Broad-band ambient noise characterization by joint use of cross-correlation and MUSIC algorithm

M Peruzzetto, A Kazantsev, K Luu, J-P Métaxian, F Huguet, Hervé Chauris

Geophysical Journal International

Geophys. J. Int. (2018) **215,** 760–779
Advance Access publication 2018 July 27
GJI Seismology

doi: 10.1093/gji/ggy311

Broad-band ambient noise characterization by joint use of cross-correlation and MUSIC algorithm

M. Peruzzetto,[1,2,3] A. Kazantsev,[1,2] K. Luu,[2] J.-P. Métaxian,[3,4] F. Huguet[1] and H. Chauris[2]

[1]*Storengy, Engie Group, Paris* 92274*, France. E-mail: peruzzetto@ipgp.fr*
[2]*MINES ParisTech, PSL – Research University, Centre de Géosciences, Fontainebleau* 77300*, France*
[3]*Institut de Physique du Globe de Paris, Seismology team, University Paris Diderot, Sorbonne Paris Cité, UMR* 7154 *CNRS, Paris* 75238*, France*
[4]*ISTerre, IRD* R219*, CNRS, Université de Savoie Mont Blanc, Le Bourget-du-Lac* 73376*, France*

Accepted 2018 July 26. Received 2018 July 3; in original form 2018 February 28

SUMMARY

Several days of passive seismic broad-band recordings (vertical component) from a dense 3 × 6 km array installed near Chémery (France), with about 100 seismometers, are analysed for wavefield characterization between 0.1 and 3 Hz. Backazimuth is determined by using the Multiple Signal Characterization (MUSIC) algorithm at frequencies below 1 Hz, and non-coherent cross-correlation beamforming above 1 Hz, since the latter is less sensitive to aliasing issues. A novel method of determining the wavefield velocity is introduced, consisting of processing a cross-correlation common-offset gather by the MUSIC algorithm. The fundamental and three higher modes of Rayleigh waves (R0, R1, R2 and R3) are identified under 1 Hz. Above 1.5 Hz, the L_g phase is detected, while R0 and R1 are also present. Roughly between 1 and 1.5 Hz, a quicker phase, probably P_g, is detected. Both P_g and L_g are dominant during night time, suggesting they have a natural origin, which is also consistent with their backazimuth pointing towards the Atlantic. Large scale 2-D spectral-element simulations using deep- and shallow-water ocean sources confirm the possibility of the L_g phase excitation. Thus, even above 1 Hz, natural sources can explain the major part of the ambient noise energy during quiet time periods.

Key words: Numerical modelling; Guided waves; Seismic interferometry; Seismic noise; Surface waves and free oscillations; Wave propagation.

1 INTRODUCTION

Ambient seismic noise applications are of growing interest in various contexts (Larose *et al.* 2015), boosting the development of numerous analysis methods. After the pioneering works by Douze (1964, 1967), Bonnefoy-Claudet *et al.* (2006a) reviewed the techniques for the ambient wavefield composition, while Koper *et al.* (2010) compared data from 18 seismic arrays around the world to reveal some general trends. The frequency band of interest for most industrial applications (ambient noise tomography, H/V ratio studies for seismic hazard, direct hydrocarbon indicators, etc.) is roughly between 0.1 and 5 Hz. The wavefield at the peak frequencies of the primary and secondary microseisms (approximately 0.07 and 0.14 Hz, respectively) is usually dominated by the fundamental mode associated to Rayleigh and Love waves (Bonnefoy-Claudet *et al.* 2006a). For higher frequencies between 0.2 and 1 Hz, higher modes of Rayleigh and Love waves (Bonnefoy-Claudet *et al.* 2006a; Riahi *et al.* 2013a; Lehujeur *et al.* 2017a), as well as regional body waves (Poli *et al.* 2012) or teleseismic body waves (Pratt *et al.* 2017) may dominate the wavefield. Frequencies above 1 Hz are generally assumed to be dominated by artificial noise, with sources at the surface generating mostly Rayleigh waves of fundamental mode (Bonnefoy-Claudet *et al.* 2006b). However, Koper *et al.* (2010) highlighted L_g phase could also be predominant at these frequencies. The crust/mantle contact might thus play a major role in the short-period ambient noise composition (Kennett 1986).

Applications usually focus on one particular type of wave as representative of the ambient noise. The most widespread approach is to extract the surface wave (Rayleigh and/or Love) dispersion curves and to invert for a vertical shear wave velocity profile. Array methods such as SPAC (spatial auto-correlation, Aki 1957), FK (frequency-wavenumber analysis, Capon *et al.* 1967) or High-Resolution FK (HRFK, Capon 1969) were traditionally used for this purpose. The relatively recent development of the ambient noise cross-correlation allows to extract the medium's Green's function between two passive seismic receivers (e.g. Shapiro & Campillo 2004). Dispersion curves can thus be estimated for each station pair of 2-D receiver arrays, allowing to invert for a smooth 3-D shear wave velocity cube (e.g. Brenguier *et al.* 2007; Mordret *et al.* 2013). For some data sets, cross-correlation based techniques also recover the body-wave part of the Green's function (e.g. Roux *et al.* 2005).

Nakata *et al.* (2015) used the direct diving body waves for a 3-D traveltime tomography, while Draganov *et al.* (2007) and Ruigrok *et al.* (2011) managed to image continuous structural interfaces by extracting reflected *P* waves from the cross-correlations. Teleseismic *P* waves were used by Landés *et al.* (2010), Obrebski *et al.* (2013) and Pratt *et al.* (2017) to track hurricanes and other meteorological perturbations affecting the oceans. Furthermore, heterogeneities within the Earth continuously scatter the ballistic surface and body waves, generating coda waves at later arrival times. They can be used to determine the mean properties of the medium. Sens-Schönfelder & Wegler (2006) presented an application for water-table monitoring.

Three-component data offer the possibility to compute the spectral ratio of the horizontal to vertical (H/V) displacements . The H/V technique was introduced by Nakamura (1989) to derive *SH*-wave resonance and an estimate of the contact depth between the bedrock and the overburden from the H/V ratio peak frequency, despite debatable hypotheses (Bonnefoy-Claudet *et al.* 2006b).

Dangel *et al.* (2003) observed an amplification of the vertical component as well as a low H/V ratio for frequencies ranging from 1 to 5 Hz above hydrocarbon reservoirs. This resulted into a debate, some authors claiming that the amplification could be used as a direct hydrocarbon indicator ('DHI', see Lambert *et al.* 2009), while others objected that it was due to either higher artificial noise level near the hydrocarbon extraction facilities or shallow geological structural effects (e.g. Green & Greenhalgh 2010,and references herein). A quantitative modelling of the amplification of the vertical component would require a detailed knowledge of the ambient noise content over the frequency band of interest. This is the main objective of this work.

We propose here a methodology based on a joint use of a high-resolution array method [Multiple Signal Characterization (MUSIC); Schmidt 1986; Goldstein & Archuleta 1987) and ambient cross-correlation on the vertical data recorded by a dense array of broad-band seismometers. In this way, the dispersion curves and the backazimuths of the wavefield can be determined over a wide frequency band. Applying MUSIC to the windowed recorded signals provides a better resolution at low frequencies compared to FK or HRFK (Cornou 2002), provided the number of incoming waves is inferred correctly. For higher frequencies, cross-correlation based beamforming techniques (e.g. Mordret *et al.* 2013; Nakata *et al.* 2015; Lehujeur *et al.* 2017b) allow to obtain reliable slowness for interstation distances beyond the wavelength (aliasing limit). This is achieved by aligning interstation correlations instead of aligning raw waveforms. We suggest the combination of these methods provides a better insight into the wavefield composition over a wider frequency range than the one associated with a specific approach.

We first present the data set recorded in Central France (Section 2) and then the methodology (Section 3). In Section 4 we focus on the real data set application. We discuss the evidence for the presence of Rayleigh wave modes and two crustal phases P_g and L_g into the [0.1–3] Hz frequency band, and the possibility to invert for a 1-D shear wave velocity profile. In Section 5, we use the spectral-element method to simulate the wavefield and to distinguish between artificial local surface sources and natural distant deep- or shallow-water sources (DWS and SWS, respectively), based on the wavefield content determined by our methodology.

2 DATA

The study site is located inside the Paris sedimentary basin. Successive sedimentary layers are mainly composed of limestones, sandstones and clays. The reservoir itself is a layer of Triassic sandstones at a depth of approximately 1130 m (blue contour lines in Fig. 1 a and green box in Fig. 1b). The thickness of the layer varies from 30 to 60 m. The depth of the basement is unknown at the reservoir location. It was neither reached by any of the wells nor seen in active seismic data, because the targeted reservoir layers were much shallower. The deepest exploration well in the area (CS01, see location in Fig. 1a), drilled in 1959 by Mobil Repga, stopped at 2680 m, where Permo-Triassic sediments were still dominating (Fig. 1b). Such a deep basement in this area is consistent with the reported presence of a Permian basin (Gély & Hanot 2014,see the map enclosed in their book) .

About 100 broad-band Trillium 40 s seismometers (nanometrics) were deployed in April 2010 and November 2010 with various geometries and a 100 Hz sampling frequency. Spectrograms reconstructed over the whole time period from the stations successively located near the centre of the study area (within the dashed green rectangle in Fig. 1a) are presented in Fig. 2. As expected, above 1 Hz, they exhibit lower power spectral density (PSD) during night time and on Sundays, as well as during lunch pauses at noon. These features, typical of anthropogenic noise, motivate a separate analysis of day- and night-time periods, because the dominating wavefield sources might alternate between natural and artificial. In our study, we specifically focus on the time periods when the array had a rectangular configuration. From 20 to 24 April, and from 4 to 11 November, most of the sensors were operational between 3 p.m. and 6 a.m. local time (UTC+1), and were shifted by about 250 m every day. There was an almost uniform spacing of 500 m between the sensors installed over a 3 km × 6 km area located above the Underground Gas Storage (UGS). The deployment geometry for 20 April is shown in Fig. 1(a). We used 3 hr of continuous recording, either 1–4 a.m. or 3–6 p.m. local time, respectively, referred to as 'night time' and 'day time', and processed separately. Such 3 hr intervals were concatenated for all the available days (four in April and eight in November). Riahi *et al.* (2013a) already analysed our data set using a three-component beamforming algorithm into a narrower frequency band [0.4–1.1] Hz. Both results will be compared in Section 5.

3 METHODOLOGY

For both cross-correlation and MUSIC approaches, recordings were first split into elementary time windows of about 328 s (2^{15} samples), overlapping by 50 per cent. Each window was tapered in time domain and signals were then filtered between 0.1 and 3 Hz. Tapered square cosine windows with a transition width of 10 per cent of the total length of the taper were used in both time and frequency domains.

3.1 Aliasing and resolution limits

We compute the array response function (ARF) following, among others, Krim & Viberg (1996), Cornou (2002) and Foti *et al.* (2011). An example of ARF for the array configuration we used in our real data application (Fig. 1a) is shown in Fig. 3(a). The estimated aliasing limit (minimal unambiguously detectable wavelength) is twice the shortest interstation distance (here, $\lambda_{min} = 2d_{min} = 1000$ m). The resolution limit (maximum detectable wavelength) can be estimated as the full width of the ARF's central lobe at half maximum (in our case, $\lambda_{max} = 4000$ m). Beyond this wavelength, the array resolution is not good enough for distinguishing a real incident wave from a wave with an infinite apparent velocity. It should be stressed that

(a) Array deployment 20/04/2010 (b) CS01 simplified log

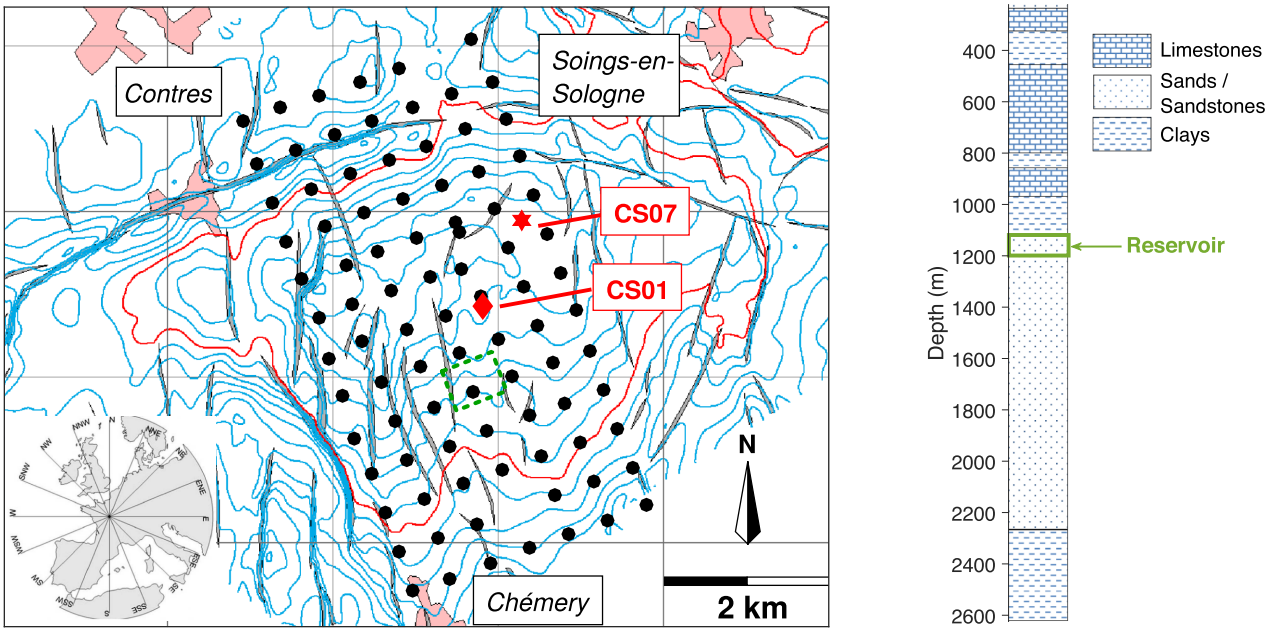
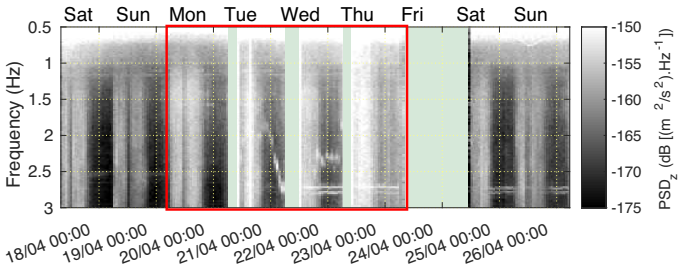


Figure 1. (a) Seismic network. Black dots mark the positions of the sensors deployed on 20 April. Reservoir's contour lines are shown in blue, and its maximum possible extent is given by the red curve. The black lines patch known faults. Urban areas are shaded in pink. The positions of wells CS01 and CS07 are shown with red markers. The green dashed line delimits the area where sensors were taken to compute the spectrograms in Fig. 2. The site location map in the left-bottom corner was readily taken from Riahi *et al.* (2013a). (b) Simplified fundamental log at well CS01. The reservoir is shown by the green box. Log data were provided by Storengy.

