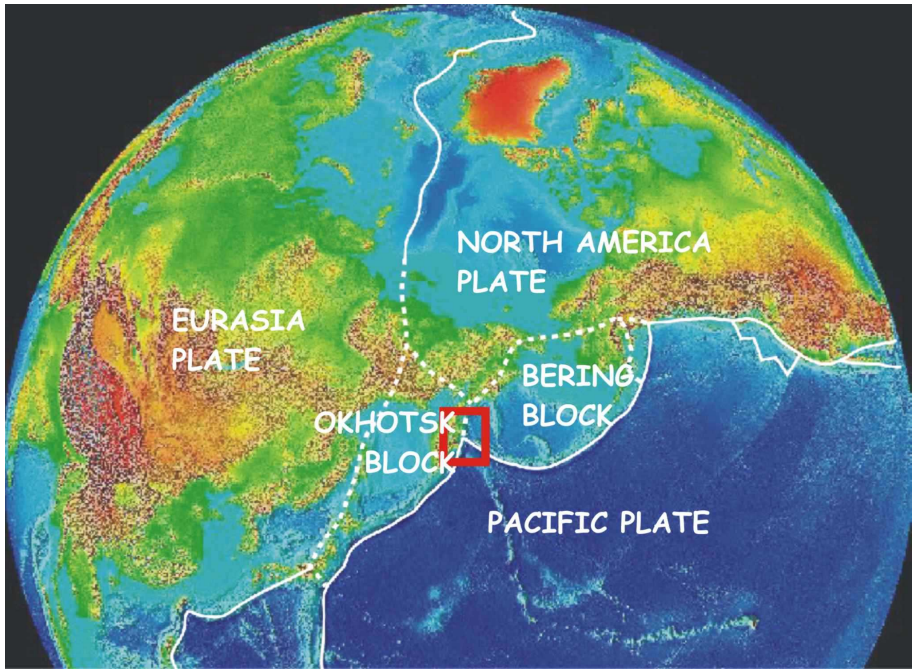
EGU General Assembly 2020

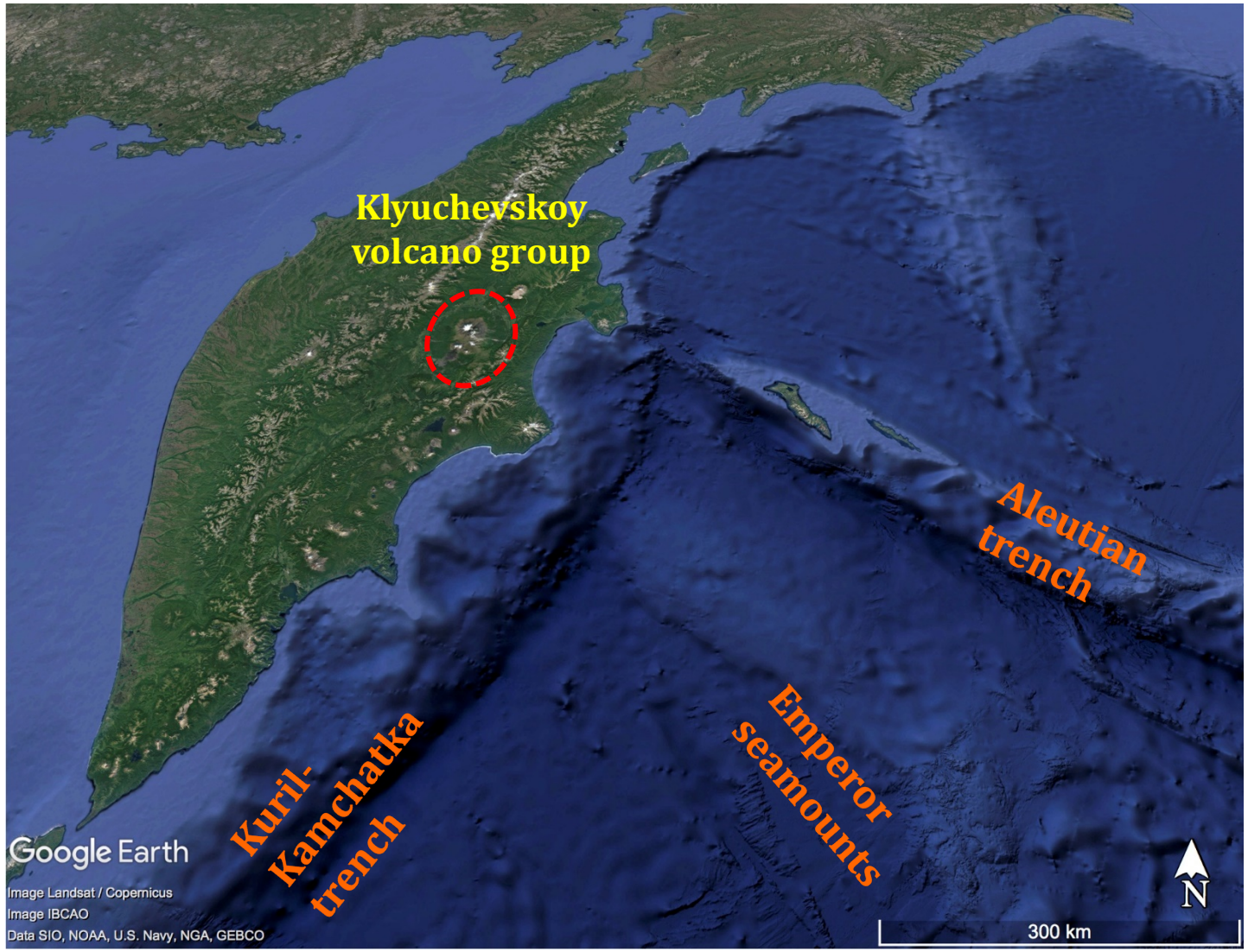
Classification of seismic events in Kamchatka (Russia) with different machine learning techniques

Natalia Galina, Nikolai Shapiro, Dmitriy Droznin, Leonard Seydoux

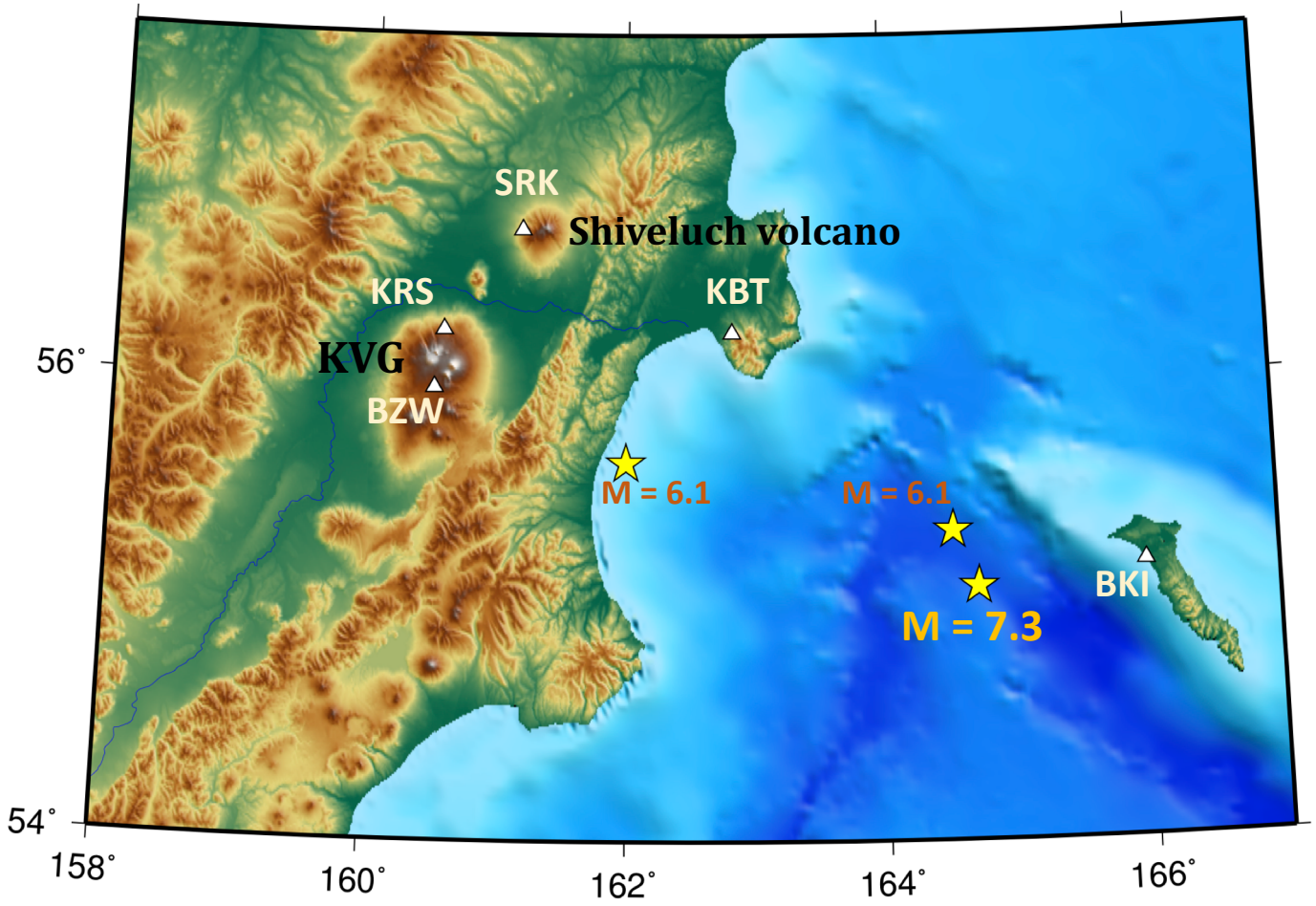
Kamchatka: present day tectonics



Kamchatka has a unique tectonic setting, it is an active subduction zone that exhibits intense seismic and volcanic activities.



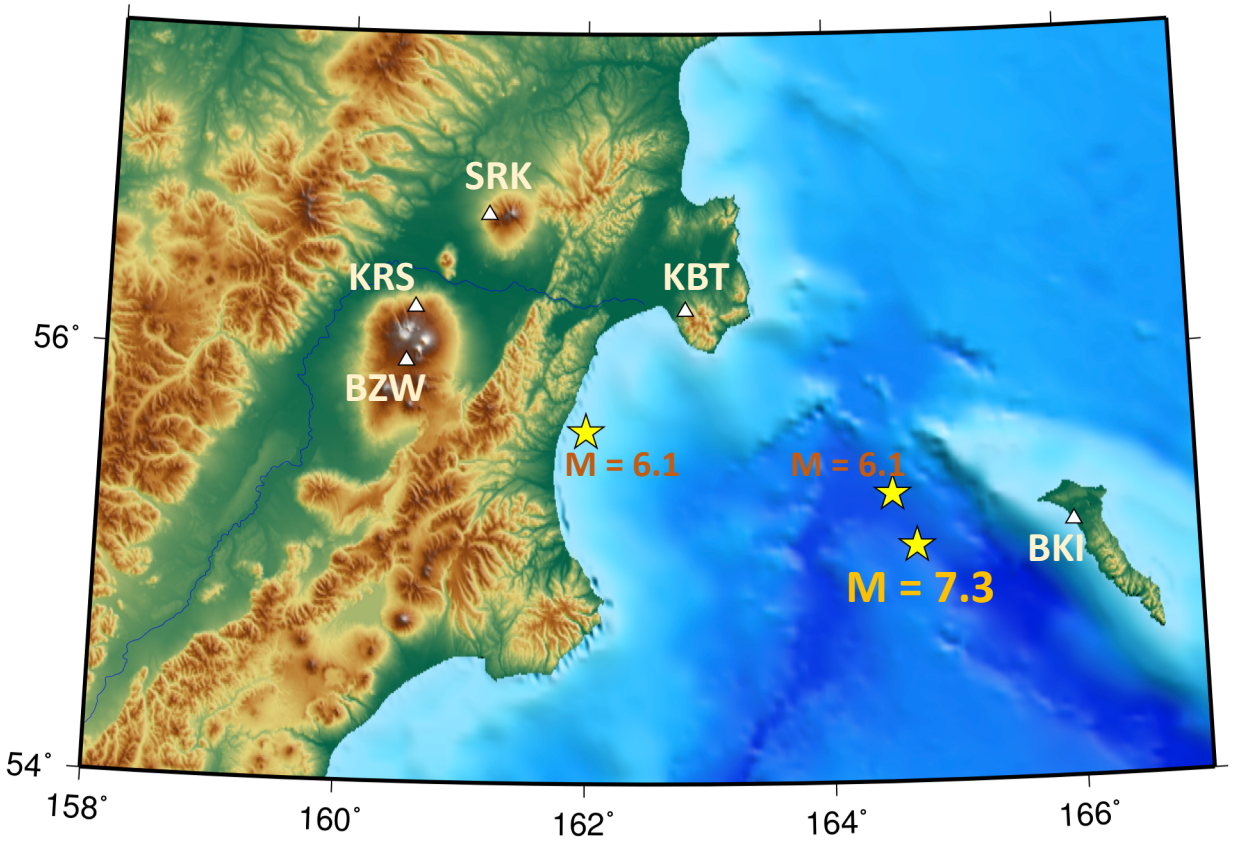
Tectonic and volcanic earthquakes are often nearly simultaneously recorded at the same station

