UNIVERSITY OF CALIFORNIA SAN DIEGO

Urban seismic noise identified with deep embedded clustering using a dense array in Long Beach, CA

A thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Earth Sciences

by

Dylan Snover

Committee in charge:

Peter Gerstoft, Chair
Duncan Agnew
Christopher Johnson
Peter Shearer

2020

The thesis of Dylan Snover is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

Chair

University of California San Diego

2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTAL VIDEOS

S1. Timelapse video showing the cluster assignments for each station in the array during a vibroseismic truck survey on March 6, 2011 at 11:04:00 for 40 s. The vibroseis trucks are located in the northeast quadrant of the array near the Long Beach Airport

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor, Peter Gerstoft, who gave me the opportunity to work on this project. I would like to thank Christopher Johnson who provided guidance and insight throughout my research. I wish to acknowledge the help provided by Michael Bianco in finalizing my research.

I also wish to extend my special thanks to my family and friends who provided endless support over the past two years. Lastly, I'd like the thank my friends at IGPP and SIO in general for all the good times we shared.

All chapters of this thesis, including the abstract and introduction, are coauthored by Johnson, Christopher W., Bianco, Michael J., and Gerstoft, Peter. The content of this thesis has been submitted for publication of the material as it may appear in Snover, D., Johnson, C. W., Bianco, M. J.; Gertsoft, P. (2020). Deep clustering to identify sources of urban seismic noise in Long Beach, CA. Seismologic Research Letters. The thesis author was the primary investigator and author of this material.

ABSTRACT OF THE THESIS

Urban seismic noise identified with deep embedded clustering using a dense array in Long Beach, CA

by

Dylan Snover

Master of Science in Earth Sciences

University of California San Diego, 2020

Peter Gerstoft, Chair

Ambient seismic noise consists of emergent and impulsive signals generated by natural and anthropogenic sources. Developing techniques to identify specific cultural noise signals will benefit studies performing seismic imaging from continuous records. We examine spectrograms of urban cultural noise from a spatially dense seismic array located in Long Beach, California. The spectral features of the waveforms are used to develop a self-supervised clustering model for differentiating cultural noise into separable types of signals. We use 161 hours of seismic data from 5200 geophones that contain impulsive signals originating from human activity. The model uses convolutional autoencoders, a self-supervised machine learning technique, to learn

latent features from spectrograms produced from the data. The latent features are evaluated using a deep clustering algorithm to separate the noise signals into different classes. We evaluate the separation of data and analyze the classes to identify the likely sources of the signals present in the data. To interpret the model performance we examine the time-frequency domain features of the signals and the spatiotemporal evolution observed for each class. We demonstrate autoencoder feature extraction used with probabilistic clustering algorithms is a useful approach to characterize seismic noise and identify signals in the data with a low signal-to-noise ratio.

INTRODUCTION

Urban seismic recordings contain human-generated noise that reduces the signal-to-noise ratio of naturally occurring tectonic signals. Cultural (i.e., anthropogenic) noise decreases the ability to detect microseismic events that contain valuable information about fault mechanics (Inbal, Clayton, et al., 2015; Inbal, Cristea-Platon, et al., 2018). The ability to identify and locate sources of microseismicity is limited without the proper characterization of emergent and impulsive noise signals that are common in continuous seismic data (Meng, Ben-Zion, and Johnson, 2019). Recent advances in machine learning have enabled researchers to detect local and regional earthquakes with a signal to noise ratio near 1 (e.g. Mousavi, Zhu, Ellsworth, et al., 2019; Ross, Meier, Hauksson, and Heaton, 2018). In densely populated urban environments the signal-to-noise ratio is typically not high enough to detect the weakest microseismicity (Inbal, Cristea-Platon, et al., 2018). Methods to properly characterize cultural noise are needed to further describe the entire seismic record and detect more low magnitude tectonic events.

Advances in seismic data collection and the availability of open source machine learning software are producing novel approaches to solve classical seismology problems (Bergen et al., 2019; Kong et al., 2019). Notable are techniques for identifying waveform phase arrivals (Ross, Meier, Hauksson, and Heaton, 2018), event localization and detection (Mousavi, Zhu, Sheng, et al., 2019; Perol et al., 2018), earthquake source discrimination (Li, Meier, et al., 2018; Mousavi, Zhu, Ellsworth, et al., 2019), seismic tomography (Bianco and Gerstoft, 2018; Bianco, Gerstoft, et al., 2019), first-motion polarity (Ross, Meier, and Hauksson, 2018), and magnitude estimates (Mousavi and Beroza, 2020). Along with these applications, advances in seismic noise characterization have improved discriminative capabilities and increased catalogue of ambient noise sources (Brenguier et al., 2019; Chamarczuk et al., 2019; Diaz et al., 2017; Groos and Ritter, 2009; Meng and Ben-Zion, 2018a; Meng, Ben-Zion, and Johnson, 2019). Riahi and Gerstoft (2015) analyze spatiotemporal features of seismic noise to identify railroad, airport, and highway traffic. Riahi and Gerstoft (2017) explore graph clustering to classify seismic noise

signatures of helicopters and oil production facilities. Meng and Ben-Zion (2018b) observe frequent air traffic events with Doppler effects corresponding to airplanes and helicopters passing above a dense seismic array. Johnson, Meng, et al. (2019) use dense array data to characterize seismic signals generated by wind interacting with trees and structures located on the surface. Johnson, Vernon, et al. (2019) investigate the contribution of atmospheric processes to the frequency modulation of instrumental noise in geophones. Applying machine learning techniques will help further characterize seismic noise sources and improve our ability to identify them in continuous seismic recordings.

We utilize an unsupervised deep learning technique called deep embedded clustering (DEC) to characterize noise in a dense urban environment. We use DEC to separate cultural noise signals into groups based on salient features extracted from spectrograms of the seismic data. Differentiating cultural noise signatures is cast as a clustering problem. Clustering the raw spectrograms is computationally expensive and ineffective due to the high dimensional feature space (Dizaji et al., 2017; Guo et al., 2017; Mousavi, Zhu, Ellsworth, et al., 2019). The DEC approach uses convolutional autoencoders to map high-dimensional inputs to a lower dimensional latent representation using an encoder and decoder model. During training, the encoder extracts salient features of the spectrograms by compressing the time-frequency domain input to the latent space. The decoder reconstructs the input spectrogram using the model-learned latent features that generalize the data (Guo et al., 2017). After training the CAE model, the latent space samples are clustered with the goal of associating the groups with physical sources in the urban environment. We seek to distinguish between 3 primary classes of anthropogenic noise recorded by a dense seismic array in Long Beach, California: (1) Vibrator truck signals that were generated actively as part as a surface wave tomography survey, (2) vehicle traffic generated along major interstates within the array, and (3) incoming and departing air traffic from the Long Beach Airport.

Chapter 1

Study Area

The Long Beach array was deployed for an oil and gas survey in Long Beach, California, USA. The array consisted of ˜5200 vertical component geophones arranged in a 7x10 km grid with an average inter-geophone spacing of 120 m (Fig. 1.1). The instruments recorded vertical ground motion continuously at 500 Hz from January to June 2011. The data were resampled to 250 Hz during postprocessing. Long Beach, California is located in south Los Angeles County with a population of 450,000 residents and contains dense commercial and industrial activity including several major highways, an airport, and an actively producing oil field. The Port of Long Beach is a busy container hub, hosting a coal export terminal, an inter-modal freight transportation station, and a naval shipyard. The cultural noise sources, along with other non-seismic noise sources, such as wind and instrument noise, contribute to the near continuous noise signals in the data. In addition to the sources of cultural noise, the data contains vibrator truck signals that occur in 1-min intervals during daytime hours between 6–9 March for an active source tomography survey. The vibroseismic signal is a linear frequency modulated sweep from 20–80 Hz over a span of 40 s. The array data is used in a variety of studies including seismic tomography studies (Bianco, Gerstoft, et al., 2019; Lin et al., 2013; Nakata et al., 2015), ground motion and event detection studies (Inbal, Ampuero, et al., 2016; Inbal, Clayton, et al., 2015; Li, Peng, et al., 2018; Schmandt

and Clayton, 2013), and ambient noise analysis (Bowden et al., 2015; Riahi and Gerstoft, 2015; Riahi and Gerstoft, 2017).