(a) Power Spectral Density in April



(b) Power Spectral Density in November

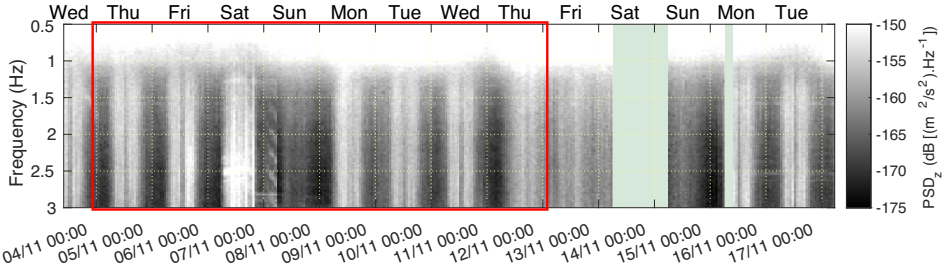


Figure 2. Power Spectral Density of the vertical component velocity for the data recorded in April 2010 (a) and November 2010 (b). They are computed using one station within the green rectangle in Fig. 1(a). The exact station location is not necessarily the same from one day to another. Times when no stations were recording in the chosen area are highlighted in pale green. The red rectangles delimit the time period when rectangular arrays were used, which corresponds to the time period analysed in this work.

these limits are suitable for standard FK analysis. In the following, we will see that high-resolution methods sometimes allow to obtain reliable results beyond these theoretical limitations, as already mentioned by Foti *et al.* (2011).

3.2 Cross-correlation

Each tapered time window was pre-processed following Bensen *et al.* (2007), including spectral whitening and one-bit normalization to filter out earthquakes. Correlations between pre-processed signals

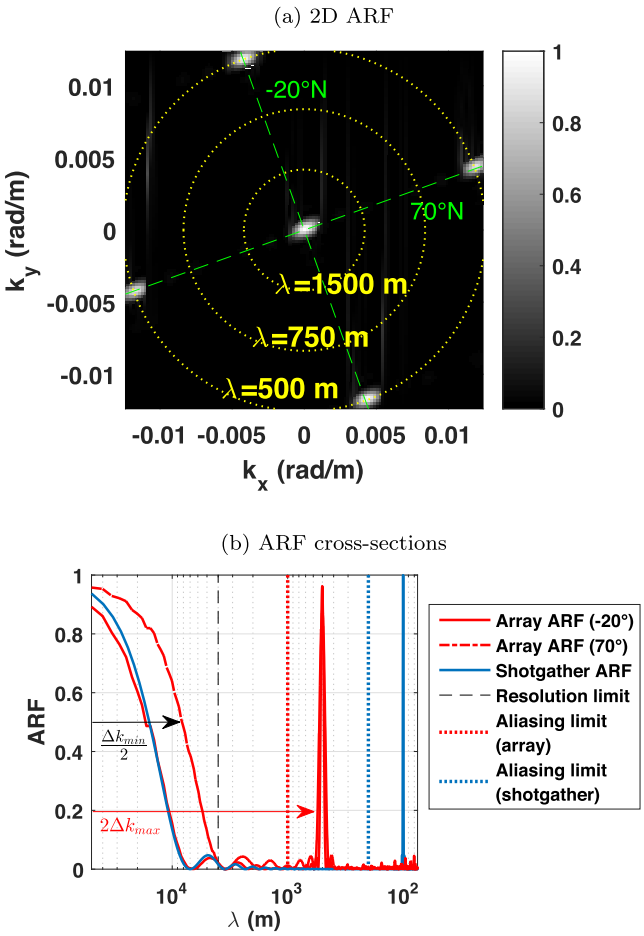


Figure 3. (a) Array Response Function (ARF) of the rectangular array deployed on 20 April. Yellow circles indicate wavelengths in the wavevector plane. (b) ARF along the two main directions of the array (red plain and dash-dotted curves), compared to the 1-D ARF of the virtual shot gather obtained from the cross-correlations (blue curve). The black dash-dotted line represents the theoretical maximum detectable wavelength (resolution limit). The red and blue dotted lines represent the minimum detectable wavelength (aliasing limit) for the 2-D array and the shot gather, respectively.

s_i and s_j recorded at stations i and j were averaged over the available L time windows, yielding

$$\Gamma_{ij}(t) = \frac{1}{L}\sum_{l=1}^{L}\int s_i^{(l)}(\tau)s_j^{(l)}(\tau+t)\mathrm{d}\tau. \quad (1)$$

In case of evenly distributed sources, $\Gamma_{ij}(t)$ approaches the Green's function between the two stations to within an amplitude factor (Boschi *et al.* 2013). For an uneven distribution, the causal part is correctly reconstructed for the station pairs aligned with the major source direction, while it is biased for station pairs with other orientations (Lehujeur *et al.* 2017a). In the latter case, the cross-correlations keep an imprint of the source azimuthal distribution, and can be used for measuring this distribution inside a routine beamforming procedure (Section 3.2.1). On the contrary, if the purpose is to measure the Green's function of the medium to derive dispersion curves, the uneven source distribution must be corrected for. This is done by averaging cross-correlations over similar station pairs (Section 3.2.2).

3.2.1 *Cross-correlation beamforming (CC-beam)*

CC-beam can be carried out either in the frequency domain (Ruigrok *et al.* 2017) or in the time domain. In our study, we choose the latter solution since it allows an easy implementation of the non-coherent beamforming (see below). Assuming a single plane wave with slowness vector $\boldsymbol{s}$, the time delay between two stations is

$$\tau_{ij}(\boldsymbol{s}) = \boldsymbol{s}\cdot\boldsymbol{r_{ij}}, \quad (2)$$

where $\boldsymbol{r_{ij}}$ is the position vector going from station i to station j. $\Gamma_{ij}(t)$ is maximum at $t=\tau_{ij}(\boldsymbol{s})$ for any couple (i,j). Thus, the sum $\sum_{i>j}\Gamma_{ij}(\tau_{ij}(\boldsymbol{s}))$ is maximum if $\boldsymbol{s}$ is actually the slowness of the wave propagating across the array (Rost & Thomas 2002). In practice, the signal-to-noise ratio (SNR) in the cross-correlation can be quite low at frequencies above 1 Hz because of strong scattering. Under such conditions, it is more robust to sum envelopes of the correlations:

$$D_{\text{CC-beam}}(\boldsymbol{s}) = \sum_{i>j}\left|\mathcal{H}\left\{\Gamma_{ij}\right\}(\tau_{ij}(\boldsymbol{s}))\right|, \quad (3)$$

where $\mathcal{H}$ is the Hilbert transform. This will be called *non-coherent* beamforming in the following. If several uncorrelated waves propagate across the array, $D_{\text{CC-beam}}$ will exhibit several peaks at the corresponding slowness vectors. In order to obtain slowness vectors into different frequency bands, correlations were filtered with a 0.2 Hz wide-tapered square cosine windows centred on the frequency of interest. Examples are shown in Figs 4 (a) and (c). Such beamformers yield an estimate of the backazimuth distribution.

This method can also be used for group velocity dispersion curve estimation (see Appendix A, Fig. A1). However the grid search over the slowness space at each requested frequency is rather CPU consuming and more importantly the resolution at low frequency is very low. This motivates another way of exploiting the cross-correlations for the dispersion curve retrieval, which will be addressed in the next paragraph.

3.2.2 *Common-offset stacking*

Most of the time, sources are not uniformly distributed around onshore arrays, preventing a proper reconstruction of the medium Green's function from cross-correlations. This problem can be efficiently fixed if we accept to drop the azimuthal information. In the theoretical derivation (e.g. Boschi *et al.* 2013), the azimuthal integration over the source distribution is indeed equivalent to an azimuthal integration over the station pair's orientation. If enough station pairs with close offsets (i.e. distance between stations) and different orientations are available, the uniform source distribution condition is fulfilled for their average correlation. As already done by Poli *et al.* (2012), Mordret *et al.* (2013) and Nakata *et al.* (2015) and others, we divided all the available station pairs into 100 m offset bins, and averaged the cross-correlations for station pairs inside each bin. To correct for the longer NW–SE array extent, we also divided each offset bin into 10° azimuthal bins, and computed the average cross-correlation as

$$\Gamma(r,t) = \frac{\sqrt{r}}{\sum w_{ij}}\sum_{|\boldsymbol{r_{ij}}|\in[r,r+\Delta r]} w_{ij}\Gamma_{ij}(t), \quad (4)$$

with $w_{ij}=1/N_\theta(i,j)$ being a weight inversely proportional to $N_\theta(i,j)$ the number of pairs in the azimuthal bin containing the couple (i,j), and $\sqrt{r}$ a correction term for surface wave geometrical spreading. This yields a 2-D shot gather with one trace per offset and a spacing equal to the offset bin width Δr. Examples of such shot gathers are

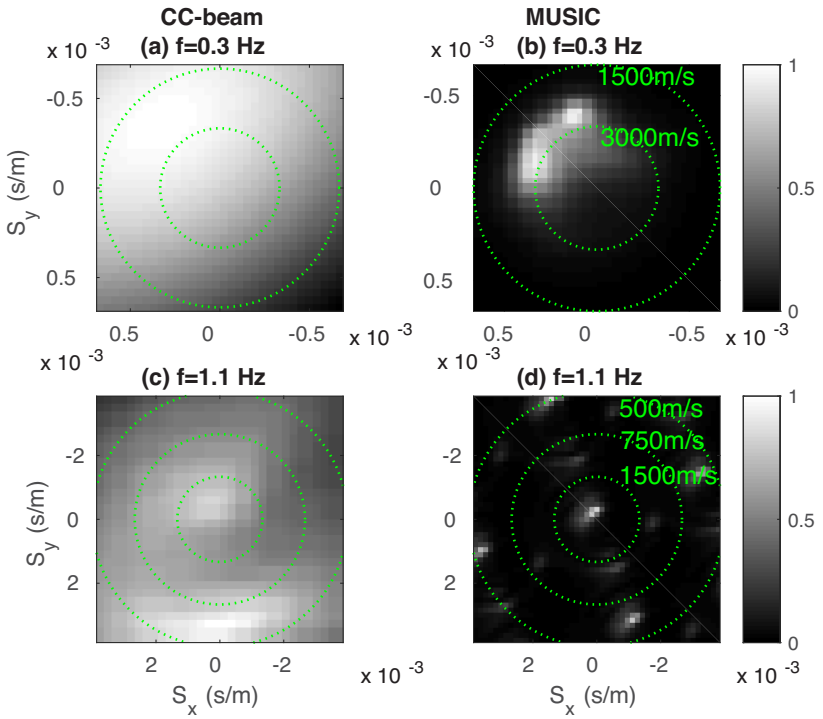


Figure 4. Compared beamformers for CC-beam and MUSIC (20–23 April). (a) CC-beam (0.3 Hz), (b) MUSIC (0.3 Hz), (c) CC-beam (1.1 Hz) and (d) MUSIC (1.1 Hz). Slowness axes represent group slowness for CC-beam and phase slowness for MUSIC. The green circles show several velocity values, which are labelled in the right column.

shown in Figs 5 (a) and (b). Δr is often smaller than the typical interstation distance, which shifts the aliasing limit towards smaller wavelengths (Fig. 3b): $\lambda_{\min} = 2\Delta r$. We assume the resolution limit does not change with respect to the 2-D ARF defined in Section 3.1. In order to obtain the dispersion curves, the shot gather can now be processed as a linear array, using standard FK processing (FK-CC, Figs 5 c and d), or, more interestingly, using the MUSIC algorithm.

3.3 MUSIC

The MUSIC algorithm separates recorded data into signal and noise subspaces (Schmidt 1986; Goldstein & Archuleta 1987). To achieve this separation, the data are first gathered into the cross-spectral matrix (CSM) $R(f)$ with elements given by

$$R_{ij}(f) = \langle S_i(f)\overline{S_j(f)}\rangle, \tag{5}$$

where $S_i(f)$ and $S_j(f)$ are the Fourier transforms of a one-component signal at stations i and j, $\bar{}$ is the complex conjugate and $\langle\,\rangle$ stands for an averaging operation, performed over adjacent frequencies (spectral smoothing), different time windows (temporal smoothing), similar subarrays or similar station pairs (spatial smoothing).

Given *a priori* knowledge of the number Q of plane waves to be detected, the MUSIC functional D_{M} minimizes the projection of the data onto the noise subspace (see Appendix B). It is difficult in practice to determine Q when analysing seismic noise. Besides, in real noisy data, one plane wave can be 'spread' over several eigenvalues. Several techniques were proposed for an automated detection of Q (see the review in Cornou 2002). Here we developed a slightly different method, more robust for the purpose of pure ambient noise analysis. It is described in detail in Appendix B. The idea is to use the slope break in the logarithmic eigenvalue decay in case of a strong SNR, and a simple eigenvalue magnitude criterion in case of a low SNR.

From the definition (5) of the CSM, it will only be of rank 1 (which means one non-zero eigenvalue) if the averaging operation is not performed. The smoothing operation in eq. (5) is thus necessary for the derivation presented in Appendix B to hold (see Bokelmann & Baisch 1999; Cornou 2002). The traditional technique for *spatial smoothing* is to average CSMs obtained from identical subarrays, which results into a smaller CSM with increased rank (Shan *et al.* 1985). However, this technique can only be applied to almost perfectly regular arrays. An alternative method consists in averaging individual terms in the CSM corresponding to nearly identical station couples (Bokelmann & Baisch 1999). We used either option for spatial smoothing, depending on the type of the data to which MUSIC was applied.

