Cluster analysis of Very-Long-Period events at Stromboli Volcano

Lukas Layer, Antonietta Esposito, Flora Giudicepietro, Rosario Peluso

Content

1. Overview over Stromboli volcano and the 2019 eruptions
2. Feature extraction for time series data
3. Visualization of high dimensional data with Self-Organizing-Maps and the t-SNE algorithm
4. Clustering of time series with K-means and assessment of the clustering quality with Silhouette scores
5. Discussion of the results

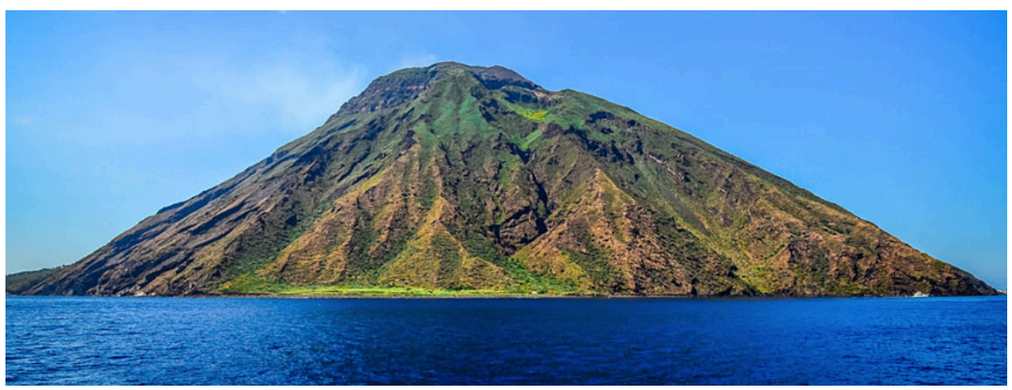
Stromboli Volcano

- **Open conduit volcano** with persistent explosive activity located in the Mediterranean Sea

- Activity characterized by **small explosions** occurring **repetitively** with a frequency ranging from five to more than 20 events per hour

- **Very-long-period (VLP) signals** associated with the Strombolian explosions

- **Rare** occurrence of **paroxysms** (violent explosions that produce eruptive columns often accompanied by pyroclastic flows) ~ **18 paroxysms over 110 years**



Paroxysms in 2019

- On **July 3, 2019** during a period of apparently moderate activity a **paroxysmal explosion** occurred

- The paroxysm gave rise to an **eruptive column** more than 5 km high and to a **pyroclastic flow**

- On **August 28**, **2019** a second **paroxysmal explosion** occurred similar to that of July 3 with another pyroclastic flow

- **Routinely monitored parameters by the INGV did not predict the July 3 paroxysm resulting in one fatality and some injuries**

The quest for new precursors

- To **reduce the risk of fatalities in the future** a wide range of parameters is analyzed to understand whether any of them show **anomalies** prior to the paroxysmal period to monitor them in the future

- Excellent analysis of various **promising parameters** recently published: *"Geophysical precursors of the July-August 2019 paroxysmal eruptive phase and their implications for Stromboli volcano (Italy) monitoring." (*Giudicepietro, F. *et al. Sci Rep* **10,** 10296 (2020). https://doi.org/10.1038/s41598-020-67220-1)

Very-Long-Period signal waveforms

- The first broadband seismic recordings have shown **differences in the waveforms** of the Very-Long-Period **VLP signals** associated with the Strombolian explosions (Chouet et al. 2003)

- Connection between the **VLP waveform** and the **physical state of the volcano** has been suggested (Esposito et al. 2008)

→ Cluster analysis of VLP waveforms (time series) in the paroxysmal period

Dataset

- The **Instituto Nazionale di Geofisica e Vulcanologia (INGV)** operates a permanent **broadband network** on Stromboli with thermal and infrared cameras

- The network consists of **7 digital stations** equipped with three-component **seismometers** and signals are recorded at a **sampling rate** of **50 samples/s**

- Signals that exceed a certain **amplitude threshold are selected** and signals from other sources (e.g. earthquakes) are removed manually

- **Final selection** consists of **~20.000 VLP events** in the period **15/05/19 to 19/09/19**

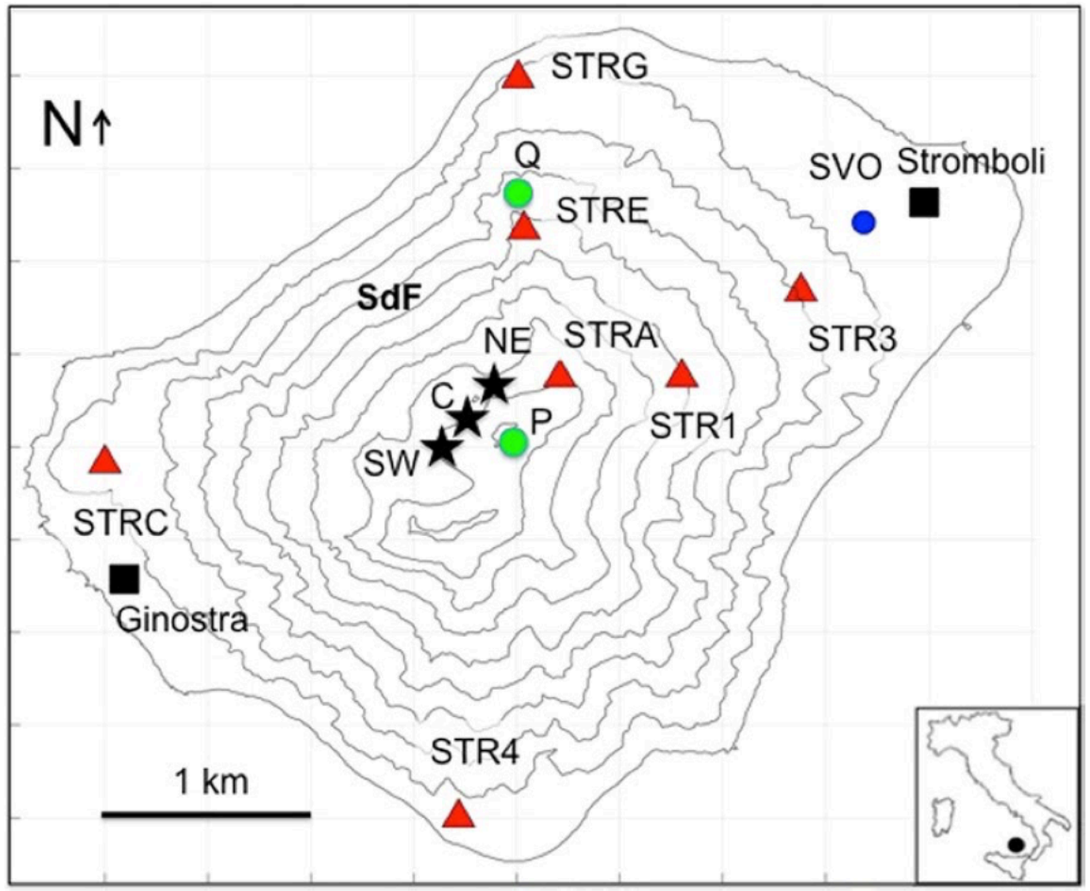
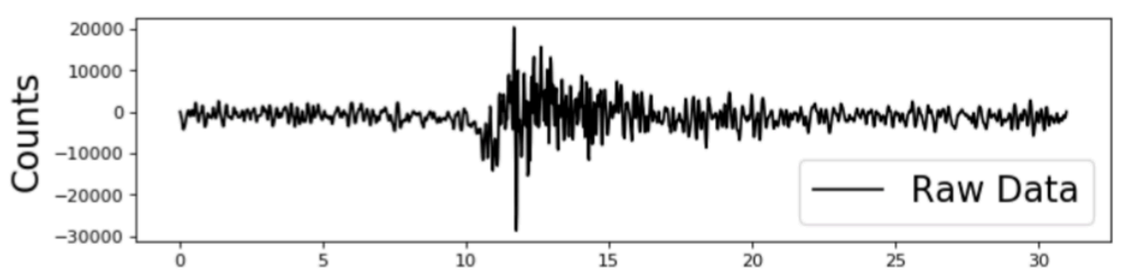


