Characterizing Seismic Events in an Industrial Corridor of the Chicago Area

Ann Mariam Thomas*[1], Omkar Ranadive[1], and Suzan van der Lee[1]

Abstract

Urban and industrial environments present a major challenge for seismic event detection and classification. Anthropogenic events can closely resemble tectonic signals and reduce their signal-to-noise ratios, leading to misclassifications of earthquakes and tremors. By detecting and characterizing these anthropogenic events, we can improve detection algorithms and motivate new applications in seismic monitoring and imaging. In this study, we develop a workflow to detect and cluster anomalous seismic events, recorded by a broadband seismic station installed in a unique industrial corridor of the Chicago area. Our workflow consists of (1) a power spectral density (PSD) misfit detector to detect anomalous events in continuous data and (2) a *k*-means clustering model to generate clusters of anomalous events. We discuss our workflow development, where we select parameters and generalizable model features to maximize the coherence and interpretability of generated clusters. When applied to two years of continuous seismic data, our workflow successfully identified several classes of events, including surface quarry blasts, underground blasts, and machinery operations. Using our results, we created a labeled data set of 1000+ man-made seismic events in the Chicago area. Our study demonstrates how a simple PSD detector and clustering model can be used to efficiently mine through a noisy multiyear data set and create an event catalog.

Cite this article as Thomas, A. M., O. Ranadive, and S. van der Lee (2025). Characterizing Seismic Events in an Industrial Corridor of the Chicago Area, *Seismol. Res. Lett.* **XX**, 1–12, doi: 10.1785/0220250109.

Introduction

Urban and industrial environments are characterized by dynamic, high-amplitude seismic noise, which can pose a serious challenge for earthquake detection and discrimination. Persistent anthropogenic noise—seismic noise from human-generated activities like traffic or construction—can decrease the signal-to-noise ratios of seismic events, leading to missed detections of earthquakes and other signals of interest. Impulsive anthropogenic events (e.g., quarry blasts, train signals) have also been falsely detected as earthquakes or tremors (e.g., Li *et al.*, 2018; Maher *et al.*, 2024). With rapid urbanization and industrialization, these detection challenges within built environments are growing in number and complexity. In response, recent studies have developed new algorithms to remove persistent anthropogenic noise from seismic data (e.g., Karasözen and West, 2022; Yang *et al.*, 2022). However, due to the broad spectrum and diversity of anthropogenic noise, further research is needed to develop robust algorithms that can accommodate diverse detection environments. Developers and other stakeholders have recommended expanding labeled data sets with dynamic anthropogenic noise, to develop and refine such algorithms (Mousavi *et al.*, 2020; Maher *et al.*, 2024).

With a growing number of studies that address the challenge of anthropogenic noise, there is also a growing body of research that harnesses this "noise" for innovative seismological applications. Recent studies have shown that urban seismometers can offer a cost-effective, less-invasive means of monitoring vehicular and even foot traffic (e.g., Díaz *et al.*, 2017; Jakkampudi *et al.*, 2020; Li *et al.*, 2023). Traffic noise has also shown great potential as a passive source for near-surface imaging (e.g., Quiros *et al.*, 2016; Zhang *et al.*, 2019). Within education and outreach, urban seismic events—especially those from major cultural activities like concerts and sports games—can capture mass media interest, showcasing the field of seismology to a wider audience (e.g., Díaz *et al.*, 2017, 2020; Caplan-Auerbach *et al.*, 2024).

These diverse applications of anthropogenic noise motivate exploratory research in built environments. Whether the aim is to harness or remove anthropogenic noise, it is valuable to identify the seismic signatures of dominant man-made signals. Previous studies have characterized seismic data within densely populated cities such as Long Beach (Riahi and Gerstoft, 2015; Snover *et al.*, 2020), Bucharest (Ritter *et al.*, 2005), Auckland (Boese *et al.*, 2015), and London (Green

1. Northwestern University, Earth and Planetary Sciences, Evanston, Illinois, U.S.A., https://orcid.org/0000-0001-6271-1738 (AMT); https://orcid.org/0000-0003-1884-1185 (SL)

*Corresponding author: annthomas2025@u.northwestern.edu

Figure 1. Map of southwest Chicago and suburbs, showing the locations of station NW.HQIL (star at 41.80° N, 87.85° W) and local noise sources. Figure was created with QGIS using ESRI Satellite Imagery. The color version of this figure is available only in the electronic edition.

et al., 2016). In this study, we explore seismic data from a broadband seismometer (NW.HQIL) installed in a unique industrial corridor of the Chicago area. The station was installed shortly after a magnitude 3.2 earthquake in 2013, near its epicenter to search for aftershocks and was operational until December 2019. Its noisy location—within a few kilometers of dense residential districts, quarrying operations, a flood-control reservoir, major highways, railroads, a sports stadium, and a commercial airport (Fig. 1)—generates thousands of transient events in just a few days of data. This combination of urban and industrial features makes station HQIL a unique site for exploratory research.

Manually exploring the rich, heterogeneous data of this station is time-consuming and challenging, especially when attempting to find patterns and groups of repeating events. Here, we present a semiautomated approach to identify clusters of anomalous seismic events in continuous, three-component data. Our approach is divided into two stages: (1) applying a power spectral density (PSD) misfit detector to identify anomalous events in continuous data and (2) applying an unsupervised learning model (*k*-means clustering) to group the anomalous events detected in stage 1. Using the results of our clustering analysis, we created a labeled data set of 1262 man-made signals recorded in HQIL data, placing a special emphasis on impulsive events that have earthquake-like characteristics.

PSD Misfit Detector

Our two-stage methodology is outlined in the concept map of Figure 2. The workflow begins by applying a PSD misfit detector to detect anomalous events in continuous data. A PSD misfit detector measures the difference (or misfit) between the PSD of a given time window and the PSD of background noise. "Detected" anomalous events are windows with PSD misfits above a selected threshold. Unlike the short-term average/long-term average (STA/LTA) ratio method (Allen, 1978), which is commonly used for event subselection in machine learning workflows, the PSD misfit detector can detect anomalous events with low signal-to-noise ratios (SNRs) (Vaezi and Van der Baan, 2015). PSD-based approaches are particularly beneficial in urban and industrial areas where persistent man-made noise (e.g., noise from construction and machinery operation) can decrease the SNR of anomalous events. PSD misfit approaches have been shown to detect a wide range of events, including weak microseismic events (Vaezi and Van der Baan, 2015) and industrial operations (Guenaga *et al.*, 2021).

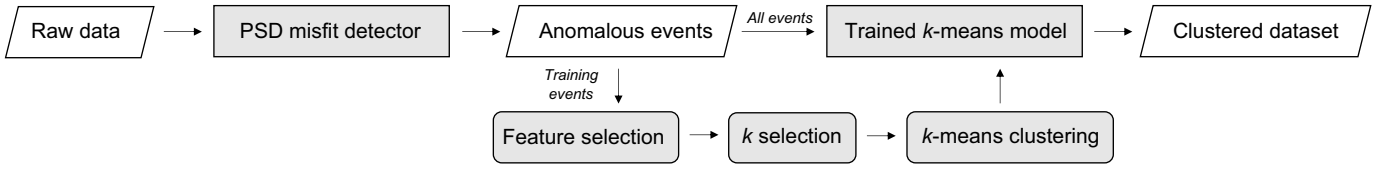


Figure 2. Concept map of study methodology for detecting and clustering anomalous events in continuous, single-station data. PSD, power spectral density.