3.3.1 2-D array

Though the 2-D arrays in our data set were quite regular, we preferred the second alternative for *spatial smoothing* (i.e. Bokelmann & Baisch 1999) in order to make the method easily repeatable on less regular arrays. The station pairs were divided into 100 m offset bins ranging from 0 to 7000 m, and 5° azimuth bins ranging from 0° to 180°. CSM values were then averaged inside each bin. *Spectral smoothing* was also performed over a 0.1 Hz interval centred on the frequency of interest. After smoothing and diagonalizing the CSM, D_{M} can be estimated from eq. (B5) and mapped over the 2-D wave vector space, or equivalently the slowness vector space, provided $\boldsymbol{k} = 2\pi f \boldsymbol{s}$. For a given frequency, such slowness maps are computed for each time window and then averaged, yielding the final beamformer (see examples shown in Figs 4 b and d). In the

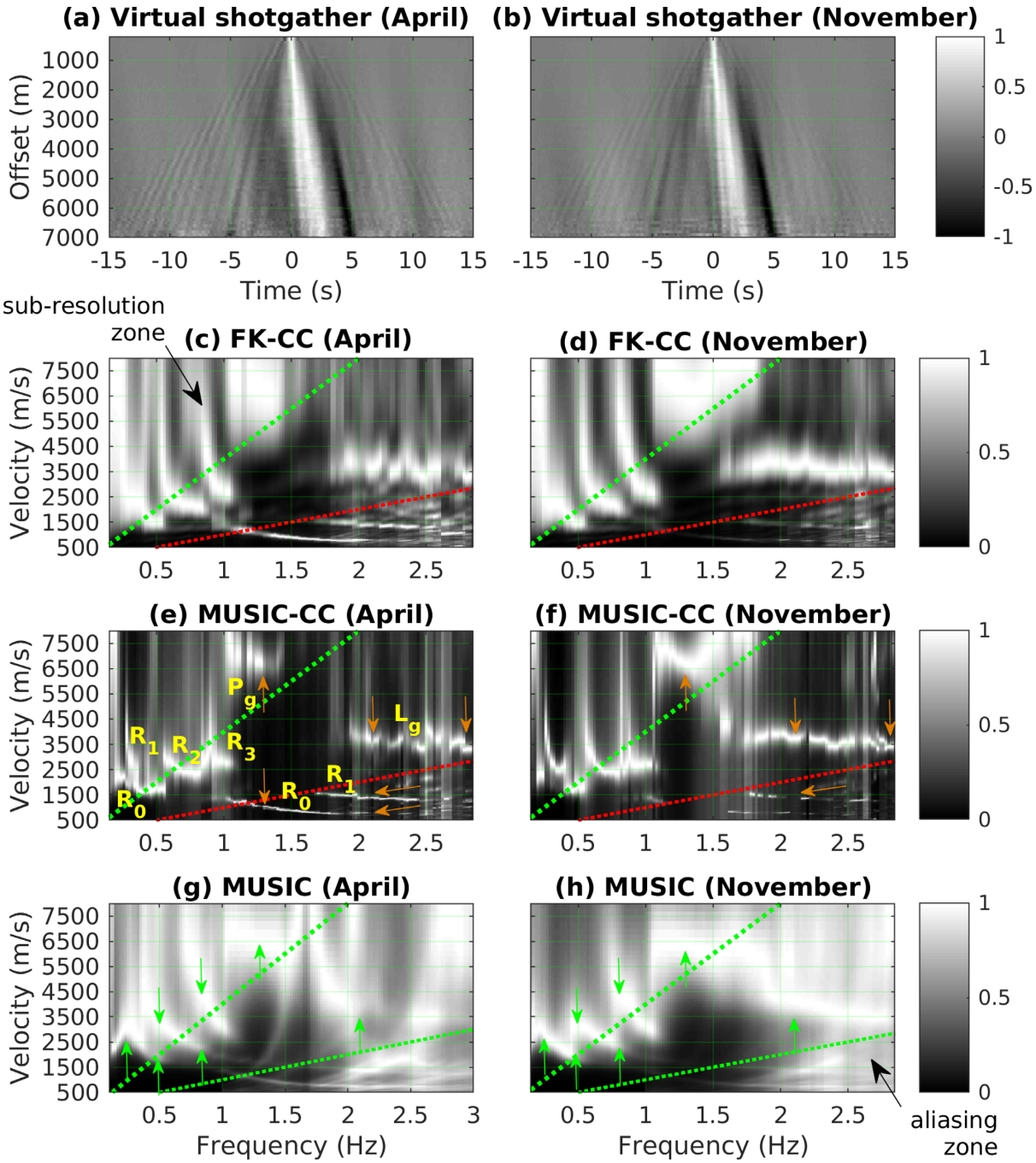


Figure 5. Night-time dispersion study in April (left column) and November (right column). (a and b) Virtual shot gathers from cross-correlation; (c and d) FK-CC (standard FK processing of the virtual shot gathers); (e and f) MUSIC-CC (MUSIC processing of the virtual shot gathers). (g and h) 'direct' MUSIC processing of the signal recorded by the 2-D array (without cross-correlation). MUSIC-CC yields the best results: for instance, compare panels (e)–(c) and panels (e)–(g). Green dotted lines mark the theoretical resolution and aliasing limits of the processed arrays. The red dotted line reported in (c)–(f) is the aliasing limit of the 2-D array. The yellow labels identify different detected modes. The green (resp. orange) arrows are plotted at frequencies for which beamformers are shown in Fig. 6 (resp. Fig. 7); they point towards the modes expected in the beamformers at these frequencies.

following, we refer to the method described in this subsection as 'direct' MUSIC, since the algorithm is applied directly to the recorded waveforms. While the resolution at low frequency (0.3 Hz, Fig. 4b) is much better compared to CC-beam described into the previous section (Fig. 4a), the MUSIC beamformer is severely aliased at higher frequencies (for instance at 1.1 Hz, the minimum velocity of the non-aliased zone is 1100 m s^{-1}, Fig. 4d). On the contrary, the CC-beam beamformer (here 1.1 Hz, Fig. 4c), though poorly resolved, does not encounter aliasing. This is because it operates on the time delay of wave packet propagation and not the phase delay. The comparison suggests a strong complementarity between both approaches for backazimuth determination. The MUSIC beamformer can finally be integrated over azimuth for each frequency. Concatenating the resulting 1-D curves yields a dispersion plot in (f, s) or (f, v_ϕ) domain (in Figs 5 g and h). The validity of such a dispersion plot between the aliasing and the resolution limits is also

confirmed in a simple synthetic test performed in Appendix C. The consistency between phase and group velocity dispersion curves derived respectively by MUSIC and CC-beam is verified in Appendix A. However, neither of these two methods applied individually is as efficient for dispersion curve determination as the MUSIC algorithm applied to a cross-correlation common-offset gather. The latter approach presented in the next section.

3.3.2 Linear array with regular spacing

MUSIC was applied to virtual shot gathers obtained by cross-correlation (MUSIC-CC). Such a shot gather is equivalent to a linear array with a constant spacing equal to the width of the offset bin used during the common-offset stacking of the cross-correlations (Section 3.2.2). Since it is very easy to define identical subarrays in this case, we used the traditional *spatial smoothing* described in Shan *et al.* (1985). CSMs were thus averaged over 20 subarrays. *Spectral smoothing* was again implemented with a 0.1 Hz smoothing interval around the frequency of interest. Here D_M can be directly plotted on a 1-D slowness or phase velocity axis as the virtual array is linear. Concatenating such curves obtained for successive frequencies again yields a dispersion plot in (f, s) or (f, v_ϕ) domain. Such a dispersion plot (see examples in Figs 5 e and f) exhibits a better resolution at low frequencies and less aliasing at high frequencies compared to 'direct' MUSIC (Figs 5 g and h). MUSIC-CC's resolution is also better than for the standard FK processing of the virtual shot gather (FK-CC, Figs 5 c and d).

4 RESULTS

In this section, we present the results obtained with our methodology on the data acquired above a UGS near Chémery (Central France, Fig. 1a) in April 2010 and November 2010. The UGS was at its maximum filling level in November. These experiments originally aimed at a time-lapse observation of a low-frequency amplitude anomaly above the UGS as a DHI. We compare the dispersion plots in the frequency band [0.3–3] Hz obtained by the different methods from Section 3 (FK-CC, MUSIC-CC and 'direct' MUSIC). Several propagating modes are identified, and their backazimuths are studied using MUSIC and CC-beam approaches, outside and inside the aliasing zone, respectively. Differences in the wavefield composition between day and night period are investigated. Finally, the dispersion curves identified for different Rayleigh wave modes are jointly inverted for a 1-D shear wave profile in order to check the compatibility of their identification with the local geology.

4.1 Dispersion plots

Cross-correlations are computed as described in Section 3.2.2 between 0.1 and 3 Hz. Dispersion plots are derived into the frequency range [0.15–2.85] Hz, which is slightly smaller than the initial range because the MUSIC-based approaches use spectral smoothing over a 0.1 Hz interval. Results are presented in Fig. 5, where the left column is for April data and the right column for November data.

The shot gathers computed for April and November (Figs 5 a and b) look very similar, confirming the relative stability of the wavefield. The standard FK processing of the shot gathers (FK-CC, Figs 5 c and d) exhibits three distinct dispersive patches below 1 Hz which it would be natural to interpret as three Rayleigh wave modes. However, several modes might be mixed into one apparent patch in case of an insufficient resolution. The velocity of the patches remains below 3500 m s^{-1} most of the time, but exhibits sharp peaks up to 8000 m s^{-1}, which are unrealistic for a surface wave dispersion curve. Compared to FK-CC, MUSIC-CC (Figs 5 e and f) exhibits a qualitatively similar pattern with a better resolution. Two distinct modes can now be distinguished below 0.4 Hz, where FK-CC just identified one wide patch. Separating these two low-frequency modes is crucial for the V_S profile inversion in depth.

Above 1 Hz, in April, two clear dispersive modes are detected below 2000 m s^{-1}, identified as the Rayleigh fundamental mode and the first overtone. A quick non-dispersive phase ($v_\phi \approx 7000$ m s^{-1}) is observed between 1 and 1.5 Hz in April, and between 1 and 1.8 Hz in November, which we interpret as P_g. It is indeed too quick for a guided S wave, and too slow for a teleseismic arrival (e.g. Obrebski *et al.* 2013). Finally, above 2 Hz, an apparently non-dispersive phase propagating slightly quicker than 3500 m s^{-1} is detected. As Koper *et al.* (2010), we interpret it as L_g, which we will further discuss in Section 5. All the identified modes are labelled in yellow in Fig. 5(e).

The dispersion plot obtained from the direct application of the MUSIC algorithm to the noise recorded by the 2-D array (Figs 5 g and h), exhibits a lower resolution compared to MUSIC-CC (Figs 5 e and f) and suffers from artefacts inside the aliasing zone. Branches due to the waves from the aliasing zone can also contaminate the aliasing-free zone, as for example, the steeply increasing branch between 1.3 and 1.7 Hz in Fig. 5(g). Below 0.4 Hz, only one modal branch is visible, with roughly the average phase velocity of the two modes $R0$ and $R1$ identified in Fig. 5(e). This behaviour is expected for MUSIC applied to noisy multimodal waveforms with random emission times and azimuths, as shown for synthetic data in Appendix C. Thus, MUSIC-CC appears to be the best suited method for dispersion curves retrieval at all frequencies (for instance, for April night-time data, compare Fig. 5 e to Fig. 5c, and Fig. 5 e to Fig. 5g). The automated signal-subspace determination plays a major role in the efficiency of the method (see Appendix B for details).

4.2 Backazimuths

As explained in Section 3, 'direct' MUSIC is used to extract beam patterns outside the aliasing zone (Fig. 6), while non-coherent CC-beam is used inside it (Fig. 7). All the plotted beamformers are normalized between 0 and 1.

The knowledge of the dispersion plots (Fig. 5) allows to better assess the mode labels in the beamformers, benefitting from a continuous representation over a wide frequency range. Green arrows in Figs 5(g) and (h) indicate the modes which are expected to show up in the MUSIC beamformers, at the frequencies for which they are plotted in Fig. 6, namely 0.25, 0.5, 0.8, 1.3 and 2.1 Hz. The arrows below 1 Hz correspond to several dispersive Rayleigh wave modes, while the arrows at 1.3 and 2.1 Hz correspond to the non-dispersive P_g and L_g phases, respectively. In the same way, brown arrows in Figs 5(e) and (f) indicate the expected modes for the CC-beam images in Fig. 7 at 1.3, 2.1 and 2.8 Hz. Phase velocities from Fig. 5 were converted into group velocities for the dispersive modes R0 and R1, in order to place the arrows at the right distance from the centre in Fig. 7. The conversion procedure is described in Appendix A. If a mode expected from the dispersion plot is not found in the beamformer, its label is put inside brackets. If an unexpected mode is identified, it is labelled with brackets and without any arrow.

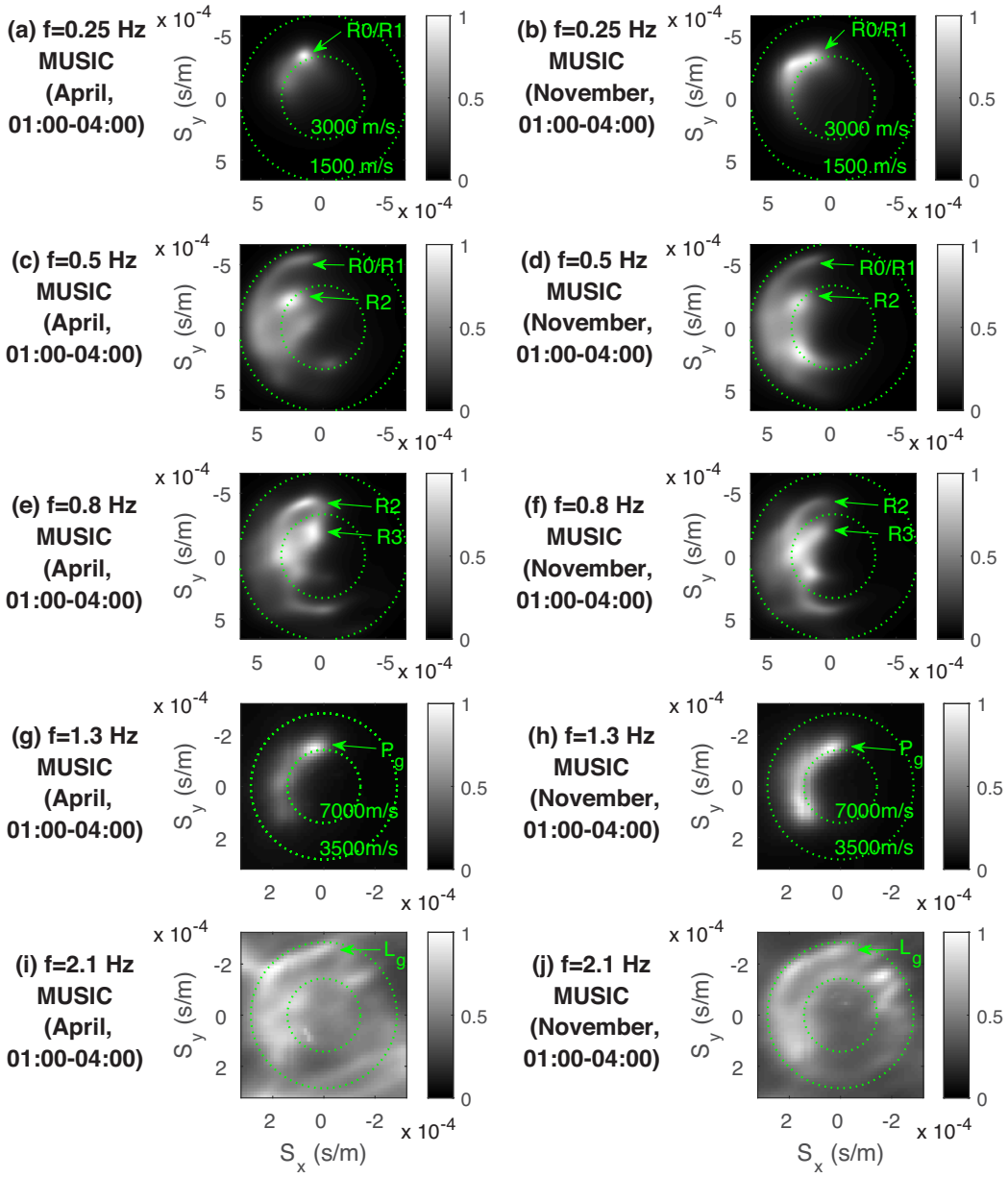


Figure 6. Night-time backazimuths in April (left column) and November (right column), provided by MUSIC outside the aliasing zone. Analysed frequencies are 0.25 Hz (a and b), 0.5 Hz (c and d), 0.8 Hz (d and f), 1.3 Hz (g and h) and 2.1 Hz (i and j). The green circles indicate some velocity values; if they are not labelled, the velocity values are those of the subplot above. The mode labels are those from Fig. 5(e). The green arrows in the left (resp. right) column are those from Fig. 5(g; resp. 5h).