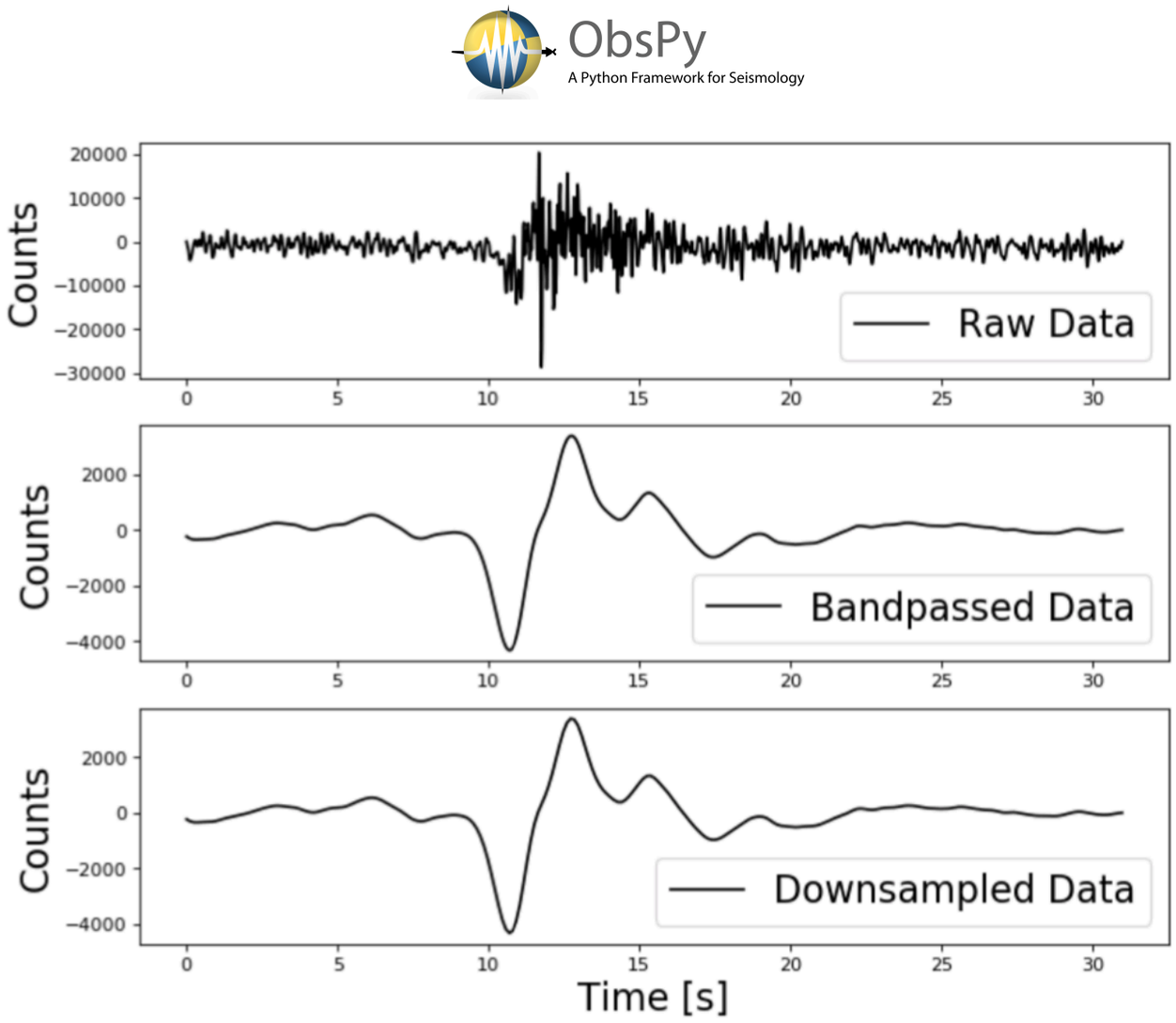
Figure taken from: https://doi.org/10.1038/s41598-020-67220-1



Preprocessing of the signals

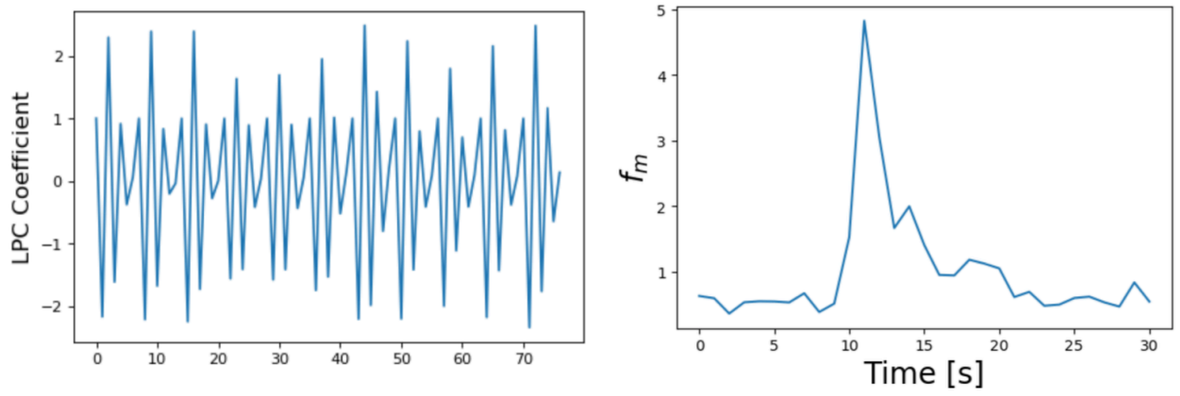
Filtering of the VLP signals

- Uses the python framework for seismology ***obspy***
- **Bandpass filter** in the VLP-frequency band (0.05–0.5 Hz)
- **Decimation** (downsampling) of the resulting signal by a factor of 6



