1816.
- 362 Jenkins, W. F., Gerstoft, P., Bianco, M. J., & Bromirski, P. D. (2021). Unsupervised
 363 deep clustering of seismic data: Monitoring the ross ice shelf, antarctica. *Journal of*
 364 *Geophysical Research: Solid Earth*, e2021JB021716.
- 365 Johnson, C. W., Ben-Zion, Y., Meng, H., & Vernon, F. (2020). Identifying different classes
 366 of seismic noise signals using unsupervised learning. *Geophysical Research Letters*,
 367 47(15), e2020GL088353.
- 368 Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3), 241–254.
- 369 Köhler, A., Ohrnberger, M., & Scherbaum, F. (2010). Unsupervised pattern recognition in
 370 continuous seismic wavefield records using self-organizing maps. *Geophysical Journal*
 371 *International*, 182(3), 1619–1630.
- 372 Köhler, A., & Weidle, C. (2019). Potentials and pitfalls of permafrost active layer monitoring
 373 using the hvsr method: a case study in svalbard. *Earth Surface Dynamics*, 7(1), 1–16.

- 374 Lindner, F., Wassermann, J., & Igel, H. (2021). Seasonal freeze-thaw cycles and permafrost
 375 degradation on mt. zugspitze (german/austrian alps) revealed by single-station seismic
 376 monitoring. *Earth and Space Science Open Archive ESSOAr*.
- 377 Miao, Y., Shi, Y., Zhuang, H., Wang, S., Liu, H., & Yu, X. (2019). Influence of seasonal
 378 frozen soil on near-surface shear wave velocity in eastern Hokkaido, Japan. *Geophysical
 379 Research Letters*, *46*(16), 9497 – 9508.
- 380 Mousavi, S. M., Zhu, W., Ellsworth, W., & Beroza, G. (2019). Unsupervised clustering of
 381 seismic signals using deep convolutional autoencoders. *IEEE Geoscience and Remote
 382 Sensing Letters*, *16*(11), 1693–1697.
- 383 Obara, K. (2002). Nonvolcanic deep tremor associated with subduction in southwest japan.
 384 *Science*, *296*(5573), 1679–1681.
- 385 Outcalt, S. I., Nelson, F. E., & Hinkel, K. M. (1990). The zero-curtain effect: Heat and
 386 mass transfer across an isothermal region in freezing soil. *Water Resources Research*,
 387 *26*(7), 1509–1516.
- 388 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duch-
 389 esnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine
 390 Learning Research*, *12*, 2825–2830.
- 391 Pfohl, A., Warren, L. M., Sit, S., & Brudzinski, M. (2015). Search for tectonic tremor
 392 on the central north anatolian fault, turkey. *Bulletin of the Seismological Society of
 393 America*, *105*(3), 1779–1786.
- 394 Sens-Schönfelder, C., & Wegler, U. (2006). Passive image interferometry and seasonal
 395 variations of seismic velocities at merapi volcano, indonesia. *Geophysical research
 396 letters*, *33*(21).
- 397 Seydoux, L., Balestrieri, R., Poli, P., De Hoop, M., Campillo, M., & Baraniuk, R. (2020).
 398 Clustering earthquake signals and background noises in continuous seismic data with
 399 unsupervised deep learning. *Nature communications*, *11*(1), 1–12.
- 400 Snover, D., Johnson, C. W., Bianco, M. J., & Gerstoft, P. (2020). Deep clustering to identify
 401 sources of urban seismic noise in long beach, california. *Seismological Research Letters*.
- 402 Steinmann, R., Hadziioannou, C., & Larose, E. (2021). Effect of centimetric freezing
 403 of the near subsurface on rayleigh and love wave velocity in ambient seismic noise
 404 correlations. *Geophysical Journal International*, *224*(1), 626–636.
- 405 Steinmann, R., Seydoux, L., Beaucé, E., & Campillo, M. (2022). Hierarchical exploration of
 406 continuous seismograms with unsupervised learning. *Journal of Geophysical Research:
 407 Solid Earth*, *127*(1), e2021JB022455.
- 408 Steinmann, R., Hadziioannou, C., & Larose, E. (2020, August). *Data of seismic urban
 409 noise in the city of Hamburg, Germany 2018*. Zenodo. Retrieved from [https://
 410 doi.org/10.5281/zenodo.3992631](https://doi.org/10.5281/zenodo.3992631) doi: 10.5281/zenodo.3992631
- 411 Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D.,
 412 ... SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific
 413 Computing in Python. *Nature Methods*, *17*, 261–272. doi: 10.1038/s41592-019-0686-2
- 414 Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of
 415 the American statistical association*, *58*(301), 236–244.