4.2.1 Outside aliasing zone (MUSIC)

In both April and November, a continuous change of the dominant backazimuth can be observed over frequencies. At low frequencies (0.25 Hz), the dominant direction is NNW (Figs 6 a and b), consistent with the well-known secondary microseim generation zone in the Northern Atlantic (Ardhuin *et al.* 2011; Stutzmann *et al.* 2012). At 0.5 and 0.8 Hz (Figs 6 c–f), backazimuths cover directions from N to SSW, suggesting a generation area along the Atlantic coast. The P_g phase (1–1.5 Hz, Figs 6 g and h) exhibits a remarkably regular beamformer in November, covering backazimuths from N to SW. The same distribution is observed in April with a strong peak for the N direction. The L_g phase (above 2 Hz, Figs 6 i and j) appears with a noisier beamformer (especially in April), probably because of higher scattering and contamination by branches coming from the aliasing zone (see Fig. 5g). The backazimuth for both April and November ranges from NE to SW, suggesting a contribution from the Scandinavian Northern coast, which is also known for being a secondary microseism generation zone (Essen *et al.* 2003).

4.2.2 Inside aliasing zone (CC-beam)

Compared to the previous paragraph, a radical change of the backazimuth distribution is observed in the aliasing zone above 1 Hz (Fig. 7). In April, the SE–S directions are dominant at all the plotted frequencies (1.3, 2.1 and 2.8 Hz, see Figs 7a, c and e, respectively).

Only the R0 mode is detected by CC-beam, also when it is unexpected from the dispersion plot (Figs 7 d–f). R1 is never detected,

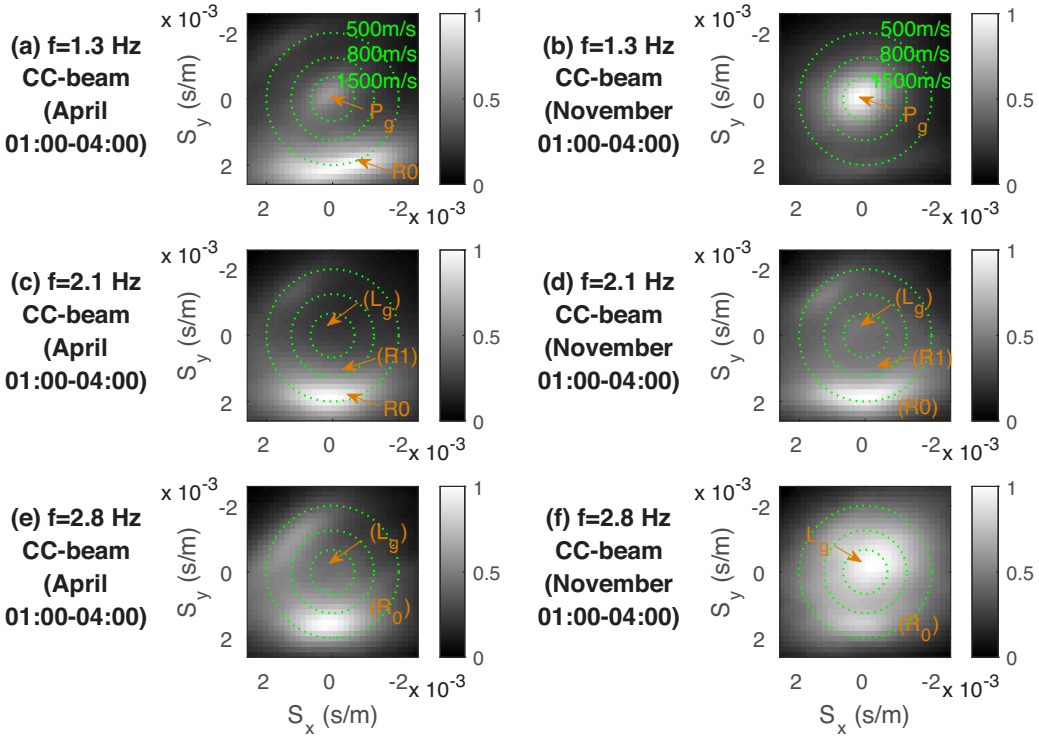


Figure 7. Night-time backazimuths in April (left column) and November (right column), provided by CC-beam inside the aliasing zone. Analysed frequencies are 1.3 Hz (a and b), 2.1 Hz (c and d) and 2.8 Hz (e and f).

although expected in Figs 7(c) and (d). This is probably due to the fact that the R1 mode does not form an independent wave packet in the cross-correlations, making it undetectable by a non-coherent beamforming approach.

Though the quick non-dispersive phases do not lie in the aliasing zone, they should still be seen in the beamformers. They are better visible in November (P_g in Fig. 7 b and L_g in Fig. 7f) compared to April (only a weak P_g peak in Fig. 7a). This confirms those phases are more energetic in November, as expected from Figs 5(g) and (h). However, it is still unclear to us why L_g is completely missed by the CC-beam method in Figs 7(c)–(e): while R1 is mixed with R0 and thus not detected, the L_g phase does not seem to be mixed with another phase.

4.3 Day/night variation

Day-time virtual shot gathers and the corresponding MUSIC-CC dispersion plots are shown in Fig. 8. It is striking to observe that the quick P_g and L_g phases, dominant at night (Figs 5 e and f), almost completely disappear during the day, overshaded by the R0 and R1 modes (Figs 8 c and d). This suggests a different origin for P_g/L_g and R0/R1 above 1 Hz, which will be discussed in Section 5. This analysis is completed by the day-time backazimuth visualization below and above 1 Hz (Figs 9 and 10, respectively). While the diurnal variation is minimal below 1 Hz (compare Figs 6 and 9), it is quite strong above 1 Hz (compare Figs 7 and 10). During the day, the backazimuth distribution above 1 Hz clusters around NNW and S directions, while only the southern directions dominate during the night. Also, as expected from Figs 8(c) and (d), the quick non-dispersive phases P_g and L_g are almost totally overshaded by the slow dispersive phases (R0 in Fig. 10).

4.4 Inversion for V_S

In order to check the compatibility of the mode identification with the available geological knowledge, and to investigate the potential for exploration purposes, we inverted Rayleigh wave dispersion for a vertical *S*-wave velocity profile. Dispersion curves were manually picked from the MUSIC-CC dispersion plots for April (Fig. 5e, all modes except R0 between 2.1 and 2.35 Hz) and for November (Fig. 5f, R0 between 2.1 and 2.35 Hz). The picked curves are shown with circles in Fig. 11(a). Only R0, R1 at high frequency (1.7–2.5 Hz) and R2 were used for inversion, since we were less confident about picking R1 at low frequency and R3 (large uncertainty and narrow spectral extent). Since the L_g phase velocity (approximately 3500 m s^{-1}, red dotted line in Fig. 11a) is close to V_S in the crust, the latter value was constrained between 3400 and 3600 m s^{-1}. Theoretical dispersion curves were estimated via the Thomson–Haskell method (Thomson 1950; Haskell 1953). The picked dispersion curves were inverted using a Competitive Particle Swarm Optimiser (Luu *et al.* 2018) with 20 independent inversions. For each inversion, we used a swarm size of 50 and a maximum number of iterations of 500. We parametrized the velocity model with nine layers, and inverted for the *S*-wave velocity V_S, the V_P/V_S ratio and the thickness of each layer. The mean V_P/V_S profile obtained is close to two in all the layers with high uncertainties as the forward problem is rather insensitive to V_P. This value is however consistent with ratios obtained from V_P and V_S logs available for a nearby Storengy gas storage facility (Céré la Ronde) with a similar sedimentary layering structure. In the following, a constant value of $V_P/V_S = 2$ is thus assumed. A subset of the best models, the mean model and the 68 per cent confidence interval are shown in Fig. 11(b). The mean V_S model is consistent with the sonic log available from the CS07 well (drilled in 1967 by Gaz de France, see well location in Fig. 1a) between 230 and 1190 m (V_P converted to

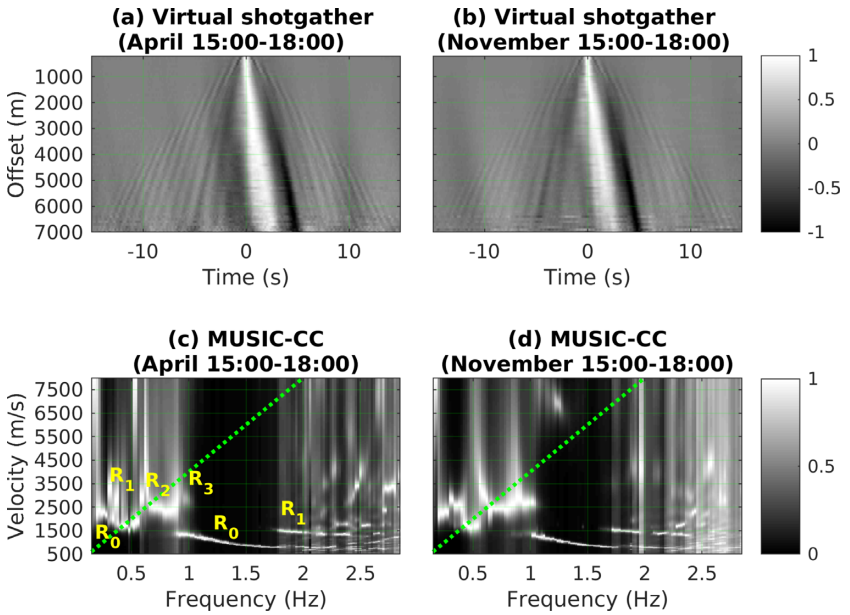


Figure 8. Day-time dispersion study in April (left column) and November (right column). (a and b) Virtual shot gathers from cross-correlation; (c and d) MUSIC-CC (MUSIC processing of the shot gather). The green lines mark the theoretical resolution limit of the virtual shot gather.

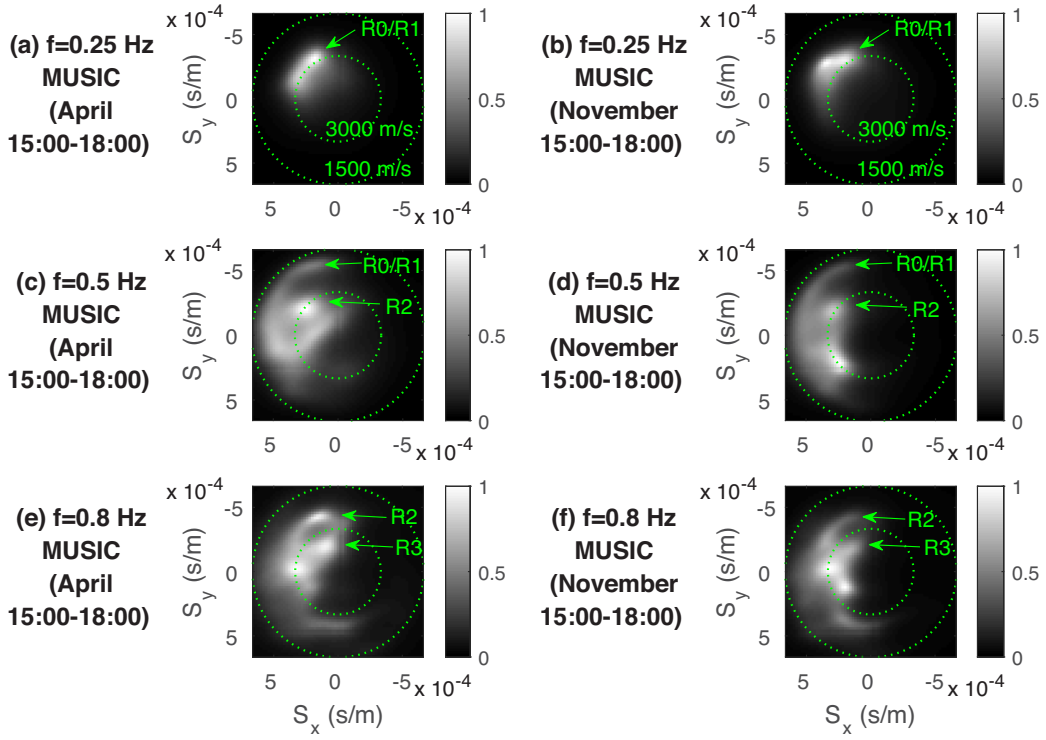


Figure 9. Same as Fig. 6, but for day time and below 1 Hz.

V_S assuming $V_P/V_S = 2$). The results of the inversion indicate the presence of a quick layer between 550 and 800 m. The basement is found at approximately 4 km depth. Since our result is an averaged measure over the array area, it can be only qualitatively compared to individual log data. The quick layer probably corresponds to the thick limestone level visible on Fig. 1(b) between 450 and 800 m. On the other hand, the basement depth below 4000 m is consistent with the fact it was not seen in the CS01 log, as the well drilling stopped at 2623 m. Besides, the theoretical dispersion curves for the first five modes estimated for the best model are shown with plain lines in Fig. 11(a). The modes R0, R1 at high frequency and R2 exhibit a good fit with the picked dispersion curves. R1 at low frequency and R3 are also compatible with the dispersion plot, although they were not used for the inversion.

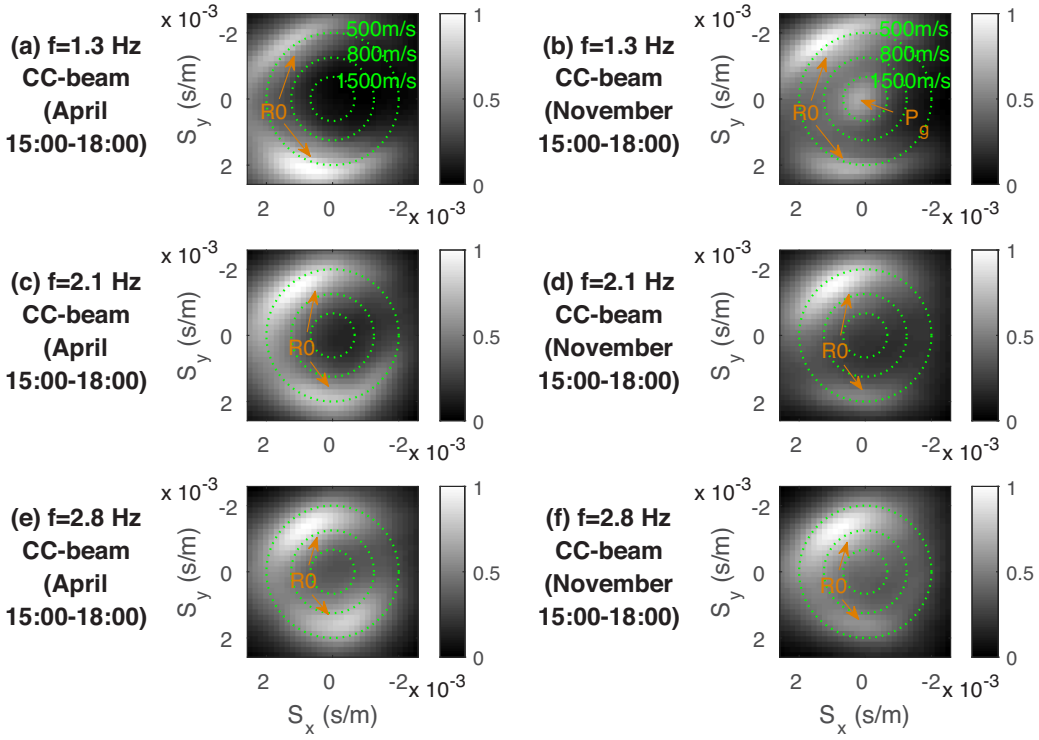


Figure 10. Same as Fig. 7, but for day time. Only the modes detected in the beamformer are labelled.