Additional features

- Extraction of **spectral information** with **Linear Predictive Coding**
- Extraction of **time domain information** by parametrizing the **amplitude**



Auto Encoders

LSTM Auto Encoder

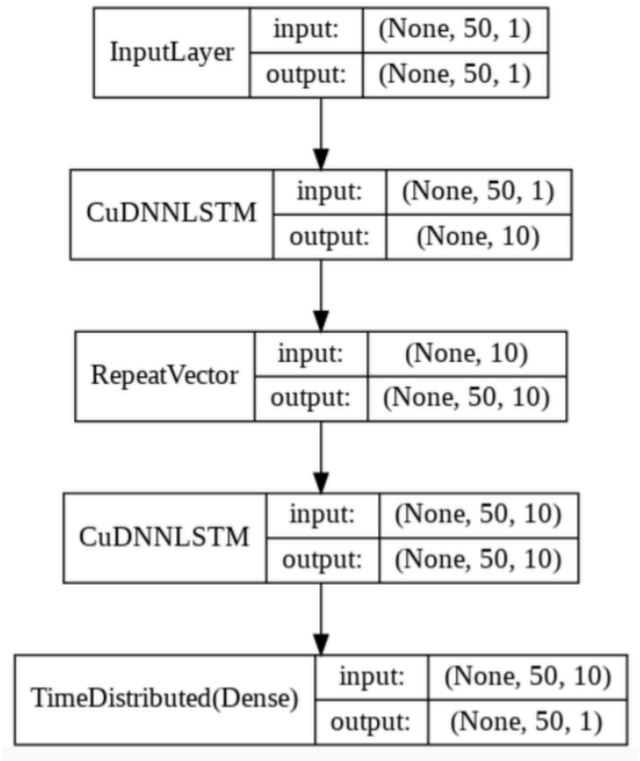
Keras

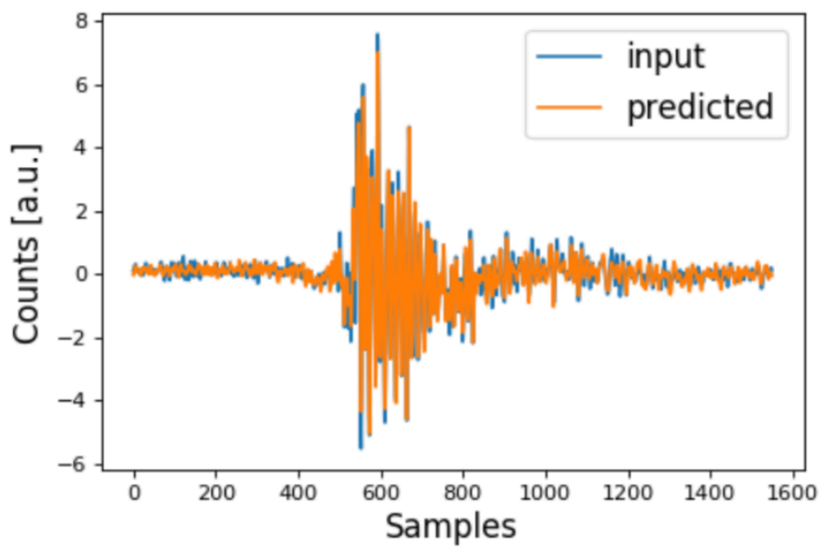
- Neural network used to learn efficient **data codings** in an **unsupervised** manner

- Consists of **encoder** that maps the input into a **code** and a **decoder** that maps the code to a reconstruction of the original input

- Training the network to **ignore** signal “**noise**” and **preserve** only the **most relevant aspects** of the data

- **LSTMs** are predestined for **sequential data** and thus promising for **time series**





Visualization of high dimensional vectors with Self Organizing Maps (SOM)

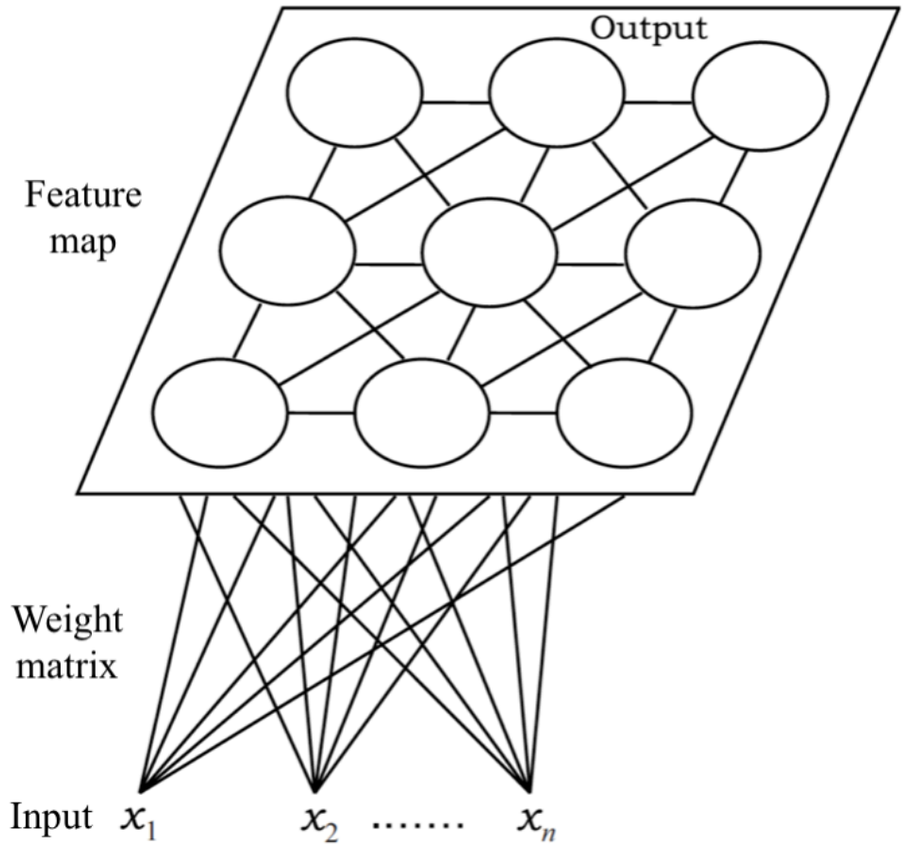
Self Organizing Maps

- Type of **neural network** that is trained using **unsupervised** learning to produce a low-dimensional discretized representation of the training samples → **dimensionality reduction**

- The **visible part** of a self-organizing map is the **map space**, which consists of components called **nodes**