Methodology

Our PSD misfit approach is based on the methodology of Vaezi and Van der Baan (2015) with some modifications. We first computed PSDs of overlapping (50%) 10 s windows in continuous HQIL data using the Welch method (Welch, 1967). Waveforms were processed with the following procedure prior to PSD computation: filling gaps via linear interpolation, linear detrending, and response correction to convert from counts to ground acceleration. Note that the gap-filling step was performed prior to this study when saving the raw data as local files. This step can be omitted with minimal impact on the results. Using a methodology similar to McNamara and Buland (2004), PSDs of each 10 s window were computed by taking the average PSD of 13 overlapping (50%) subsegments of 2.5 s. This subsegmentation and averaging procedure reduce the variance of each power measurement, providing a more stable PSD estimate of each 10 s window. We then computed the weighted difference $u[f]$ between the PSD of the 10 s window, PSD$[f]$, and the PSD of a "background noise" window, $\mathrm{PSD}_{\mathrm{bg}}[f]$, using the following equation:

$$u[f] = \frac{\mathrm{PSD}[f] - \mathrm{PSD}_{\mathrm{bg}}[f]}{\sigma_{\mathrm{bg}}[f]}, \tag{1}$$

in which σ_{bg} is a standard deviation estimate of the "background noise" PSD at a given frequency f. The Dynamic background noise section describes how we estimated $\mathrm{PSD}_{\mathrm{bg}}[f]$ and $\sigma_{\mathrm{bg}}[f]$ for HQIL noise.

Following the methodology of Vaezi and Van der Baan (2015), we set $u(f)$-values below 1 as zero:

$$u'[f] = \begin{cases} u[f] & \text{if } u[f] > 1 \\ 0 & \text{if } u[f] \le 1. \end{cases} \tag{2}$$

Finally, we computed the PSD misfit Λ of each 10 s window:

$$\Lambda = \frac{1}{N}\sum_{f=0}^{N} u'[f], \tag{3}$$

in which N is the number of all discrete frequencies that comprose the PSD estimate PSD$[f]$. In summary, our PSD misfit Λ is the average value of the adjusted difference $u'[f]$ between the PSD of a 10 s window and the PSD of a "background noise" window.

Dynamic background noise

Selecting an appropriate background noise PSD for urban/industrial environments is particularly challenging due to their strongly heterogeneous and time-dependent noise. Noise levels typically peak during daytime hours of the workweek and decrease during evening and night hours. Because of this dynamic nature of man-made noise, we implemented a dynamic or time-dependent background noise window for our misfit detector. Depending on the following four times of day, a different background noise PSD, $\mathrm{PSD}_{\mathrm{bg}}[f]$, and standard deviation estimate, $\sigma_{\mathrm{bg}}[f]$, will be selected:

1. Night: 12:00 a.m.–6:00 a.m.
2. Morning: 06:00 a.m.–12:00 p.m.
3. Afternoon: 12:00 p.m.–06:00 p.m.
4. Evening: 06:00 p.m.–12:00 a.m.

All times are with respect to the local Chicago time (US/central time zone). Compared to a static or time-constant background noise window, a dynamic PSD is better suited to detect anomalous events during low-noise periods (e.g., nighttime).

To estimate the four background noise PSDs and their standard deviations, we randomly selected 50 days in the time period of July 2014 to December 2019, excluding select days with poor or incomplete data. We then computed 1 hr PSDs for each station component using the same methodology of McNamara and Buland (2004), omitting the smoothing procedure. For each of the four time periods, we computed the average and standard deviation of all 1 hr PSDs that occurred with that period. The average PSD and standard deviation are then smoothed using the 1/8 octave smoothing procedure implemented by McNamara and Buland (2004) and decimated to have the same frequency resolution of the 10 s PSDs. The processed averages and standard deviations became the background noise PSDs, $\mathrm{PSD}_{\mathrm{bg}}[f]$, and the standard deviation estimates, $\sigma_{\mathrm{bg}}[f]$, for our misfit detector. Figure 3 shows the background noise PSDs, $\mathrm{PSD}_{\mathrm{bg}}[f]$, we used for all three station components and four times of day.

Threshold selection

Our dynamic PSD misfit detector was applied to overlapping (50%) 10 s windows that span about two years of HQIL data (June 2017 to June 2019). After manually inspecting a subset of 10 s windows and comparing their PSD misfit values, we

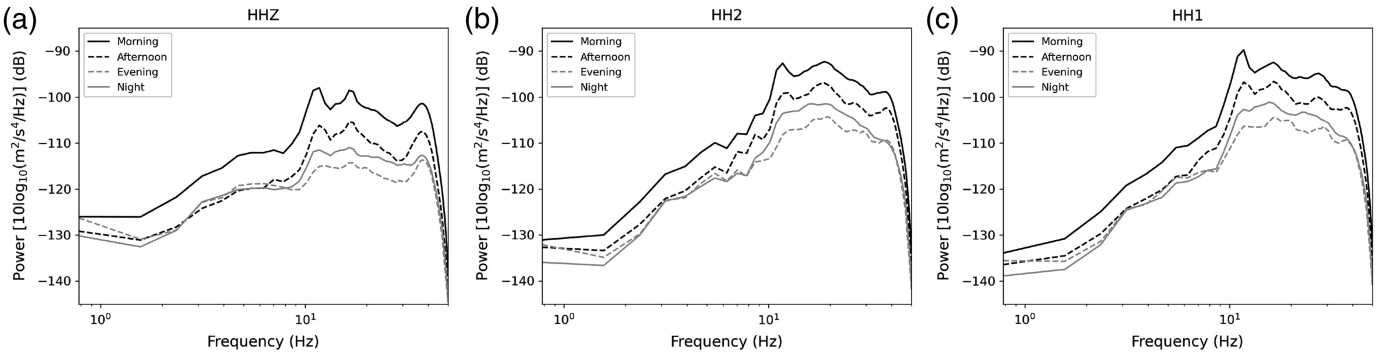


Figure 3. Background noise power spectral densities PSD_{bg} used in the PSD misfit detector for each component of station HQIL: (a) HHZ, (b) HH1, and (c) HH2. PSD_{bg} are estimated for four times of day: night (12:00 a.m.–06:00 a.m.), morning (06:00 a.m.–12:00 p.m.), afternoon (12:00 p.m.–06:00 p.m.), and evening (06:00 p.m.–12:00 a.m.). All times are with respect to local Chicago time (US/central).

decided on a misfit threshold of 1.0 standard deviation; if the PSD misfit of any component of a 10 s window is above 1.0 standard deviation, then it is detected as an anomalous event and passed to the clustering model. With this threshold, about 5% of all overlapping windows in the two-year data set are detected. As will be discussed in the Application and Analysis section, the PSD misfit detector identified a diverse range of events in our two-year data set, including both narrowband and broadband anomalies spanning a wide range of SNRs.