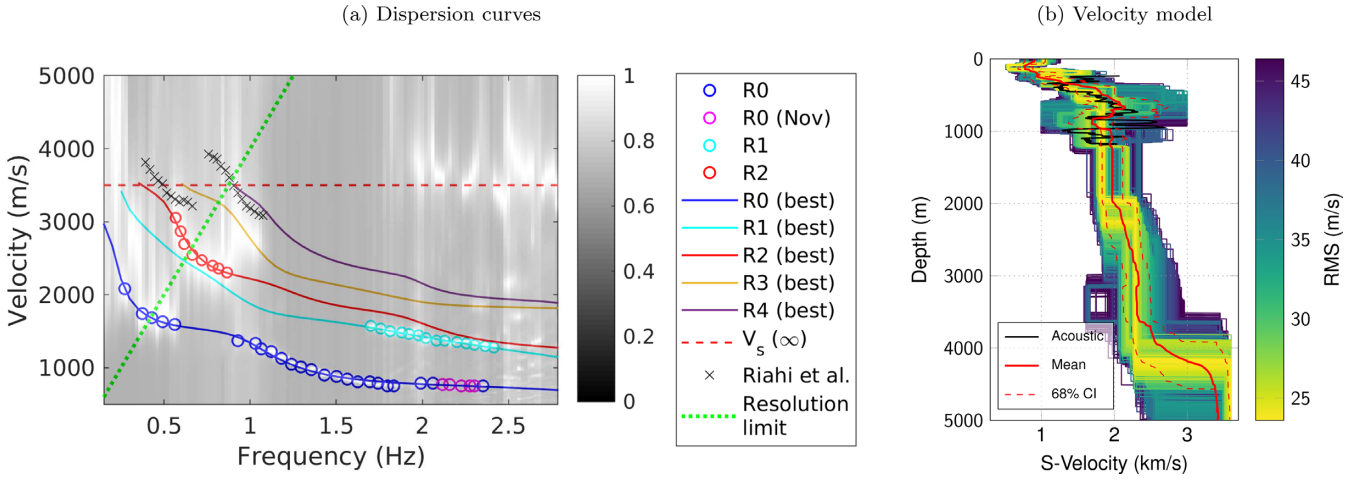


Figure 11. Dispersion curves picking and inversion. (a) Dispersion curves overlaid onto the MUSIC-CC dispersion plot for April (1–4 a.m.); circles : picked curves; plain pines : forward modelled dispersion curves for the best model; dashed line : high-frequency velocity limit for the L_g phase, equal to V_S in the crust, considered as an infinite half-space during the inversion; crosses : dispersion curves from Riahi *et al.* (2013a). (b) Inverted mean velocity profile (red) along with the acoustic log provided by Storengy (black). The velocity models sampled by the different runs are represented in the background with the colourscale indicating their RMS value. The red dashed lines delimit the 68 per cent confidence interval.

5 DISCUSSION

5.1 Ambient noise sources

5.1.1 Anthropogenic or natural?

Below 1 Hz both modal composition and backazimuth distribution remained stable between day and night periods (Figs 6 and 9). This suggests an exclusively natural origin for these waves with source zones located in the Atlantic. Above 1 Hz, though natural sources can show diurnal variations, ambient noise amplitude often follows the human activity cycles, usually weaker at night, on week-ends and on holidays (e.g. Lehujeur *et al.* 2015). Such cycles can be spotted in Fig. 2: above 1 Hz the power spectral density is clearly lower during the night, at midday and on week-ends (e.g. Fig. 2b, see Sunday, 2010 November 7). At these frequencies, results from Section 4.3 showed a strong variation of the noise content between day and night in terms of both modal content and back-arriving azimuths. The Rayleigh wave fundamental mode originates from the south during night time (Fig. 7). During day time, its backazimuth distribution clusters around NNW and S directions (Fig. 10). The day-time backazimuth distribution might be due to the noise

from the cities of Blois and Tours for the northwestern directions, and the A85 motorway close to the southern end of the seismic network. Interpreting the night-time distribution is more ambiguous. Rayleigh waves originating from the south at frequencies near 1 Hz were also observed by Lehujeur *et al.* (2017a) near Strasbourg (Eastern France), which they interpreted as microseisms arriving from the Mediterranean sea. However, we cannot exclude the existence of industrial facilities with persistent night-time activity or a night-time traffic on the A85.

As P_g and L_g phases are not visible during the day, we can infer they are not linked to anthropogenic activity. Taking into account the backazimuths of the two quick phases (Figs 6 g–j), the strongest source zones (from N to SW) are located within the Atlantic Ocean direction. The beamformer for the L_g phase is more wide and diffuse than for P_g, suggesting other generation zones (such as the Scandinavian Northern coastline) and/or strong scattering.

It is interesting to note that the observations made by Riahi *et al.* (2013b), who analysed differences between day and night time cross-correlations on the Jonas field in the USA, could also be interpreted in terms of alternating natural and anthropogenic noise domination.

5.1.2 Excitation of L_g and higher modes

So far, we are in line with the conclusions of Koper *et al.* (2010), according to which L_g is of natural (oceanic) origin, based on backazimuth analysis from several arrays around the world. The seismic wave generation mechanism via nonlinear ocean waves interactions (Longuet-Higgins 1950) is indeed valid beyond the traditional secondary microseism band. For instance, the ambient noise amplitude above 1 Hz was well explained in terms of oceanic short-wave interaction for some mid-ocean islands (Gimbert & Tsai 2015). However, a major issue is to know if the L_g phase is likely to be excited by such sources, knowing that this wave cannot propagate inside the oceanic crust (Zhang & Lay 1995). This is because of the nature of the L_g phase, which is a superposition of high-order overtones of both Rayleigh and Love waves (Kennett 1986). These overtones are due to the crust/mantle interface (Moho) and their number increases with the thickness of the crust. The too thin oceanic crust does not offer enough available overtones for L_g to develop. We used numerical simulation in order to check whether it is realistic to observe an L_g phase arrival on the continent from oceanic sources. Simple elastic 2-D spectral-element simulations of the wave propagation from the ocean towards the continent across a passive margin representative of the French Atlantic coast were performed. We used exactly the same profile of the margin as Gualtieri *et al.* (2015). Compared to their work, we added the mantle below the crust, in order to enable the existence of the L_g phase. The approximate mantle depth was taken from Artemieva & Thybo (2013). The model is sketched in Fig. 12(a). In terms of seismic velocities, no difference was made between continental and oceanic crust. Sedimentary layers were added below the array in order to mimic the Paris basin. The maximum basin depth was of 3 km, which is shallower than the results from the dispersion curves inversion, but more representative of the Paris Basin. The wavefield generated at the array by deep- or shallow-water natural oceanic sources (respectively called DWS and SWS), modelled separately, was mixed with the one generated by local surface sources (LSS), reproducing human activity. For each source type, 500 vertical point-force sources were used, with dominant frequency, emission time and position randomly chosen within defined ranges. Details about seismic source implementation and wavefield mixture are given in Appendix D.

The resulting dispersion plots for DWS and SWS are shown in Figs 12(c) and (d), respectively. The theoretical dispersion curves for the first 30 modes are overlaid. They were estimated using *Computer Programs for Seismology* software (Herrmann 2013) for the profile 2 in Fig. 12(a). The dispersion plot for DWS exhibits patches of energy into the L_g phase band, where the higher modes are concentrated. Such patches are present up to 3 Hz. At some frequencies, individual higher modes are excited (R1/R2 below 1 Hz, R10/R11 between 2 and 2.6 Hz). There is a gap into the R0 branch between 0.5 and 0.7 Hz, where higher modes dominate. A similar pattern is observed for the real data (see Fig. 5). On the other hand, the dispersion plot for SWS exhibits a continuous R0 excitation over the entire frequency range. R1 is excited between 0.8 and 1.4 Hz, and between 1.7 and 2.4 Hz. No higher modes above R1 are excited. From these observations, we can confirm the possible existence of the modes detected into the real data set, except for P_g. While the first two modes R0 and R1 can be excited by any type of sources (local and oceanic), the L_g phase and higher modes are specific to distant DWS. From the wavefield snapshots not shown here, we saw that the L_g phase excitation took place as the wavefield reached the continental margin. Before reaching the margin, the wavefield propagates as a superposition of several individual modes specific to an oceanic environment with a water layer on its top, as described by Gualtieri *et al.* (2015). On the contrary, coastal shallow water merely excites the fundamental mode. Such different excitation could probably be explained through modal summation, as done by Gualtieri *et al.* (2015), but taking the mantle into account. This, however, is out of the scope of our work.

P_g was not observed in the numerical simulations. This can be due to the fact we neglected viscous attenuation, which is higher in shallow layers, and thus tends to enhance phases propagating in depth, as does P_g. Unfortunately, the CPU cost of a large scale visco-elastic simulation with high-frequency content was prohibitive. Hence, we are unable to discriminate whether P_g is as a signature of DWS or SWS. Another explanation is the simple homogeneous model used to represent the crust: introducing lower velocities layers could help guide P_g waves near the surface. Besides, the L_g phase can be spotted down to 0.5 Hz in the simulations, while it disappears below 2 Hz in the data (Figs 5 e and f). This might be due to a weaker modal density at low frequencies (Fig. 12b), down to some threshold frequency beneath which the weakening L_g phase gets overshaded by stronger phases.

An interesting development would be to find a robust way to determine the true amplitudes of the P_g and L_g phases, and to look for correlations with particular source zones, provided an ocean wave model for short waves analogous to WAVEWATCH III (Tolman 1991), as done for example by Essen *et al.* (2003). However, reliably measuring of the nonlinear interaction term for short waves remains a challenge (Peureux & Ardhuin 2016).

5.2 Methodology and inversion

5.2.1 Backazimuth retrieval

Compared to Riahi *et al.* (2013a), we were able to extend the frequency band of the wavefield analysis towards higher frequencies by using incoherent beamforming with the envelope of cross-correlations. For frequencies 0.2, 0.54 and 0.81 Hz, where the beamformers were visualized in their work, our backazimuth distribution

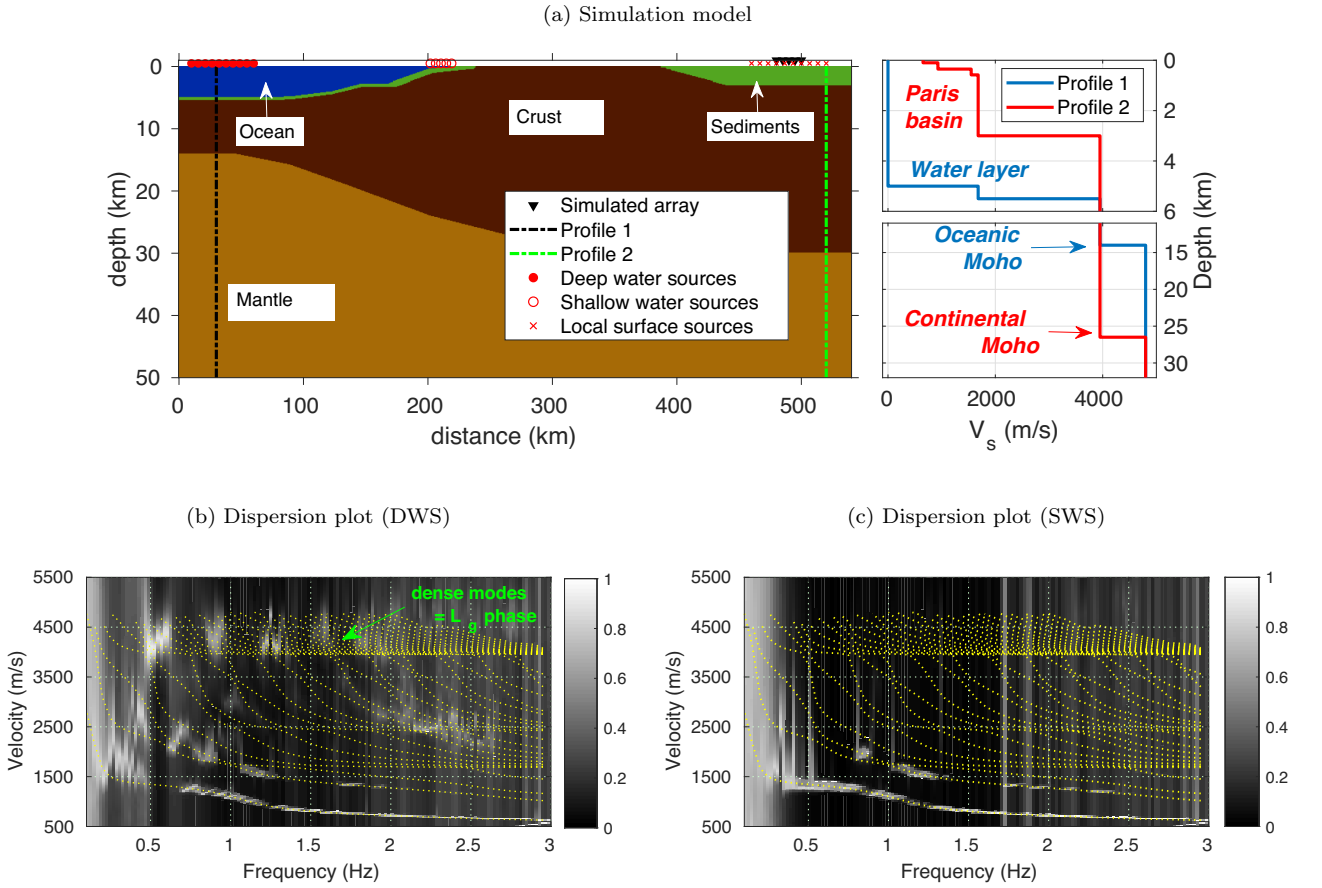


Figure 12. Spectral-element simulation results. (a) 2-D Model used for simulations. Red filled and unfilled dots, respectively, represent deep- and shallow-water oceanic sources. Red crosses represent local anthropogenic sources. Black and red dash-dotted lines indicate the oceanic and continental profiles for which medium properties are presented in Table D1. Black triangles show the location of the simulated seismic array. (b) Dispersion plot obtained by applying MUSIC on 30 independent realizations of the wavefield with local and deep-water sources. (c) Same as (b), but with local and shallow-water sources.

is similar to theirs (NW below 0.3 Hz and wide range NNW–S at 0.5 and 0.8 Hz, see our Fig. 6 and their fig. 6). Nevertheless, the beamformers we derive at 0.5 and 0.8 Hz (Figs 6 e–h) distinguish between two modes at each frequency, while their beamformer at 0.54 Hz displays a single patch for Rayleigh waves. At 0.81 Hz, their dominant identification matches what we identify as R3 with similar backazimuths. Another small patch of prograde Rayleigh waves, which is visible in their fig. 6 at $f = 0.81$ Hz, is compatible with our mode R2, but with limited backazimuth match, as they only find a southerly direction, while we have a wider backazimuth distribution.