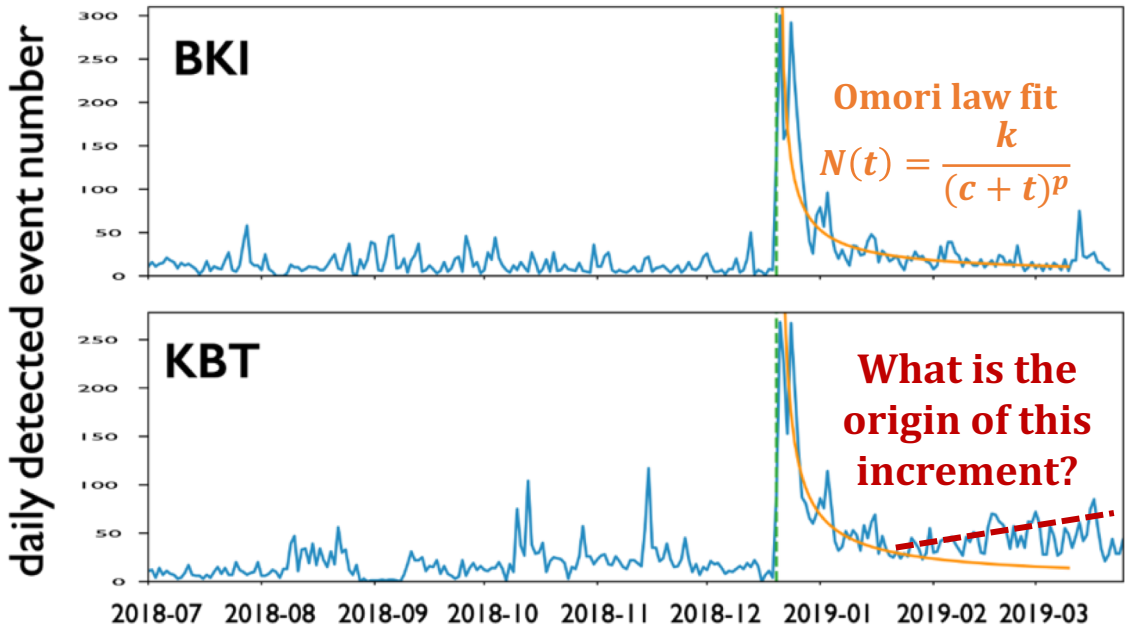


Here, we consider seismograms recorded between December 2018 and April 2019. During this time period when a M=7.3 earthquake followed by an aftershock sequence occurred nearly simultaneously with a strong eruption of Shiveluch volcano.

Detected earthquakes

July 2018 – April 2019





M=7.3 Earthquake : 2018-12-20 17:01:55 (UTC)

Shiveluch eruption : started on 2018-12-22 ???

Plots by Droznin D.

So, in this work we will study data from KBT station that recorded both tectonic and volcanic events and try to find a reason of increment in earthquakes number, i.e. was it connected to volcanic unrests or other tectonic activity

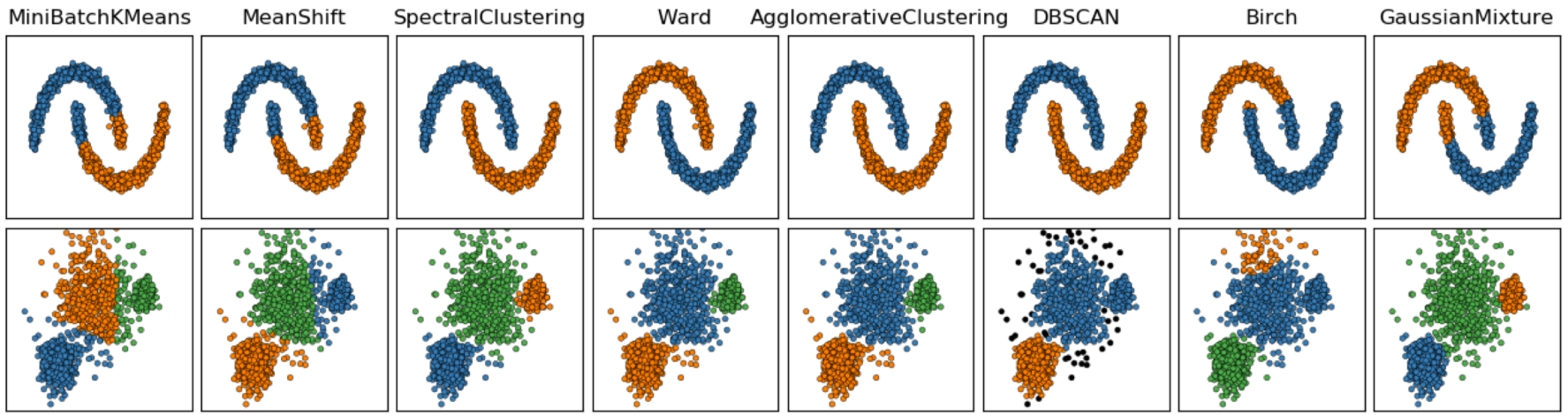
In the problem of dividing activity into two types we used both unsupervised and supervised methods of machine learning and several representations of seismic signals:

- regular features (signal duration, amplitude, peak frequency, etc.)
- smoothed and unsmoothed spectra of signal
- scattering coefficients (result of wavelet transform of a signal)

Clustering

is automatic grouping of similar objects into sets
and is the class of unsupervised machine learning methods

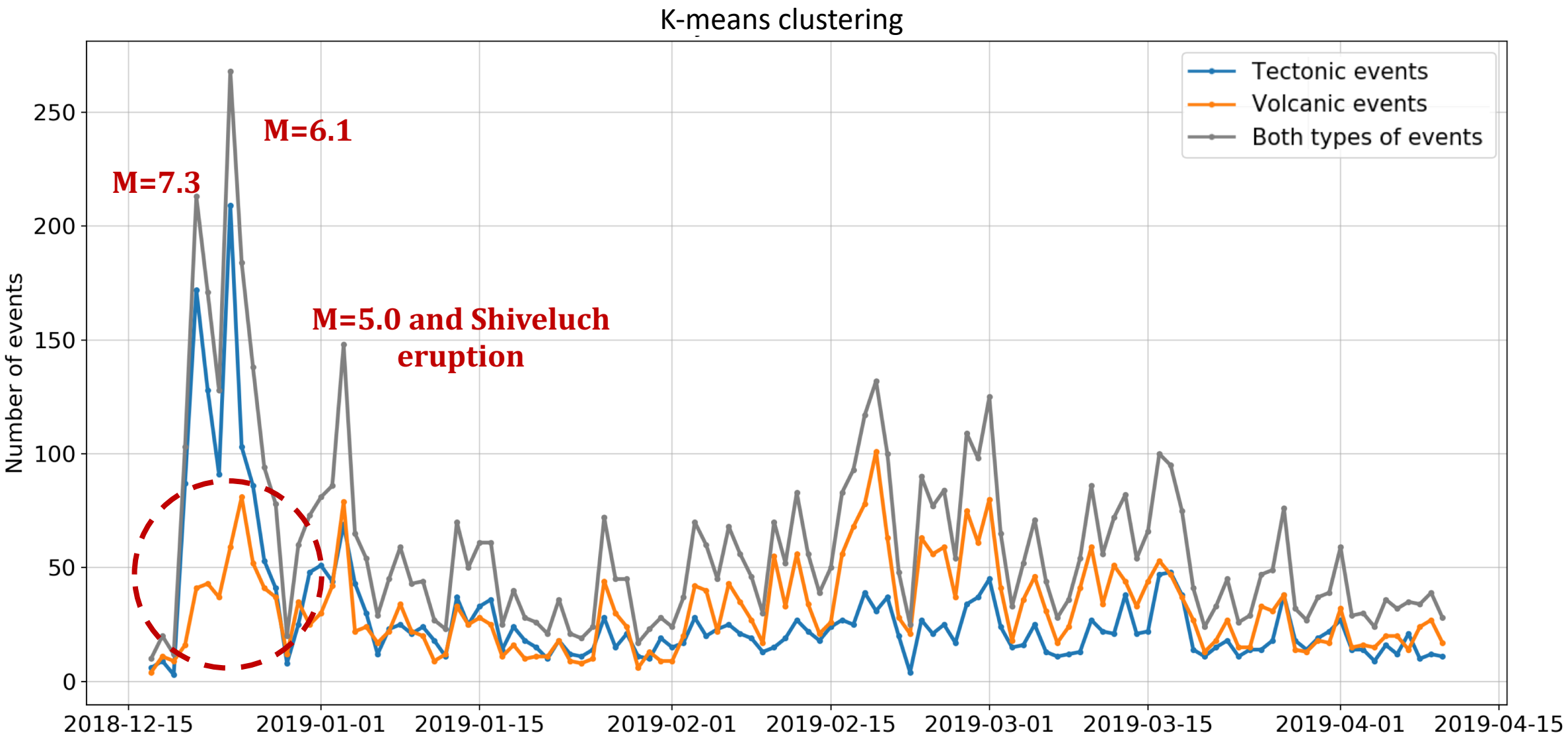
One can see that it is an ambiguous problem, and the result varies with the chosen method



from scikit-learn.org

Here we chose simple methods: K-means and Agglomerative clustering

Clustering using smoothed spectra



Shiveluch activation

December	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Шивелуч	Ж	Ж	Ж	О	О	О	О	О	О	О	О	О	О	О	О	Ж	О	О	О	О	О	К	О	О	К	К	К	К	К	К	О