K-means Clustering

After detecting anomalous events using the PSD detector, we applied the *k*-means clustering algorithm to identify groups (clusters) of repeating anomalous events of the same source. The *k*-means algorithm (also known as Lloyd's algorithm, Lloyd, 1982) is an unsupervised learning algorithm that groups unlabeled data into k number of clusters. The algorithm aims to assign cluster labels that minimize the following criterion known as inertia J:

$$J = \sum_{k=1}^{k} \sum_{i=1}^{n} \|(x_i - \mu_k)\|^2, \tag{4}$$

in which k is the number of clusters, n is the number of data points, x_i is the ith data point, and u_k is the kth centroid. The algorithm begins by setting an initial guess of cluster centroids and assigning each data sample with the cluster label of its closest centroid. Cluster centroids are then reset to be the average of all assigned data samples in that cluster. This process of assigning cluster labels and resetting cluster centroids is repeated until the centroids are stable (Watt *et al.*, 2016; Yuan and Yang, 2019). With this simple iterative procedure, the *k*-means algorithm is a powerful and widely used tool in seismology for exploratory data analysis (Johnson *et al.*, 2020; Mousavi and Beroza, 2023; Saadia and Fotopoulos, 2023). It is critical to carefully select model features and a *k*-value for the algorithm to produce meaningful and well-separated clusters.

Feature selection

Table 1 details the features we explored and selected for our final clustering model. The majority of the features are statistical and are motivated by our previous work on earthquake detection in the Chicago area (Thomas *et al.*, 2023). For each 10 s window detected by the PSD misfit detector, we computed its STA/LTA (Allen, 1978) ratio using the 10 s detection as the short-term (ST) window and a preceding 40 s window as the long-term (LT) window. We also computed the skewness and kurtosis of the 10 s window. Each statistical feature was computed on waveforms that were processed with the following procedure: filling gaps via linear interpolation, linear detrending, and division by the stage zero sensitivity to convert to units of velocity. Additional features include the previously computed PSD misfit values, the hour of day, and the day of the week corresponding to each 10 s detection. Excluding the two temporal features (hour and day of the 10 s window), each feature is computed on each of the three channel components (HHZ, HH1, and HH2).

Our feature selection was informed by the observed temporal characteristics of urban and industrial noise in the Chicago area. As expected, man-made noise dominates the seismic record during daytime hours of the weekday. Events like quarry blasts consistently occur at the same few hours of the early afternoon, likely due to a routine blasting schedule. Events of different noise sources can also be differentiated by their event duration. Compared to blastlike events, construction noise and machinery operation tend to be longer in duration (≥10 s), resulting in a distinct skewness and kurtosis value for a 10 s window. Using a combination of statistical and time-based features, we aimed to maximize the separation between noise events of different sources.

From this initial list of 14 total features, we selected a subset of 10 model features using a simple correlation analysis. Because highly correlated features can introduce unnecessary complexity into a clustering algorithm, we aimed to remove redundant features from the initial list. For three weeks of

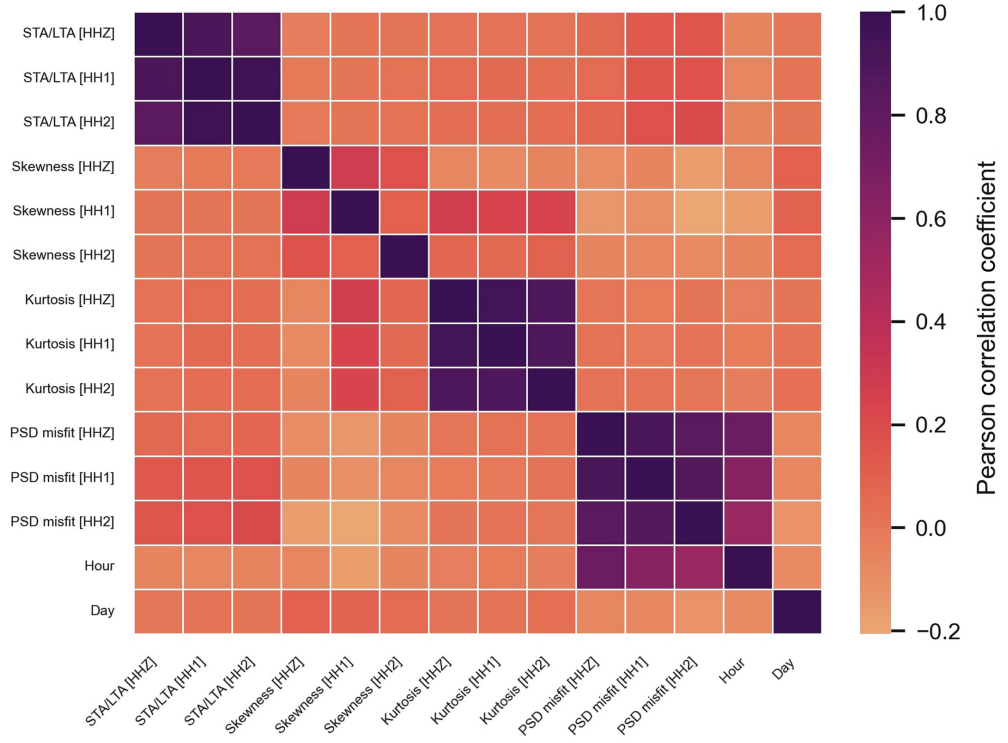


Figure 4. Correlation matrix of the 14 features, detailed in Table 1, corresponding to anomalous events detected in 3 weeks of HQIL data. The following features were omitted in the final model to eliminate any feature pairs with correlations greater than 0.85: STA/LTA [HH1], PSD misfit [HH1], Kurtosis [HHZ], and Kurtosis [HH2]. The color version of this figure is available only in the electronic edition.

HQIL data (1–21 August 2017), we applied the PSD misfit detector and computed all 14 features for the detected events. We then computed the correlation coefficient of each pair of features. Figure 4 shows the correlation matrix of the 14 features computed for the three-week data set. We removed four features to eliminate any feature pairs with correlations greater than 0.85: a threshold determined through empirical testing to balance redundancy reduction and feature retention. See Table 1 for all features included in the final model.

