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

Rene Steinmann[1], Léonard Simon Seydoux[2], and Michel Campillo[1]

[1]Université Grenoble Alpes
[2]Massachusetts Institute of Technology

November 21, 2022

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.

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

René Steinmann[1], Léonard Seydoux[1,2] and Michel Campillo[1]

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

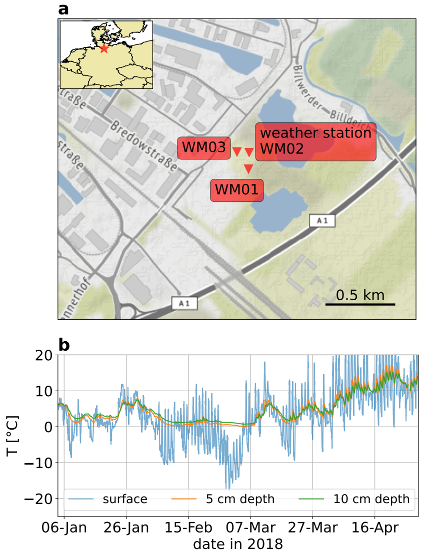

Figure 1. Temperature and seismic stations used in the study. (**a**) Map of the measuring site in Hamburg, Germany, with the three broadband and three-component seismic sensors WM01, WM02, and WM03. (**b**) Temperature time series measured at the surface, 5 cm and 10 cm depth close to station WM02 with a sampling period of 10 min.

al., 2020; Jenkins et al., 2021; Steinmann et al., 2022). In the most common approach, characteristics — often called features — are calculated for a sliding window. Then, clustering algorithms perform a similarity measurement within the set of characteristics and assign a cluster to each window. Until now, the applications showed that this approach mainly identifies families of signals related to source processes such as geothermal activity (Holtzman et al., 2018), different types of anthropogenic activity (Snover et al., 2020), seismic background activity (C. W. Johnson et al., 2020) or precursory signals of a landslide (Seydoux et al., 2020). To our knowledge, medium changes have been disregarded so far in this task.

In the present study, we make the first attempts towards inferring not only source processes but also medium changes from continuous single station seismograms in a data-driven fashion.

2 A thin ground frost layer visible in temperature data and seismic velocity variations

The study site is located in the city of Hamburg, Germany (Figure 1a). Besides the three broadband sensors WM01, WM02, and WM03, the site includes various meteorological sensors near station WM02. At 5 cm, 10 cm, 80 cm, and 120 cm depth and at the surface, temperature sensors deliver a measurement every 10 min. Figure 1b depicts the temperature time series at the surface, 5 cm, and 10 cm depth from January 4 to April 30 in 2018. Until the end of March, the air temperature ranges between $-20\,^{\circ}\mathrm{C}$ and $20\,^{\circ}\mathrm{C}$ indicating a continuous freezing and thawing of the near-surface. In particular, the end of February is a cold period with freezing air temperature during daytime and nighttime. However, at 5 cm and 10 cm depth, the sensors do not reach below $0\,^{\circ}\mathrm{C}$ and do not follow the air temperature as they do later in March. This is known as the zero-curtain effect: the phase change from water to ice in the soil releases latent heat, which causes the freezing process to slow down (Outcalt et al., 1990). This implies that the ground frost is not deeper than 5 cm during the coldest period.

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

3 Seismic pattern detection with hierarchical waveform clustering

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

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

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

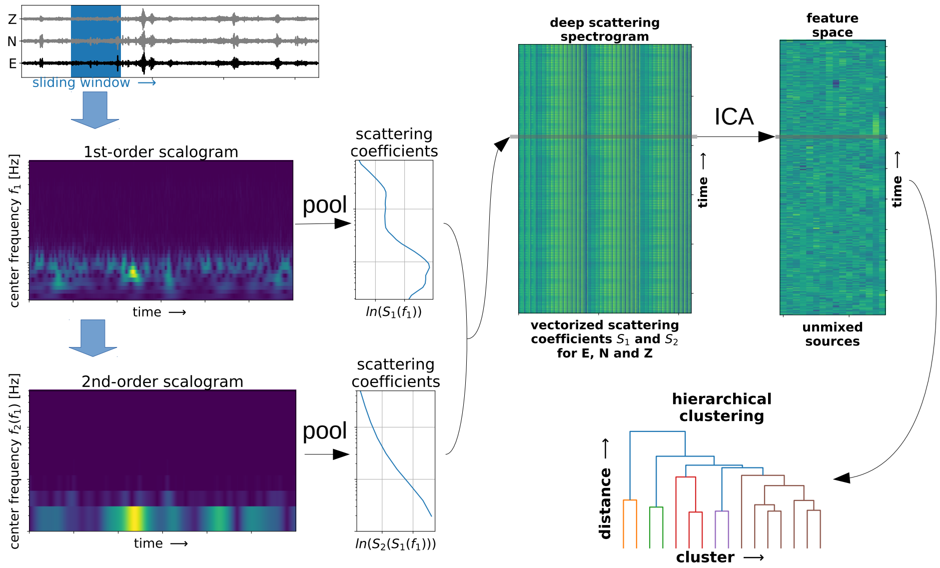

Figure 2. Sketch of the hierarchical waveform clustering approach. A two-layer scattering network with wavelet transforms, modulus and pooling operations calculates the deep scattering spectrogram. An independent component analysis (ICA) extracts the most relevant features, which are used for hierarchical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