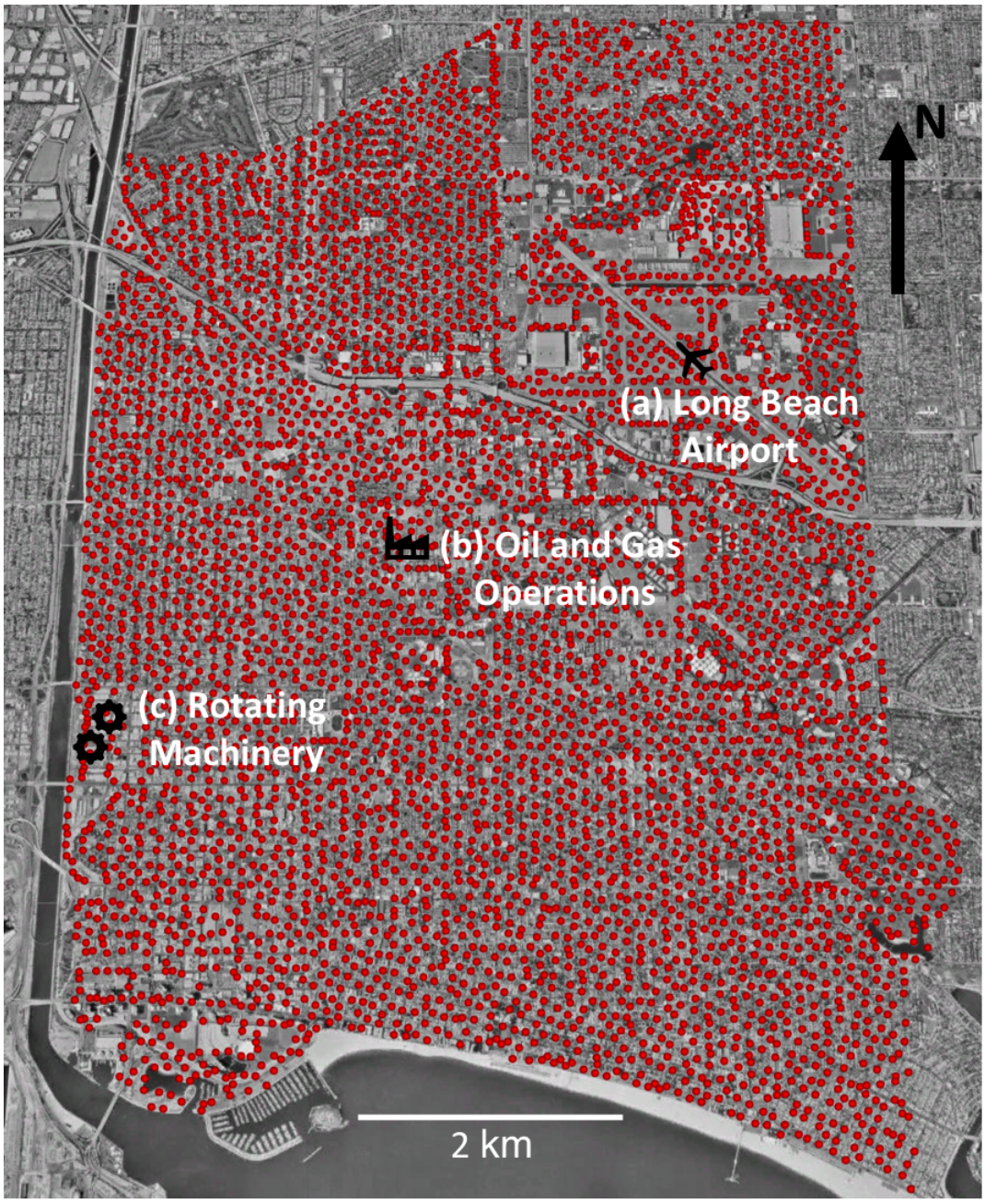

Figure 1.1: Dense seismic array in Long Beach, California shown on satellite imagery (GoogleEarth v7.3.2.5776 accessed 6 April 2020). Shown are 5341 geophones location in the 7×10 km study area. The locations of (a) The Long Beach Airport and (b) active oil and gas operations and (c) rotating machinery are shown for reference.

Chapter 2

Seismic Data

The data are continuous seismic waveforms from all operating geophones between 6 March 2011 00:00:00 UTC to 12 March 2011 16:59:59 UTC for a total of 161 hours. Training data are assembled by randomly selecting stations, days, and 1-hour intervals between 12:00:00 and 23:00:00 on 7–12 March 2011, excluding 6 March 2011. We include more stations by using only the first 10 min of each hour, which increases the spatial coverage and signal diversity. Each 10 min interval is windowed into 4 s slices with 3 s overlap, which gives 597 waveforms. The amplitude of each 4 s waveform is normalized to ± 1 and a Short-time Fourier transform is applied using Kaiser windowing with a length of 0.5 s (125 samples) and 90% overlap. The resulting spectrograms contain 5,166 spectral amplitude features discretized by 126 frequency and 41 time intervals. The procedure is repeated to produce ˜2,000,000 spectrograms for training. The number of training examples is selected to maximize signal diversity while limiting the training data to 100 GB considering the available computational resources. The spectrograms are split into training (80%) and validation (20%). One full day, 6 March 2011, is excluded from the training and used in testing; the day includes the vibrator trucks generating surface wave signals between 6:00:00 and 13:00:00 (Li, Peng, et al., 2018). The test data contains the recordings between 11:00:00–11:10:00 for every geophone to include all locations. The training and test data were

standardized to ensure the spectrograms have comparable scales to effectively train and test the model coefficients. Using all spectrograms, we standardize each spectral amplitude (each pixel in the input spectrogram) by removing the mean and dividing by the standard deviation based on that pixel's index in all spectrograms. The result is an amplitude spectrum with unit variance for each frequency at each time interval. The spectrograms are cropped to contain 4800 (originally 5166) features that include 120 frequency bins (1–120 Hz) and 40 time bins (4 s). The higher frequencies, 121–126 Hz, contain lower amplitudes and the last time interval is redundant in the overlapping windows. Cropping the image to an even size ensures the CAE model reconstruction can produce the same dimensions as the input.

Chapter 3

Seismic noise cluster analysis

The model training consists of two optimization procedures: (1) train the CAE model and (2) train the DEC model that incorporates a clustering layer and continues training the CAE in parallel. While training step (2), the latent space from the encoder is passed to both the CAE decoder and the clustering layer to reconstruct the input and generate cluster assignments. In step (2) both the CAE weights and cluster centroids are updated after each epoch. The initial training of the CAE in step (1) develops a reconstruction model optimizing salient features. Assuming the initial latent space obtained from step (1) is well approximated and the network is well trained, the CAE weights are further refined using the clustering layer in step (2).

3.1 Convolutional Autoencoders

Autoencoding (Goodfellow et al., 2016) is a learning technique that compresses the input to a reduced dimensional representation and attempts to reconstruct the input with the highest accuracy. The design is an unsupervised technique to extract key features of the input. CAEs utilize convolutional operations to produce a compressed representation of the input. The layers that reduce the dimension to a latent space is the encoder. The decoder performs the reverse process, and reconstructs the input. The encoder consists of four convolution layers followed by a

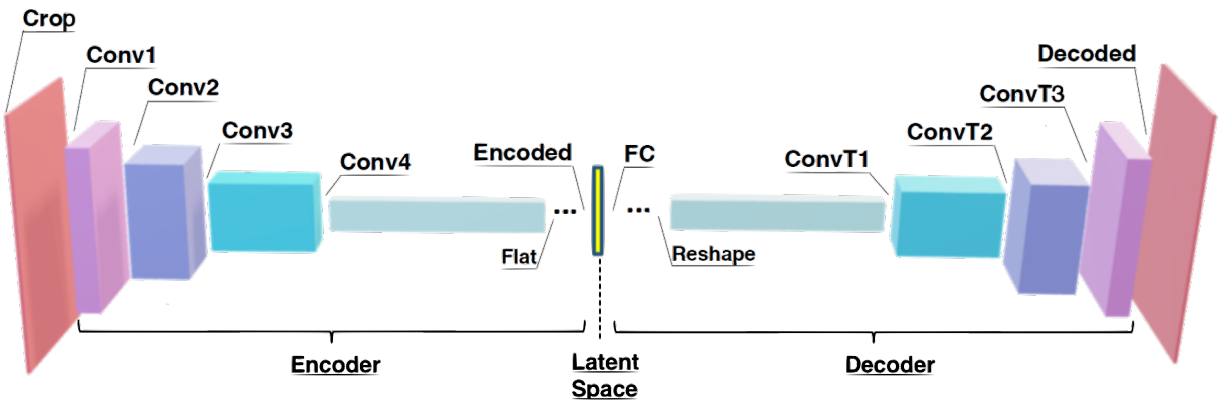

Figure 3.1: The convolutional autoencoder (CAE) architecture shown with the layer output dimensions are shown as [height, width, depth]. The encoder, which compresses the input images to the 14 dimensional latent space. The weights of the fully connected neural network are the input to the clustering layer. The decoder reconstructs the input from the latent features. A summary of the parameters used in the CAE architecture are in Table 3.1.