Number of clusters

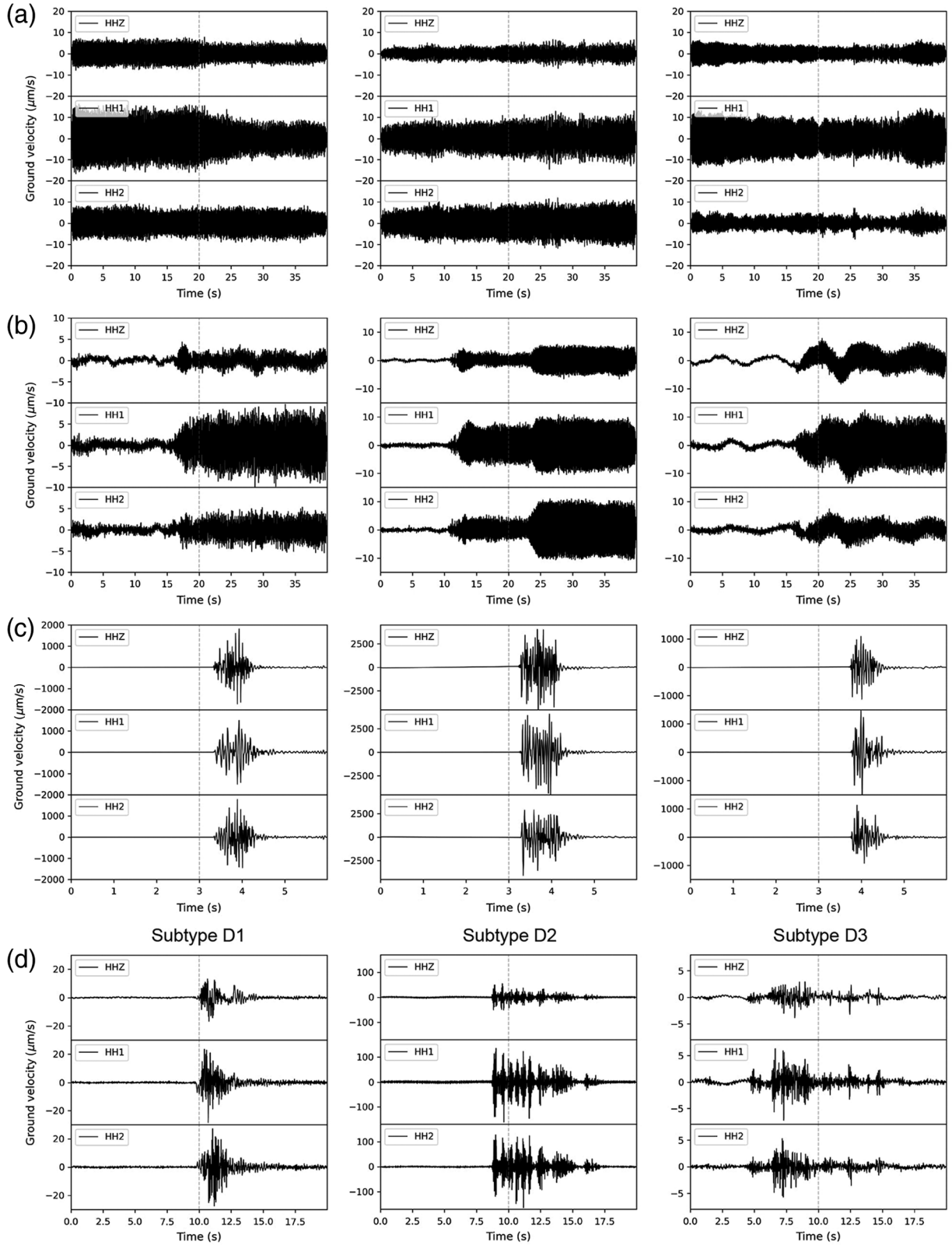
We explored both quantitative and qualitative methods to determine the number of clusters k to use in our model. For events detected by the PSD misfit detector in three weeks of HQIL data (1–21 August 2017), we applied the k-means clustering algorithm for k-values ranging from 2 to 20. We implemented Lloyd's k-means algorithm using the Python package *scikit-learn* and its default hyperparameters (greedy k-means++ algorithm for cluster centroid initialization, 300 maximum iterations) (Pedregosa *et al.*, 2011).

As a quantitative method, we computed the silhouette coefficient—a measure of intracluster similarity and intercluster separation (Yuan and Yang, 2019)—for each k-value. We also applied the elbow method, which identifies the optimal k-value as the value that corresponds to an "elbow" or an inflection

TABLE 1
Features Explored and Selected for the Clustering Model

Feature	Description	Equation	HHZ	HH1	HH2
STA/LTA	Ratio of the average absolute amplitude in the 10 s short-term (ST) window to the average absolute amplitude in a preceding 40 s long-term (LT) window	$\frac{\frac{1}{N_S}\sum_{n=N_L}^{N_S+N_L-1}\lvert y[n]\rvert}{\frac{1}{N_L}\sum_{n=0}^{N_L-1}\lvert y[n]\rvert}$	Yes	No	Yes
Skewness	Skewness of the amplitude distribution in the 10 s window	$\sqrt{N_S}\,\frac{\sum_{n=N_L}^{N_S+N_L-1}(y[n]-\bar{y})^3}{\left(\sum_{n=N_L}^{N_S+N_L-1}(y[n]-\bar{y})^2\right)^{3/2}}$	Yes	Yes	Yes
Kurtosis	Kurtosis of the amplitude distribution in the 10 s window	$N_S\,\frac{\sum_{n=N_L}^{N_S+N_L-1}(y[n]-\bar{y})^4}{\left(\sum_{n=N_L}^{N_S+N_L-1}(y[n]-\bar{y})^2\right)^2}$	No	Yes	No
PSD misfit	Misfit between the power spectral density (PSD) of the 10 s window and a background noise PSD	See equation (3)	Yes	No	Yes
Hour	Hour of day in local time corresponding to the 10 s window. Value is an integer within the interval [0, 23].	N/A	Yes		
Day	Day of the week corresponding to the 10 s window. Value is an integer within the interval [1,7], in which 1 is Monday and 7 is Sunday.	N/A	Yes		

The discrete time series $y[n]$ includes amplitudes of the 10 s window detected by the PSD misfit detector and a preceding 40 s window. N_S is the number of samples within the 10 s window and N_L is the number of samples within the 40 s window. STA/LTA, short-term average/long-term average.



point within an inertia versus k plot (Yuan and Yang, 2019). Both methods suggested k-values below 7.

As our qualitative approach, we visualized a subset of events in every cluster for each tested k-value. We found that a k-value of 12 consistently produced a cluster of signals that indicated an operational change (i.e., machinery turning ON and

Figure 5. (a–d) Examples of HQIL anomalous events in the four event types (A–D) described in Table 2. The gray dashed lines indicates the start of the 10 s window, detected as an anomalous event by the PSD misfit detector. Note the changes in times and amplitude scaling for each example.