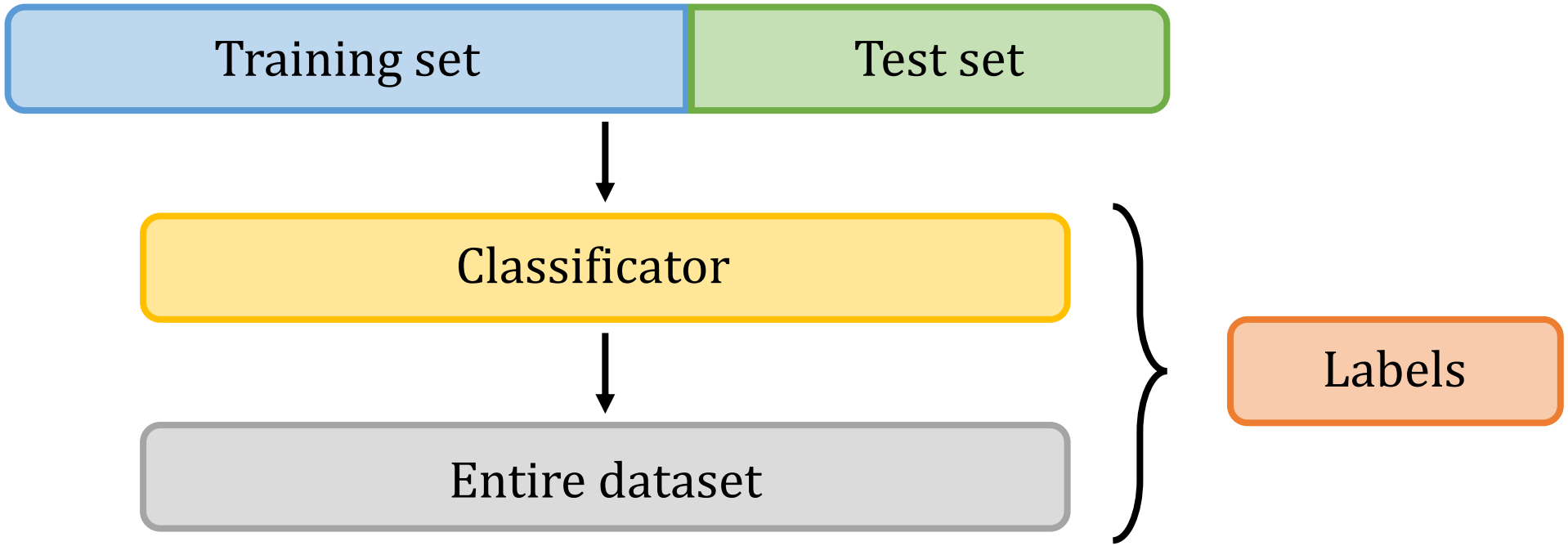
Data on the volcanic activity from emsd.ru

Classification

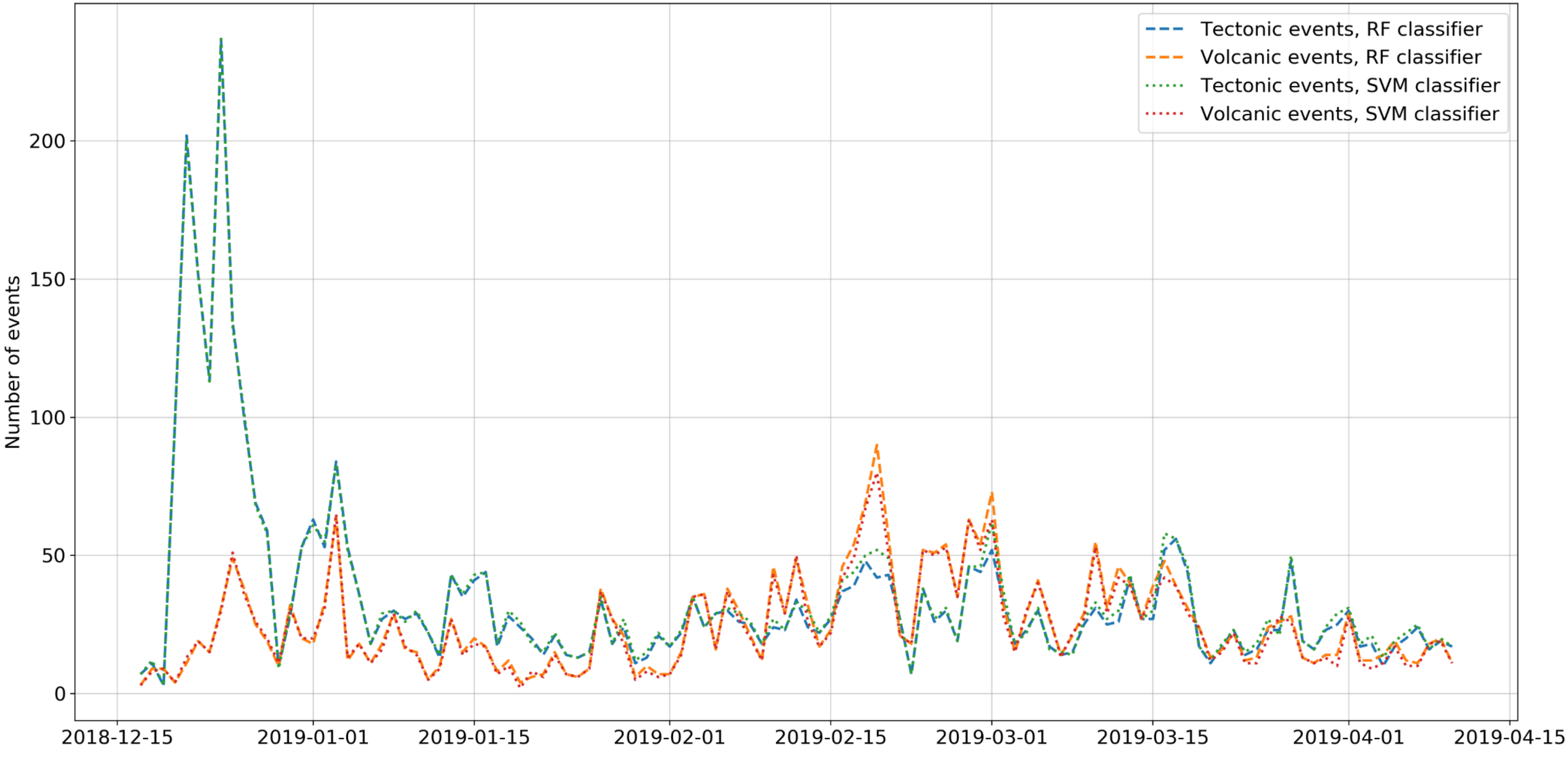
is identifying to which category an object belongs to
and is the class of supervised machine learning methods.

In this work we used next algorithms: SVM and Random Forest

Labeled data: 902 tectonic and 273 volcanic events



Classification: RF vs. SVM



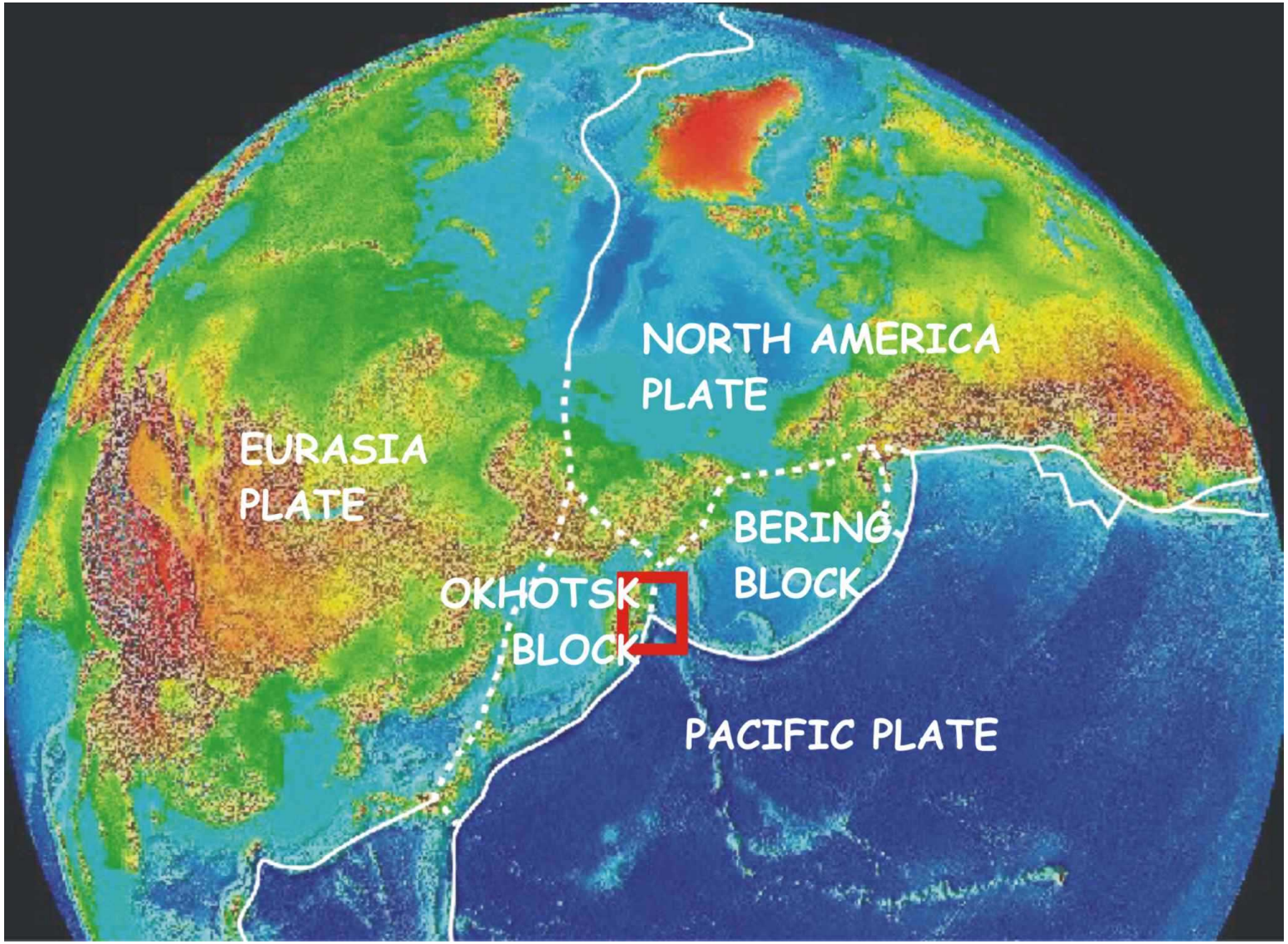
Conclusions

- The results of manual processing showed that, regardless of the signal representations used, supervised algorithms provide better results: tectonic earthquakes are less often classified as volcanic.
- Results are quite stable relatively different classifiers and their main parameters
- Deviation of the aftershocks distribution from the Omori law cannot be explained only by volcanic activity