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

4 Cluster of signals occurs during ground frost

We show a truncated dendrogram of the continuous three-component seismogram recorded at station WM01 from January to April 2018 in Figure 3a, using a truncation distance to end up with 16 clusters in this case. A data point in the feature space represents 10 min of continuous waveform data without overlap. Moreover, the feature space contains 16 unmixed sources, as a trade-off between keeping enough information and low dimensionality (see Text S2 and Figure S1 in the supplementary material). Note that finding a cluster related to ground frost effects is an exploratory task where we do not know where such a cluster would appear in the dendrogram nor if it even exists. As suggested in Steinmann et al. (2022), we extract a few large clusters at a high distance threshold to overview the whole dataset. We can then focus on certain branches in the dendrogram and extract subclusters hierarchically to get a more detailed cluster analysis if needed. In our case, we extract five clusters (hereafter denoted A, B, C, D, and E) at a distance threshold of 0.9 (Figure 3a). In the following lines, we will interpret the clusters and assign meaningful labels with certain inherent clusters properties such as the normalized cumulative detections in time (Figure 3b–f), the number of detections per hour during the day (Figure 3g–k), the number of detections per weekday (Figure 3l–p), and the first-order scattering coefficients averaged for each input channel (Figure 3q–u). In particular, the normalized cumulative detections in time can help identify a cluster related to the presence of ground frost since the temperature time series indicate the periods of freezing air temperature.

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

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

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

Cluster D activates mainly during two periods (Figure 3e). At the beginning of February, it accumulates 25 % of its size followed by a slight pause. Then, at the end of February and beginning of March it detects the remaining 75 % of its total size. The detection periods occur during the coldest temperatures recorded at 5 cm depth. Therefore, cluster D most likely groups seismic signals related to ground frost. Cluster D detects during all hours and all days. However, slightly more detections appear during the weekend and nighttime (Figure 3j and i). There are probably two effects that explain this behavior. Firstly, due to colder temperatures, ground frost occurs predominantly at night and so do the associated seismic signals (Figure 1b). Secondly, due to anthropogenic activity, the seismic wavefield in an urban environment changes significantly between day and night and weekdays and weekends. Thus, the changing wavefield modulates the signature of the ground frost recorded by continuous seismograms. For instance, a seismogram containing seismic signals generated by road traffic during ground frost could be found in cluster B or D. Indeed, inside cluster B, we can identify subcluster B.1 as anthropogenic seismic signals effected by the ground frost (see Figure 3a and Figure S2 in the supplementary materials). This points out a limitation of clustering: a seismogram containing multiple types of signals is assigned to a single cluster, which oversimplifies the nature of the data and has been already noted by Steinmann et al. (2022). The averaged first-order scattering coefficients show no clear and distinct pattern (Figure 3t). Cluster D seems different from Cluster A and B due to lower scattering coefficients for higher frequencies. However, it is unclear how cluster D differs from clusters C and E. We can note that the averaged first-order scattering coefficients do 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