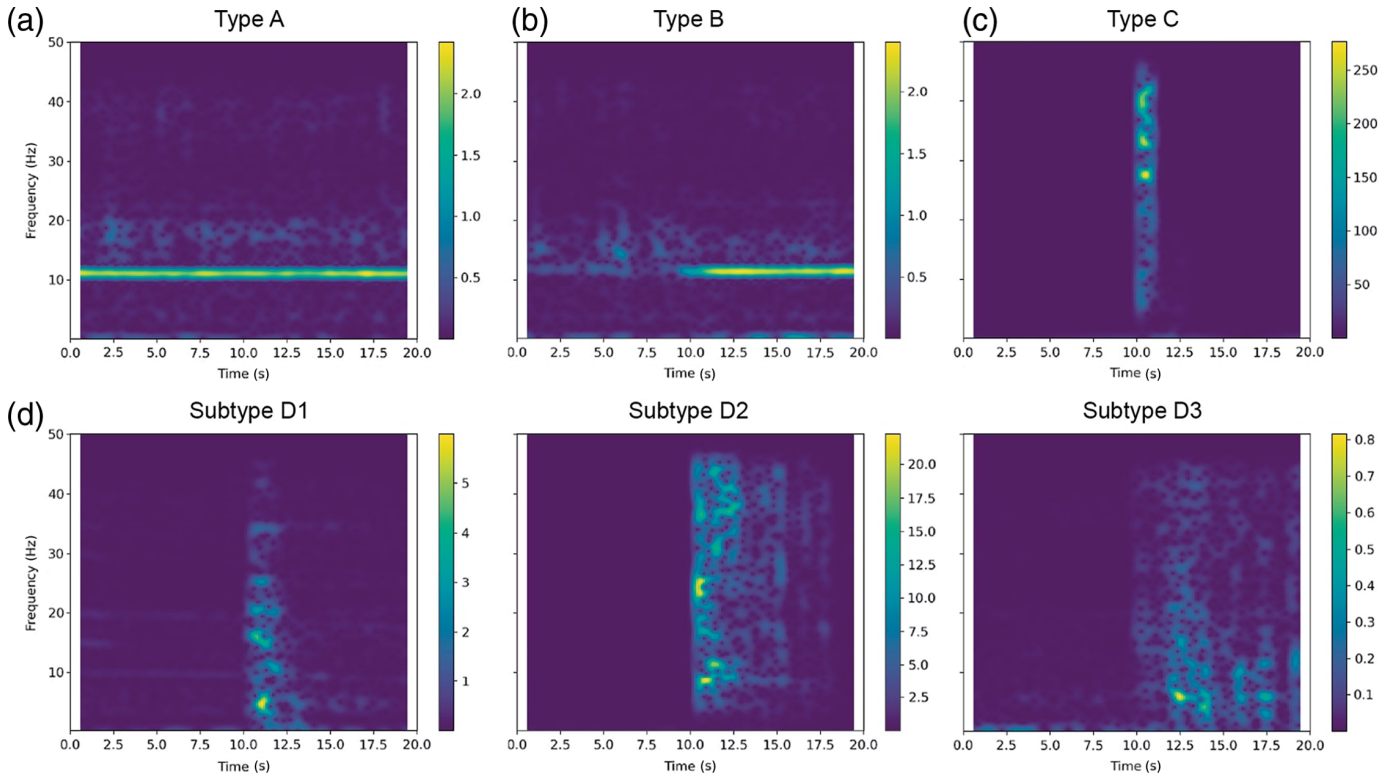


Figure 6. Examples of vertical-component spectrograms for event types (a–d) A–D described in Table 2. Each spectrogram shows 10 s before and after the event onset. Raw waveforms were linearly detrended and converted from counts to ground velocity by dividing by the stage-zero sensitivity. Color scales are in units of micrometers per second. The color version of this figure is available only in the electronic edition.

producing a sudden change in relative amplitude, see Fig. 5b). These signals are of special interest because they have been commonly misclassified as earthquakes by earthquake detection algorithms; they constituted about 20% of EQTransformer (Mousavi *et al.*, 2020) detections for station HQIL in 2015 (Thomas *et al.*, 2023). For lower *k*-values, these transitional signals were combined into large clusters (>3000 events) of high-amplitude long-duration events.

Selecting a *k*-value of 12 resulted in a silhouette score of 0.36, which indicates fair clustering structure but strong overlapping. This overlapping quality was evident in our visualization analysis because a few small clusters (<30 events) contained similar blast events. Although choosing a smaller *k* would reduce this overlapping and produce a higher silhouette coefficient, we decided to compromise on this consequence in exchange for a strong, distinct cluster of transitional signals. It is less time-consuming to manage multiple small clusters, which ideally should be combined, compared to sifting through large clusters to find a few signals of interest. Hence, we finalized our workflow with a *k*-value of 12, as supported by our qualitative approach.

Model fitting

Our final *k*-means clustering model was trained on 21,293 anomalous events detected by the PSD misfit detector in 3 weeks of HQIL data (1–21 August 2017). We used the 10 features listed in Table 1 and a *k*-value of 12. We trained the final model using *scikit-learn* (Pedregosa *et al.*, 2011) and its default parameters for *k*-means clustering (*k*-means++ initialization of cluster centroids, Lloyd's algorithm). Hereafter, we refer to this model as the pretrained model.

Application and Analysis

We applied the pretrained clustering model to 654,410 ten-second windows of anomalous events detected by the PSD misfit detector in two years of HQIL data (June 2017–June 2019). Applying the pretrained model to both the 3-week training data set and the full two-year data set produced similar clusters and cluster distributions. During workflow development, we also evaluated the clustering performance of a *k*-means model trained on the full two-year data set. Our decision to use the pretrained model in the final workflow is described in detail in the Discussion section.

To characterize each cluster and identify potential noise sources, we analyzed the average spectra and a subset of waveforms in each cluster of the two-year data set. We also examined their temporal occurrences using heat maps. This analysis identified four major event types that characterize one or multiple clusters in our data set. Table 2 provides descriptions of the four event types and their distribution in both the

TABLE 2
Description, Potential Sources, and Cluster Distribution of Detected Event Types

Type	Description	Potential Source(s)	Cluster Label(s)	Distribution	
				Training	All Data
A	Long-duration (>10 s) man-made noise, dominated by 11 Hz energy (characteristic of eight-pole motors). Predominantly occurs during weekday daytime hours.	Operation of industrial machinery (e.g., conveyor belt with an eight-pole motor)	0,2,5,7,10	98.5%	98.6%
B	Transitional signals (turning ON signals) with similar waveform and spectral features as type A events. Signals typically have high STA/LTA ratios.	Turning ON of type A machinery	11	1.35%	1.20%
C	High-amplitude, short-duration (<10 s) blasts, dominated by high-frequency energy (>30 Hz). Signals typically have a maximum amplitude on the order of 0.001 m/s across all components. Predominantly occurs between 11:00 a.m. and 02:00 p.m. local time on weekdays.	Surface blasts at the nearest quarry (<1 km from HQIL)	1,4,6,8,9	0.06%	0.10%
D	Low-amplitude blast-like events of mixed sources. Signals have a maximum amplitude that are about 1–3 orders of magnitude lower than type C signals. Occurs during a wide range of hours, but predominantly from 12:00 p.m. to 02:00 p.m. local time.	Quarry blasts with source distances ≥1 km from station HQIL (D1), underground explosions at the nearest quarry (D2), wind interaction with local structures (D3)	3	0.06%	0.11%

STA/LTA, short-term average/long-term average.

training set (1–21 August 2017) and the entire two-year data set. Figures 5 and 6 show example waveforms and spectrograms, respectively, of each event type.

Our visual analysis of the two-year data set revealed a moderately strong degree of cluster coherence. With the exception of cluster 3 (event type D), events within a cluster generally had similar spectral features and likely originated from the same source (e.g., industrial machinery, surface quarry blasts). However, some cluster overlapping was observed. When visualizing a randomized subset of events, about 70% of the events in each cluster matched the description given in Table 2; the remaining events were either misclassified or consisted of signals not clearly associated with any of the event types described in Table 2.

As discussed in the Number of clusters section, selecting a high *k*-value of 12 resulted in multiple clusters containing the same event type. These events were likely grouped into different clusters based on their temporal occurrence. For example, clusters 0 and 2 contain events with similar waveforms and spectra, indicating a similar source or event type. However, cluster 0 events predominately occur between 12:00 and 03:00 p.m. local time while cluster 2 events predominately occur between 06:00 and 08:00 p.m. local time. These temporal distinctions are of minor significance to our analysis and discussion. Therefore, we will proceed by referring to events based on their event types rather than their assigned cluster labels.

As detailed in Table 2, type A events are characterized by monochromatic waveforms, long durations (>10 s), and a peak frequency at 11 Hz, indicative of an eight-pole motor (Groos and Ritter, 2009). A potential source is the operation of heavy industrial machinery, such as a conveyor belt, at the nearest quarry operation. Type B events are similar in spectral and temporal features to type A events, but they are transitional signals (i.e., turning ON of machinery). Type C events are the highest-amplitude events recorded in station HQIL. They are short-duration (<10 s) events, dominated by energy at frequencies >30 Hz. From previous correspondence with local quarry representatives and residents, a few type C events have been confirmed as surface blasts from the nearest quarry operation to HQIL.