- Each **node** is associated with a **weight vector** that has the same dimension as each input vector

- **Training** consists in moving **weight vectors** toward the input data - reducing a **distance metric** - without spoiling the **topology** induced from the map space



Application of SOM to the VLP waveforms

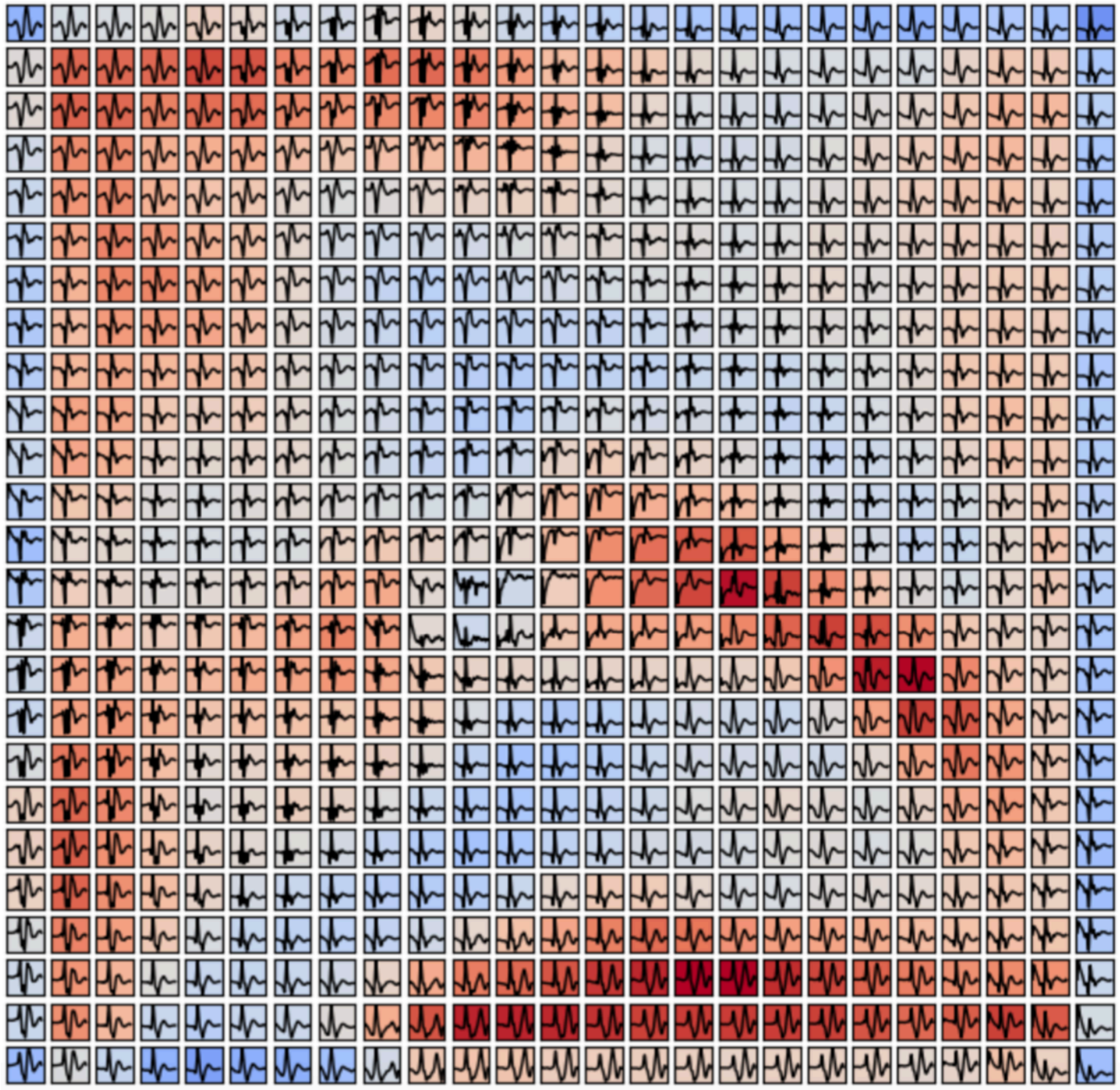
- **25x25 rectangular SOM** is used with **Euclidean distance** as distance metric implemented in python package ***minisom***

- Output of the SOM shows the **different signal forms** and the **distances** between neighboring nodes

- **Similar waveforms** are **close** in the map space

- Also the **frequencies** of signals in the nodes can be analyzed

Distance matrix of the 25x25 SOM



Blue = close to neighbors / Red = distant to neighbors

Visualization of the VLP waveform signals with the t-SNE algorithm

t-SNE

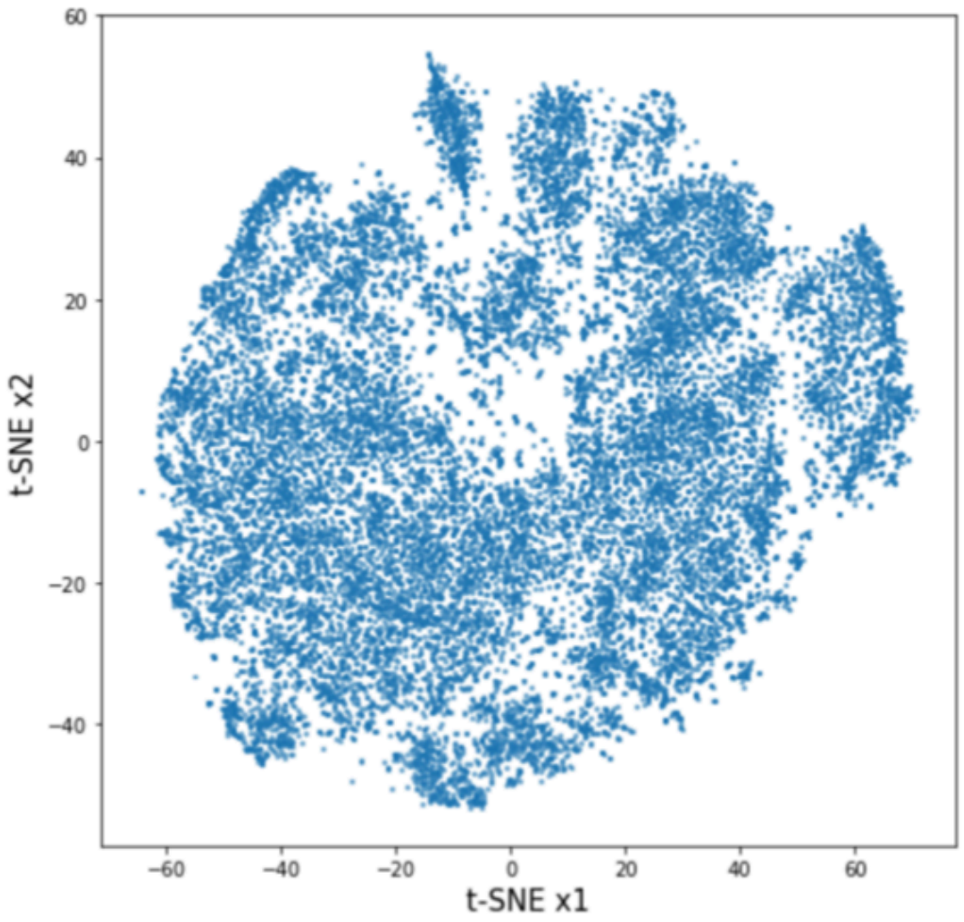
- **Nonlinear dimensionality reduction** for embedding high-dimensional data in a low dimensional space