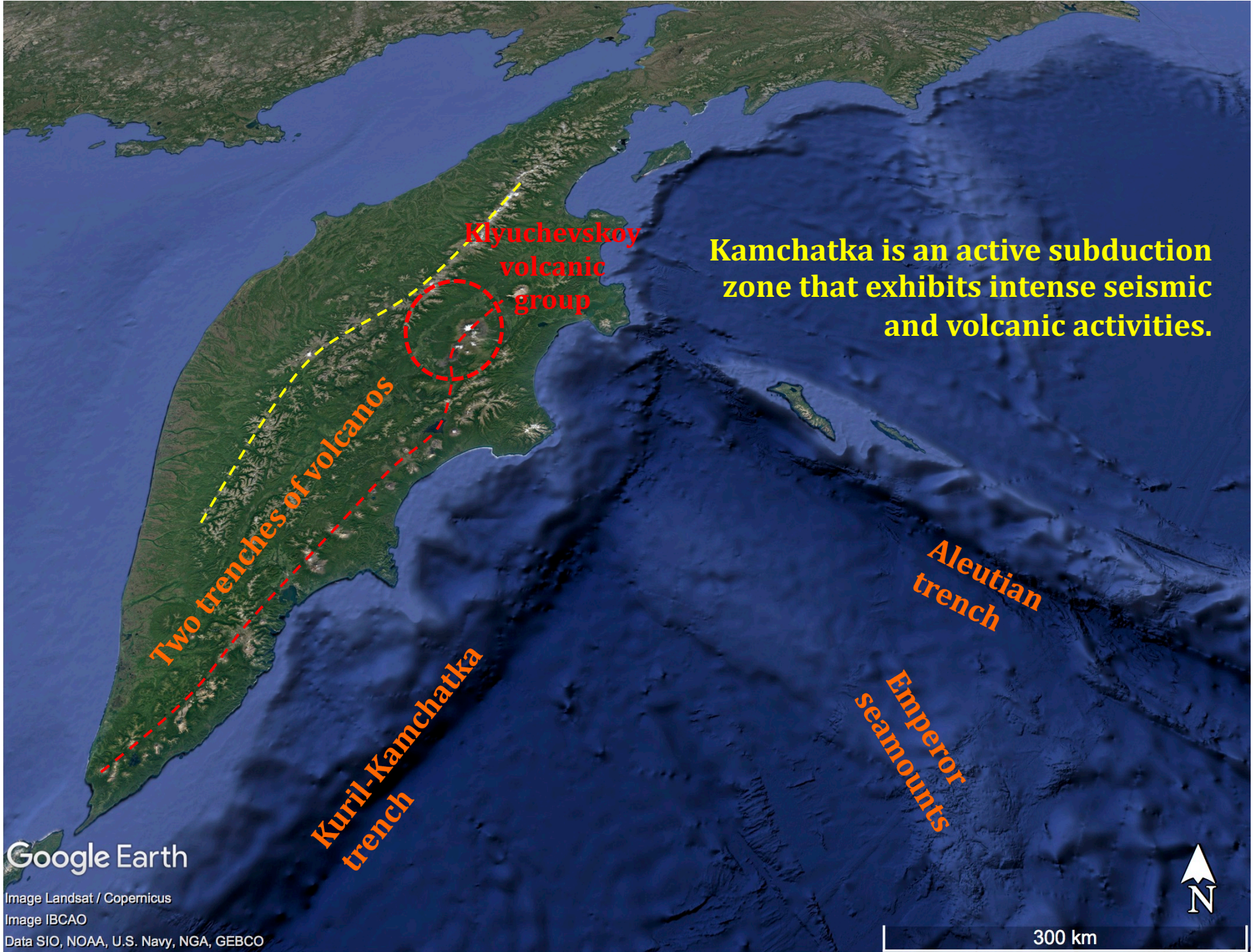
Thank you for your attention!

Appendix:
Detailed explanations

Kamchatka: present day tectonics



Kamchatka has a unique tectonic setting

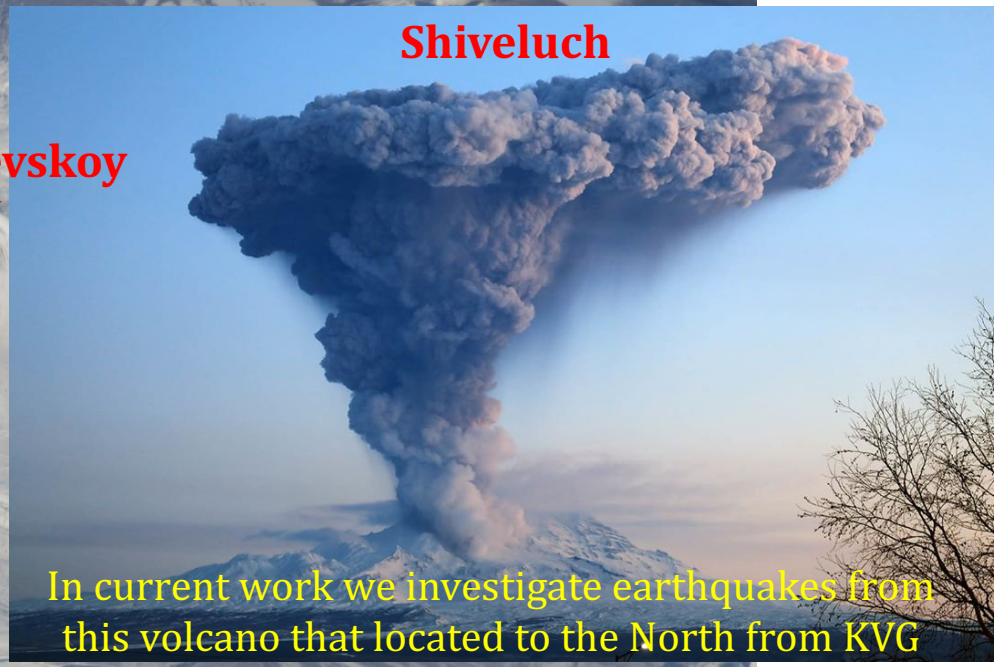

Kyuchevskoy volcanic group
Two trenches of volcanos
Kuril-Kamchatka trench
Aleutian trench
Emperor seamounts
Kamchatka is an active subduction zone that exhibits intense seismic and volcanic activities.
Google Earth
Image Landsat / Copernicus
Image IBCAO
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
N
300 km

Klyuchevskoy volcanic group (KVG)

Klyuchevskoy

Bezymyanny

Tolbachik



In current work we investigate earthquakes from this volcano that located to the North from KVG

- Largest cluster of subduction volcanoes in the world
- 13 strato-volcanoes with different compositions and eruption styles
- 3 volcanoes active at present
- Magma production rate ~$1m^3$/sec (comparable to Hawaii)
- Possible connection with Hawaii-Emperor seamounts

Kamchatka seismicity and catalogs

01.01.2018 – 01.04.2019 (~450 days)

9603	catalogued regional earthquakes (locations + magnitudes)
14574	catalogued volcanic earthquakes (locations + magnitudes)
24768	volcanic earthquakes on Shiveluch
9890	volcanic earthquakes on Klyuchevskoy
5683	volcanic earthquakes on Tolbachik
603	volcanic earthquakes on Bezymianny

Kamchatka seismicity and catalogs

On average: 20 regional and 90 volcanic (30 located) earthquakes per day

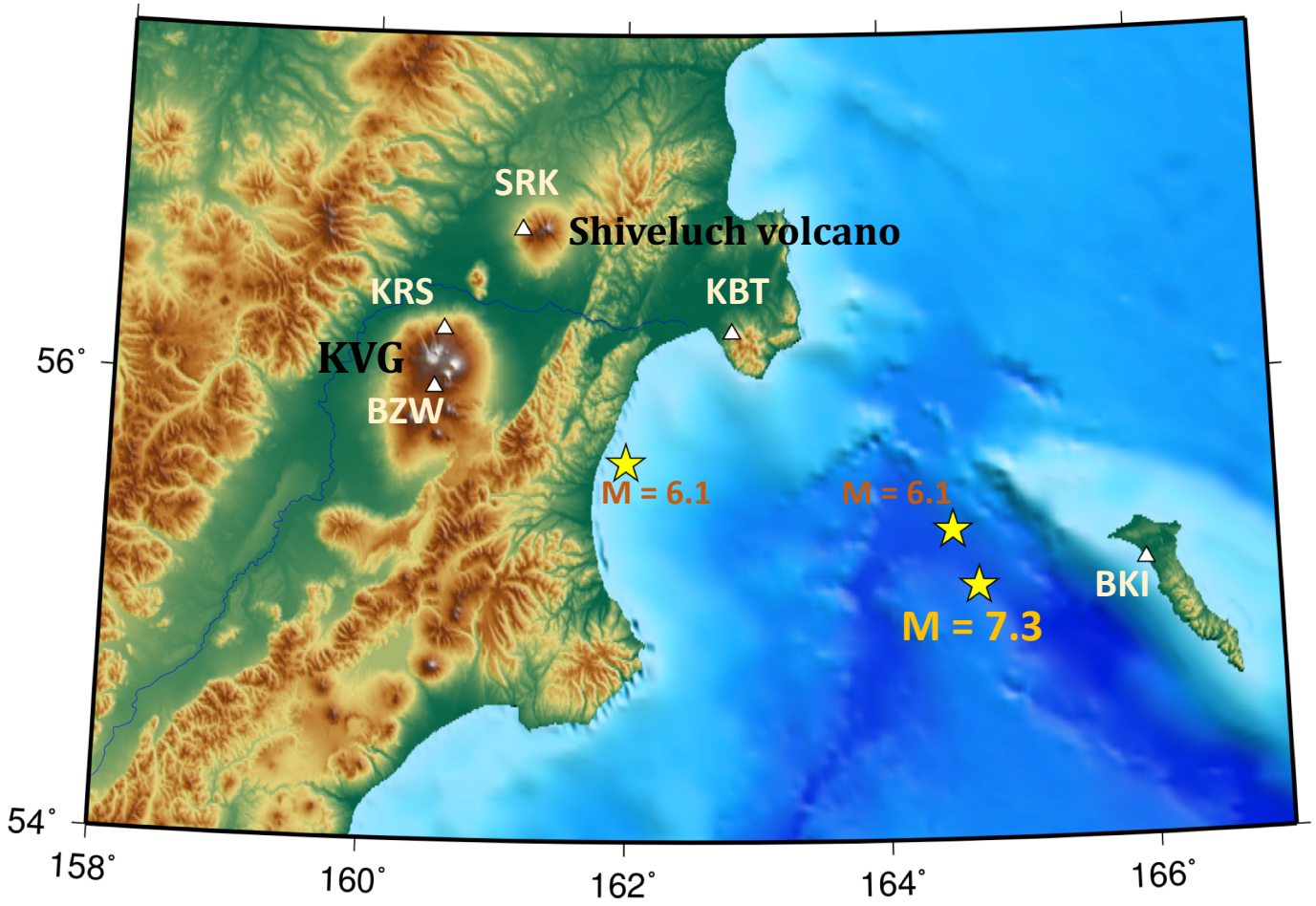
This activity strongly increases during eruptions and aftershock sequences

(100-ds events per day)

1. Amount of data is too big for manual processing
2. The situation with volcanic events is complicated due to their low-amplitudes of wave arrivals (so it is impossible to locate them)

These reasons led us to investigating of the automatic seismograms processing algorithm with detection of events and their further classification (seismic or volcanic class at first stage)

Tectonic and volcanic earthquakes are often nearly simultaneously recorded at the same station



Here, we consider seismograms recorded between December 2018 and April 2019. During this time period when a M=7.3 earthquake followed by an aftershock sequence occurred nearly simultaneously with a strong eruption of Shiveluch volcano.