Type D events consist of blastlike events with diverse waveforms, which are generally lower in amplitude and longer in duration than type C blasts. They are of special interest because many type D events have waveform features that are similar to local earthquakes. Although there are at least five different types of event waveforms (and thus sources) within type D, the majority of D events can be grouped into two subtypes: subtype D1) waveforms of a 3–4 s duration that are dominated by frequencies below 20 Hz (Fig. 5d, left panel) and subtype D2) waveforms of a 5–10 s duration and peak frequencies above 20 Hz (Fig. 5d, center panel). D1 and D2 events make up about 17% and 46% of all D events, respectively. Both D1

and D2 events occur predominately during early afternoon hours of weekdays.

D1 events have similar waveforms and timing to type C blasts but are longer in duration and lower in amplitude, suggesting they may also be quarry blasts but from a farther distance. Many D1 events show signs of air-wave arrivals that occur between 2.5 and 3.5 s after the first arrival, indicating source distances of about 1 km. There are two quarry operations within a 5 km radius of HQIL. Type C events are surface blasts from the nearest operation and D1 events are likely surface blasts from the second operation.

Compared to D1 events, D2 events have a more complex seismic signature. Generally, D2 events are composed of 5–15 bursts with variable time intervals between bursts. From previous correspondence with local quarry representatives, one D2 event was confirmed as an underground blast at the nearest quarry operation. To further test this potential source of D2 events, we applied our PSD misfit detector and pretrained model to the 6-month period preceding the quarry's official transition to underground blasting. When analyzing cluster 3 (type D) events, we did not detect any events that strongly resembled D2 events, indicating that underground explosions are the likely source for D2 events.

Although few in number, we also detected 11 events (D3 events, Fig. 5d, right panel) that have features that look like *P*- and *S*-wave arrivals. All D3 detections occurred during late afternoon or early evening hours. This temporal trend and their high-frequency content suggests that D3 events are not earthquakes. D3 events closely resemble wind-generated signals characterized in Johnson *et al.* (2019, see their fig. 9). The earthquake-like signals of the 2019 study are generated from wind interaction with local vegetation and structures (Johnson *et al.*, 2019). An independent investigation is needed to verify that D3 events are indeed wind-generated.

Labeled Data Set

Using the results of our clustering analysis, we created a labeled data set of 1262 events recorded by station HQIL. Links to metadata and three-component seismograms (stored as miniSEED files) are provided in Data and Resources. Each miniSEED file is 20 s in duration, containing 10 s before and after the onset of each event. The data set contains all type C and type D1-3 events detected during our two-year study period. There are 176 type C, 75 type D1, 199 type D2, and 12 type D3 events. We also include a subset of type A and type B events (400 events each). Note that our workflow detected thousands of types A and B events in our two-year data set. However, due to the manual inspection procedure described subsequently, we only included a subset of these events.

For each event in the data set, we manually inspected their waveforms and spectrograms to (1) confirm that the event type matches the description given in Table 2, (2) check for any hidden transient event signals in both the 10 s detection window and the preceding 10 s window, and (3) manually pick the onset time for each event. Because we used overlapping (50%) 10 s windows, the window start time of an event may not coincide with its real onset (or first wave arrival) (see Fig 5). Onset times are not applicable for type A events, so their "onset times" are the same as their window start times.

Discussion

Our motivation and methodology for seismic exploration in built environments are similar to those of Saadia and Fotopoulos (2023) and Chai *et al.* (2025), who also apply an unsupervised learning approach to an urban or industrial setting. In this study, we develop and apply our workflow to a longer, multiyear data set in a unique urban and industrial environment. Unlike previous studies, we place a special focus on impulsive, blast- and earthquake-like signals and create a labeled data set that can be used to improve future earthquake detection and discrimination methods.

In designing our workflow, our guiding philosophy was to start with the simplest approach and to iteratively add complexity until we were satisfied with the clustering performance on three weeks of continuous HQIL data. In particular, we visually inspected the waveforms of clustered events after each addition or modification to confirm that they produced a visually discernable improvement in cluster coherence and interpretability. As an example, early iterations of our workflow did not have a detection stage: we initially applied *k*-means to all overlapping windows in continuous HQIL data. Implementing the PSD misfit detector prior to clustering produced one of the most substantial improvements in our results. Without this detection stage, *k*-means primarily generated small clusters of high-amplitude surface blasts (type C), with smaller-amplitude blastlike events (type D) and transitional signals (type B) hidden in large clusters of mixed waveforms.

In line with our guiding philosophy, we began and ultimately finalized our workflow with one of the simplest clustering algorithms: *k*-means. After including a detection stage and expanding our feature space with higher-order statistics (skewness and kurtosis), the *k*-means algorithm consistently generated coherent, interpretable clusters of diverse anomalous events in our data set. In the past decade, more advanced clustering methods (e.g., HDBSCAN as used by Chai *et al.*, 2025) have been developed to provide a more robust approach for exploratory data analysis (e.g., Campello *et al.*, 2013). Implementing such clustering algorithms may further improve our clustering performance and alleviate some of our major challenges (*k*-selection and type D subclustering). Comparing the implementation and performance of different unsupervised learning approaches is a promising direction for our future work. Our labeled data set and the complex built environment of station HQIL offers a unique testbed for such comparisons.

As mentioned in the Application and Analysis section, we explored two options for clustering our two-year data

set: (1) applying the pretrained model (model trained on the three-week data set) and (2) training a new *k*-means model on the full two-year data set using the same parameters as the pretrained model. Option 2 was theoretically ideal because it incorporates the full variability of anomalous events in the two-year data set. However, when we visualized the clusters of option 2, we found that the transitional signals (type B) were hidden in large clusters of type A events. This is understandable because the features differences between type A and type B events are relatively small compared to the differences between these events and higher-amplitude blastlike events (types C and D). It is possible that a different *k*-value or set of initial centroids may generate a well-separated cluster of type B events for a model trained on the two-year data set. We finalized our workflow with the pretrained model because it produced clusters of similar coherence to the clusters of the 3-week training data set. This analysis indicates that our workflow may be highly sensitive to clustering parameters, particularly the *k*-value. At the same time, it also demonstrates the success of our parameter selection process, which produced a clustering model that generalizes well to a larger data set. When adapting our workflow to a different built environment, we strongly recommend following a similar procedure of qualitatively assessing clustering performance across different *k*-values. We also suggest exploring local news sources and collaborating with local entities to ensure that no major disruptions or transitions (e.g., moving from surface to underground mining) occurred during the period used to train the clustering model. If prior knowledge of local operations is limited, it will be beneficial to tune and train the clustering model on a larger, randomized subset of days within your desired study period.

Another promising area for workflow improvement and comparison is with respect to the dynamic PSD misfit detector. Our approach accommodates the most prominent source of variation in HQIL data: diurnal fluctuations. However, weekday–weekend and seasonal variations also impact HQIL noise, although to a lesser extent. Incorporating these additional variations into the PSD misfit detector may produce a more diverse range of detections (e.g., sports stadium events, aircraft signals, etc.), particularly on low-noise days such as Sundays. During our workflow development, we considered using separate background PSDs for weekdays and weekends. However, we ultimately omitted this distinction after observing that the local quarry frequently operates on Saturdays.

Our relatively simple workflow produced coherent, interpretable clusters in a noisy two-year data set. With only one exception, the majority of events in each cluster shared similar waveforms and thus the same source. Cluster 3 (type D) events needed to be manually subdivided for analysis. It is possible to automate this process by training another clustering model only on type D events. Incorporating frequency-based features into the original clustering model may also achieve cluster separation for type D events.