- First, t-SNE constructs a **probability distribution over pairs of high-dimensional objects** such that higher probabilities are assigned to similar objects

- Second, t-SNE defines a **similar probability distribution** over the points in the **low-dimensional map**, and it **minimizes** the **KL divergence** between the two distributions

- **Similar objects** are modeled by **nearby points** and **dissimilar objects** are modeled by **distant points** with high probability

t-SNE applied to the VLP waveforms



Structures are visible

Clustering of the VLP waveforms with K-Means

K-Means

- Clusters data by trying to **separate samples groups** of equal variance, **minimizing** a criterion called **within-cluster sum-of-squares**

- **Number** of clusters need to be **specified**

Silhouette analysis

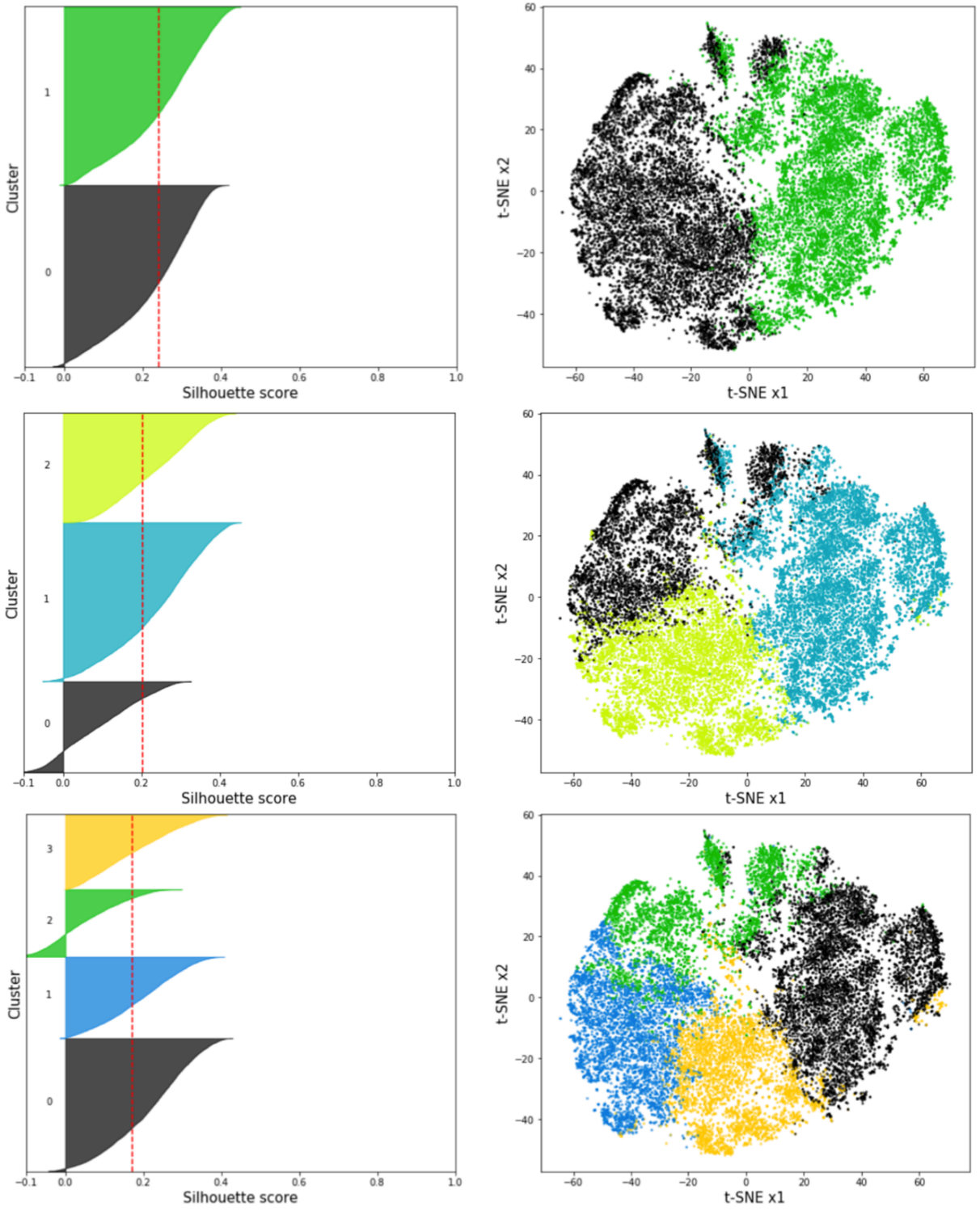
- **Optimal number** of clusters is **unknown** → performance has to be evaluated using the model itself

- **Silhouette scores** measure the **separation distance** between the resulting **clusters,** where **higher scores** relate to a model with **better defined clusters**

- Coefficients **near +1** → sample is **far away** from the neighboring cluster

- Coefficients **near 0** → sample **close to the decision boundary** between two neighboring clusters,

- Coefficients with negative values (max -1) → samples might have been **assigned to the wrong cluster**