fully connected neural network to map the features to a 14 variable latent space (Fig. 3.1 , Table 3.1). The decoder is a fully connected neural network followed by four transposed convolutional layers to return to the original input dimensions. For dimension reduction we find more amplitude information is retained when using a stride of 2 for the convolution operations which reduces the dimensions by 75% in each layer, rather than applying a max-pooling operation (see Table 3.1). We use a rectified linear unit (ReLU) activation function for the convolutional and fully connected layers. The final reconstruction layer uses a linear activation function to retrain spectral amplitudes. The model design has 87,247 trainable parameters (Table 3.1). The reconstruction loss function is the mean-squared-error (MSE) between the input and reconstruction. The Adam optimizer is used for training with learning rate $3\text{x}10^{-4}$ and batch size 512 samples. The CAE is initially trained for 155 epochs. To prevent over-fitting the training is stopped early when the validation loss does not improve for 10 consecutive epochs. The best model is used in the subsequent cluster analysis.

Table 3.1: CAE architecture. The input and output shapes of each layer are given in the format specified by the [height, width, depth] of the feature space or in the format [number of features] as in the layers Flatten, Encoded, FC, Reshape. The kernel size specifies the shape of the two dimension filters applied to the image in the format [height, width].

Layer Name	Layer Type	Input shape	# Filters	Kernel Size	Stride	Activation	Output Shape	Trainable Parameters
Crop	Cropping	[126, 41, 1]	-	-	-	-	[120. 40, 1]	0
Conv1	Convolution	[120, 40, 1]	8	[5, 5]	2	ReLU	[60, 20, 8]	208
Conv2	Convolution	[60, 20, 8]	16	[5, 5]	2	ReLU	[30, 10, 16]	3,216
Conv3	Convolution	[30, 10, 16]	32	[3, 3]	2	ReLU	[15, 5, 32]	4,640
Conv4	Convolution	[15, 5, 32]	64	[3, 3]	2	ReLU	[7, 2, 64]	18,496
Flat	Flatten	[7, 2, 64]	-	-	-	-	[896]	0
Encoded	Fully Connected	[896]	-	-	-	ReLU	[14]	12,558
FC	Fully Connected	[14]	-	-	-	ReLU	[896]	13,440
Reshape	Reshape	[896]	-	-	-	-	[7, 2, 64]	0
ConvT1	Transposed Conv.	[7, 2, 64]	32	[3, 3]	2	ReLU	[15, 5, 32]	18,464
ConvT2	Transposed Conv.	[15, 5, 32]	16	[5, 5]	2	ReLU	[30, 10, 16]	12,816
ConvT3	Transposed Conv.	[30, 10, 16]	8	[5, 5]	2	ReLU	[60, 20, 8]	3,208
Decoded	Transposed Conv.	[60, 20, 8]	1	[5, 5]	2	Linear	[120,40,1]	201
							Total Trainable Parameters	**87,247**

3.2 Deep embedded cluster analysis

A Deep Embedded Clustering (DEC) model (Guo et al., 2017) is implemented to cluster the seismic spectrograms. The full model architecture is designed with a CAE backbone. The CAE encoder maps the spectrograms to the latent space. The latent space is used as input by two model components, the decoder and clustering layer. The DEC is trained using both the reconstruction loss contributed by the CAE and the clustering loss contributed by an additional clustering layer (see Eq. 3.5). During the DEC training routine, the output of the encoder is passed to both the CAE decoder and the clustering layer. DEC uses a weighted sum of the reconstruction error and clustering error as a loss function (Dizaji et al., 2017; Guo et al., 2017; Xie et al., 2016).

Clustering is an unsupervised learning method to separate data into a specified number of groups using features of the data. We implement the K-means algorithm (Lloyd, 1982) to identify clusters in the CAE latent space. The K-means algorithm takes N observations of a D-dimensional variable and separates the observations into K clusters. The first step is to randomly initialize K cluster centroids, μ_k, in D-dimensional space. The nearest cluster centroid (in Euclidian space)

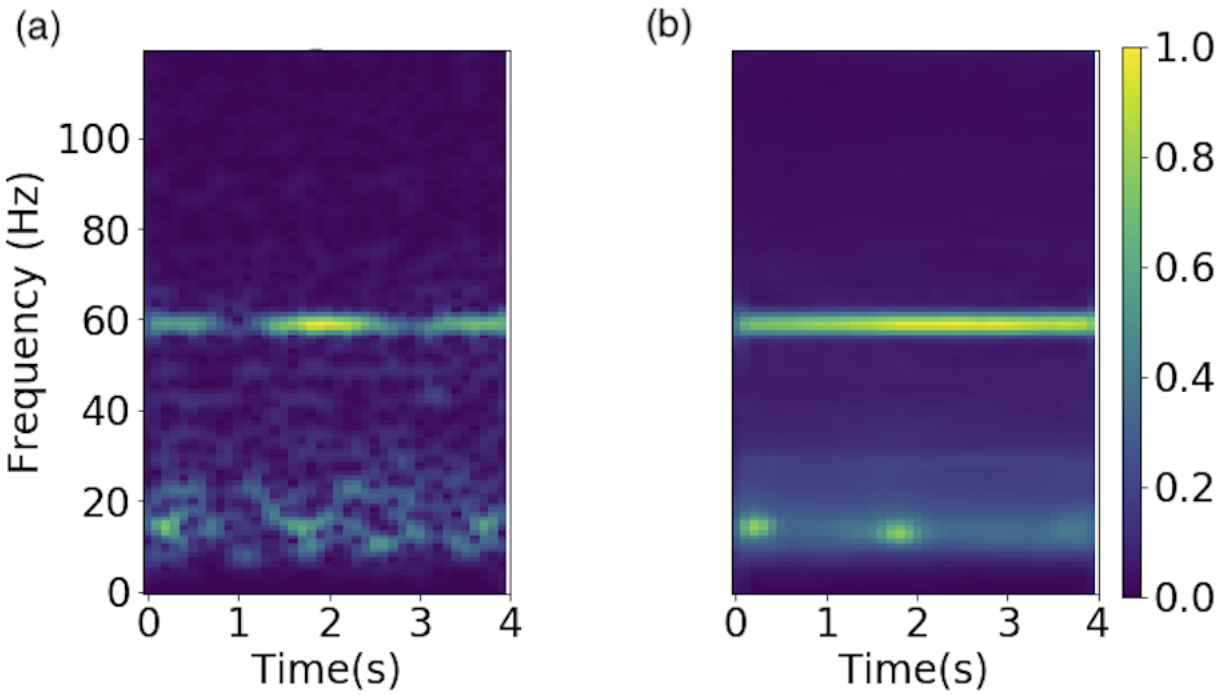

Figure 3.2: A reconstructed spectrogram example using the CAE training before the DEC training. (a) The input spectrogram contains 4800 features in 120 frequency and 40 time bins. (b) The reconstruction generated by the CAE after compressing to 14 latent features and exhibits some feature smoothing but demonstrates that the CAE preserves the structure and amplitude of the input.

is identified for each of the $N = \sim 2,000,000$ latent samples. The method uses hard assignment classifications, i.e. each example is assigned to only one cluster. The centroid positions are updated iteratively to minimize the sum of squares distance within each cluster (Eq. 3.1). Over 100 trials the model with the lowest sum of squares error is chosen as the optimal K-means model. The sum-of-squares distance, J is calculated as

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||\mathbf{x}_n - \boldsymbol{\mu}_k||^2 \tag{3.1}$$

here $\mathbf{x}_n$ is the nth observation of the variable $\mathbf{x}$ in $D = 14$ dimensional space, $\boldsymbol{\mu}_k$ is the kth centroid (also in $D = 14$ dimensions), and r_{nk} is a binary indicator that specifies which observations are assigned to a specific cluster.

The number of clusters is chosen by compromising between the number of clusters and the total sum of squares error. We varied the number of clusters, K, between 2–20 and found a smaller rate of decrease for $K > 5$ clusters. We chose to cluster the data into $K = 5$ classes when training the DEC that account for the 3 primary signals we attempt to identify (vibroseismic, highway traffic, and airport traffic) and two additional classes for unknown sources.