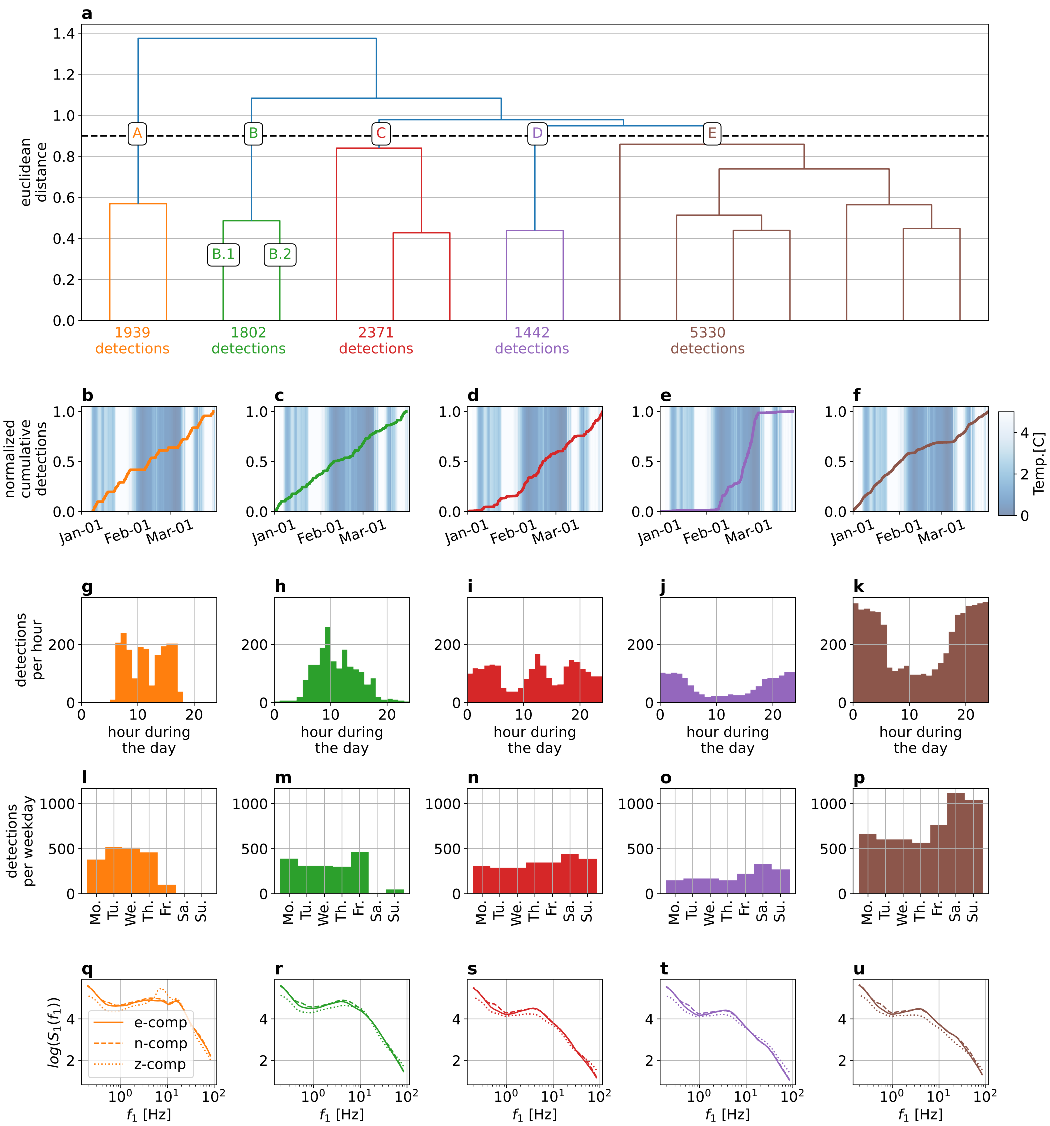

Figure 3. Results of seismic data clustering from the three-component broadband station WM01 between 1 January to 1 April 2018. (**a**) dendrogram with a truncation distance set to obtain 16 clusters. (**b–f**) normalized cumulative detection. (**g–k**) daily occurrence. (**l–p**) weekly occurrence.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

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

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

6 Conclusion

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

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

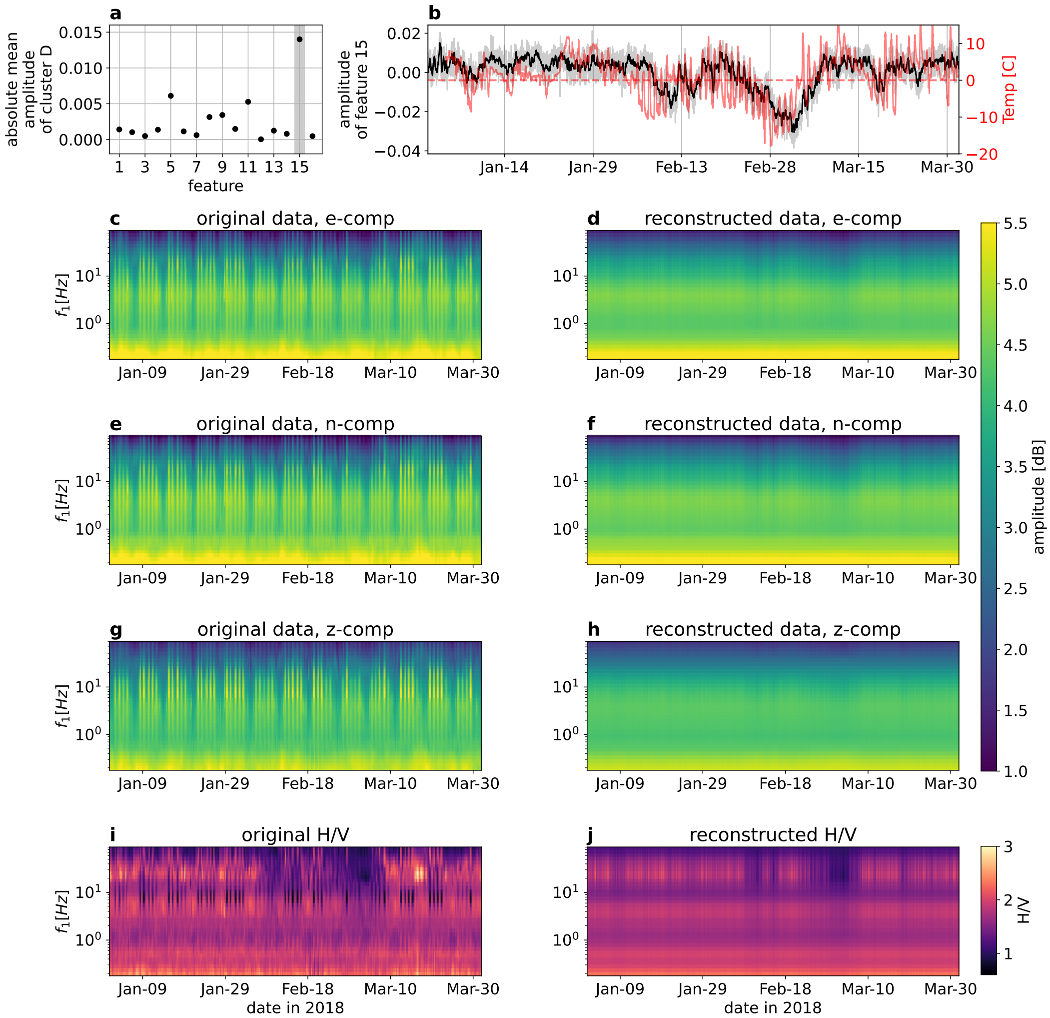

Figure 4. **The signature of freezing** (**a**) coordinates of the centroid of cluster D in the eight-dimensional feature space. (**b**) feature 15 as a smoothed time-series (black) compared to the temperature time-series recorded above ground (red). The orignal feature without smoothing is represented 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.