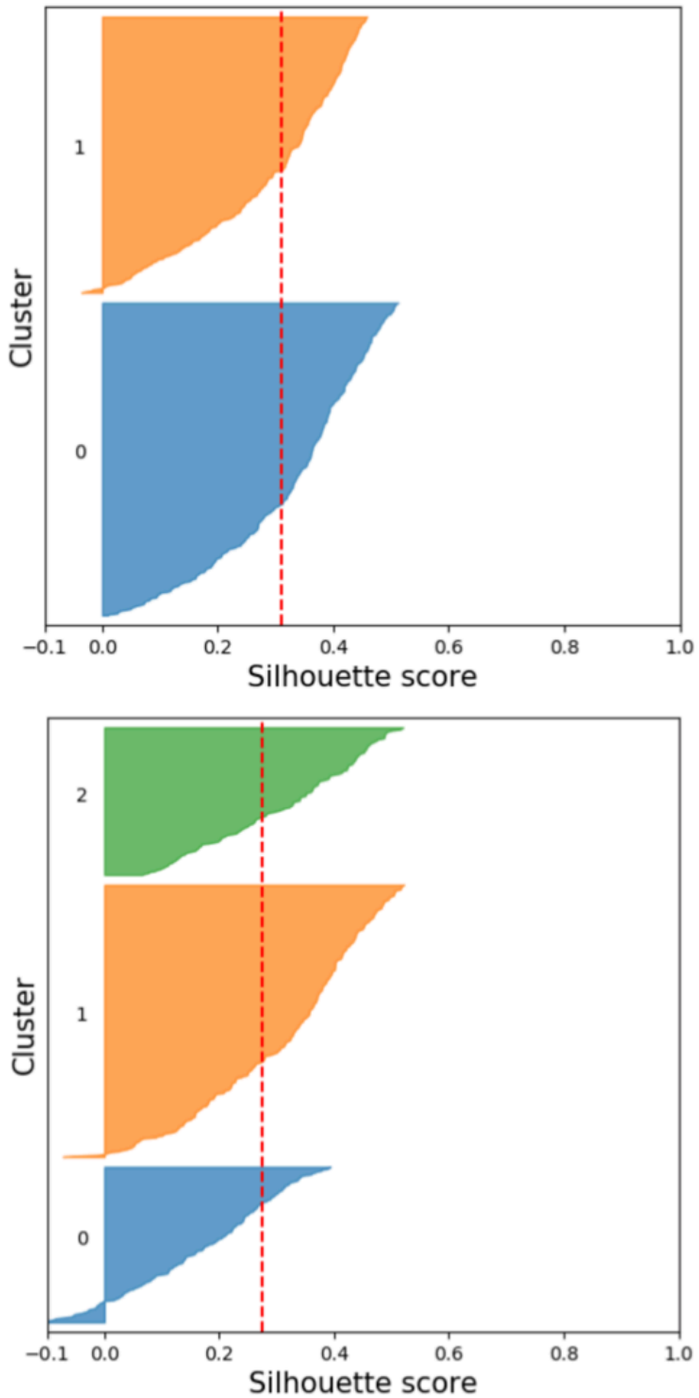
→ Best average silhouette score obtained for two clusters

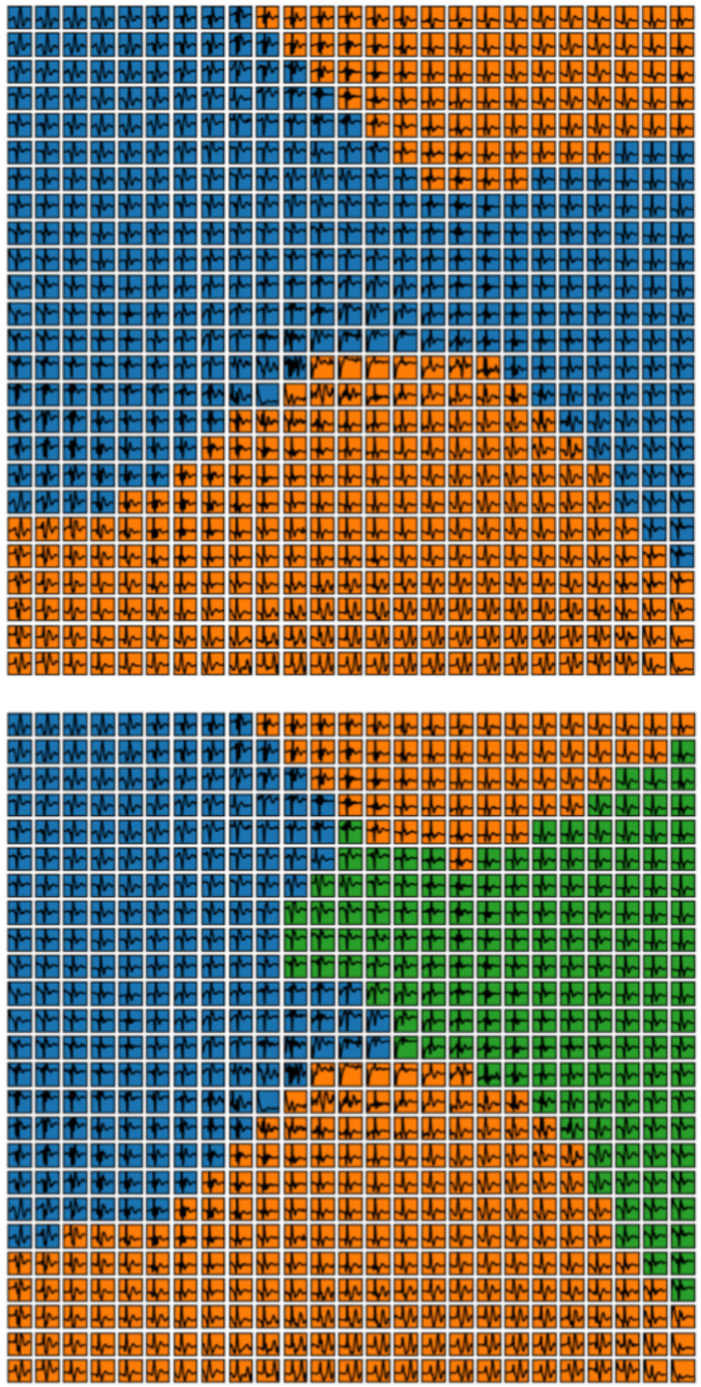
Clustering of the VLP waveform SOM with K-means

- K-means can also be applied to the **weight vectors** of the trained **SOMs → less noise** can be advantageous

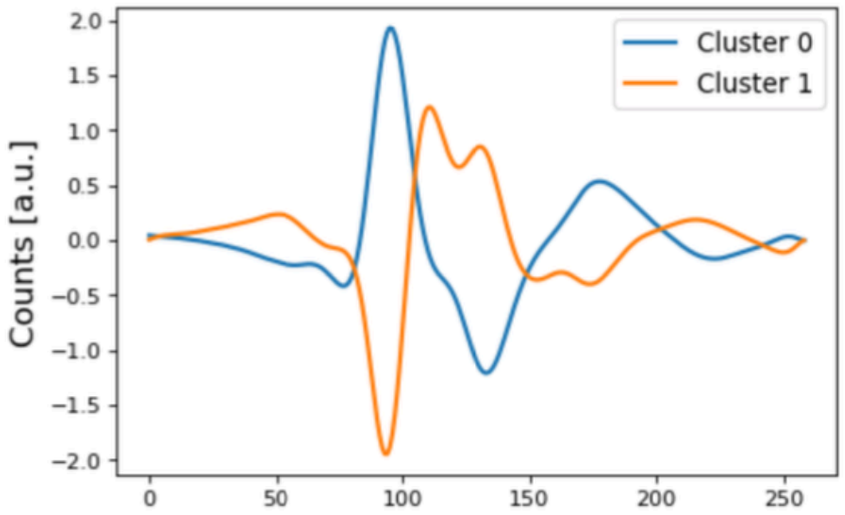
- **Analysis** of the silhouette scores is **repeated**