The K-means algorithm for $K = 5$ clusters is applied to the latent space data to establish the initial centroid positions from the initial CAE training, i.e. the coordinates of each cluster center in the feature space. The clustering layer in the DEC model is initialized with these centroids. The DEC assumes that the initial centroid locations are close to optimal and refines the centroid positions and the distribution of data assigned to each cluster. The algorithm minimizes the KL Divergence between the target probability distribution P and the sampled probability distribution Q sampled across observations n and clusters k with values p_{nk} and q_{nk}. Ideally, the target distribution P should be a sum of delta functions at the true unknown cluster centroids positions $\boldsymbol{\mu}_k$, but this is unknown and an empirical target distribution is used (Xie et al., 2016). The KL divergence (Eq. 3.2), is used as the loss function for the clustering layer in the DEC model (Guo et al., 2017; Xie et al., 2016).

$$L_C = KL(P||Q) = \sum_{n}^{N}\sum_{k}^{K} p_{nk} \log\left(\frac{p_{nk}}{q_{nk}}\right), \tag{3.2}$$

where q_{nk} is the sampled distribution, p_{nk} is the target distribution, $N = 2,000,000$, and $K = 5$.

The sampled distribution is approximated as the Student's t-distribution and represents the probability that a latent space sample will be assigned to a cluster (Maaten and Hinton, 2008; Xie et al., 2016)

$$q_{nk} = \frac{(1 + ||\mathbf{x}_n - \boldsymbol{\mu}_k||^2)^{-1}}{\sum_k (1 + ||\mathbf{x}_n - \boldsymbol{\mu}_k||^2)^{-1}}. \tag{3.3}$$

The target distribution is generated by weighting the high-confidence samples found in (3.3), i.e. those with high probability of assignment to a certain cluster, more than low confidence

samples.

$$p_{nk} = \frac{q_{nk}^2 / \sum_n q_{nk}}{\sum_k \left(q_{nk}^2 / \sum_n q_{nk} \right)}. \tag{3.4}$$

In an optimal model, all samples in a cluster are high-confidence assignments and the target distribution would be a delta function at $\boldsymbol{\mu}_k$. Based on experiments, however, (Eq. 3.4) has been found to be more robust (Xie et al., 2016). During training the DEC updates the encoder weights, the decoder weights, and the cluster centroid positions. By iteratively updating the parameters, the DEC maximizes the separability between clusters to improve the distribution of data assigned to each cluster centroid (Guo et al., 2017; Xie et al., 2016).

Training the clustering layer of the DEC algorithm is a fine-tuning process that balances the contribution of the clustering loss (Eq. 3.2) with the reconstruction loss(Guo et al., 2017; Mrabah et al., 2019). The CAE and the clustering algorithm are jointly trained using a weighted sum of the reconstruction loss and the clustering loss (Eq. 3.5). We train the DEC model for 50 epochs, updating the soft assignment distribution, the target distribution, and the total loss function 10 times per epoch.

$$L = L_R + \lambda L_C \tag{3.5}$$

Where L_R is the MSE loss function, L_C is the KL-Divergence (3.2), and λ is the regularization parameter. The total loss function preserves the encoders ability to extract salient features. The regularization term helps prevent overfitting by ensuring the latent space correctly reconstructs samples (Chazan et al., 2018; Mrabah et al., 2019). We use $\lambda = .11$ to balance the contributions of the reconstruction loss and the clustering loss to the total loss function. This ensures the clustering loss does not distort the latent space features.

Chapter 4

Results

4.1 CAE and DEC model training

A well trained CAE will produce reconstructions in which most of the amplitude information is retained and the local structure of the original image is preserved (i.e. signals occur in the same location in the reconstructions). The trained CAE has a MSE loss of 0.25. The results from the test data are evaluated by visual inspection of the reconstructed spectrograms (Fig. 3.2). The results show feature smoothing in the reconstructions, which is expected from lossy compression when reconstructing a 4800-dimensional image from a 14-dimensional latent space. For all spectrograms inspected we find most of the amplitude information is correctly represented and the local structure is well preserved. Training the DEC further refines the CAE model weights while assigning a cluster label. Similar reconstruction performance is observed after the DEC model has been trained and applied to unseen test data. Fig. 4.1 shows the DEC model reconstructions for multiple types of signals in the test data. All characteristics from the original images are observed in the same location in the reconstructions suggesting the latent space is successfully generalizing the inputs to a set of unique latent space variables.

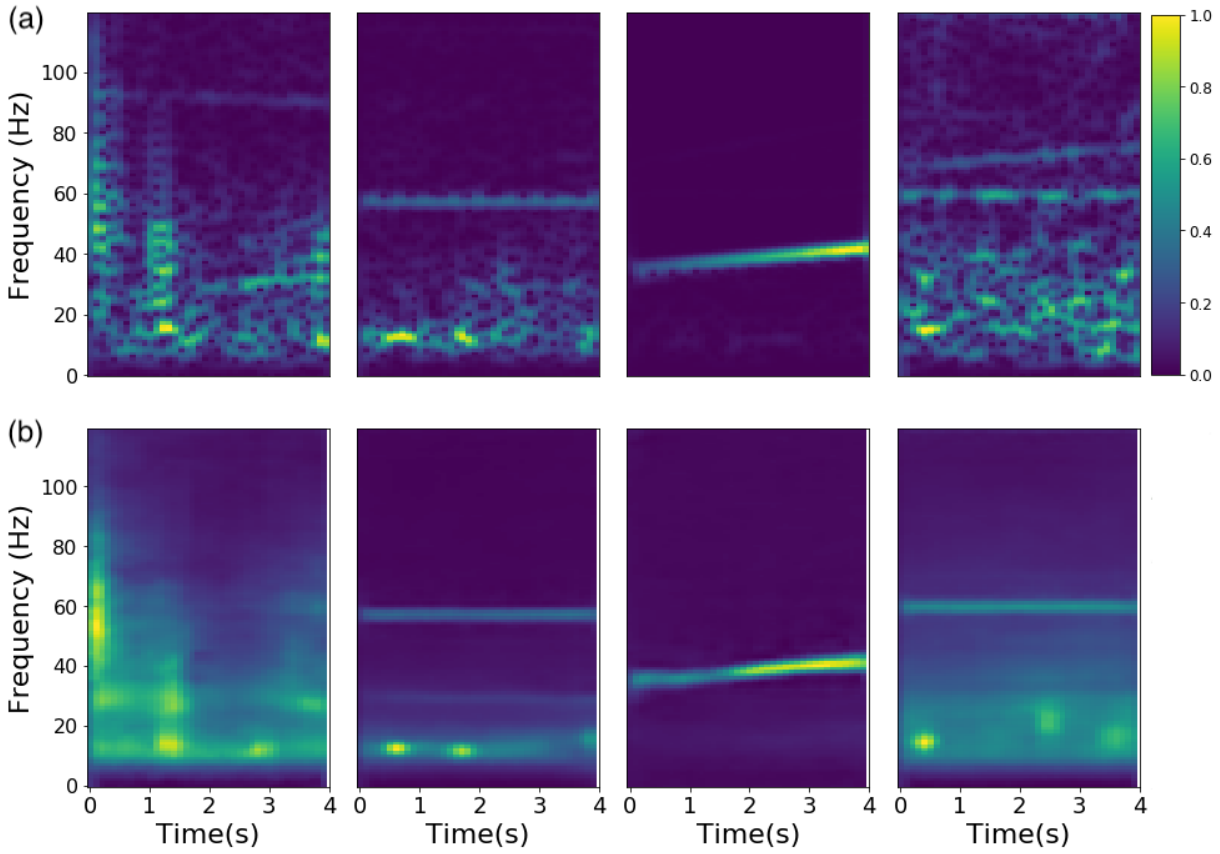

Figure 4.1: Spectrogram reconstructions after DEC model training. (a) Four input spectrograms each containing different types of noise signals in the data. The amplitudes are normalized to the range [0,1]. (b) The corresponding reconstructions that retain the amplitude and local structure. The examples shown are selected from 11:00:00–11:10:00 on March 6, 2011.

4.2 Cluster Analysis

The DEC model develops classes of noise for data that did not have existing labels. We evaluate the clustering performance by inspecting the predicted label assignments through 3 visualization techniques.