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

7 Open Research

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

Acknowledgments

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

References

Allen, R. V. (1978). Automatic earthquake recognition and timing from single traces. *Bulletin of the seismological society of America*, *68*(5), 1521–1532.

Andén, J., & Mallat, S. (2014). Deep scattering spectrum. *IEEE Transactions on Signal Processing*, *62*(16), 4114–4128.

Bellman, R. (1966). Dynamic programming. *Science*, *153*(3731), 34–37.

Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., & Wassermann, J. (2010). Obspy: A python toolbox for seismology. *Seismological Research Letters*, *81*(3), 530–533.

Bocchini, G., Martínez-Garzón, P., Harrington, R., & Bohnhoff, M. (2021). Does deep tectonic tremor occur in the central-eastern mediterranean basin? *Journal of Geophysical Research: Solid Earth*, *126*(1), 2020JB020448.

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, *36*(3), 287–314.

Estivill-Castro, V. (2002). Why so many clustering algorithms: a position paper. *ACM SIGKDD explorations newsletter*, *4*(1), 65–75.

Holtzman, B. K., Paté, A., Paisley, J., Waldhauser, F., & Repetto, D. (2018). Machine learning reveals cyclic changes in seismic source spectra in geysers geothermal field. *Science advances*, *4*(5), eaao2929.

James, S., Knox, H., Abbott, R., Panning, M., & Screaton, E. (2019). Insights into permafrost and seasonal active-layer dynamics from ambient seismic noise monitoring. *Journal of Geophysical Research: Earth Surface*, *124*(7), 1798–1816.

Jenkins, W. F., Gerstoft, P., Bianco, M. J., & Bromirski, P. D. (2021). Unsupervised deep clustering of seismic data: Monitoring the ross ice shelf, antarctica. *Journal of Geophysical Research: Solid Earth*, e2021JB021716.

Johnson, C. W., Ben-Zion, Y., Meng, H., & Vernon, F. (2020). Identifying different classes of seismic noise signals using unsupervised learning. *Geophysical Research Letters*, *47*(15), e2020GL088353.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*(3), 241–254.

Köhler, A., Ohrnberger, M., & Scherbaum, F. (2010). Unsupervised pattern recognition in continuous seismic wavefield records using self-organizing maps. *Geophysical Journal International*, *182*(3), 1619–1630.

Köhler, A., & Weidle, C. (2019). Potentials and pitfalls of permafrost active layer monitoring using the hvsr method: a case study in svalbard. *Earth Surface Dynamics*, *7*(1), 1–16.

Lindner, F., Wassermann, J., & Igel, H. (2021). Seasonal freeze-thaw cycles and permafrost degradation on mt. zugspitze (german/austrian alps) revealed by single-station seismic monitoring. *Earth and Space Science Open Archive ESSOAr*.

Miao, Y., Shi, Y., Zhuang, H., Wang, S., Liu, H., & Yu, X. (2019). Influence of seasonal frozen soil on near-surface shear wave velocity in eastern Hokkaido, Japan. *Geophysical Research Letters*, *46*(16), 9497 – 9508.

Mousavi, S. M., Zhu, W., Ellsworth, W., & Beroza, G. (2019). Unsupervised clustering of seismic signals using deep convolutional autoencoders. *IEEE Geoscience and Remote Sensing Letters*, *16*(11), 1693–1697.

Obara, K. (2002). Nonvolcanic deep tremor associated with subduction in southwest japan. *Science*, *296*(5573), 1679–1681.

Outcalt, S. I., Nelson, F. E., & Hinkel, K. M. (1990). The zero-curtain effect: Heat and mass transfer across an isothermal region in freezing soil. *Water Resources Research*, *26*(7), 1509–1516.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pfohl, A., Warren, L. M., Sit, S., & Brudzinski, M. (2015). Search for tectonic tremor on the central north anatolian fault, turkey. *Bulletin of the Seismological Society of America*, *105*(3), 1779–1786.

Sens-Schönfelder, C., & Wegler, U. (2006). Passive image interferometry and seasonal variations of seismic velocities at merapi volcano, indonesia. *Geophysical research letters*, *33*(21).

Seydoux, L., Balestriero, R., Poli, P., De Hoop, M., Campillo, M., & Baraniuk, R. (2020). Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning. *Nature communications*, *11*(1), 1–12.

Snover, D., Johnson, C. W., Bianco, M. J., & Gerstoft, P. (2020). Deep clustering to identify sources of urban seismic noise in long beach, california. *Seismological Research Letters*.