- Same result as before: the **highest average score** is obtained with **two clusters**





Time evolution of the two main clusters

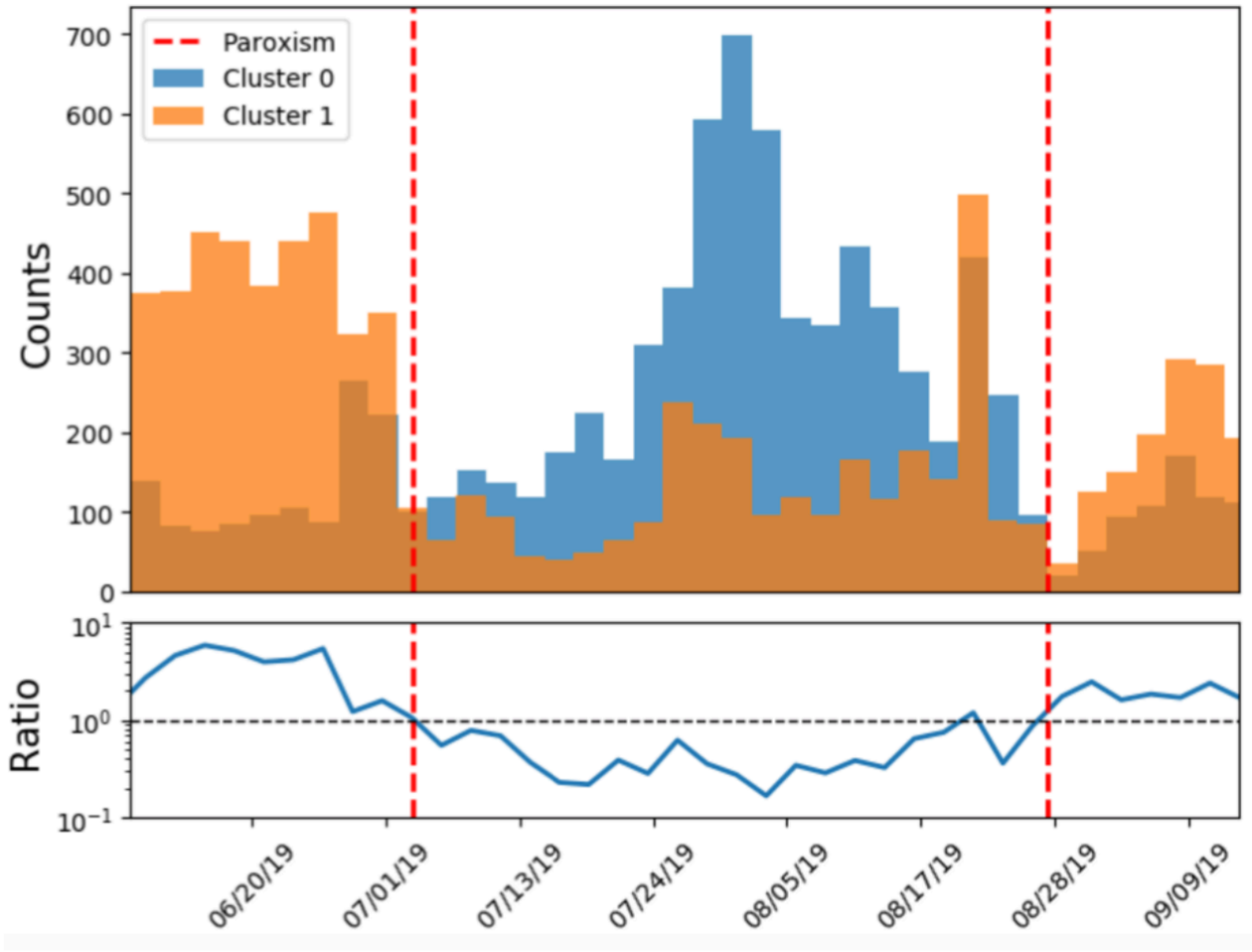


Visible difference in the shape of the centroids of the two clusters

- **Time stamps** of the signals allow to display the **time evolution** of the **signal counts** in the two main clusters

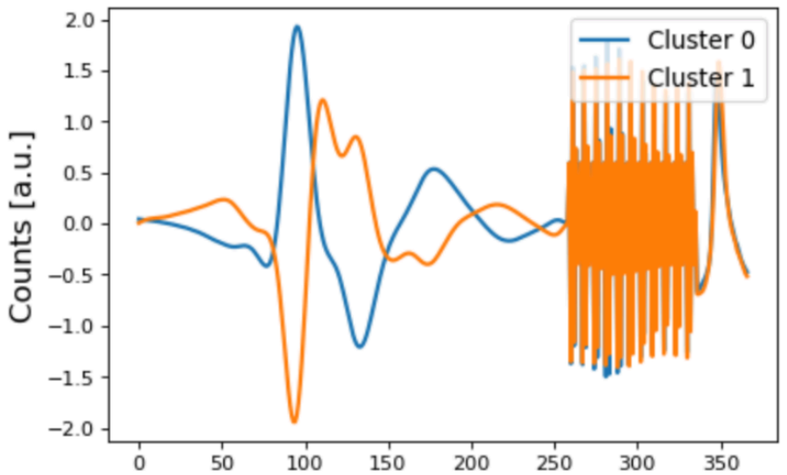
- **Strong imbalance of the two clusters** in the analyzed period of the paroxysms