Despite the heterogeneous setting of station HQIL (Fig. 1), the majority of detected events were related to the nearest quarrying operation. This is understandable as the quarry is the closest noise source and their blasts produce the largest shaking events in the area. In addition to different types of blasts, our workflow detected low-amplitude machinery operations, notably type A events, which do not have a sudden amplitude change in the detected 10 s window or a preceding 40 s window. These types of events would not have been detected by strictly amplitude-based methods like the STA/LTA ratio method (Allen, 1978), which require high relative amplitudes with respect to a preceding "noise" window. Using a PSD detector in our workflow, we were able to detect and characterize a diverse range of operations at the local quarrying facility. Thus, similar to Chai *et al.* (2025), our study demonstrates how seismometers can be a cost-effective and multiuse tool for industrial facility monitoring. When coupled with additional sensors, our workflow can be adapted to identify and locate anomalous operations in an industrial setting.

Finally, by building a labeled data set of 1262 events with their potential sources, we demonstrate how our workflow can be used to extract interpretable clusters of signals in a noisy data set and create an event catalog. With the persistent challenge of earthquake and tremor detection in built environments, it can be critical to create/expand catalogs of dynamic nontectonic events, such as the Exotic Seismic Events Catalog (Bahavar *et al.*, 2019). Nontectonic catalogs can be used to train or refine detection algorithms to better accommodate man-made events with quake- or tremorlike features. They can also be used to identify the sources of unknown events and potentially motivate new applications in urban and environmental seismology. Because our workflow is based on a single-station approach, we acknowledge that our simple data set is limited in its attributes and capacity. We cannot provide reliable source locations for our events. In the future, we hope to install a small network of sensors in the same environment and apply our workflow to create a robust catalog of located events. We plan to add these located events to the Exotic Seismic Events Catalog (Bahavar *et al.*, 2019) for the community to easily access and visualize.

Conclusions

To explore the seismic landscape of a unique urban and industrial environment of the Chicago area, we developed a workflow to detect and cluster anomalous events in continuous data from a single broadband seismometer. Our workflow uses a PSD misfit detector and a *k*-means clustering model for detecting and clustering, respectively. The workflow utilizes simple and intuitive features, which can be applied to other built environments. When applied to 2 yr of continuous single-station data in the Chicago area, we detected and clustered over 650,000 windows of anomalous events. With one exception, our workflow produced coherent

clusters of events with similar waveforms and likely the same source. One cluster needed to be manually subdivided into individual subclusters for event characterization. By analyzing the temporal and spectral features of each cluster and subcluster, we successfully identified several sources for our detections, including industrial machinery operations, surface quarry blasts, underground blasts, and earthquake-like signals that are potentially wind-generated. Using our results, we created a labeled data set of 1262 anomalous events. Expanding seismic catalogs to include dynamic nontectonic events is critical for improving earthquake detection and discrimination in noisy built environments. Our study demonstrates how a simple two-stage workflow can be used to efficiently mine a noisy multiyear data set and build a labeled catalog of dynamic nontectonic events.

Data and Resources

The code used in this study is available on GitHub at https://github.com/am-thomas/SeismoExplore-Chicago and has been archived on Zenodo at https://zenodo.org/records/15047293. Both websites were last updated on March 2025. Metadata and waveforms of their labeled data set can be accessed on Zenodo at https://zenodo.org/records/15047874 (last modified May 2025). HQIL waveforms and metadata are publicly available; they were accessed using the facilities of EarthScope Consortium. These services are funded through the National Science Foundation's Seismological Facility for the Advancement of Geoscience (SAGE) Award under Cooperative Agreement EAR-1724509. Their code is written in Python and utilizes tools from NumPy (Harris *et al.*, 2020), SciPy (Virtanen *et al.*, 2020), ObsPy (Beyreuther *et al.*, 2010), and Scikit-learn (Pedregosa *et al.*, 2011). Visualizations were created using Matplotlib (Hunter, 2021) and Seaborn (Waskom, 2021).

Declaration of Competing Interests

The authors acknowledge that there are no conflicts of interest recorded.

Acknowledgments

The authors would like to express their gratitude to the local quarry representatives for allowing the broadband station HQIL to be installed on their land and for valuable discussion. The authors also thank the EarthScope Primary Instrument Center (EPIC, formerly The Incorporated Research Institutions for Seismology [IRIS] Program for the Array Seismic Studies of the Continental Lithosphere [PASSCAL] Instrument Center) for loaning auxiliary and primary instrumentation. For their analysis, the authors thank Christopher W. Johnson for his feedback during the Fall 2024 Meeting of the American Geophysical Union, particularly in suggesting that some of their unknown earthquake-like events may be wind-generated. The authors are also grateful to Caio Ciardelli for his helpful feedback and suggestions during the development of their workflow and presentation of results. This research was supported in part through the computational resources and staff contributions provided for the Quest high-performance computing facility at Northwestern University which is jointly supported by the Office of the Provost, the Office for Research, and Northwestern University Information Technology.

References

Allen, R. V. (1978). Automatic earthquake recognition and timing from single traces, *Bull. Seismol. Soc. Am.* **68,** no. 5, 1521–1532, doi: 10.1785/BSSA0680051521.

Bahavar, M., K. E. Allstadt, M. Van Fossen, S. D. Malone, and C. Trabant (2019). Exotic seismic events catalog (ESEC) data product, *Seismol. Res. Lett.* **90,** no. 3, 1355–1363, doi: 10.1785/0220180402.

Beyreuther, M., R. Barsch, L. Krischer, T. Megies, Y. Behr, and J. Wassermann (2010). ObsPy: A python toolbox for seismology, *Seismol. Res. Lett.* **81,** no. 3, 530–533, doi: 10.1785/gssrl.81.3.530.

Boese, C. M., L. Wotherspoon, M. Alvarez, and P. Malin (2015). Analysis of anthropogenic and natural noise from multilevel borehole seismometers in an urban environment, Auckland, New Zealand, *Bull. Seismol. Soc. Am.* **105,** no. 1, 285–299, doi: 10.1785/0120130288.

Caplan-Auerbach, J., K. Marczewski, and G. S. Bullock (2024). Beast quake (Taylor's version): Analysis of seismic signals recorded during two Taylor swift concerts in Seattle, July 2023, *GSA Today* **34,** 4–10, doi: 10.1130/GSATG589A.1.

Campello, R. J. G. B., D. Moulavi, and J. Sander (2013). Density-based clustering based on hierarchical density estimates, in *Advances in Knowledge Discovery and Data Mining*, J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu (Editors), Springer, Berlin, Heidelberg, 160–172, ISBN 978-3-642-37456-2, doi: 10.1007/978-3-642-37456-2_14.

Chai, C., O. Marcillo, M. Maceira, J. Park, S. Arrowsmith, J. O. Thomas, and J. Cunningham (2025). Exploring continuous seismic data at an industry facility using unsupervised machine learning, *Seism. Rec.* **5,** no. 1, 64–72, doi: 10.1785/0320240046.

Díaz, J., M. Ruiz, P. S. Sánchez-Pastor, and P. Romero (2017). Urban seismology: On the origin of earth vibrations within a city, *Sci. Rep.* **7,** no. 1, 15296, doi: 10.1038/s41598-017-15499-y.

Díaz, J., M. Schimmel, M. Ruiz, and R. Carbonell (2020). Seismometers within cities: A tool to connect earth sciences and society, *Front. Earth Sci.* **8,** doi: 10.3389/feart.2020.00009.

Green, D. N., I. D. Bastow, B. Dashwood, and S. E. J. Nippress (2016). Characterizing broadband seismic noise in central London, *Seismol. Res. Lett.* **88,** no. 1, 113–124, doi: 10.1785/0220160128.