Steinmann, R., Hadziioannou, C., & Larose, E. (2021). Effect of centimetric freezing of the near subsurface on rayleigh and love wave velocity in ambient seismic noise correlations. *Geophysical Journal International*, *224*(1), 626–636.

Steinmann, R., Seydoux, L., Beaucé, E., & Campillo, M. (2022). Hierarchical exploration of continuous seismograms with unsupervised learning. *Journal of Geophysical Research: Solid Earth*, *127*(1), e2021JB022455.

Steinmannn, R., Hadziioannou, C., & Larose, E. (2020, August). *Data of seismic urban noise in the city of Hamburg, Germany 2018.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.3992631` doi: 10.5281/zenodo.3992631

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. doi: 10.1038/s41592-019-0686-2

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, *58*(301), 236–244.

Supporting Information for "AI-based unmixing of medium and source signatures from seismograms: ground freezing patterns"

René Steinmann[1], Léonard Seydoux[1,2]and Michel Campillo[1]

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

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

Introduction

The seismic data is sampled with 200 Hz. Because the data was retrieved manually from the field, three data gaps of ca. 3 h occur in the dataset. Before applying the hierarchical waveform clustering, the data was demeaned and high-pass filtered with a corner frequency of 0.1 Hz. The data gaps were filled with zeroes. However, the scattering coefficients of the data gaps were removed before the feature selection. The supporting information provides details about:

1. the design of the deep scattering network (Text S1)

2. the number of releveant features retrieved with an ICA (Text S2 and Figure S1)

3. the cumulative detections for subcluster B.1, B.2 and the combination of cluster D and E (Figure S2)

4. the HVSR models with and without a thin layer of ground frost (Text S3 and S4, Table S1, and Figure S3 and S4)

Text S1: Design of deep scattering network

We design a deep scattering network with 36 complex-valued Gabor wavelets in the first layer and 9 Gabor wavelets in the second layer. A modulus operation retrieves real-valued scalograms. The first layer creates 36 scattering coefficients and the second layer creates 324 (as from 36×9) scattering coefficients per sliding window and component. The center frequencies of the first-layer wavelets range from 0.2 to 89 Hz and the center frequencies of the second layer wavelets range from 0.2 to 50 Hz. The number of wavelets was chosen specifically to cover a wide range of frequencies above the oceanic microseism. The upper frequency of the first layer is bounded by the sampling frequency of 200 Hz. The center frequencies are spaced logarithmically with four wavelets per octave in the first layer and one wavelet per octave in the second layer. The sliding window is set to 10 min to mimic the time resolution of the temperature data. In contrast to Steinmann, Seydoux, Beaucé, and Campillo (2022), we apply average pooling instead of maximum pooling to the first and second layer scalograms since we are not searching for transient signals but changes in the ambient seismic wavefield.

Text S2: Extracting the most relevant features

After calculating the deep scattering spectrogram, we apply an ICA to retrieve the most relevant features. The ICA model can be written as:

$$\mathbf{x} = \mathbf{s}\mathbf{A}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^{N \times F}$ are the N observations of dimension F, $\mathbf{A} \in \mathbb{R}^{F \times C}$ is the mixing matrix, and $\mathbf{s} \in \mathbb{R}^{C \times N}$ are the unmixed sources. Equation 1 considers the observations $\mathbf{x}$ as a linear combination of the independent sources $\mathbf{s}$, with the mixing weights gathered in $\mathbf{A}$. In our case, $\mathbf{x}$ are the whitened scattering coefficients. Setting the number of features is an exploratory task that can be seen as a trade-off between keeping the dimensionality low for clustering and retaining the most crucial data information. We use the reconstruction loss $\epsilon(C)$ between the original data $\mathbf{x}$ and the reconstructed data $\hat{\mathbf{x}}^{(C)}$, based on the C independent components, as a guideline for choosing an optimal number for C. The reconstruction loss is defined as following:

$$\epsilon(C) = \frac{\sum_{i=0}^{N} |x_i - \hat{x}_i^{(C)}|}{N}. \tag{2}$$

Figure S1 depicts the reconstruction loss $\epsilon(C)$ for an increasing number of independent components C. The reconstruction loss decreases rapidly with the first 14 components. With more than 14 components, the rate of error decrease becomes smaller and almost linear. However, a small jump occurs from 14 to 16 components. Therefore, 16 independent components, marking a kink in the reconstruction error curve, seem like a good choice to us and are the basis for building the linkage matrix for the dendrogram.