Indicates a connection between the state of the volcano and the VLP signals



Clustering with additional information

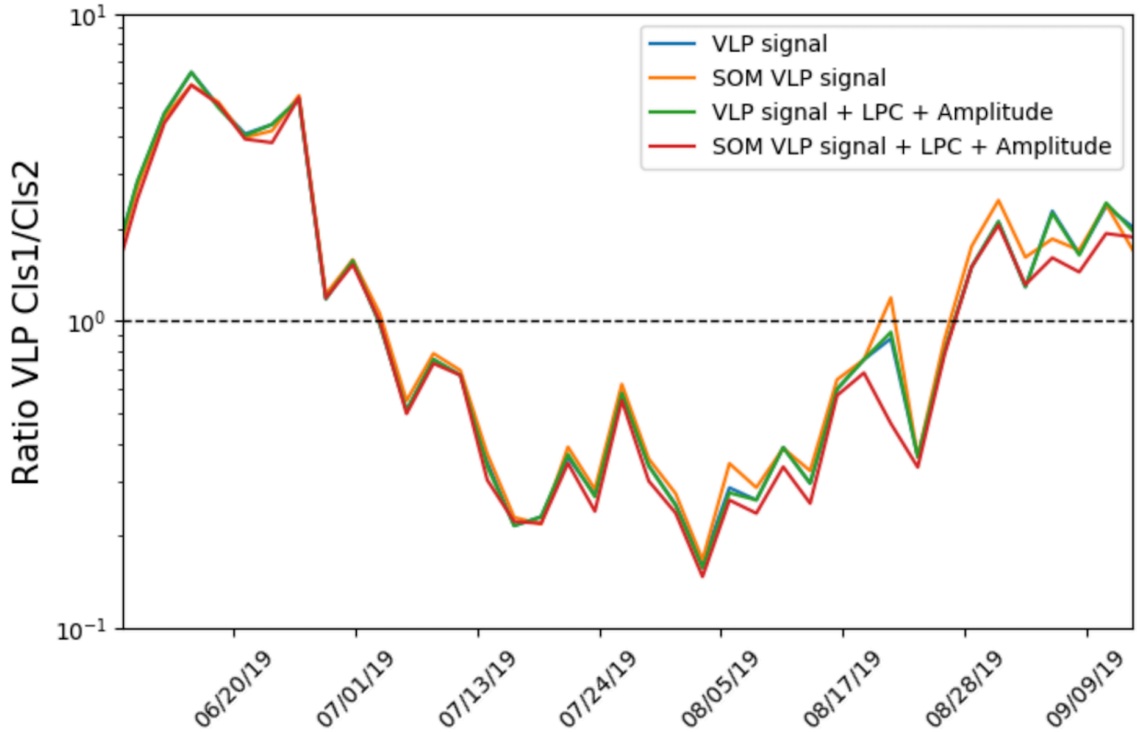
- Goal: add information to obtain **better defined clusters**
- Different **features** are **scaled** to the global standard deviation to **preserve the relative size** between the samples
- **LPC coefficients** and the **parametrized amplitude** / latent representation of the **Auto Encoder** are added



Comparison of the clustering results

- **Repeating the analysis** with additional information gives very **similar results** → dominating feature is the VLP waveform
- Obtained **ratio** of the two main clusters is **similar** for clustering directly with K-Means and first applying a SOM

→ Ratio of the two main clusters is stable



Summary

Conclusions

- **Visualization** and **Clustering analysis** of **time series** with unsupervised Machine Learning methods is a powerful tool to define **parameters** that **characterize the state of a system**

- The **ratio of the two main cluster** of the Very-Long-Period **waveforms of Stromboli** may constitute an **important parameter** to **characterize** the **state** of the volcano that can be **monitored in real time**

Outlook

- Analysis of a **larger time period** is in progress to understand whether the VLP waveforms have the potential to **predict anomalous behavior** of the volcano

- Analysis of **signals from controlled experiments** to better understand how the results of **clustering** and **visualization** with Machine Learning are **linked to the physical conditions** of the system


Thank you!

BACKUP

LPC and amplitude parametrization

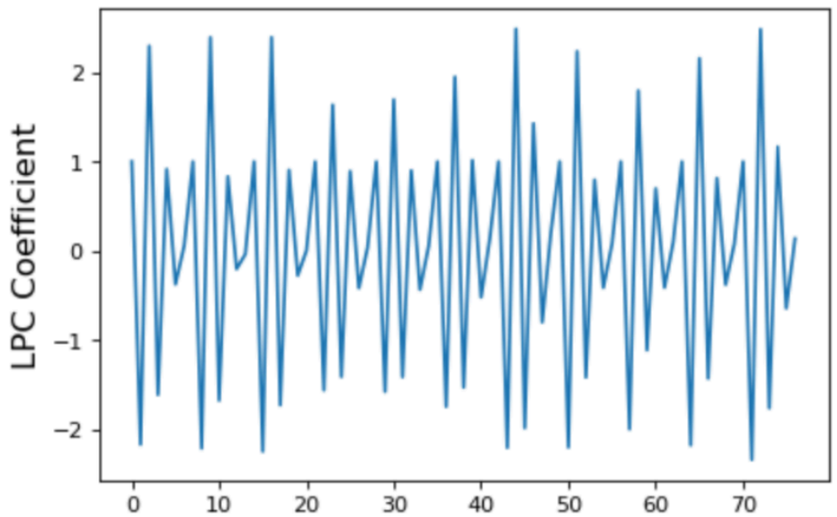
Linear predictive coding

$$\hat{x}(n) = \sum_{i=1}^{p} a_i x(n-i)$$

- **LPC** tries to **predict** a signal **sample** by means of a **linear combination** of **previous** signal **samples** → extracts **spectral information** (e.g. heavily used in speech analysis)

- **Converts** segments of a real time signal into a small set of **predictor coefficients {a_i}**

- **6 LPC coefficients** are extracted in 256-sample **windows**, with an **overlap** of 128 samples with the python audio library ***librosa***

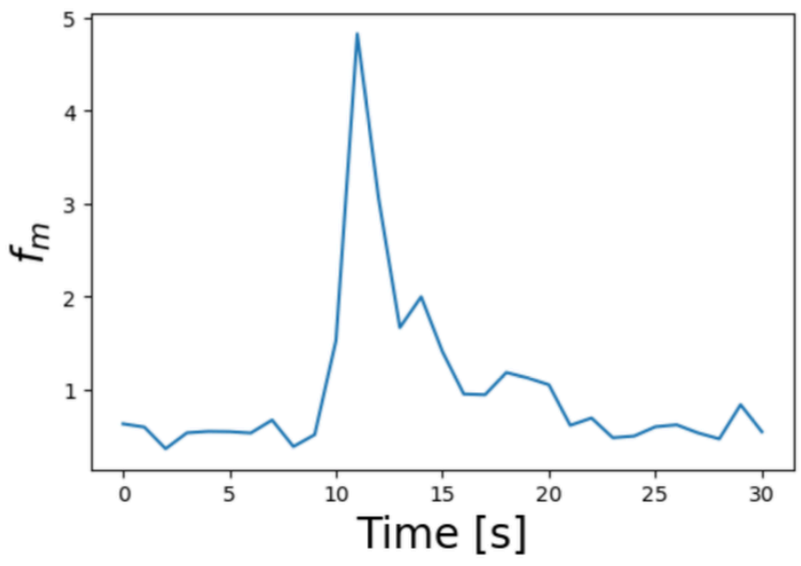


Amplitude parametrization

$$f_m = \frac{(max(A_{i,m}) - min(A_{i,m})) \cdot N}{\sum_{m=0}^{N}(max(A_{i,m}) - min(A_{i,m}))}$$

- **Time-domain information** is added by **parametrizing** the **amplitude** of the signal

- The function f_m is defined as the **normalized difference between the maximum and minimum signal amplitudes** within a 1 second window



Silhouette coefficient

- Calculated using the mean intra-cluster distance *a* and the mean nearest-cluster distance *b* for each sample,

- *a* is the mean distance between a sample and all other points in the same class

- *b* is the mean distance between a sample and all other points in the next nearest cluster

- The Silhouette Coefficient for a single sample is:

$$s = \frac{b - a}{max(a, b)}$$