First, to visualize the distribution of the data we use the t-distributed Stochastic Neighbor Embedding (t-SNE), a data visualization method to map the high dimensional latent space data to 2 dimensions (Maaten and Hinton, 2008). The t-SNE transformation is used only to visualize the high dimensional space and we do not interpret the t-SNE transformed spatial dimensions, only the clusters represented after the transformation. The t-SNE results are shown using the

latent space data before and after the DEC model is applied (Fig. 4.2). The labels in Fig 4.2a correspond to the cluster assignments used to initialize the DEC Model. The distribution of data prior to DEC indicates a large central cluster with multiple small groups extending beyond the center mass and resembles a 2D normal distribution. In Fig 4.2b the cluster assignments are those generated from the trained DEC model after the centroid locations and latent space data has been refined. The distribution shows the increased separation between clusters and higher density within most of the clusters. The DEC assigned clusters with labels 0, 1, 3, and 4 are examples of well separated clusters in which the majority of data are densely grouped. The data assigned to Label 2 is more distributed and is seen in several smaller groups suggesting more clusters could be resolved using the DEC model.

Second, we visualize spectrograms assigned to each cluster to evaluate what features are contained in each group (Fig. 4.3). Shown are 5 spectrograms for each of the 5 labels that occur during a known vibroseismic sweep at 11:00:24 on 6 March 2011. Label 0 contains time varying frequency modulations on the order of 1 second from 20–40 Hz. Label 1 contains a signal with increasing frequency from 15–25 Hz. Labels 2, 3, and 4 show distributed energy from 0–20 Hz that can be observed in other groups that contain a more dominant signal. Label 2 contains energy in higher frequencies and Label 4 is mostly in the ¡20 Hz range. While some patterns can be inferred from the visualization, each cluster may contain additional signals from what is presented. The examples shown are good representations of the assignments made in the cluster analysis.

Third, the cluster results are shown as a time series to highlight the spatiotemporal evolution (See Video S1 in the electronic supplement to this article). For each 1-second interval from 11:00:00 to 11:10:00 on 6 March 2011 we show the cluster assignments for every station in the array. We observe how the classification of a station changes through time and identify areas where the label is time invariant. This provides insight to identifying the source of signals in areas with localized energy and aids in visualizing how certain signals propagate through the

study area. In Fig. 4.4 we show 4 examples of a vibroseismic sweep (located near the Long Beach Airport) in which the signals are classified into different clusters. Label 2 (green dots) is consistently observed in the southern portion of the array with localized ares of Label 4 (red dots) persisting through time. In Fig. 5.1a, the area around the airport is classified as Label 1 and is associated with the fundamental signal from the sweep of the vibroseismic truck, as shown with the 10 second high amplitude signals from 5–30 Hz (Fig. 5.1b). We observe two notable groups of stations where the label has minimal variability (Fig. 4.4). The stations centered around (38.80, -118.18) are assigned Label 0 (shown in Fig. 4.4a-d and Fig. 5.1a in black) at each instance and are located in an area of active oil and gas extraction and storage suggesting industrial equipment is separable in this area. We show a 30 s spectrogram generated from the signal recorded by one station within this grouping (station 1042-5052) which shows a frequency modulated signal that oscillates between 28 Hz and 38 Hz with a period of 5 s, possibly from mechanical pumps used for extracting oil and gas (Fig. 5.1c). Another group of stations with consistent labeling is a column in the southwest quadrant of the array that are classified as label 4 (Fig. 4.4 & Fig 5.1a). The 30 s spectrogram generated from station 1027-5002 in this group show a constant signal at ˜25 Hz with overtones at ˜38 Hz and ˜45 Hz (Fig. 5.1d). The hum-like signals are possibly from rotating machinery producing noise signals at several different frequencies.

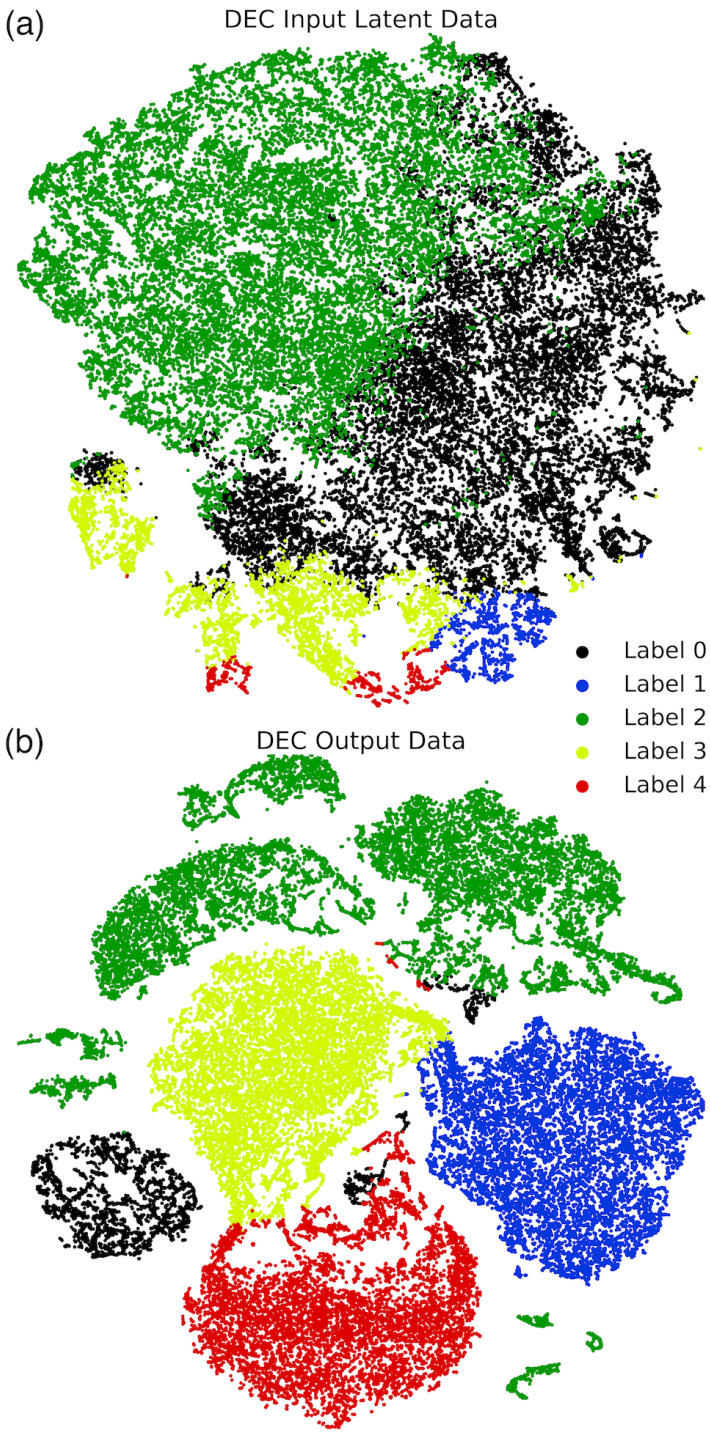

Figure 4.2: Data visualization using t-SNE plots. (a) The distribution of the of the latent space input prior to the DEC model being applied. The label assignments are those generated by K-means to initialize the DEC model. (b) The distribution of the latent data after DEC has been applied, the model weights updated, and the cluster centroids locations refined.