Text S3: Inverting for a 1D velocity model

To forward model the effect of ground frost on the HVSR, we need a 1D velocity model with the shear wave velocity v_s, the compressional wave velocity v_p, the thickness of the layer h and the density ρ. Steinmann, Hadziioannou, and Larose (2021) provides a 1D velocity model to a depth of less than 30 m based on a shear wave refraction profile. The forward modelled HVSR based on this velocity model together with the observed HVSR at the three stations at 15 April 2018 are shown in Figure S3. We chose this day for an HVSR measurement for two reasons. Firstly, the time of the year and the temperature data suggest that we do not have any ground frost (Figure 1a). Secondly, it is a Sunday and, thus, we have better conditions for an equipartitionned wavefield without anthropogenic activity (Figure 3). It is clear that the modelled HVSR does not fit the observations. Since the two resonance peaks below 1 Hz do not appear in the modelled HVSR, it appears that the velocity model is not deep enough. To update the velocity model, we invert the HVSR measurements based on the diffusive field assumption (Piña-Flores et al., 2016). We invert for a three-layer model with the observed HVSR between 0.1 and 1 Hz to fit the two resonance peaks. The higher frequency content seems unreliable, since the variations between the stations are too large given the fact that they are only 100 m apart (see map in Figure 1b). These variations at higher frequencies can be the result of different installation types. WM01 and WM02 are placed on a concrete slab while WM03 is inside a shed. We constrain the range of possible shear wave velocity of the first layer with the values given in Steinmann et al. (2021). The updated and deeper velocity model fits better the observations and, thus, is utilized for modelling the effect of the ground frost. The values of the updated model are presented in Table S1.

Text S4: Modelling the effect of a frozen surface on the HVSR

We model the effect of ground frost on the HVSR based on a 1D velocity model and diffuse wavefield assumption (García-Jerez et al., 2016). Firstly, we derive a 1D velocity model from the inversion of H/V measurements (Piña-Flores et al., 2016) and constraints from a shear wave refraction profile (Steinmann et al., 2021). To evaluate the effect of ground frost, we insert a centimeter thick high-velocity layer at the surface of the 1D model. Different thicknesses and shear wave velocities account for different scenarios of the ground frost. The shear wave velocity of the ground frost depends strongly on the temperature and composition of the soil. A silt-clay mixture with a high water content as in our case can reach the eight-fold of its shear wave velocity with temperatures below

$-8\,^{\circ}$C (Miao et al., 2019). Through the shear wave velocity and a constant Poisson's ratio of 0.33 (Zimmerman & King, 1986), we define the compressional wave velocity. We neglect changes in the density and set it to $2000\,\mathrm{kg\,m^{-3}}$ for all layers.

Figure S4 shows the HVSR for different scenarios of ground frost and different number of considered surface waves modes. All models confirm the qualitative observation that the HVSR experiences a broadband decrease above 1 Hz due to a layer of ground frost with a certain thickness and increased shear wave velocity. Apart from the broadband decrease at higher frequencies, the two resonance peaks below 1 Hz do not seem to be effected. With increasing thickness and shear wave velocity the decrease is more pronounced and the maximum decrease moves to lower frequencies. Note that both parameters show a similar effect on the HVSR. Thus, it is difficult to disentangle the two effects in actual observations. We observe this scenario at the end of February and beginning of March marking the coldest and also the longest period of freezing air temperature (Figure 1b). During that time, the horizontal component and the HVSR experience the strongest decrease. However, we cannot say if an increasing thickness or decreasing temperature dominates the process. The number of surface modes considered in the wavefield has also an effect on the pattern of decrease. It has already been shown that large stiffness contrasts or reversal of velocity layers – that is high-velocity layer over low-velocity layer – can cause modal energy pertubation and dominant higher modes (O'Neill & Matsuoka, 2005). Freezing the soil from the surface downwards causes a reversal of velocity layers and might lead to modal energy pertubation. The broadband high-frequent HVSR decrease and its dependence on the number of modes suggest that this effect occurs. This would be important to consider when passive image interferometry is used for monitoring permafrosts. Dominant higher modes could appear on cross-correlograms during times of refreezing in autumn and corrupt measurements of velocity variations. A proper wavefield analysis would be needed to understand this process better, however, it is out of the scope of this work and, thus, subject to future research.

Overall, the model brings interesting insights to our observations retrieved from the seismic data. The observations and model agree qualitatively on a broadband high-frequent HVSR decrease due to grounfrost. The decrease is more pronounced for deeper and colder ground frost. Moreover, the model shows that it is difficult to entangle the interaction between the thickness and temperature of the ground frost and surface wave modes present in the wavefield. It is also clear that the HVSR ratio of the seismic data contains many different source and medium effects (Figure 4i) and, thus, the diffusive wavefield assumption is not valid for the data. This highlights the strength of our data-driven approach, which isolated a pattern in the continuous seismograms related to the freezing and thawing process despite all the other source and medium effects affecting the data.

References

García-Jerez, A., Piña-Flores, J., Sánchez-Sesma, F. J., Luzón, F., & Perton, M. (2016). A computer code for forward calculation and inversion of the h/v spectral ratio under the diffuse field assumption. *Computers & geosciences*, *97*, 67–78.

Miao, Y., Shi, Y., Zhuang, H., Wang, S., Liu, H., & Yu, X. (2019). Influence of seasonal frozen soil on near-surface shear wave velocity in eastern Hokkaido, Japan. *Geophysical Research Letters*, *46*(16), 9497 – 9508.

O'Neill, A., & Matsuoka, T. (2005). Dominant higher surface-wave modes and possible inversion pitfalls. *Journal of Environmental & Engineering Geophysics*, *10*(2), 185–

201.

Piña-Flores, J., Perton, M., García-Jerez, A., Carmona, E., Luzón, F., Molina-Villegas, J. C., & Sánchez-Sesma, F. J. (2016). The inversion of spectral ratio h/v in a layered system using the diffuse field assumption (dfa). *Geophysical Journal International*, ggw416.