An intrinsic limitation arises in the statistical analysis carried out by Riahi *et al.* (2013a) or Koper *et al.* (2010), where only one maximum is picked from the beamformer per time window. If the wavefield is steady enough over time, the dominant phases will almost always be picked and the histogram of detections (e.g. fig. 6 in Riahi *et al.* 2013a) will systematically miss weaker phases. This might be the reason why we observe more modes compared to Riahi *et al.* (2013a). On the other hand, we were not able to separate different polarization states with our one-component method, contrary to their three-component beamforming approach. Other non-traditional array methods, such as the CLEAN-PSF algorithm (e.g. Gal *et al.* 2016), remove iteratively the contribution of main seismic phases from the beamformer, ie their point spread function (PSF), which enables an efficient recovery of weaker phases.

5.2.2 *Dispersion curves retrieval*

We implemented dispersion curves retrieval both directly from the recorded waveforms (MUSIC, Figs 5 g and h) and from the cross-correlations (FK-CC: Figs 5 c and d/MUSIC-CC: Figs 5 e and f). We found that MUSIC-CC was the most appropriate method, since it works in both subresolution and aliasing zones. This is in keeping with the conclusions drawn in the synthetic study by Gouédard *et al.* (2008): using cross-correlations enables to widen the aliasing-free zone, while high-resolution methods (HRFK was used in Gouédard *et al.* 2008) yield better results beyond the resolution limit. The benefit was crucial for the inversion. Indeed, only MUSIC-CC was able to resolve R0 and R1 below 0.5 Hz. Other methods exhibited a wide patch including both modes. A natural identification of this patch as R0 would yield a model with an erroneous (too shallow) basement position. The implementation of a stable automated detection of the signal-subspace dimension for the MUSIC algorithm greatly improved the results, as shown in Appendix B.

Dispersion curves from Riahi *et al.* (2013a) were also reported in Fig. 11(a, black crosses). Their identification by the authors as higher modes is confirmed by our results. Still, the phase velocities picked by Riahi *et al.* (2013a) exceed the L_g phase high-frequency velocity limit (3500 m s^{-1}). One possible explanation is that the higher modes are influenced by the mantle, as suggested by the theoretical dispersion curves in Fig. 12(b). Otherwise, velocities can be biased as they lie beyond the theoretical resolution limit

for standard beamforming, or close to it. Velocity overestimation into the subresolution zone was also highlighted by Gouédard *et al.* (2008).

Finally, it should be pointed out that the dispersion curves from Riahi *et al.* (2013a) are representative of the isotropic part of the wavefield, which can explain some discrepancy between their higher-mode dispersion curves and ours. As they found a non-negligible level of anisotropy (up to 10 per cent), it might be useful to introduce a correction for anisotropy in our methodology. This however is beyond the scope of this paper.

5.3 Benefits of wavefield characterization

For surface-wave based methods, the wavefield characterization yields a major indication on the mode labels to be used according to the location of the pointed dispersion curve into the f–v_ϕ plane. For regionalized inversion methods based on ambient noise cross-correlation, it would allow to choose reliable station-pairs aligned with the major source direction. The knowledge of the L_g phase velocity at relatively high frequencies (above 2 Hz) yields a useful constraint on V_S into the infinite half-space during the inversion. For body-wave methods, wavefield characterization would give an indication about the frequency bands where they are most likely to be found (1–1.5 Hz for the present data set). Whatever the targeted wave-type, the knowledge of the wavefield composition evolution over time is useful for selecting the appropriate time windows to be processed. For example, using day-time records would be preferable for Rayleigh wave dispersion curves extraction above 1 Hz, both in terms of strength and source distribution, while night-time records would allow to extract body waves between 1 and 1.5 Hz. Finally, for methods with unclear wave-type assumptions, such a characterization of the incident wavefield should guide the numerical simulation. For example, Lambert *et al.* (2013) modelled the ambient wavefield as a superposition of randomly excited deep and surface sources, in order to investigate the nature of the hydrocarbon-related amplitude anomalies. Such a setup completely misses the L_g phase, which however seems to carry a significant amount of energy at the frequencies of interest (several Hz). A model similar to the one shown in Fig. 12(a) would better approach the reality, though at a higher computational cost.

6 CONCLUSION

We suggest to combine methods as following for a wide-band one-component ambient wavefield characterization with a dense array:

(1) Backazimuth retrieval: 'direct' MUSIC in the subresolution and working zones of the array and non-coherent CC-beam in the aliasing zone.

(2) Dispersion curves retrieval and mode labelling: MUSIC applied to a common-offset gather of the interstation cross-correlations (MUSIC-CC).

This methodology was applied to the vertical component recordings acquired above a UGS in Central France by a 3 × 6 km array with about 100 sensors. Between 0.1 and 3 Hz, four Rayleigh wave modes (R0, R1, R2 and R3) were identified, as well as L_g and P_g phases. The comparison of day-time and night-time recordings made it possible to distinguish between anthropogenic seismic sources (mainly R0 and R1 above 1 Hz) and natural seismic sources (higher Rayleigh modes, P_g and L_g) in deep ocean and along the Atlantic coast. Numerical simulations were carried out to confirm this interpretation. Eventually, surface wave dispersion curves were inverted to yield a V_S profile consistent with the available sonic log, and compatible with the presence of a Permian basin in this area of the Paris Basin . Given the growing interest of the scientific and industrial communities in seismic noise, our work opens interesting perspectives. We developed a robust technique to analyse data recorded by dense arrays which are ever more used in the industry. A direct application are 1-D V_S inversions, as our method helps to correctly identify and label different surface wave modes. Furthermore, having a precise knowledge of the incoming wavefield is of prior importance when studying possible links between hydrocarbon reservoirs and amplitude features at surface.

ACKNOWLEDGEMENTS

The authors are grateful to Jérôme Vergne, Maximilien Lehujeur, Nikolaï Shapiro, Eleonore Stutzmann and Fabrice Ardhuin for helpful discussions, as well as to Catherine Formento and Simon Lejart (Storengy) for their help with the well logs. They are also grateful to Storengy, an Engie company, for funding this research, and to the reviewers for their comments and advice.

REFERENCES

Aki, K., 1957. Space and time spectra of stationary stochastic waves, with special reference to microtremors, *Bull. Earthq. Res. Inst. Univ. Tokyo,* **35,** 415–456.

Ardhuin, F., Stutzmann, E., Schimmel, M. & Mangeney, A., 2011. Ocean wave sources of seismic noise, *J. geophys. Res.,* **116**(C9),

Artemieva, I.M. & Thybo, H., 2013. A seismic model for Moho and crustal structure in Europe, Greenland, and the North Atlantic region, *Tectonophysics,* **609,** 97–153.

Bensen, G.D., Ritzwoller, M.H., Barmin, M.P., Levshin, A.L., Lin, F., Moschetti, M.P., Shapiro, N.M. & Yang, Y., 2007. Processing seismic ambient noise data to obtain reliable broad-band surface wave dispersion measurements, *Geophys. J. Int.,* **169**(3), 1239–1260.

Bokelmann, G.H. & Baisch, S., 1999. Nature of narrow-band signals at 2.083 Hz, *Bull. seism. Soc. Am.,* **89**(1), 156–164.

Bonnefoy-Claudet, S., Cotton, F. & Bard, P.-Y., 2006a. The nature of noise wavefield and its applications for site effects studies, *Earth-Sci. Rev.,* **79**(3-4), 205–227.

Bonnefoy-Claudet, S., Cornou, C., Bard, P.-Y., Cotton, F., Moczo, P., Kristek, J. & Fäh, D., 2006b. H/V ratio: a tool for site effects evaluation. Results from 1-D noise simulations, *Geophys. J. Int.,* **167**(2), 827–837.

Boschi, L., Weemstra, C., Verbeke, J., Ekstrom, G., Zunino, A. & Giardini, D., 2013. On measuring surface wave phase velocity from station–station cross-correlation of ambient signal, *Geophys. J. Int.,* **192**(1), 346–358.

Brenguier, F., Shapiro, N.M., Campillo, M., Nercessian, A. & Ferrazzini, V., 2007. 3-D surface wave tomography of the Piton de la Fournaise volcano using seismic noise correlations, *Geophys. Res. Lett.,* **34**(2).

Capon, J., 1969. High-resolution frequency-wavenumber spectrum analysis, *Proc. IEEE,* **57**(8), 1408–1418.

Capon, J., Greenfield, R. & Kolker, R., 1967. Multidimensional maximum-likelihood processing of a large aperture seismic array, *Proc. IEEE,* **55**(2), 192–211.

Cornou, C., 2002. *Traitement d'antenne et imagerie sismique dans l'agglomération grenobloise (Alpes françaises) : implications pour les effets de site,* PhD thesis, Laboratoire de Géophysique Interne et Tectonophysique, Grenoble, France.

Dangel, S., Schaepman, M., Stoll, E., Carniel, R., Barzandji, O., Rode, E.-D. & Singer, J., 2003. Phenomenology of tremor-like signals observed over hydrocarbon reservoirs, *J. Volc. Geotherm. Res.,* **128**(1–3), 135–158.

Douze, E.J., 1964. Rayleigh waves in short-period seismic noise, *Bull. seism. Soc. Am.,* **54**(4), 1197–1212.

Douze, E.J., 1967. Short-period seismic noise, *Bull. seism. Soc. Am.,* **57**(1), 55–81.

Draganov, D., Wapenaar, K., Mulder, W., Singer, J. & Verdel, A., 2007. Retrieval of reflections from seismic background-noise measurements, *Geophys. Res. Lett.,* **34**(4).

Essen, H.-H., Krüger, F., Dahm, T. & Grevemeyer, I., 2003. On the generation of secondary microseisms observed in northern and central Europe, *J. geophys. Res.,* **108**(B10).

Foti, S., Parolai, S., Albarello, D. & Picozzi, M., 2011. Application of surface-wave methods for seismic site characterization, *Surv. Geophys.,* **32**(6), 777–825.

Gal, M., Reading, A., Ellingsen, S., Koper, K., Burlacu, R. & Gibbons, S., 2016. Deconvolution enhanced direction of arrival estimation using one- and three-component seismic arrays applied to ocean induced microseisms, *Geophys. J. Int.,* **206**(1), 345–359.

Gély, J.-P. & Hanot, F., 2014. *Le Bassin parisien: Un nouveau regard sur la géologie,* AGBP.

Geuzaine, C. & Remacle, J.-F., 2009. Gmsh: a 3-D finite element mesh generator with built-in pre-and post-processing facilities, *Int. J. Numer. Methods Eng.,* **79**(11), 1309–1331.

Gimbert, F. & Tsai, V.C., 2015. Predicting short-period, wind-wave-generated seismic noise in coastal regions, *Earth planet. Sci. Lett.,* **426,** 280–292.

Goldstein, P. & Archuleta, R.J., 1987. Array analysis of seismic signals, *Geophys. Res. Lett.,* **14**(1), 13–16.

Gouédard, P., Cornou, C. & Roux, P., 2008. Phase-velocity dispersion curves and small-scale geophysics using noise correlation slantstack technique, *Geophys. J. Int.,* **172**(3), 971–981.

Green, A.G. & Greenhalgh, S., 2010. Comment on 'Low-frequency microtremor anomalies at an oil and gas field in Voitsdorf, Austria' by Marc-André Lambert, Stefan Schmalholz, Erik H. Saenger and Brian Steiner, *Geophys. Prospect.* **57,** 393–411, Geophys. Prospect., 58(2), 335–339.

Gualtieri, L., Stutzmann, E., Capdeville, Y., Farra, V., Mangeney, A. & Morelli, A., 2015. On the shaping factors of the secondary microseismic wavefield, *J. geophys. Res.,* **120**(9), 6241–6262.

Haskell, N.A., 1953. The dispersion of surface waves on multilayered media, *Bull. seism. Soc. Am.,* **43**(1), 17–34.

Herrmann, R.B., 2013. Computer programs in seismology: an evolving tool for instruction and research, *Seismol. Res. Lett.,* **84**(6), 1081–1088.

Kennett, B.L.N., 1986. L_g waves and structural boundaries, *Bull. seism. Soc. Am.,* **76**(4), 1133–1141.

Komatitsch, D., Vilotte, J.-P., Vai, R., Castillo-Covarrubias, J.M. & Sánchez-Sesma, F.J., 1999. The spectral element method for elastic wave equations—application to 2-D and 3-D seismic problems, *Int. J. Numer. Methods Eng.,* **45**(9), 1139–1164.

Koper, K.D., Seats, K. & Benz, H., 2010. On the composition of Earth's short-period seismic noise field, *Bull. seism. Soc. Am.,* **100**(2), 606–617.

Krim, H. & Viberg, M., 1996. Two decades of array signal processing research: the parametric approach, *IEEE Signal Process. Mag.,* **13**(4), 67–94.

Lambert, M.-A., Schmalholz, S.M., Saenger, E.H. & Steiner, B., 2009. Low-frequency microtremor anomalies at an oil and gas field in Voitsdorf, Austria, *Geophys. Prospect.,* **57**(3), 393–411.

Lambert, M.-A., Saenger, E., Quintal, B. & Schmalholz, S., 2013. Numerical simulation of ambient seismic wavefield modification caused by pore-fluid effects in an oil reservoir, *Geophysics,* **78**(1), T41–T52.

Landés, M., Hubans, F., Shapiro, N.M., Paul, A. & Campillo, M., 2010. Origin of deep ocean microseisms by using teleseismic body waves, *J. geophys. Res.,* **115**(B5). .

Larose, E. *et al.*, 2015. Environmental seismology: what can we learn on Earth surface processes with ambient noise? *J. appl. Geophys.,* **116,** 62–74.

Lehujeur, M., Vergne, J., Schmittbuhl, J. & Maggi, A., 2015. Characterization of ambient seismic noise near a deep geothermal reservoir and implications for interferometric methods: a case study in northern Alsace, France, *Geotherm. Energy,* **3**(1).

Lehujeur, M., Vergne, J., Maggi, A. & Schmittbuhl, J., 2017a. Ambient noise tomography with non-uniform noise sources and low aperture networks: case study of deep geothermal reservoirs in northern Alsace, France, *Geophys. J. Int.,* **208**(1), 193–210.

Lehujeur, M., Vergne, J., Maggi, A. & Schmittbuhl, J., 2017b. Vertical seismic profiling using double beamforming processing of non-uniform anthropogenic seismic noise: the case study of Rittershoffen, Upper Rhine Graben, France, *Geophysics,* **82**(6), B209–B217.