Let us consider the cumulative number of events detected on different stations

Information on stations location:

BKI is far from the coast but close to main tectonic events in the observed period of time

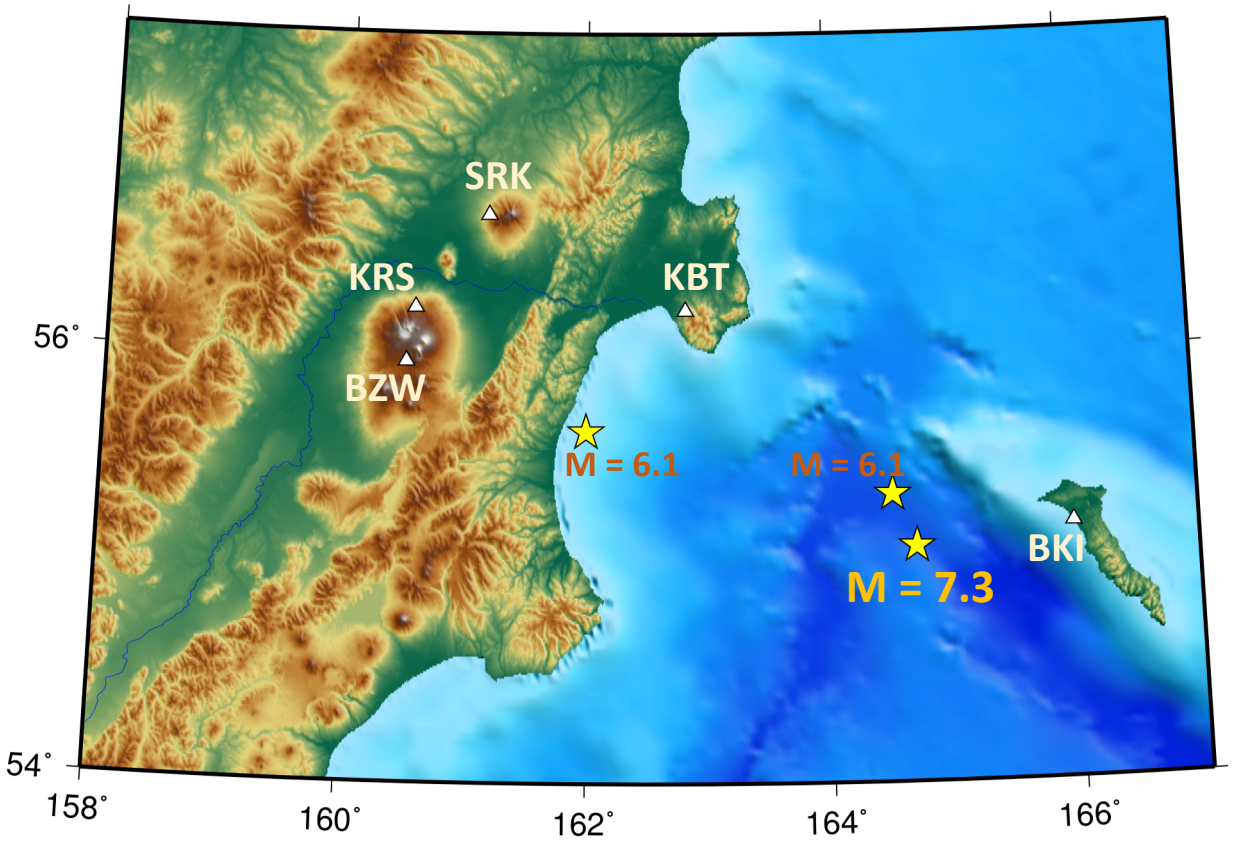
SRK is located on the slope of Shiveluch

KBT is just on the coastline and equal distanced from location of tectonic events and Shiveluch

KRS and BZW are stations in the KVG

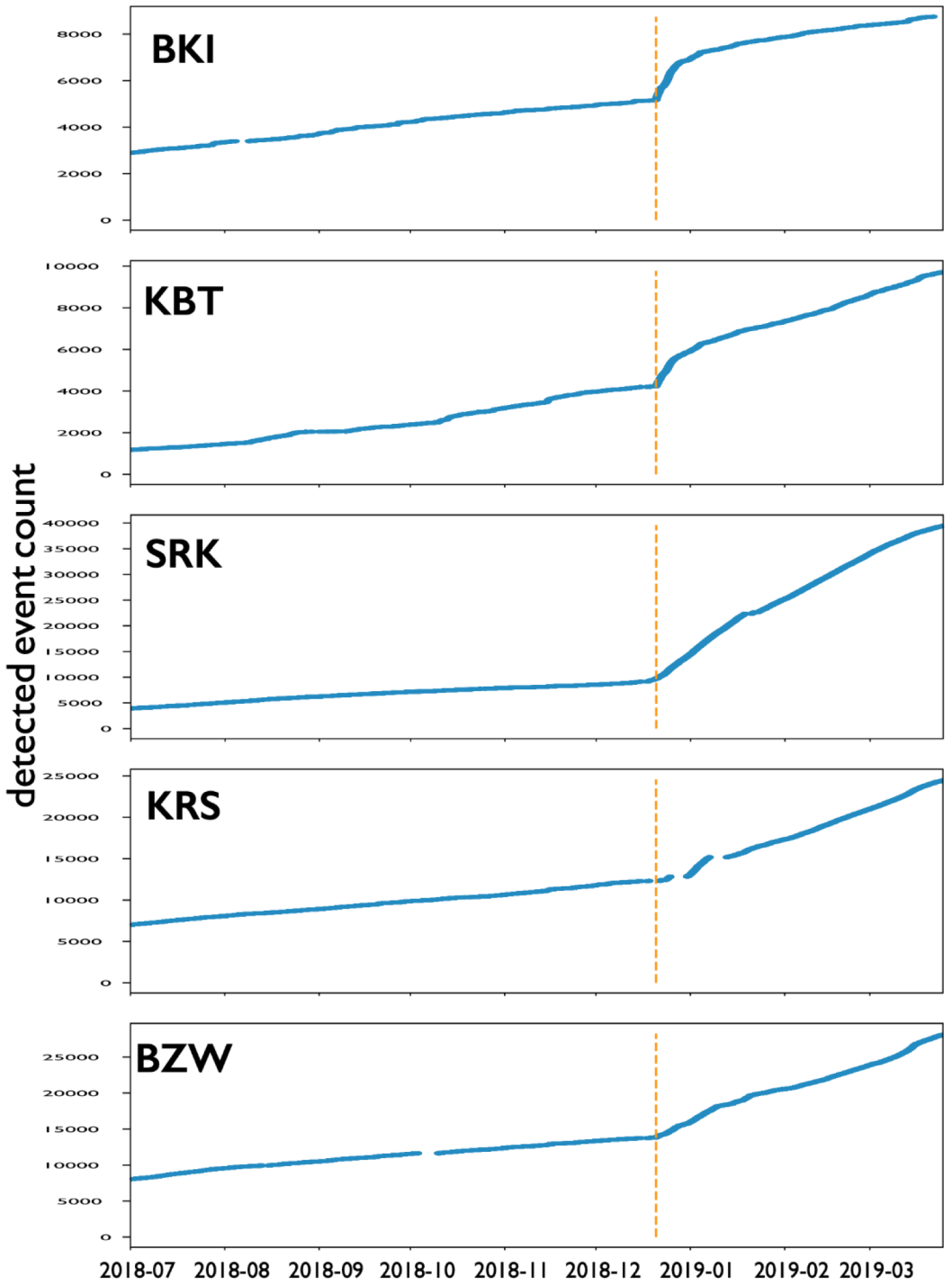
Detected earthquakes

July 2018 – April 2019



M=7.3 Earthquake : 2018-12-20 17:01:55 (UTC)

Shiveluch eruption : started on 2018-12-22 ???



Plots by Droznin D.

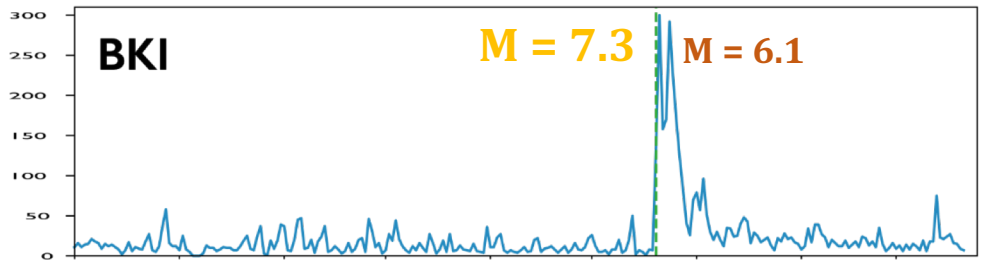
Well, cumulative plots look pretty similar with a leap at the same time.

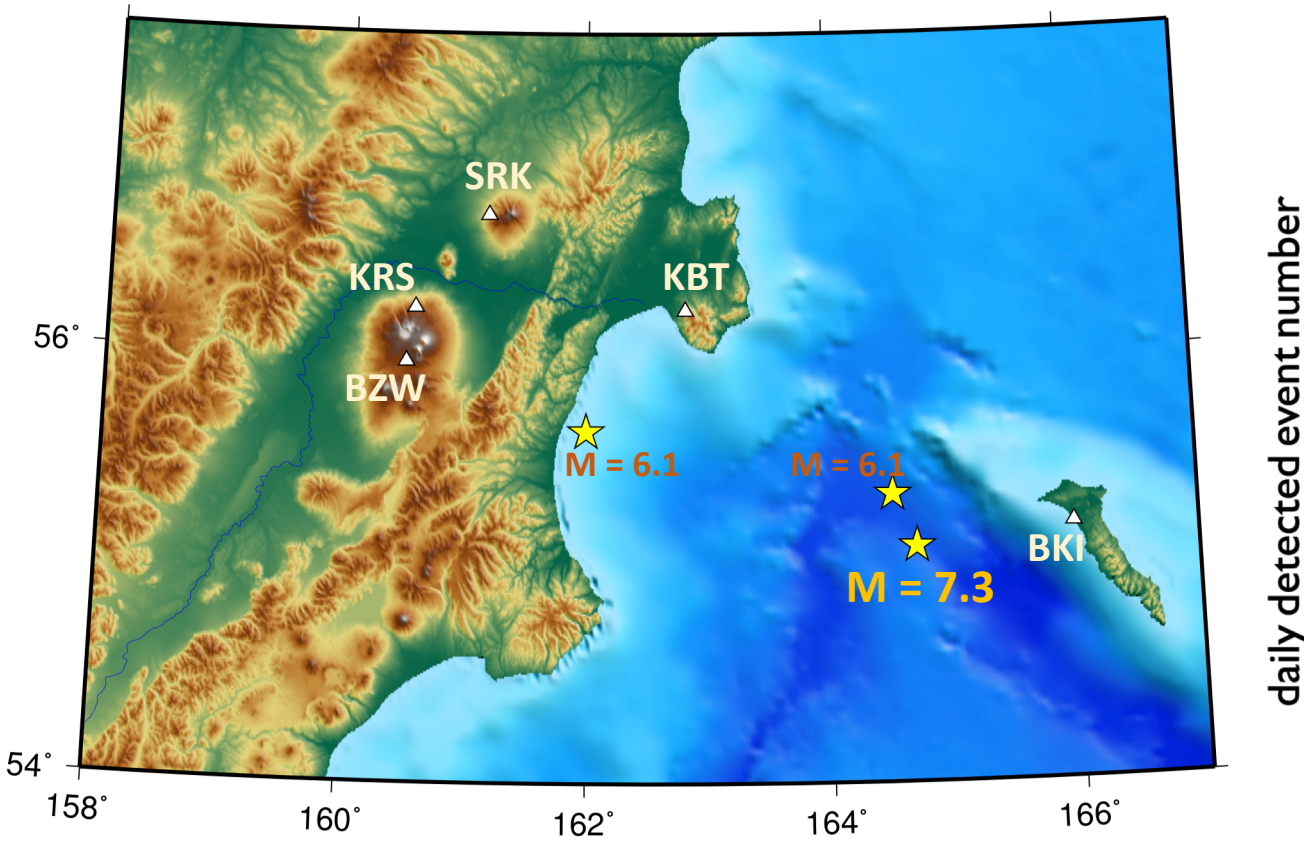
But is the reason of this leap same at all stations?