Steinmann, R., Hadziioannou, C., & Larose, E. (2021). Effect of centimetric freezing of the near subsurface on rayleigh and love wave velocity in ambient seismic noise correlations. *Geophysical Journal International*, *224*(1), 626–636.

Steinmann, R., Seydoux, L., Beaucé, E., & Campillo, M. (2022). Hierarchical exploration of continuous seismograms with unsupervised learning. *Journal of Geophysical Research: Solid Earth*, *127*(1), e2021JB022455.

Zimmerman, R. W., & King, M. S. (1986). The effect of the extent of freezing on seismic velocities in unconsolidated permafrost. *Geophysics*, *51*(6), 1285–1290.

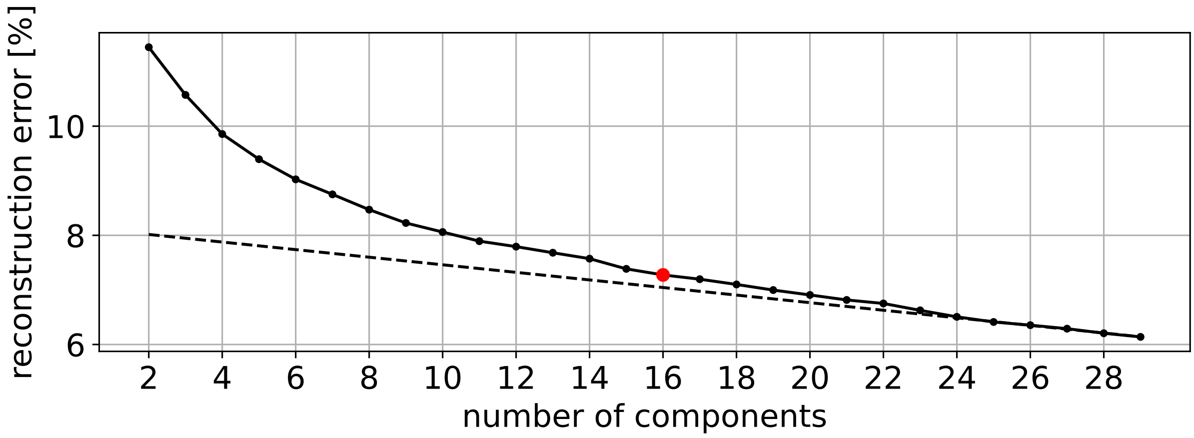

Figure S1. Reconstruction error for ICA-models with different number of independent components. The red dot marks the model we choose for further analysis. The dashed line fits a linear function based on the last seven points.

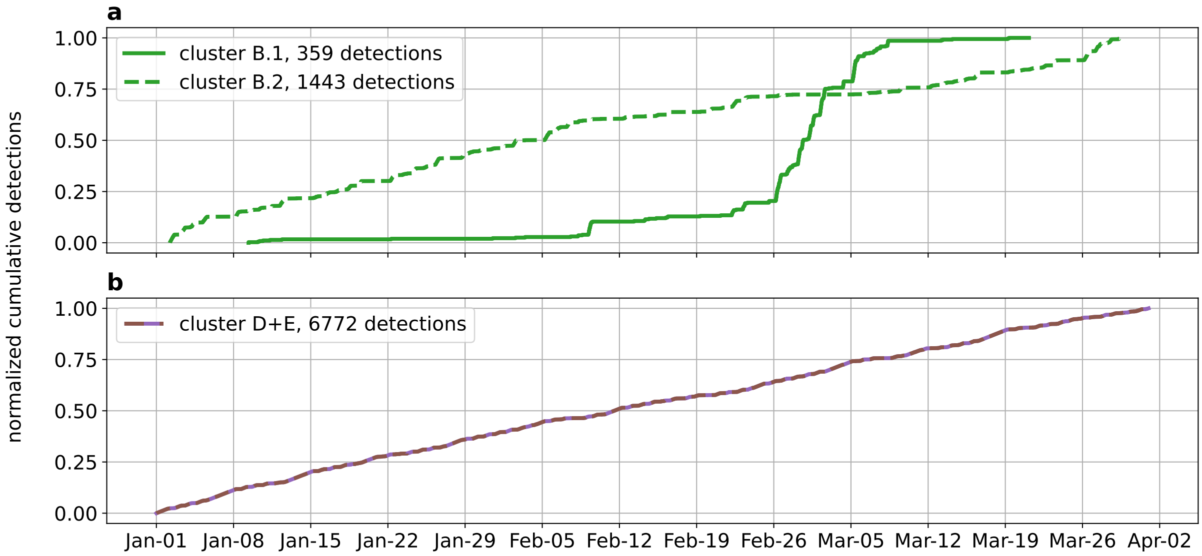

Figure S2. Normalized cumulative detections for other cluster solutions. Normalized cumulative detections for subcluster B.1 and B.2 and the cluster-combination of D and E. Note that each tick at the x-axis marks a Monday.

h [m]	v_s [m/s]	v_p [m/s]	ρ [g/cm^3]
172.82	394.54	1255.93	2000
611.60	520.96	2075.66	2000
∞	947.09	4250.25	2000

Table S1. 1D model of the subsurface at the measuring site based on the inversion of the HVSR with the diffusive field assumption