Longuet-Higgins, M.S., 1950. A theory of the origin of microseisms, *Phil. Trans. R. Soc. Lond. A,* **243**(857), 1–35.

Luu, K., Noble, M., Gesret, A., Belayouni, N. & Roux, P.-F., 2018. A parallel competitive Particle Swarm Optimization for non-linear first arrival traveltime tomography and uncertainty quantification, *Comput. Geosci.,* **113,** 81–93.

Mordret, A., Landes, M., Shapiro, N.M., Singh, S.C., Roux, P. & Barkved, O.I., 2013. Near-surface study at the Valhall oil field from ambient noise surface wave tomography, *Geophys. J. Int.,* **193**(3), 1627–1643.

Nakamura, Y., 1989. A method for dynamic characteristics estimation of subsurface using microtremor on the ground surface, *Q. Rep. RTRI,* **30,** 25–33.

Nakata, N., Chang, J.P., Lawrence, J.F. & Boué, P., 2015. Body wave extraction and tomography at Long Beach, California, with ambient-noise interferometry, *J. geophys. Res.,* **120**(2), 1159–1173.

Obrebski, M., Ardhuin, F., Stutzmann, E. & Schimmel, M., 2013. Detection of microseismic compressional (*P*) body waves aided by numerical modeling of oceanic noise sources, *J. geophys. Res.,* **118**(8), 4312–4324.

Peureux, C. & Ardhuin, F., 2016. Ocean bottom pressure records from the Cascadia array and short surface gravity waves., *J. geophys. Res.,* **121**(5), 2862–2873.

Poli, P., Pedersen, H.A. & Campillo, M., 2012. Emergence of body waves from cross-correlation of short period seismic noise, *Geophys. J. Int.,* **188**(2), 549–558.

Pratt, M.J., Wiens, D.A., Winberry, J.P., Anandakrishnan, S. & Euler, G.G., 2017. Implications of sea ice on Southern Ocean microseisms detected by a seismic array in West Antarctica, *Geophys. J. Int.,* **209**(1), 492–507.

Riahi, N., Bokelmann, G., Sala, P. & Saenger, E.H., 2013a. Time-lapse analysis of ambient surface wave anisotropy: a three-component array study above an underground gas storage, *J. geophys. Res.,* **118**(10), 5339–5351.

Riahi, N., Goertz, A., Birkelo, B. & Saenger, E.H., 2013b. A statistical strategy for ambient seismic wavefield analysis: investigating correlations to a hydrocarbon reservoir, *Geophys. J. Int.,* **192**(1), 148–162.

Rost, S. & Thomas, C., 2002. Array seismology: methods and applications, *Rev. Geophys.,* **40**(3), 2–1-2-27.

Roux, P., Sabra, K.G., Gerstoft, P., Kuperman, W.A. & Fehler, M.C., 2005. P-waves from cross-correlation of seismic noise, *Geophys. Res. Lett.,* **32**(19), L19303.

Ruigrok, E., Campman, X. & Wapenaar, K., 2011. Extraction of P-wave reflections from microseisms, *C. R. Geosci.,* **343**(8), 512–525.

Ruigrok, E., Gibbons, S. & Wapenaar, K., 2017. Cross-correlation beamforming, *J. Seismol.,* **21**(3), 495–508.

Schmidt, R.O., 1986. Multiple emitter location and signal parameter estimation, *IEEE Trans. Antennas Propag.,* **AP-34**(3), 276–280.

Sens-Schönfelder, C. & Wegler, U., 2006. Passive image interferometry and seasonal variations of seismic velocities at Merapi Volcano, Indonesia, *Geophys. Res. Lett.,* **33**(21), L21302.

Shan, T.-J., Wax, M. & Kailath, T., 1985. On spatial smoothing for direction-of-arrival estimation of coherent signals, *IEEE Trans. Acoust, Speech Signal Process.,* **33**(4), 806–811.

Shapiro, N.M. & Campillo, M., 2004. Emergence of broadband Rayleigh waves from correlations of the ambient seismic noise: correlations of the seismic noise, *Geophys. Res. Lett.,* **31**(7),

Stutzmann, E., Ardhuin, F., Schimmel, M., Mangeney, A. & Patau, G., 2012. Modelling long-term seismic noise in various environments, *Geophys. J. Int.,* **191**(2), 707–722.

Thomson, W.T., 1950. Transmission of elastic waves through a stratified solid medium, *J. Appl. Phys.,* **21**(2), 89–93.

Tolman, H.L., 1991. A third-generation model for wind waves on slowly varying, unsteady, and inhomogeneous depths and currents, *J. Phys. Oceanogr.,* **21**(6), 782–797.

Zhang, T.-R. & Lay, T., 1995. Why the L_g phase does not traverse oceanic crust, *Bull. seism. Soc. Am.*, **85**(6), 1665–1678.

APPENDIX A: GROUP AND PHASE VELOCITY

The consistency between group and phase velocities measured respectively by non-coherent CC-beam and MUSIC-CC is investigated. Knowing the phase velocity v_ϕ, the group velocity v_g can be computed as

$$v_g = v_\phi \left(1 - \frac{\omega}{v_\phi}\frac{\partial v_\phi}{\partial \omega}\right)^{-1}. \tag{A1}$$

Relation (A1) is applied to the smoothed interpolated dispersion curve of the R0 mode (blue line in Fig. A1) picked from the MUSIC-CC dispersion plot (see Fig. 11a). The result is again smoothed and plotted in Fig. A1 (red line). This is the expected group velocity from the phase velocity measurements. On the other hand, CC-beam was applied to April's night time cross-correlations between 0.7 and 2.7 Hz, with a step of 0.2 Hz. For each obtained beamformer, a Gaussian was fitted to the azimuthal section crossing the beamformer's maximum. This yields an estimate of the group velocity (Gaussian's mean) and the associated uncertainty (Gaussian's standard deviation) at each considered frequency. Results were reported on Fig. A1 (red circles with uncertainty bars). Group and phase velocity measurements are clearly consistent with each other.

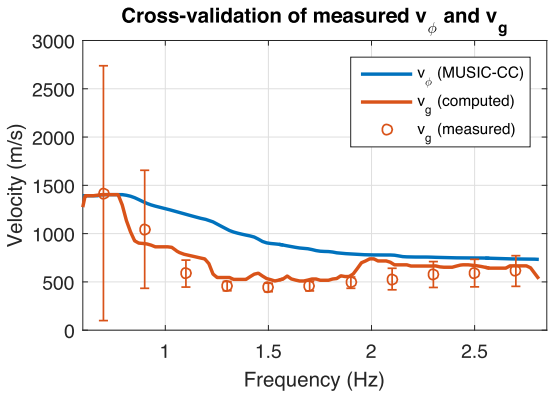


Figure A1. Phase and group velocity consistency check. Blue curve: interpolated and smoothed phase velocity dispersion curve for the fundamental mode, picked from the MUSIC-CC dispersion plot for April (night time). Red curve: smoothed theoretical group velocity dispersion curve computed from the blue curve using eq. (A1). Red circles with error bars: group velocities measured by CC-beam (April, night time), with associated uncertainties.

The same velocity conversion procedure was applied for the R1 mode in order to place the arrows in Fig. 7 at the expected group velocity for this mode. However, Fig. A1 confirms that only the fundamental mode (R0) is detected by the CC-beam approach.

APPENDIX B: MUSIC AND AUTOMATED SIGNAL-SUBSPACE DETERMINATION

In this section, we present the derivation of the MUSIC algorithm and a method to determine the dimension of the signal subspace.

We drop the frequency dependence from the notations, simply writing R instead of $R(f)$ for the CSM introduced in eq. (5). At a given frequency, if a wavefield containing Q uncorrelated plane waves of amplitudes A_q is recorded by N stations affected by white noise of intensity η (e.g. instrumental noise), the CSM can be written as

$$R_{ij} = \sum_{q=1}^{Q} |A_q|^2 e^{i\boldsymbol{k_q}\cdot(\boldsymbol{x_i}-\boldsymbol{x_j})} + \eta^2 \delta_{ij}, \tag{B1}$$

where δ is the Kronecker symbol. Assuming $N > Q$, R can be diagonalized as following:

$$R = E_s \Lambda_s E_s^\dagger + E_n \Lambda_n E_n^\dagger, \tag{B2}$$

with Λ_s being a $Q \times Q$ diagonal matrix containing the Q biggest eigenvalues, and Λ_n an $(N-Q)\times(N-Q)$ diagonal matrix containing the $(N-Q)$ remaining eigenvalues. Using simple linear algebra, Schmidt (1986) showed that $\Lambda_n = \eta^2 I$. The orthonormal eigenvectors e_l which form the noise space basis (columns of E_n) then verify

$$a(\boldsymbol{k_1})^\dagger e_l = \ldots = a(\boldsymbol{k_Q})^\dagger e_l = 0, \tag{B3}$$

with

$$a(\boldsymbol{k}) = \frac{1}{\sqrt{N}} \begin{pmatrix} e^{i\boldsymbol{k}\cdot\boldsymbol{x_i}} \\ \vdots \\ e^{i\boldsymbol{k}\cdot\boldsymbol{x_N}} \end{pmatrix}. \tag{B4}$$

That is, the vector $a(\boldsymbol{k_q})$ is orthogonal to the noise subspace generated by the columns of E_n if $\boldsymbol{k_q}$ corresponds to a plane wave which actually propagates across the array. The principle of the MUSIC algorithm is to seek such optimal $\boldsymbol{k}$ vectors by minimizing the projection of $a(\boldsymbol{k})$ on the noise subspace, and thus maximizing the so-called MUSIC *functional*

$$D_M(\boldsymbol{k}) = \frac{1}{\left|\sum_{l=1}^{N-Q} a(\boldsymbol{k})^\dagger e_l\right|^2} = \frac{1}{a(\boldsymbol{k})^\dagger E_n E_n^\dagger a(\boldsymbol{k})}. \tag{B5}$$

In comparison, the classical FK method (or standard frequency domain beamforming) differs insofar as it seeks the vector $\boldsymbol{k}$ maximizing

$$D_{FK}(\boldsymbol{k}) = a(\boldsymbol{k})^\dagger R a(\boldsymbol{k}), \tag{B6}$$

without performing any diagonalization.

Automated ways to choose the right dimension of the signal subspace for MUSIC, that is, Q, use the eigenvalues profile (Cornou 2002). One possible approach is to look for the slope break into the logarithmic decay of the eigenvalues, while another one relies on comparing the slope of the logarithmic decay to the one obtained for random noise. Both approaches require threshold parameters that depend on the SNR. Since the latter is unknown for the real ambient noise data, we developed a slightly different method which allows to address frequency ranges with high and low SNRs at once. A criterion on the eigenvalue ratio $R_i = \log \frac{\lambda_i}{\lambda_1}$, with λ_1 being the major eigenvalue, is introduced:

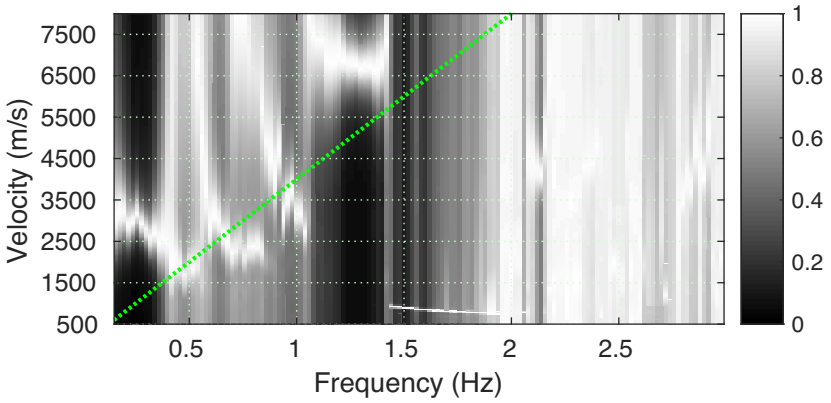
$$\begin{cases} C_R(i) = 1 \text{if} |R_i| \le n_R \\ C_R(i) = 0 \text{if} |R_i| > n_R, \end{cases} \tag{B7}$$

where n_R is a threshold parameter to be optimized. The threshold eigenvalue number is the defined as

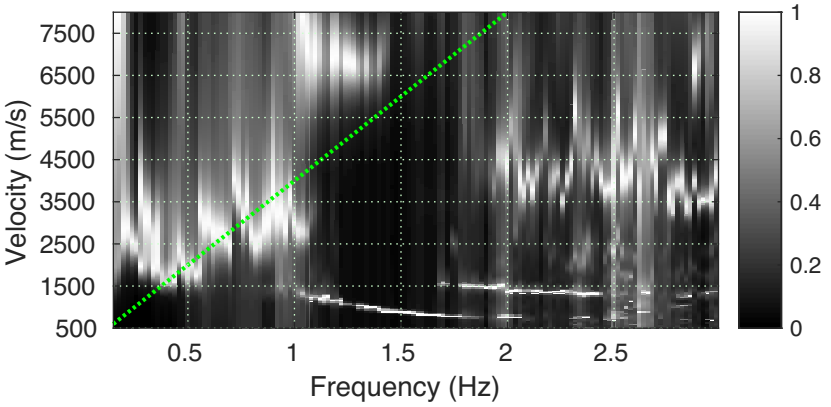
$$i_{\text{mag}} = \max\{i | C_R(i) = 1\}. \tag{B8}$$

This simply means that the first i_{mag} eigenvalues have a non-negligible magnitude compared to the major eigenvalue. On the

(a) MUSIC-CC with fixed $n_S = 1$



(b) MUSIC-CC with automated n_S



(c) Eigenvalue decay

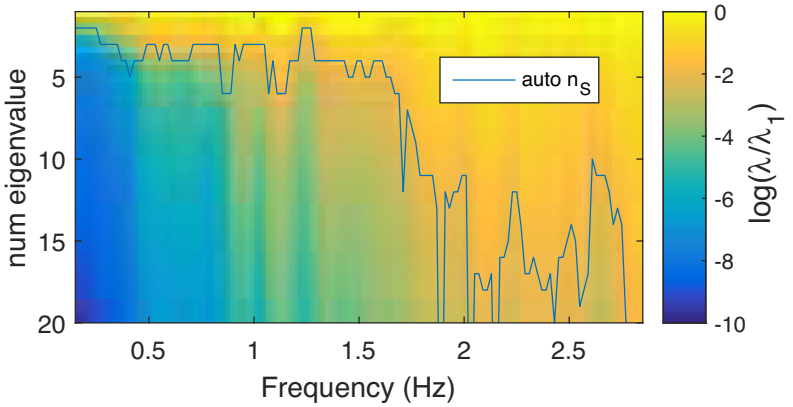


Figure B1. MUSIC-CC dispersion plot obtained for April (night-time data) with (a) fixed $n_S = 1$; (b) automated n_S determination (same as Fig. 5e). The green dotted line shows the resolution limit. (c) Eigenvalue magnitudes (logarithmic ratio to the major eigenvalue) at different frequencies (coloured image). The blue curve shows the determined n_S as function of frequency.

other hand, the local logarithmic slope of the eigenvalue decay is defined as

$$S_\lambda(i) = \tan^{-1}\left\{\log\left(\frac{\lambda_{i+1}}{\lambda_i}\right)\right\}. \tag{B9}$$

The slope break corresponds to the maximum value of the slope derivative with respect to the eigenvalue number:

$$i_{\text{slope}} = \text{argmax}\left(\left\|\frac{\partial S_\lambda}{\partial i}\right\|\right). \tag{B10}$$

We choose to define the signal subspace dimension as

$$n_s = \max(i_{\text{slope}}, i_{\text{mag}}). \tag{B11}$$

The maximum signal subspace dimension is limited by the CSM smoothing, as explained in Section 3.3. In order to take this into account, the CSM smoothing and diagonalization is first applied to pure white noise of same duration and sampling as the analyse signal. The resulting eigenvalue profile typically exhibits a strong jump, which is detected using the slope break criterion defined above. The number of the eigenvalue corresponding to the this jump ($i_{\text{slope}}^{(\text{noise})}$) is considered to be the maximum available signal-subspace dimension for the given CSM smoothing procedure.