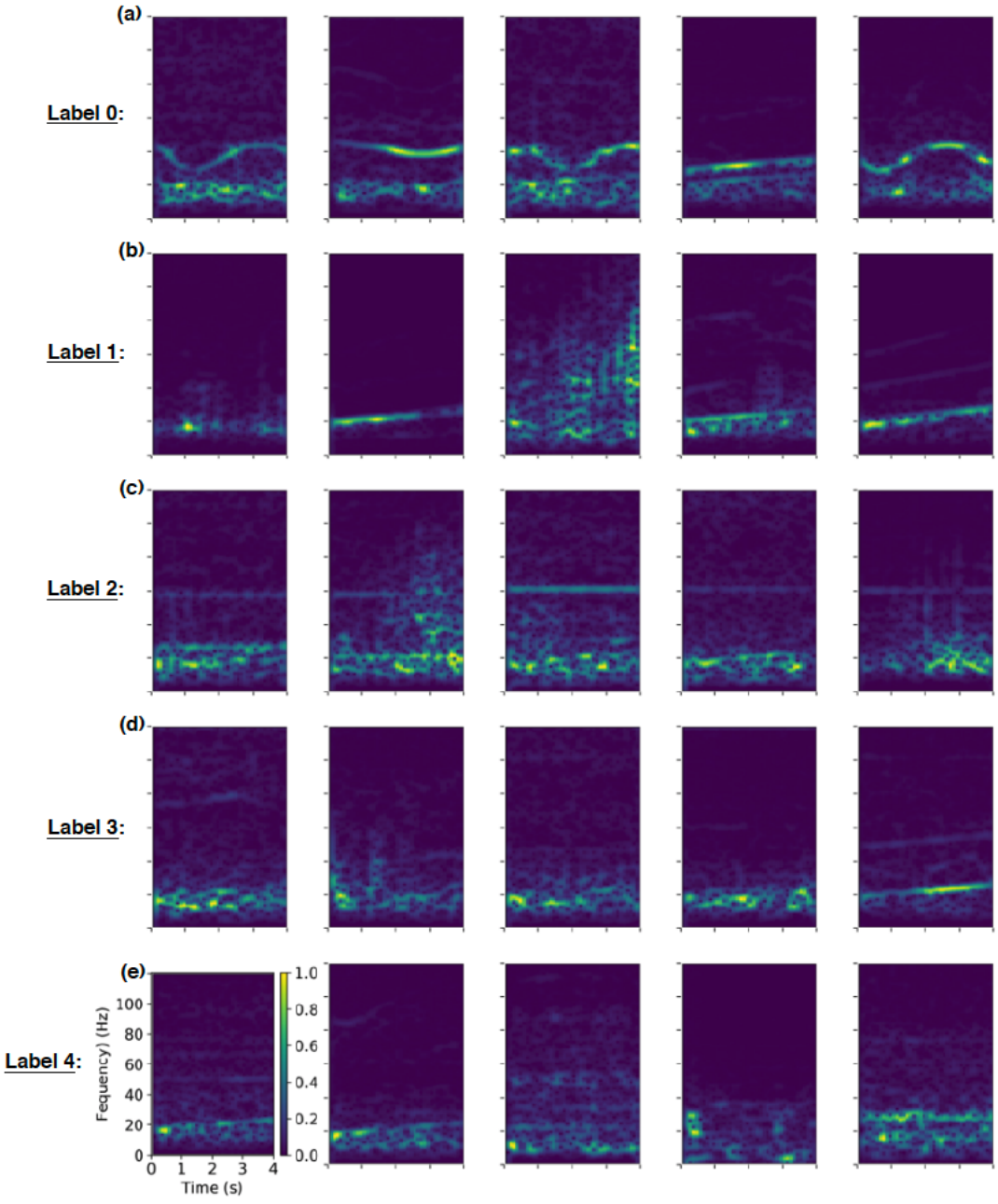

Figure 4.3: Examples of nonconsecutive 4 s spectrograms assigned to the 5 clusters using the test data from March 6, 2011 11:00:24 during a known vibroseismic sweep.All samples represent the frequencies 1–120 Hz on the y-axis and 0-4 s on the x-axis. The amplitudes are normalized between values of 0 and 1 to highlight the similar features.

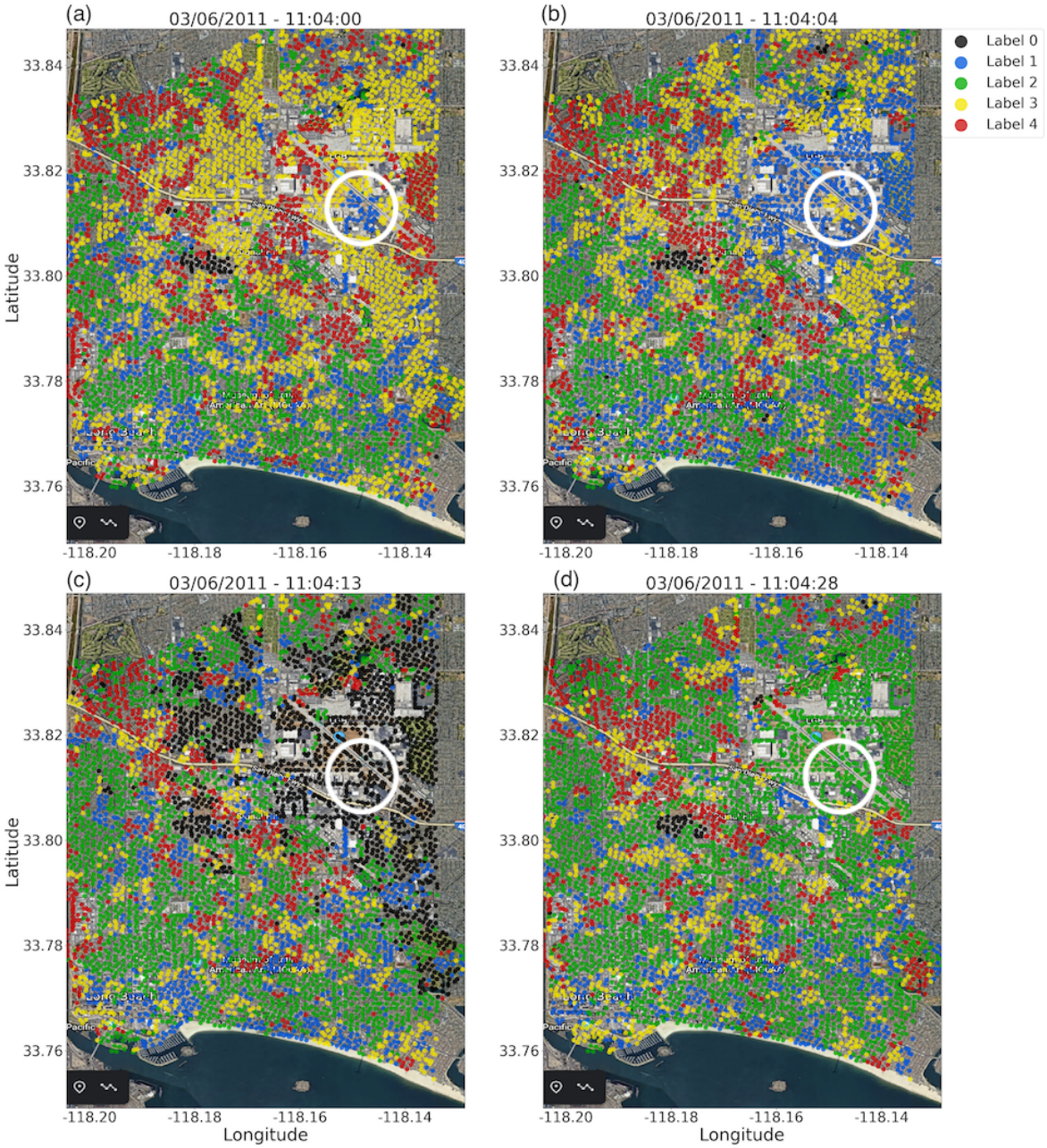

Figure 4.4: Four examples of cluster labels for the entire array during a vibroseismic truck survey on March 6, 2011 at 11:04:00for 40 s. The vibroseis trucks are located in the northeast quadrant of the array near the Long Beach Airport (denoted by the white circle). As the linear frequency modulated (LFM) signal is being generated over 40 seconds, the signal is classified under different labels beginning with (a) Label 3 at 11:04:00, (b) Label 1 at 11:04:04, (c) Label 0 at 11:04:13, and (d) Label 2 at 11:04:28.

Chapter 5

Discussion and Conclusion

5.1 Seismic noise classification

We've shown the DEC model can effectively identify and separate signals containing specific spectral characteristics that are different from the background noise. The spectrograms of the three sources we identify show a strong signal with high amplitude at specific frequencies compared to the background noise or less prominent signals originating from weaker sources (Fig. 5.1b-d). The spectrograms indicate distinct spectral patterns the time-frequency domain (i.e. frequency modulation, constant through time, etc.) and is shown best by the classification of the vibroseismic signals.