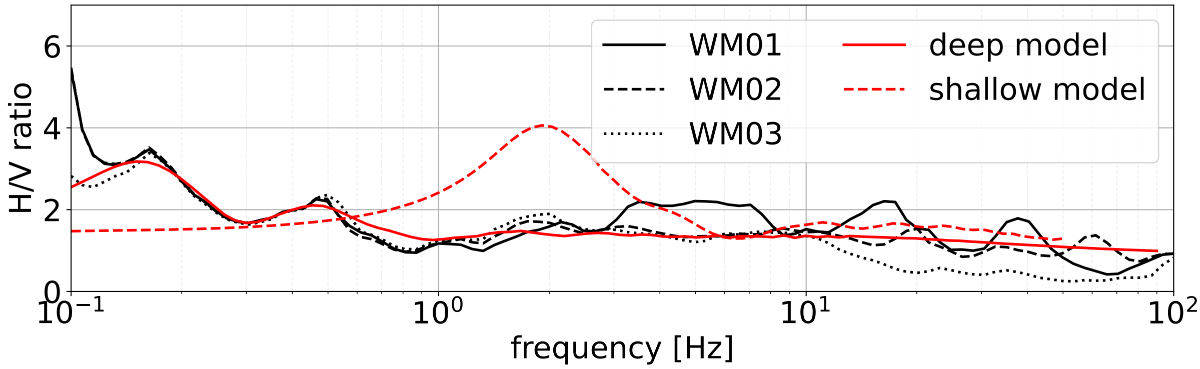

Figure S3. The observed HVSR at all three stations, the modelled HVSR based on the velocity model given in Steinmann et al. (2021) as the dashed red line and the modelled HVSR based on the inversion of the HVSR as the red solid line.

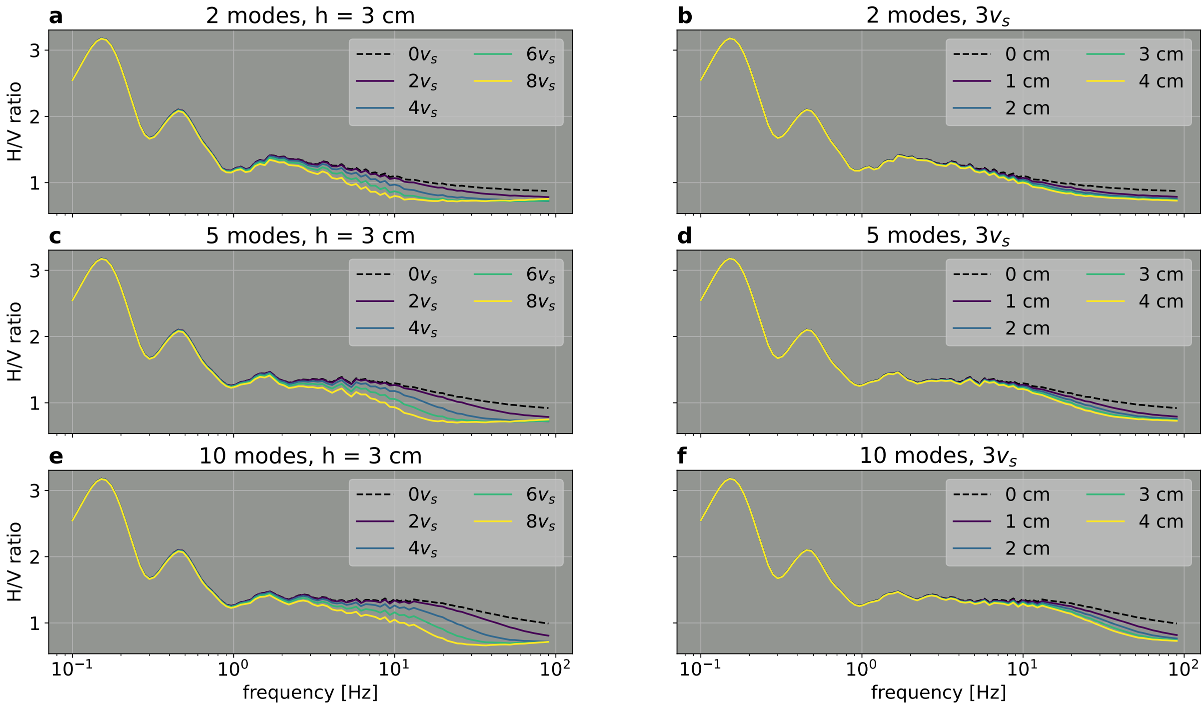

Figure S4. (**a,c,e**) The HVSR in the presence of a 3 cm thick frozen surface layer with varying shear wave velocities and varying number of Rayleigh and Love wave modes. The shear wave velocity of the frozen layer ranges between two-fold and eight-fold of the shear wave velocity of the first layer in the 1D model. The model without a frozen layer is depicted as a black dashed line. (**b,d,f**) The HVSR in the presence of a frozen surface layer with a thickness ranging from 1 to 4 cm and varying number of Rayleigh and Love wave modes. The shear wave velocity is fixed to the three-fold shear wave velocity of the first layer. The model without a frozen layer is depicted as a black dashed line.