Let us have a look at noncumulative plots…

Detected earthquakes

July 2018 – April 2019





daily detected event number

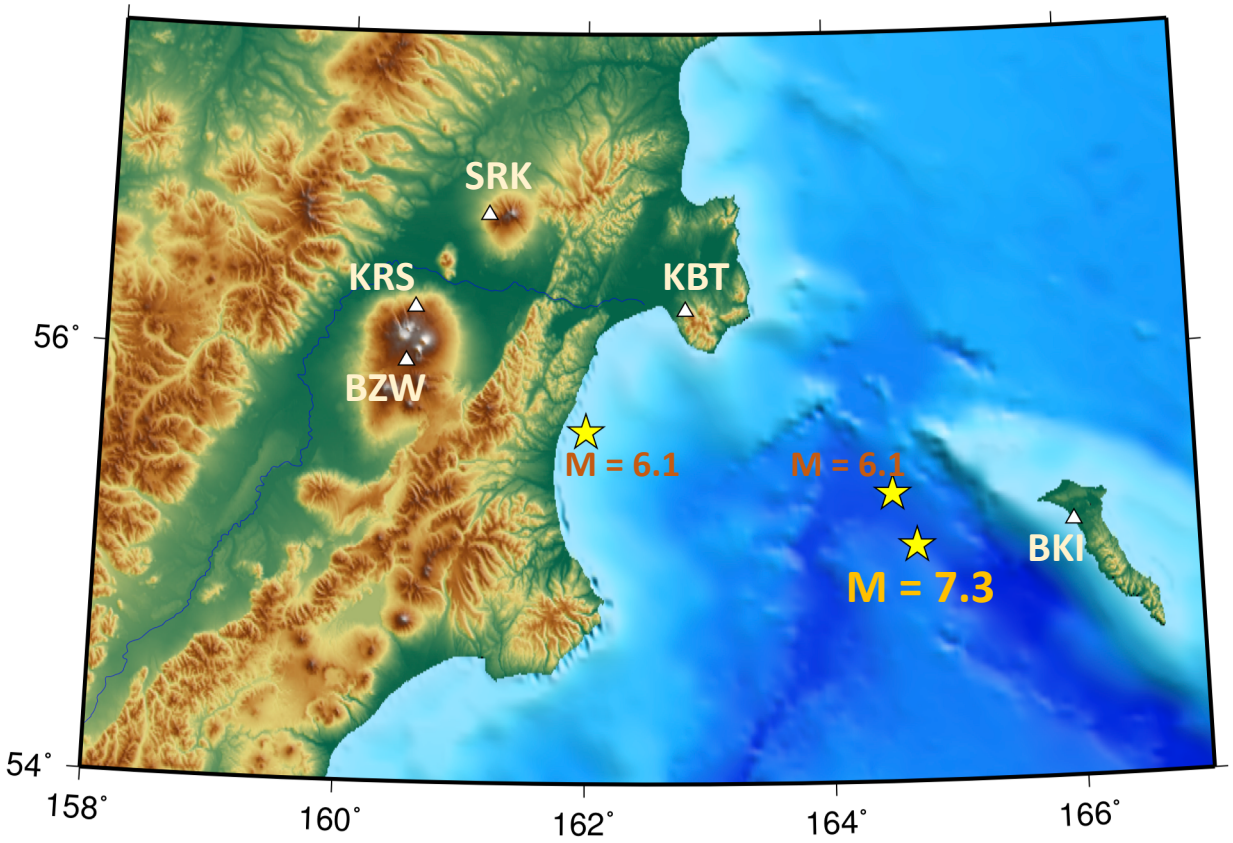
M=7.3 Earthquake : 2018-12-20 17:01:55 (UTC)

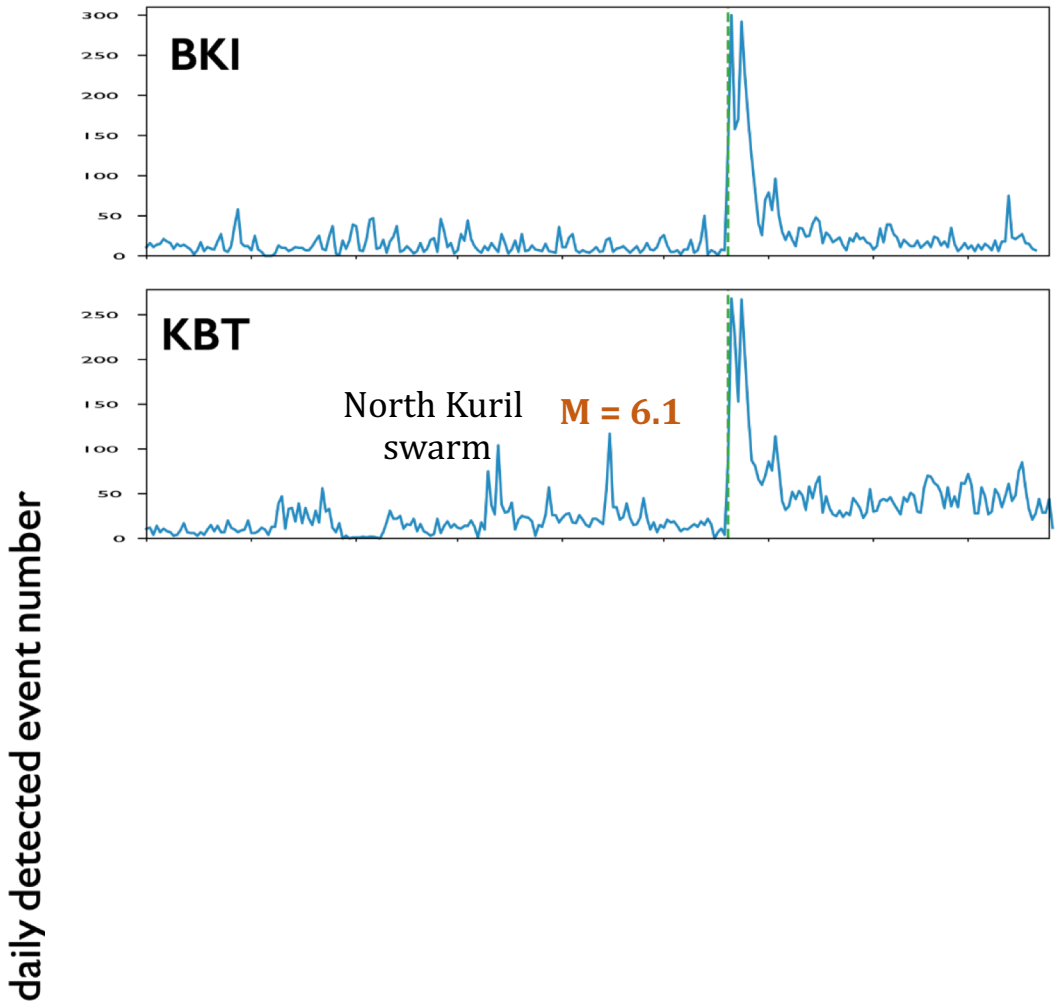
Shiveluch eruption : started on 2018-12-22 ???

Plots by Droznin D.

Detected earthquakes

July 2018 – April 2019





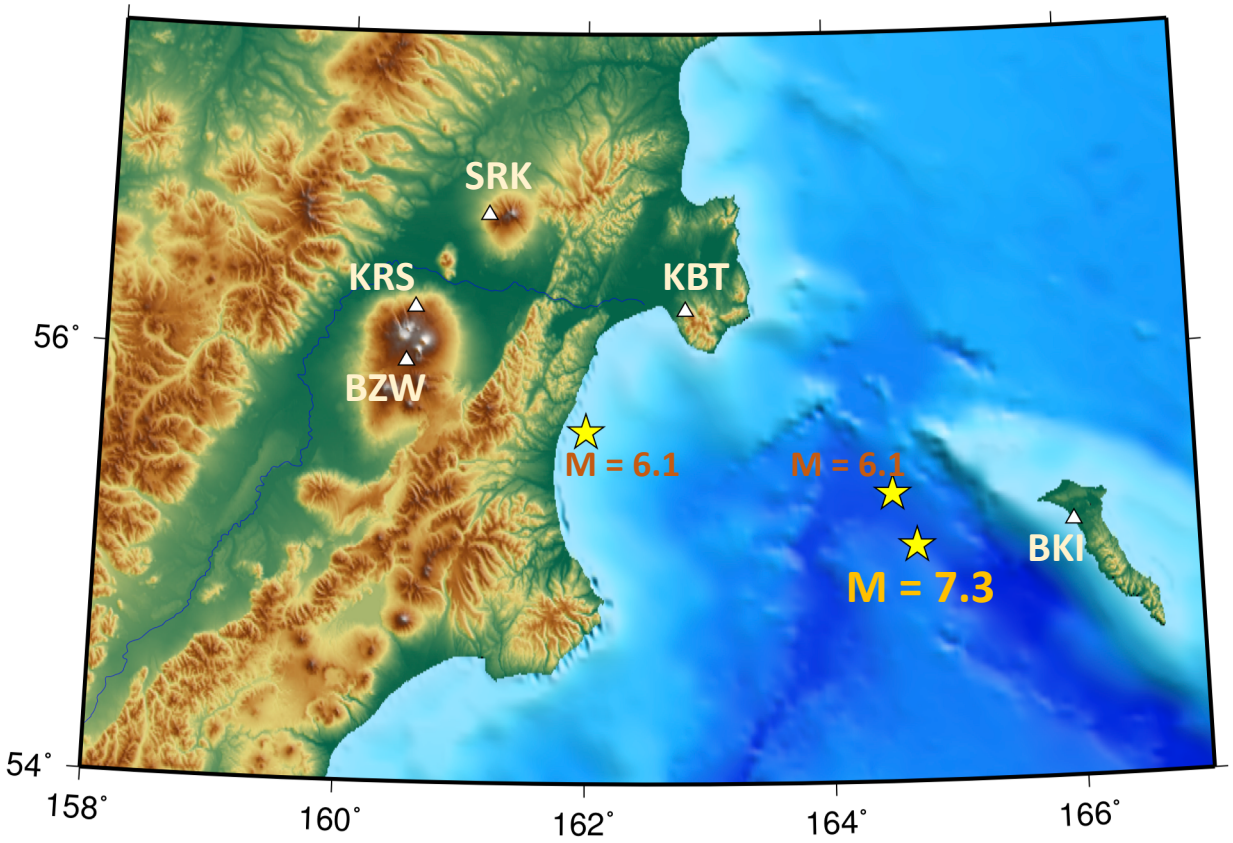
M=7.3 Earthquake : 2018-12-20 17:01:55 (UTC)

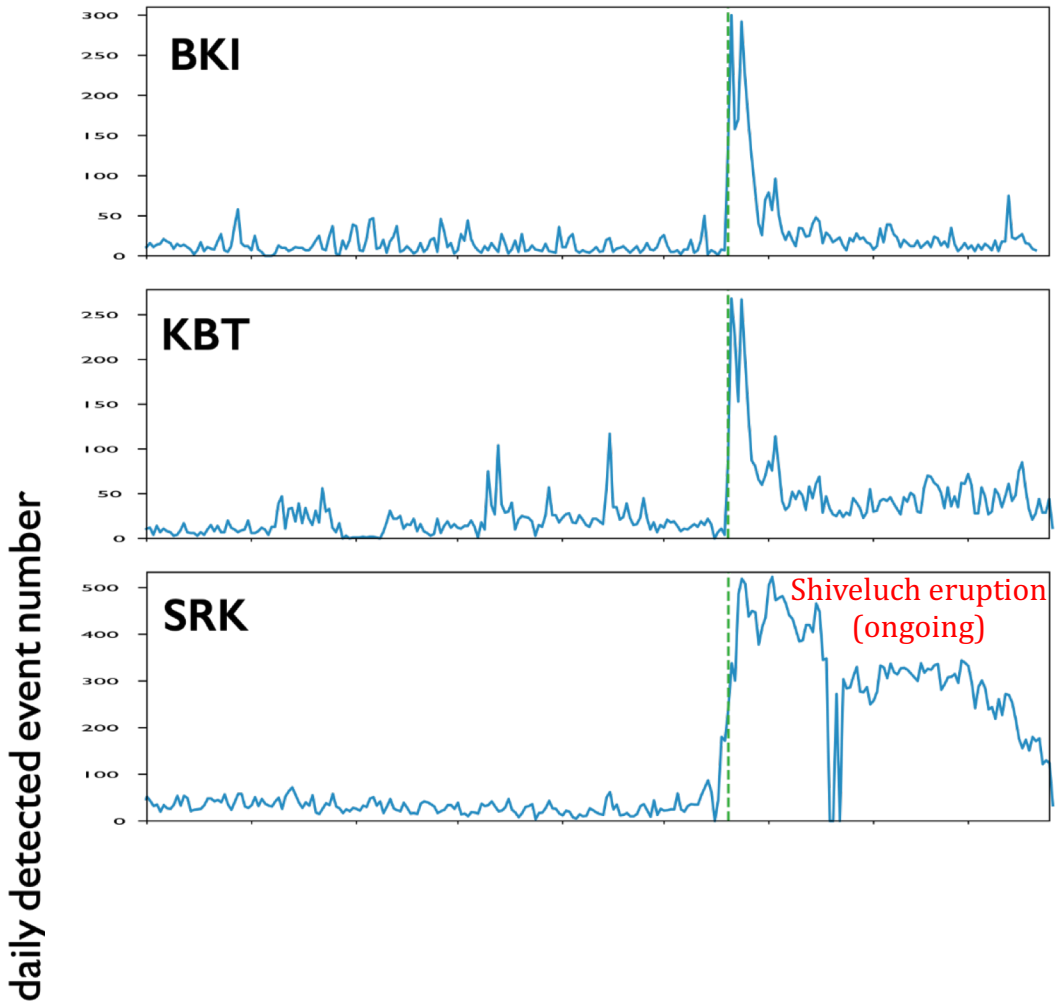
Shiveluch eruption : started on 2018-12-22 ???