The Long Beach Vibroseismic survey data has been used in tomography studies to image the subsurface structure of the Long Beach Area, specifically the Signal Hill Oil Field, and the Newport-Inglewood Fault that crosses the center of the Long Beach array (Bianco, Gerstoft, et al., 2019). Vibroseismic sweeps are well suited for this purpose since it produces the dominant signal recorded at each station during the survey. The DEC model is able to detect the vibroseismic signals due to the distinctive noise characteristics in the spectrograms. As the frequency of the vibroseismic sweep increases linearly with time, the signal is assigned different labels during

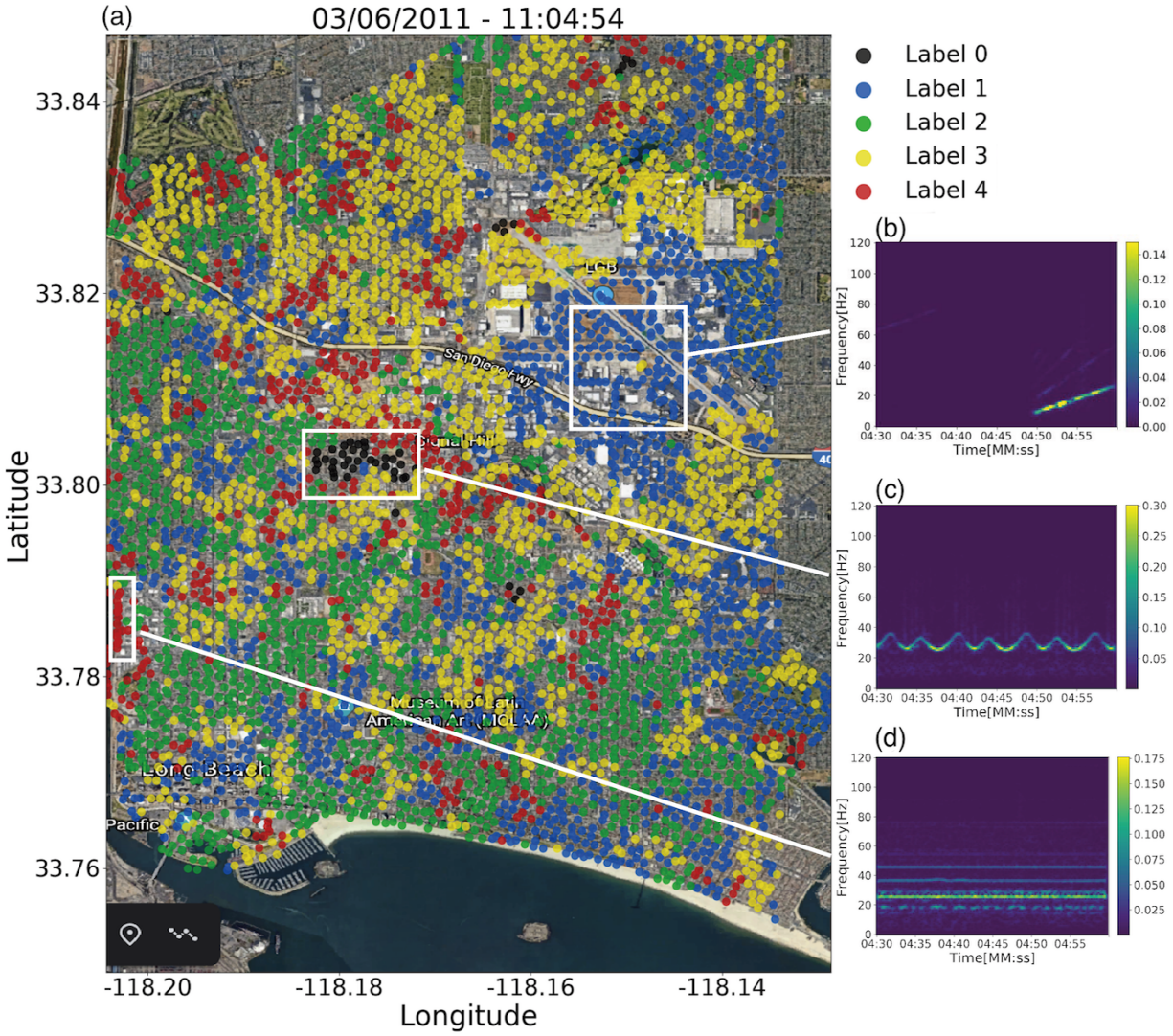

Figure 5.1: Location and spectrogram examples of different classes on 6 March 2011 between 11:04:30 and 11:05:00. (a) The Long Beach array, with color scale corresponding to label assigned to the recorded signal's spectrogram at 11:04:54. (b) 30 s spectrogram of station 1058-5059 where a vibroseismic signal was recorded. (c) 30 s spectrogram of station 1042-5052, representative of the stations grouped near pump jack activity from the oil extraction process. (d) 30 s spectrogram of station 1027-5002, representative of the stations grouped near rotating machinery.

clustering. The sequence of label assignments during a sweep is repeated in the same order each time the vibroseismic signal is active (see Video S1 in the electronic supplement to this article) and indicates the model detects and identifies the sweep as it spans a range of frequency bands, which are classified as different signals. Other sources are not detected by the current DEC model

using an unsupervised approach. The cultural noise is near continuous in the seismic waveforms and most spectrogram do not contain a dominant or distinct signal. Instead we find multiple overlapping signals from numerous sources. For example, a station located along Highway 405 continuously records the seismic signatures of multiple vehicles passing at a time. While this generates strong ground motion, the constructive interference of the signals in the spectrograms lack a distinct signature that the DEC model can recognize from the latent space. An improved model should be able to recognize any similar signal and group then as a single class.

5.2 Model performance and improvements

The model training and hyperparameter tuning process allow us to evaluate how the DEC model identifies specific types of noise characteristics while neglecting others. The goal is to identify known noise sources that potentially produce repeatable signals across the array. One reason the DEC model may not be able to classify all the noise sources is the usage of the K-means algorithm to initialize the cluster centroids. The K-means algorithm makes two assumptions when clustering the data: (1) the clusters in the data have equal variance and (2) all of the clusters will contain a similar number of observations. These assumptions are often violated when working with real seismic data (Johnson, Meng, et al., 2019; Meng, Ben-Zion, and Johnson, 2019) and are difficult to satisfy when utilizing an unsupervised approach. Indeed, there is some evidence the assumptions are not satisfied with the Long Beach data set. A large portion of the samples in our data set contain overlapping signals that are not distinguishable using features in the spectrograms. The encoder is designed to extract the salient features, but there still may not be sufficient information to distinguish between these complex samples. Evidence for this exists in the distribution of latent space features prior to applying the DEC model (Fig. 4.2a). The K-means clustering assigns Label 0 and Label 2 to much of the data and resembles a 2D normal distribution without a distinction between the two. The smaller groups that extend

from the central distribution would ideally be classified as their own labels, but instead these are assigned to several overlapping classes. The likely cause is the K-means criterion that all cluster distributions contain equal variance. Underlying assumptions exist in all clustering algorithms and are required to solve the optimization problem.

Without any prior knowledge of the data structure it is difficult to determine the best performing clustering algorithm for seismic data. The advantage of applying K-Means in this study is the algorithm allows the refinement of the cluster centroids and the latent space to improve the distribution of data assigned to a cluster. The ideal clustering algorithm for this problem would contain the following properties: (1) centroid based, (2) account for clusters of variable size and variance, (3) automatically determine the optimal number of clusters for the data, (4) and properly scale to large datasets. No clustering algorithm that we are aware of satisfies all of these criteria. Improvements are possible using different clustering algorithms that are able to better meet the necessary conditions and the ability to iteratively update using the autoencoder performance.

An important parameter that contributes to the DEC performance is the number of latent space features. The scope of this study is to separate and classify spectrograms in a lower dimension using feature extraction and dimension reduction with probabilistic clustering. Our choice of 14 features to represent the spectrograms (4800 features) is shown to produce adequate reconstructions and distinct clusters. However, some data sets may produce poor results due to the elimination of important information from the input. During the initial testing we trained the model using 300 features in the latent space and iteratively reduced to the value of 14. The choice is not arbitrary. The performance of the clustering is limited if using very high dimensional data, but the CAE reconstructions improve greatly when increasing the latent space dimensions. Using more latent features improves the reconstruction loss but results in less accurate clustering. The method requires testing the tradeoff between cluster performance and input reconstruction to produce usable results.

The inputs used only include the amplitude of the time-frequency spectrum as a single model layer. We tested using an additional layer that incorporates the phase information from the complex variable of the STFT but found no performance improvements. Additionally, we tested using the time domain waveforms and can produce well defined clusters but found the CAE reconstruction of waveforms to be less accurate with a loss of phase and amplitude information. Considering the data is signal channel vertical component we found the spectrogram approach to provide the best results. If the analysis is repeated with three channel data we suggest revisiting the time domain input and remove the STFT computation time from the workflow.