Simple synthetic tests were performed with a linear array analogous to the virtual shot gathers shown in Figs 5(a) and (b). Dispersive phases with known phase velocities and known SNR were propagated in frequency domain following the procedure described in Appendix C. A misfit function was defined for the dispersion plots and the SNR estimated by the MUSIC algorithm. A grid search was performed to determine the number K of subarrays and the frequency bandwidth Δf used to smooth the CSM (see Sections 3.3.1 and 3.3.2), and the eigenvalue magnitude threshold parameter n_R. After repeated tests with different SNR levels, the values $K = 20$, $\Delta f = 0.1$ Hz and $n_R = 2$ were chosen.

For the real-data virtual shot gather (April: night time), the dispersion plot obtained with the automated signal-subspace determination is compared to the one obtained assuming $n_S = 1$ in Fig. B1. The latter appears much noisier and does not distinguish between R0 and R1 below 0.5 Hz. R0 is also not identified between 1 and 1.5 Hz. Eigenvalue magnitudes are displayed in Fig. B1(c). The low frequencies (below 1 Hz), characterized by a high SNR, exhibit only several strong eigenvalues, and are governed by the slope break criterion. Higher frequencies (above 1 Hz), characterized by a smaller SNR, exhibit a more gradual eigenvalue decay and are governed by the eigenvalue magnitude criterion $n_R = 2$. More eigenvalues are thus kept: plane waves are indeed 'spread' over several eigenvalues, which must all be included into the signal subspace to properly retrieve the correct phase velocity.

APPENDIX C: SYNTHETIC TESTS OF THE MUSIC ALGORITHM

In this appendix, the MUSIC algorithm with automated signal-subspace dimension determination is tested on synthetic signals with known backazimuth and dispersion relation. Both MUSIC-CC (linear array: single time window) and 'direct' MUSIC (2-D array: several time windows) are implemented and compared.

Synthetic plane waves recorded by an array of sensors with positions $\boldsymbol{x_i}$ were generated in Fourier domain with respect to a fictitious

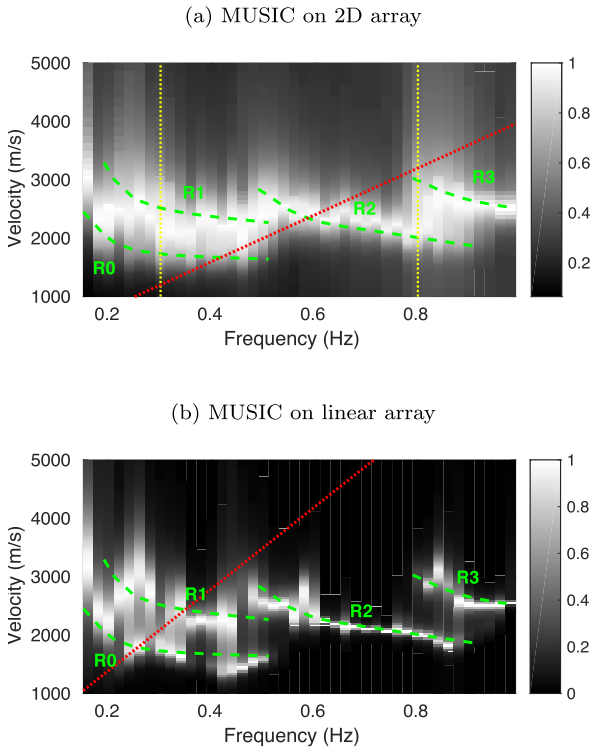


Figure C1. Dispersion plots of synthetic signals. (a) 2-D array with random source parameter distributions from Table C1: yellow dotted lines indicate the frequencies at which beamformers are plotted in Fig. C3; (b) linear array with a single shot. Red dotted lines show the resolution limit for both arrays.

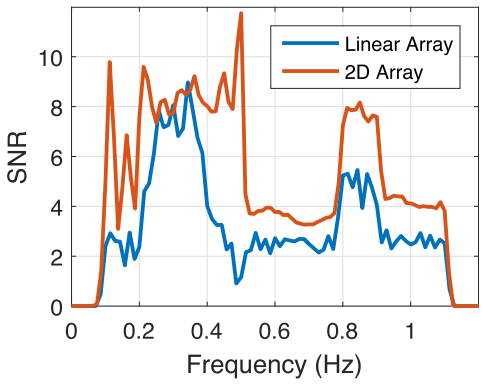


Figure C2. SNR in the simulated seismograms. Blue line: linear array; red line: 2-D array.

source[1] located at $\boldsymbol{x_s}$, taken as the time reference:

$$u(\omega, \boldsymbol{x_i}) = s(\omega)\exp\left[\mathrm{i}\omega\left(t_0 - \frac{(\boldsymbol{x_i} - \boldsymbol{x_s}).\boldsymbol{e_\theta}}{v_\phi(\omega)}\right)\right], \tag{C1}$$

where ω is the angular frequency, $s(\omega)$ is the source spectrum, t_0 is the emission time at $\boldsymbol{x_s}$, $\boldsymbol{e_\theta}$ is the unit direction vector along the wave's backazimuth θ, and $v_\phi(\omega)$ is the phase velocity dispersion relation. The fictitious source location was always chosen at a distance of 10 km from the array's centre, in the direction of the wave's backazimuth. Time domain seismograms were obtained by

[1]We use the term 'fictitious' because a strictly plane wave cannot have any localized source. Here the 'source' is only used as a reference point for fixing a backazimuth and a phase delay.

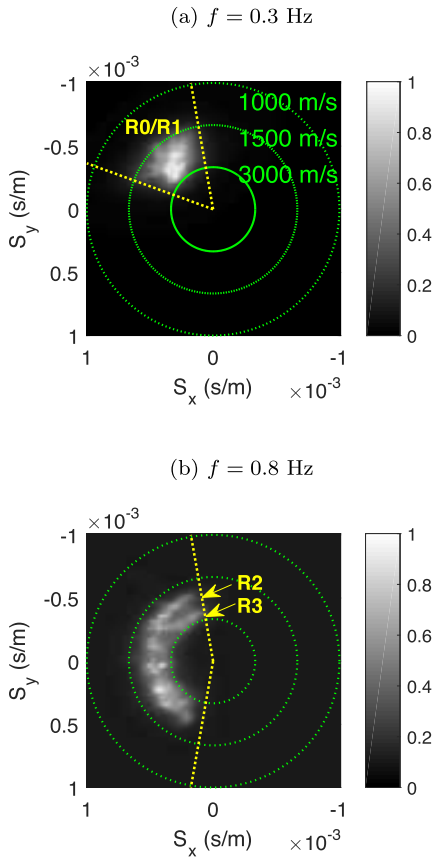


Figure C3. MUSIC beamformers for synthetic signals. (a) $f = 0.3$ Hz; (b) $f = 0.8$ Hz. Yellow dotted lines indicate the bounds of the backazimuth uniform distribution (a) for R0–R1, and (b) for R2–R3.

Table C1. Frequency and backazimuth bounds of the simulated dispersive phases for MUSIC synthetic tests

Parameter	R0	R1	R2	R3
f (Hz)	[0.05,0.5]	[0.2,0.5]	[0.5,0.9]	[0.8,1.1]
BAZ (°)	[−70;; −10]	[−70;; −10]	[−170, −10]	[−170, −10]

inverse Fourier transformation. They were then resampled at 100 Hz as for the real data set. The array configuration shown in Fig. 1(a) was used for synthetic tests. Four distinct dispersive phases (green dashed curves in Fig. C1) were simulated below 1 Hz in order to roughly reproduce those identified as R0, R1, R2 and R3 in Fig. 5(e). 50 time windows of 100 s were generated. Each time window contained signals generated by 10 fictitious sources with backazimuths following a uniform distribution, with bounds given in Table C1. Each fictitious source emitted 10 wave trains reaching the array within 100 s, with random emission times (t_0 in Table C1). The source spectrum associated to each wave train was a cosine tapered window with bounds given in Table C1. Random white noise was added to the synthetic signals with an average resulting SNR ratio between 4 and 8 (Fig. C2, red curve). The MUSIC algorithm (Section 3.3.1) was then applied to the resulting 50 time windows as described in Section 3.

C1 Dispersion curves retrieval

The mean dispersion plot (Fig. C1a) does not allow to separate R0 and R1, while R2 and R3 are retrieved with the correct phase velocities. For comparison, we propagated the same dispersive phases on a linear array of 7000 m aperture with 100 m spacing to reproduce the virtual shot gathers cross-correlations would yield. Only one wave train was excited for each dispersive phase, all of them emitted simultaneously by the same fictitious source. The latter is aligned with the array and located 200 m apart from the closest receiver. Random white noise was added to the simulated wavefield, so that the resulting SNR was close to the one used for the 2-D array (Fig. C2, blue line). The MUSIC algorithm was applied to this single noisy shot gather, as explained in Section 3.3.2. The obtained dispersion plot, shown in Fig. C1(b), was able to separate all the four dispersive phases. MUSIC applied to a single shot gather has thus a better resolution power than MUSIC applied to several time windows recorded by a 2-D array. Hence, MUSIC-CC should be preferred to the 'direct' MUSIC for dispersion curve retrieval from dense array data.

C2 Backazimuth retrieval

The MUSIC beamformers correctly retrieve the backazimuth distribution, both below and above 0.5 Hz (Figs C3 a and b, respectively). While at 0.3 Hz R0 and R1 are not separated (Fig. C3a), as expected from the dispersion plot analysis above, R2 and R3 at 0.8 Hz are only separated for backazimuths aligned with the direction of the longest array extent (NNW). Similar observations can be made on the real-data results in Figs 6(c)–(f), where the circular patterns typically become more refined in the NNW direction. As stressed by Cornou (2002), the theoretical resolution power of the MUSIC algorithm is asymptotically infinite as the SNR tends to infinity. The true resolution power depends on the SNR. The SNR is however hard to quantify, since both random and coherent noise are present in real data. Thus, the only conclusion we draw is that applying MUSIC allows to obtain reliable results beyond the theoretical resolution limit. However, we are unable to quantify the new resolution power precisely, and suggest a specially designed synthetic study similar to the present one in order to investigate the performance of a given array on given targeted seismic modes.

APPENDIX D: SPECTRAL-ELEMENT SIMULATIONS OF THE WAVEFIELD COMPOSITION

Synthetics were generated using the spectral-element based SPECFEM2D code (Komatitsch *et al.* 1999) coupled to Gmsh software used for generating a quadrangular mesh (Geuzaine & Remacle 2009). The simulations were performed into a vertical 2-D plane in order to avoid too costly computations associated with 3-D simulations. Absorbing Stacey conditions (provided option in SPECFEM2D) were applied at the left, right and bottom edges of the model. The typical quadrangle size used in different regions, was chosen so that there were at least four points per minimal S wavelength at $f = 3$ Hz, which is the maximum frequency analysed. We checked that taking a more refined grid did not affect our results significantly. The time step was set to 3×10^{-4} s, so that the stability condition was fulfilled.

Each simulated source was a vertical point-force source emitting a Ricker wavelet with dominating frequency f_0 and emission date t_0. These parameters along with the source's position followed a

Table D1. Truncation bounds used for the source parameter uniform distributions in spectral-element simulations

Parameter	DWS	SWS	LSS
x_s (km)	[10,60]	[202,219]	[460,520]
f_0 (Hz)	(log-norm)	(log-norm)	[1,3]
t_0 (s)	[20,520]	[20,440]	[5,190]

truncated uniform probability distribution. An exception was made for the oceanic sources dominating frequency, where we used a log-normal distribution centred on 0.25 Hz with a standard deviation of 0.7 Hz, truncated between 0.1 and 3 Hz. Truncation bounds of the uniform distributions used for each type of sources are listed in Table D1, with x_s being the source's horizontal location. 30 sets of 500 random realizations of these distributions were used to simulate 30 independent realizations of seismic noise generated by 500 sources of each type. The total simulation times were of 600 s for DWS, 480 s for SWS and 220 s for local surface sources (LSS), allowing the wavefield from the latest excited source to reach the array. Time intervals extracted for processing were [400–580] s for DWS, [300–480] s for SWS and [20–200] s for LSS, allowing to work with a fully developed wavefield where all possible phases are mixed.

The simulated array of receivers spans over 20 km with a 100 m spacing. It has a bigger offset compared to the real data, which allows to unambiguously resolve the low-frequency content of the simulated wavefield. Particle velocity wavefields recorded by the array, resulting from natural and anthropogenic sources, respectively written through $u_1(x, t)$ and $u_2(x, t)$, were mixed with a rate α. White random noise $\epsilon(x, t)$ was also added to the simulated data. The final signal can thus be written as

$$u(x, t) = u_1(x, t) + \alpha u_2(x, t) + \epsilon(x, t). \tag{D1}$$

The frequency-dependent amplitude ratio between u_1 and u_2 measured at the array being

$$\xi(f) = \sqrt{\frac{\langle u_1^2(x_i, f)\rangle_i}{\langle u_2^2(x_i, f)\rangle_i}}, \tag{D2}$$

where the index i runs over the 200 receivers, the relative contributions of the natural sources with respect to the anthropogenic sources and the white noise in the final signal are

$$\eta^{(12)}(f) = \sqrt{\frac{\langle u_1^2(x_i, f)\rangle_i}{\langle (\alpha u_2)^2(x_i, f)\rangle_i}} = \xi(f)/\alpha, \quad \text{and} \tag{D3}$$

$$\eta^{(1\epsilon)}(f) = \sqrt{\frac{\langle u_1^2(x_i, f)\rangle_i}{\langle \epsilon^2(x_i, f)\rangle_i}}, \tag{D4}$$

respectively. While $\xi(f)$ is fixed by the simulation, we choose α so that $\eta^{(12)}(1\ \text{Hz}) = 10$, and the white noise amplitude so that $\eta^{(1\epsilon)}(1\ \text{Hz}) = 100$. This allows to reproduce the situation where both natural and anthropogenic sources can influence the dispersion plot above 1 Hz, as in the real data. In order to get the dispersion plots in Figs 12(c) and (d), we directly applied MUSIC to the simulated linear antenna, and averaged the dispersion plots over the 30 independent realizations of the seismic noise, as described in Section 3.3.