Plots by Droznin D.

Detected earthquakes

July 2018 – April 2019





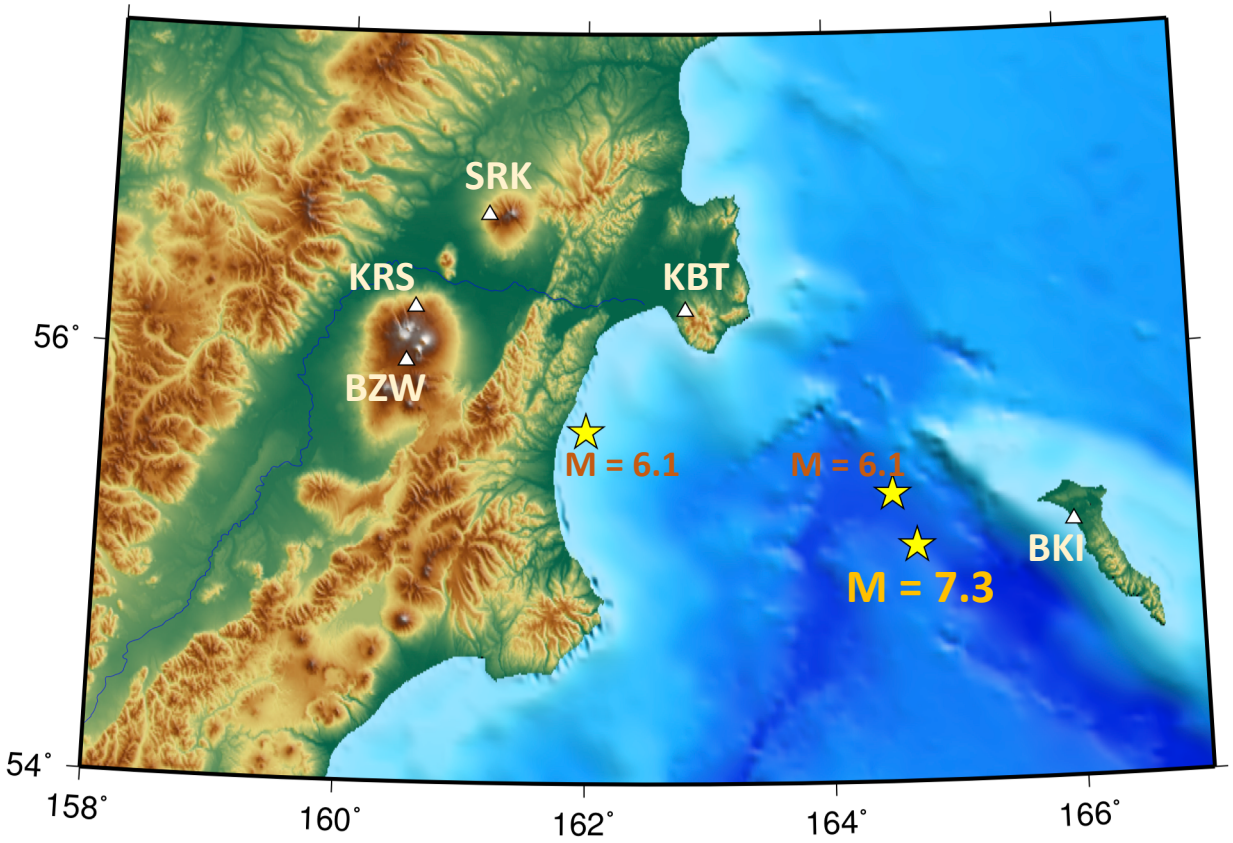
M=7.3 Earthquake : 2018-12-20 17:01:55 (UTC)

Shiveluch eruption : started on 2018-12-22 ???

Plots by Droznin D.

Detected earthquakes

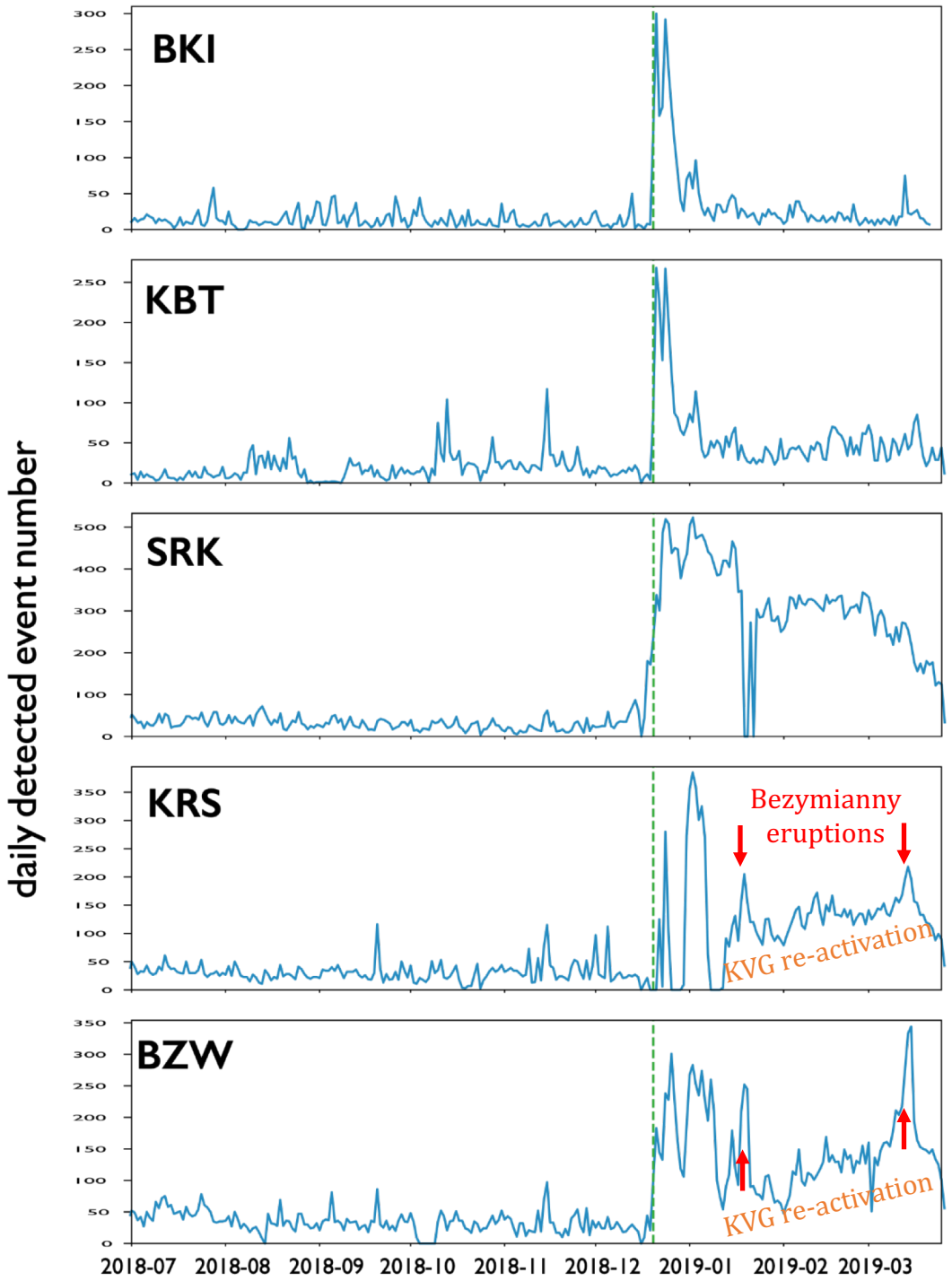
July 2018 – April 2019



M=7.3 Earthquake : 2018-12-20 17:01:55 (UTC)

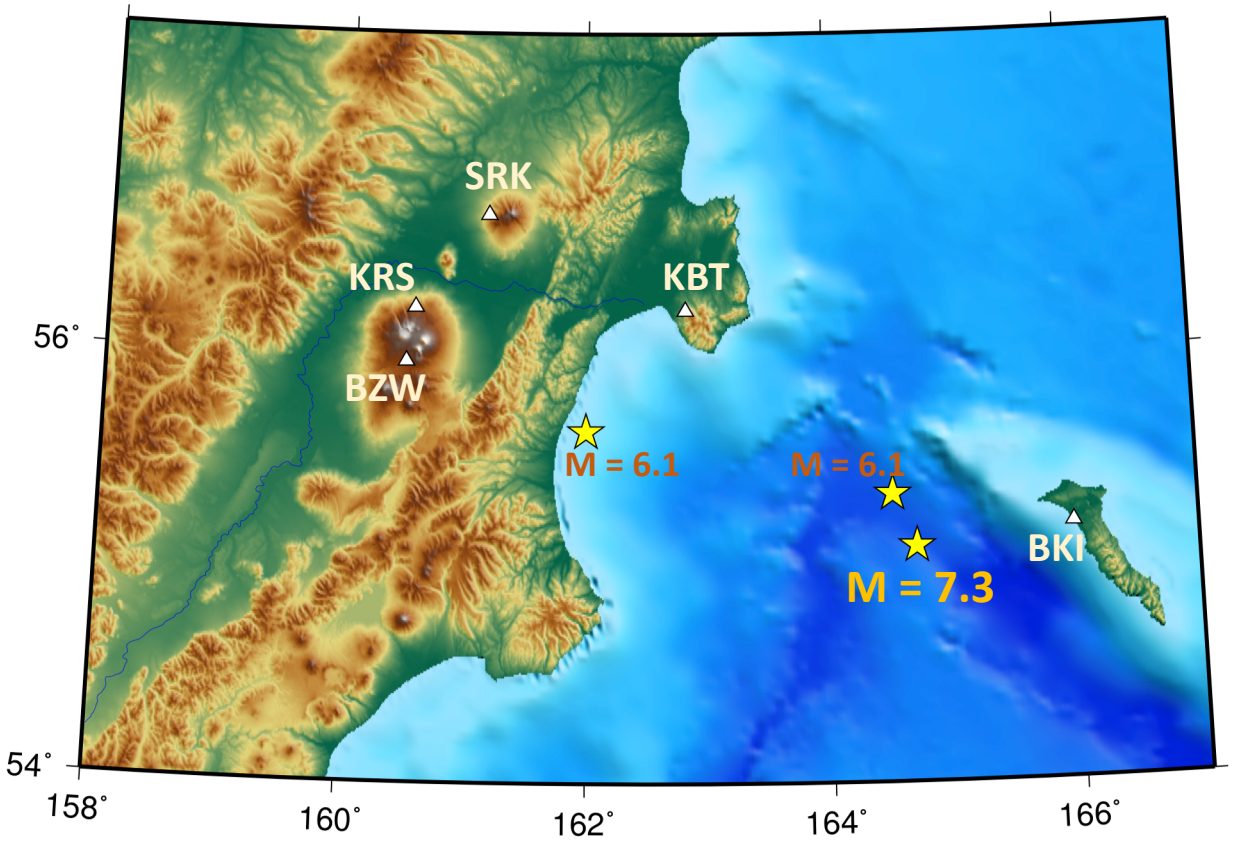
Shiveluch eruption : started on 2018-12-22 ???