All chapters of this thesis, including the abstract and introduction, are coauthored by Johnson, Christopher W., Bianco, Michael J., and Gerstoft, Peter. The content of this thesis has been submitted for publication of the material as it may appear in Snover, D., Johnson, C. W., Bianco, M. J., & Gertsoft, P. (2020). Deep clustering to identify sources of urban seismic noise in Long Beach, CA. Seismologic Research Letters. The thesis author was the primary investigator and author of this material.

References

Bergen, K. J., Johnson, P. A., de Hoop, M. V., & Beroza, G. C. (2019). Machine learning for data-driven discovery in solid Earth geoscience. *Science*, *363*(6433). https://doi.org/10.1126/science.aau0323

Bianco, M. J., & Gerstoft, P. (2018). Travel time tomography with adaptive dictionaries. *IEEE Trans. Comput. Imag.*, *4*(4), 499–511. https://doi.org/10.1109/TCI.2018.2862644

Bianco, M. J., Gerstoft, P., Olsen, K., & Lin, F.-C. (2019). High-resolution seismic tomography of Long Beach, CA using machine learning. *Sci. Rep.*, *9*, 14987. https://doi.org/10.1038/s41598-019-50381-z

Bowden, D. C., Tsai, V. C., & Lin, F. C. (2015). Site amplification, attenuation, and scattering from noise correlation amplitudes across a dense array in Long Beach, CA. *Geophys. Res. Lett.*, *42*, 1360–1367.

Brenguier, F., Boué, P., Ben-Zion, Y., Vernon, F., Johnson, C., Mordret, A., Coutant, O., Share, P., Beaucé, E., Hollis, D., & Lecocq, T. (2019). Train traffic as a powerful noise source for monitoring active faults with seismic interferometry. *Geophys. Res. Lett.*, *46*(16), 9529–9536.

Chamarczuk, M., Nishitsuji, Y., Malinowski, M., & Draganov, D. (2019). Unsupervised learning used in automatic detection and classification of ambient-noise recordings from a large-n array. *Seismol. Res. Lett.*, *91*. https://doi.org/10.1785/0220190063

Chazan, S. E., Gannot, S., & Goldberger, J. (2018). Deep clustering based on a mixture of autoencoders. *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6.

Diaz, J., Ruiz Fernandez, M., Sánchez-Pastor, P., & Romero, P. (2017). Urban seismology: On the origin of earth vibrations within a city. *Sci. Rep.*, *7*. https://doi.org/10.1038/s41598-017-15499-y

Dizaji, K. G., Herandi, A., Deng, C., Cai, W., & Huang, H. (2017). Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. *2017 IEEE International Conference on Computer Vision (ICCV)*, 5747–5756.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* [http://www.deeplearningbook.org]. MIT Press.

Groos, J. C., & Ritter, J. R. R. (2009). Time domain classification and quantification of seismic noise in an urban environment. *Geophys. J. Int.*, *179*(2), 1213–1231. https://doi.org/10.1111/j.1365-246X.2009.04343.x

Guo, X., Gao, L., Liu, X., & Yin, J. (2017). Improved deep embedded clustering with local structure preservation. *In Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, 1753–1759.

Inbal, A., Ampuero, J. P., & Clayton, R. W. (2016). Localized seismic deformation in the upper mantle revealed by dense seismic arrays. *Science*, *354*, 88–92.

Inbal, A., Clayton, R. W., & Ampuero, J.-P. (2015). Imaging widespread seismicity at midlower crustal depths beneath Long Beach, CA, with a dense seismic array: Evidence for a depth-dependent earthquake size distribution. *Geophys. Res. Lett.*, *42*(15), 6314–6323.

Inbal, A., Cristea-Platon, T., Ampuero, J., Hillers, G., Agnew, D., & Hough, S. E. (2018). Sources of long-range anthropogenic noise in Southern California and implications for tectonic tremor detection. *Bull. Seismol. Soc. Am.*, *108*(6), 3511–3527.

Johnson, C. W., Meng, H., Vernon, F., & Ben-Zion, Y. (2019). Characteristics of ground motion generated by wind interaction with trees, structures, and other surface obstacles. *J. Geophys. Res. B. Solid Earth*, *124*(8), 8519–8539. https://doi.org/10.1029/2018JB017151

Johnson, C. W., Vernon, F., Nakata, N., & Ben-Zion, Y. (2019). Atmospheric processes modulating noise in Fairfield Nodal 5 Hz geophones. *Seismol. Res. Lett.* https://doi.org/10.1785/0220180383

Kong, Q., Trugman, D., Ross, Z., Bianco, M., Meade, B., & Gerstoft, P. (2019). Machine learning in seismology: Turning data into insights. *Seismol. Res. Lett.*, *90*, 3–14.

Li, Z., Meier, M.-A., Hauksson, E., Zhan, Z., & Andrews, J. (2018). Machine learning seismic wave discrimination: Application to earthquake early warning. *Geophys. Res. Lett.*, *45*(10), 4773–4779. https://doi.org/10.1029/2018GL077870

Li, Z., Peng, Z., Hollis, D., Zhu, L., & McClellan, J. (2018). High-resolution seismic event detection using local similarity for Large-N arrays. *Sci. Rep.*, *8*(1).

Lin, F.-C., Li, D., Clayton, R. W., & Hollis, D. (2013). High-resolution 3d shallow crustal structure in Long Beach, California: Application of ambient noise tomography on a dense seismic array. *Geophysics*, *78*(4), Q45–Q56.

Lloyd, S. P. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, *28*, 129–137.

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.*, 2579–2605.

Meng, H., & Ben-Zion, Y. (2018a). Characteristics of airplanes and helicopters recorded by a dense seismic array near anza california. *J. Geophys. Res. B. Solid Earth*, *123*. https://doi.org/10.1029/2017JB015240

Meng, H., & Ben-Zion, Y. (2018b). Characteristics of Airplanes and Helicopters Recorded by a Dense Seismic Array Near Anza California. *J. Geophys. Res. B. Solid Earth*, *123*, 4783–4797.

Meng, H., Ben-Zion, Y., & Johnson, C. W. (2019). Detection of random noise and anatomy of continuous seismic waveforms in dense array data near Anza California. *Geophys. J. Int.*, *219*(3), 1463–1473.

Mousavi, S. M., & Beroza, G. C. (2020). A machine-learning approach for earthquake magnitude estimation. *Geophys. Res. Lett.*, *47*(1), e2019GL085976. https://doi.org/10.1029/2019GL085976

Mousavi, S. M., Zhu, W., Ellsworth, W., & Beroza, G. C. (2019). Unsupervised clustering of seismic signals using deep convolutional autoencoders. *Geosci. Rem. Sens. Lett. IEEE*, *16*(11), 1693–1697.

Mousavi, S. M., Zhu, W., Sheng, Y., & Beroza, G. C. (2019). Cred: A deep residual network of convolutional and recurrent units for earthquake signal detection. *Sci. Rep.*, *9*. https://doi.org/10.1038/s41598-019-45748-1

Mrabah, N., Khan, N., & Ksantini, R. (2019). Deep clustering with a dynamic autoencoder. *arXiv*, *abs/1901.07752*.

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. B. Solid Earth*, *120*(2), 1159–1173. https://doi.org/10.1002/2015JB011870

Perol, T., Gharbi, M., & Denolle, M. (2018). Convolutional neural network for earthquake detection and location. *Sci. Adv.*, *4*(2), e1700578.

Riahi, N., & Gerstoft, P. (2015). The seismic traffic footprint: Tracking trains, aircraft, and cars seismically. *Geophys. Res. Lett.*, *42*(8), 2674–2681.

Riahi, N., & Gerstoft, P. (2017). Using graph clustering to locate sources within a dense sensor array. *Signal Process.*, *132*, 110–120.

Ross, Z. E., Meier, M.-A., & Hauksson, E. (2018). P wave arrival picking and first-motion polarity determination with deep learning. *J. Geophys. Res. B. Solid Earth*, *123*(6), 5120–5129. https://doi.org/10.1029/2017JB015251

Ross, Z. E., Meier, M.-A., Hauksson, E., & Heaton, T. H. (2018). Generalized seismic phase detection with deep learning. *Bull. Seismol. Soc. Am.*, *108*(5A), 2894–2901.

Schmandt, B., & Clayton, R. W. (2013). Analysis of teleseismic p waves with a 5200-station array in Long Beach, California: Evidence for an abrupt boundary to Inner Borderland rifting. *J. Geophys. Res. B. Solid Earth*, *118*(10), 5320–5338.

Xie, J., Girshick, R., & Farhadi, A. (2016). Unsupervised deep embedding for clustering analysis, In *Proceedings of the 33rd international conference on machine learning*.