Groos, J. C., and J. R. R. Ritter (2009). Time domain classification and quantification of seismic noise in an urban environment, *Geophys. J. Int.* **179,** no. 2, 1213–1231, doi: 10.1111/j.1365-246X.2009.04343.x.

Guenaga, D. L., C. Chai, M. Maceira, O. E. Marcillo, and A. A. Velasco (2021). Seismically detecting nuclear reactor operations using a power spectral density (PSD) misfit detector, *Bull. Seismol. Soc. Am.* **111,** no. 3, 1378–1391, doi: 10.1785/0120200267.

Harris, C. R., K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, *et al.* (2020). Array programming with NumPy, *Nature* **585,** no. 7825, 357–362, doi: 10.1038/s41586-020-2649-2.

Hunter, J. D. (2021). Matplotlib: A 2d graphics environment, *Comput. Sci. Eng.* **9,** no. 3, 90–95, doi: 10.1109/MCSE.2007.55.

Jakkampudi, S., J. Shen, W. Li, A. Dev, T. Zhu, and E. R. Martin (2020). Footstep detection in urban seismic data with a convolutional neural network, *The Leading Edge* **39,** no. 9, 654–660, doi: 10.1190/tle39090654.1.

Johnson, C. W., Y. Ben-Zion, H. Meng, and F. Vernon (2020). Identifying different classes of seismic noise signals using

unsupervised learning, *Geophys. Res. Lett.* **47,** no. 15, e2020GL088353, doi: 10.1029/2020GL088353.

Johnson, C. W., H. Meng, F. Vernon, and Y. Ben-Zion (2019). Characteristics of ground motion generated by wind interaction with trees, structures, and other surface obstacles, *J. Geophys. Res.* **124,** no. 8, 8519–8539, doi: 10.1029/2018JB017151.

Karasözen, E., and M. E. West (2022). An adaptive spectral subtraction algorithm to remove persistent cultural noise, *Bull. Seismol. Soc. Am.* **112,** no. 5, 2297–2311, doi: 10.1785/0120210317.

Li, C., Z. Li, Z. Peng, C. Zhang, N. Nakata, and T. Sickbert (2018). Long-period long-duration events detected by the IRIS community wavefield demonstration experiment in Oklahoma: Tremor or train signals? *Seismol. Res. Lett.* **89,** no. 5, 1652–1659, doi: 10.1785/0220180081.

Li, Y. E., E. A. Nilot, Y. Zhao, and G. Fang (2023). Quantifying urban activities using nodal seismometers in a heterogeneous urban space, *Sensors* **23,** no. 3, 1322, doi: 10.3390/s23031322.

Lloyd, S. (1982). Least squares quantization in PCM, *IEEE Trans. Inf. Theory* **28,** no. 2, 129–137, doi: 10.1109/TIT.1982.1056489.

Maher, S. P., M. E. Glasgow, E. S. Cochran, and Z. Peng (2024). Distinguishing natural sources from anthropogenic events in seismic data, *Seismol. Res. Lett.* **96,** no. 1, 1–6, doi: 10.1785/0220240330.

McNamara, D., and R. Buland (2004). Ambient noise levels in the continental United States, *Bull. Seismol. Soc. Am.* **94,** 1517–1527, doi: 10.1785/012003001.

Mousavi, S. M., and G. C. Beroza (2023). Machine learning in earthquake seismology, *Annu. Rev. Earth Planet. Sci.* **51,** 105–129, doi: 10.1146/annurev-earth-071822-100323.

Mousavi, S. M, W. L. Ellsworth, W. Zhu, L. Y. Chuang, and G. C. Beroza (2020). Earthquake transformer—An attentive deep-learning model for simultaneous earthquake detection and phase picking, *Nat. Commun.* **11,** no. 1, 3952, doi: 10.1038/s41467-020-17591-w.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.* (2011). Scikit-learn: Machine learning in Python, *J. Machine Learn. Res.* **12,** 2825–2830, doi: 10.48550/arXiv.1201.0490.

Quiros, D. A., L. D. Brown, and D. Kim (2016). Seismic interferometry of railroad induced ground motions: body and surface wave imaging, *Geophys. J. Int.* **205,** no. 1, 301–313, doi: 10.1093/gji/ggw033.

Riahi, N., and P. Gerstoft (2015). The seismic traffic footprint: Tracking trains, aircraft, and cars seismically, *Geophys. Res. Lett.* **42,** no. 8, 2674–2681, doi: 10.1002/2015GL063558.

Ritter, J. R. R., S. F. Balan, K.-P. Bonjer, T. Diehl, T. Forbriger, G. Mărmureanu, F. Wenzel, and W. Wirth (2005). Broadband urban seismology in the Bucharest metropolitan area, *Seismol. Res. Lett.* **76,** no. 5, 574–580, doi: 10.1785/gssrl.76.5.574.

Saadia, B., and G. Fotopoulos (2023). Unsupervised clustering of ambient seismic noise in an urban environment, *Comput. Geosci.* **179,** 105432, doi: 10.1016/j.cageo.2023.105432.

Snover, D., C. W. Johnson, M. J. Bianco, and P. Gerstoft (2020). Deep clustering to identify sources of urban seismic noise in Long Beach, California, *Seismol. Res. Lett.* **92,** no. 2, 1011–1022, doi: 10.1785/0220200164.

Thomas, A. M., O. Ranadive, and S. van der Lee (2023). Towards detecting small, local earthquakes in greater Chicago using single-station data, *AGU Fall Meeting Abstracts* (S31D–0377).

Vaezi, Y., and M. Van der Baan (2015). Comparison of the STA/LTA and power spectral density methods for microseismic event detection, *Geophys. J. Int.* **203,** no. 3, 1896–1908, doi: 10.1093/gji/ggv419.

Virtanen, P., R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, *et al.* (2020). SciPy 1.0: Fundamental algorithms for scientific computing in python, *Nat. Methods* **17,** no. 3, 261–272, doi: 10.1038/s41592-019-0686-2.

Waskom, M. L. (2021). Seaborn: statistical data visualization, *J. Open Source Software* **6,** no. 60, 3021, doi: 10.21105/joss.03021.

Watt, J., R. Borhani, and A. K. Katsaggelos (2016). *Machine Learning Refined: Foundations, Algorithms, and Applications*, Cambridge University Press, First Ed., ISBN 978-1-107-12352-6.

Welch, P. (1967). The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms, *IEEE Trans. Audio Electroacoust.* **15,** no. 2, 70–73, doi: 10.1109/TAU.1967.1161901.

Yang, L., X. Liu, W. Zhu, L. Zhao, and G. C. Beroza (2022). Toward improved urban earthquake monitoring through deep-learning-based noise suppression, *Sci. Adv.* **8,** no. 15, eabl3564, doi: 10.1126/sciadv.abl3564.

Yuan, C., and H. Yang (2019). Research on k-value selection method of k-means clustering algorithm, *Multidisciplinary Scientific J.* **2,** no. 2, 226–235, doi: 10.3390/j2020016.

Zhang, Y., Y. E. Li, H. Zhang, and T. Ku (2019). Near-surface site investigation by seismic interferometry using urban traffic noise in Singapore, *Geophysics* **84,** no. 2, B169–B180, doi: 10.1190/geo2017-0798.1.

Manuscript received 26 March 2025
Published online 1 August 2025