Plots by Droznin D.



Detected earthquakes

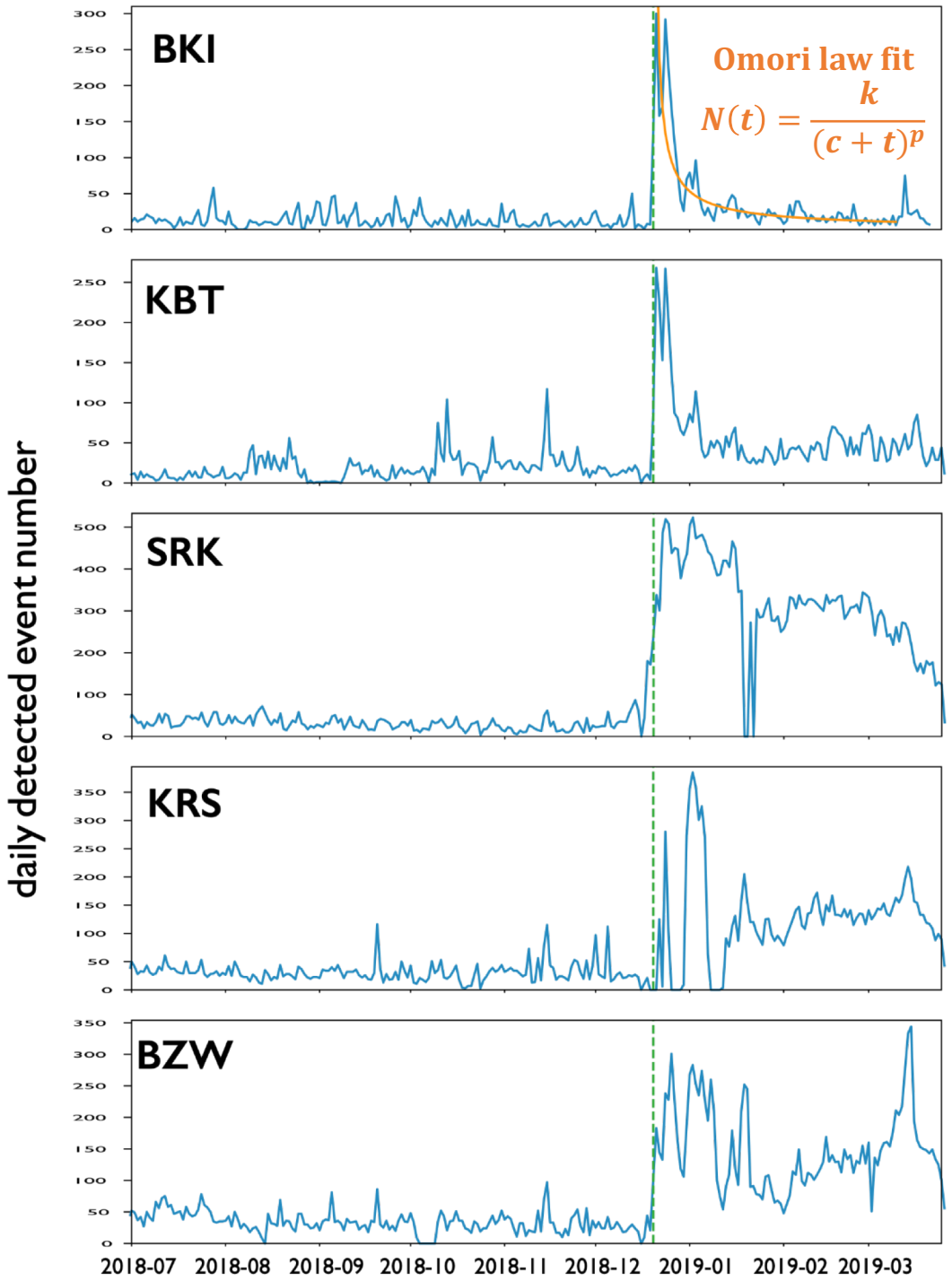
July 2018 – April 2019



M=7.3 Earthquake : 2018-12-20 17:01:55 (UTC)

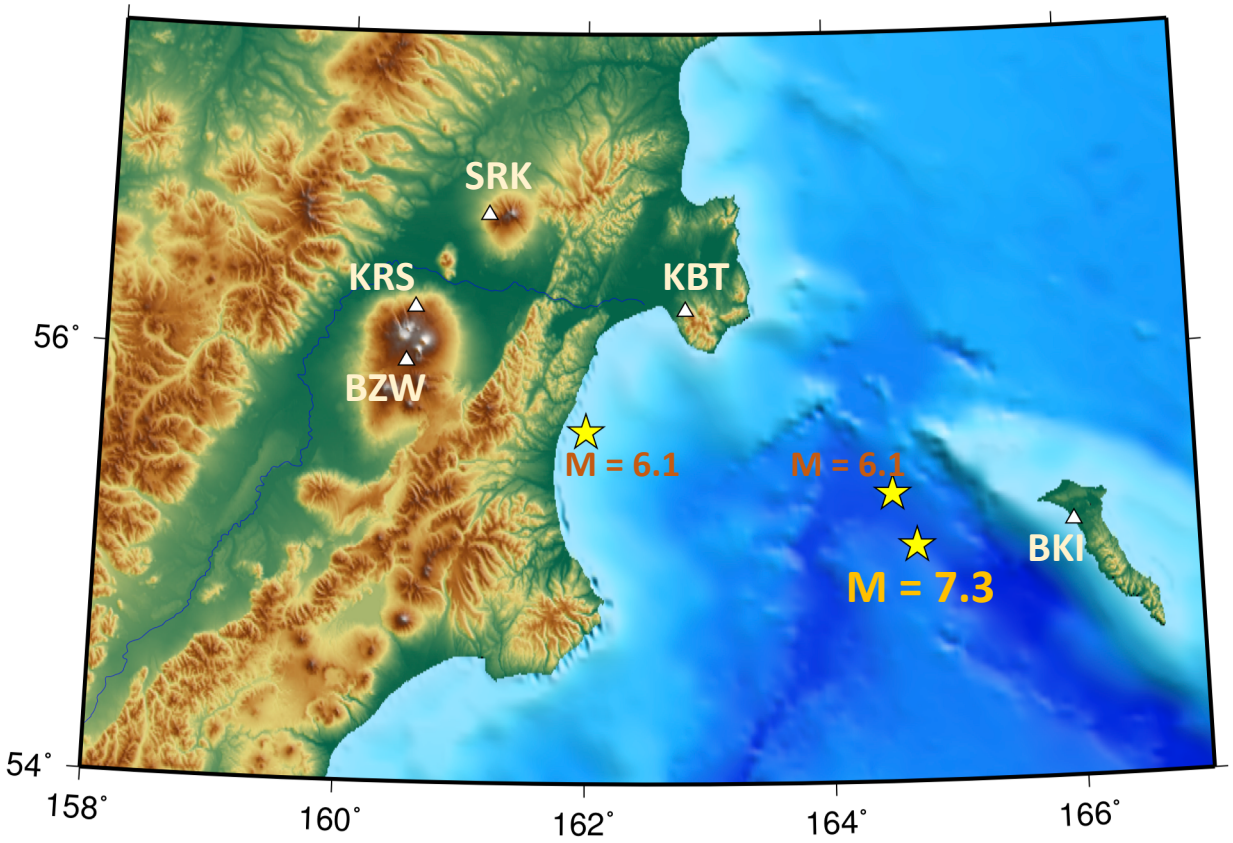
Shiveluch eruption : started on 2018-12-22 ???

Plots by Droznin D.



Detected earthquakes

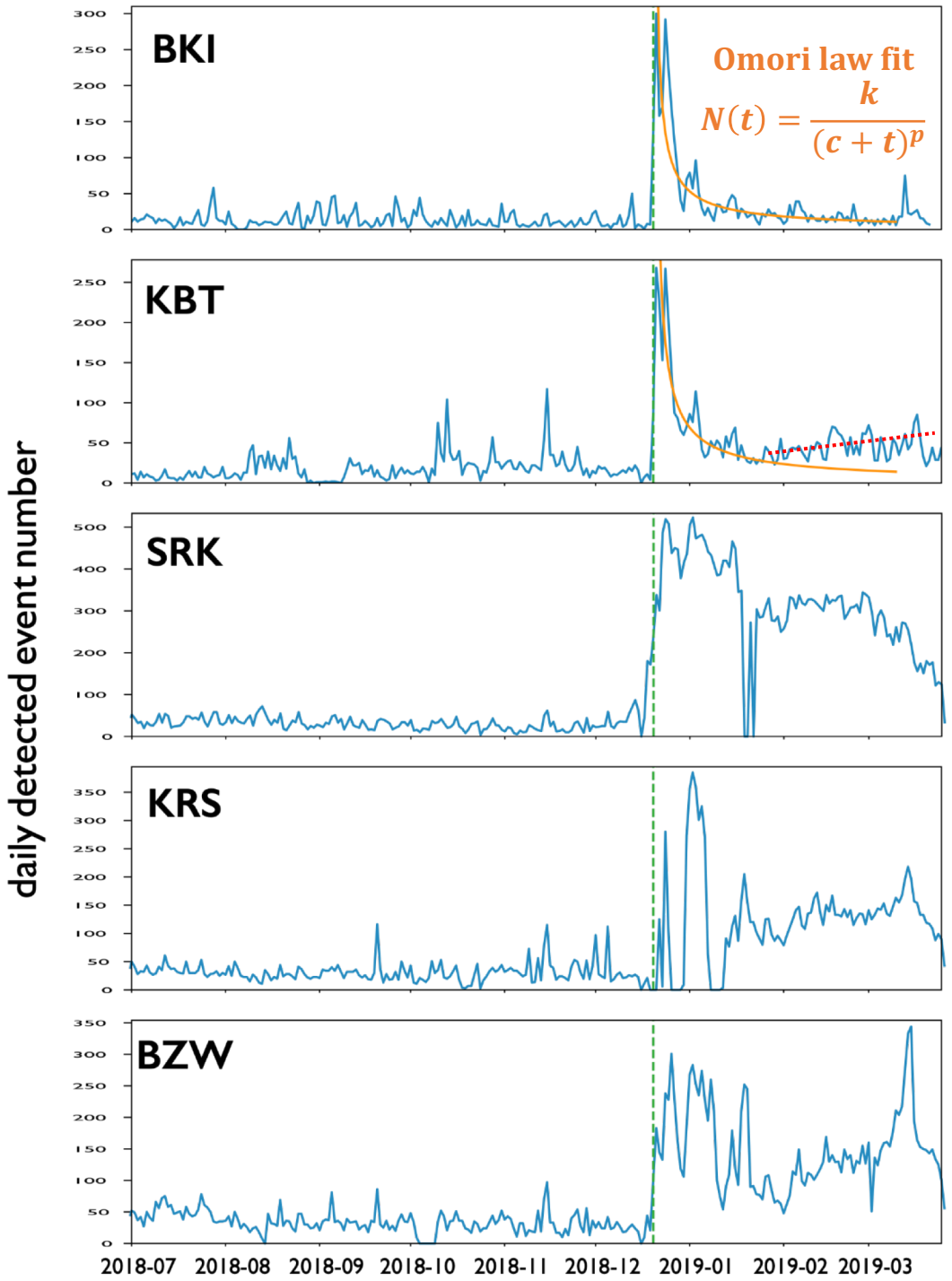
July 2018 – April 2019



M=7.3 Earthquake : 2018-12-20 17:01:55 (UTC)

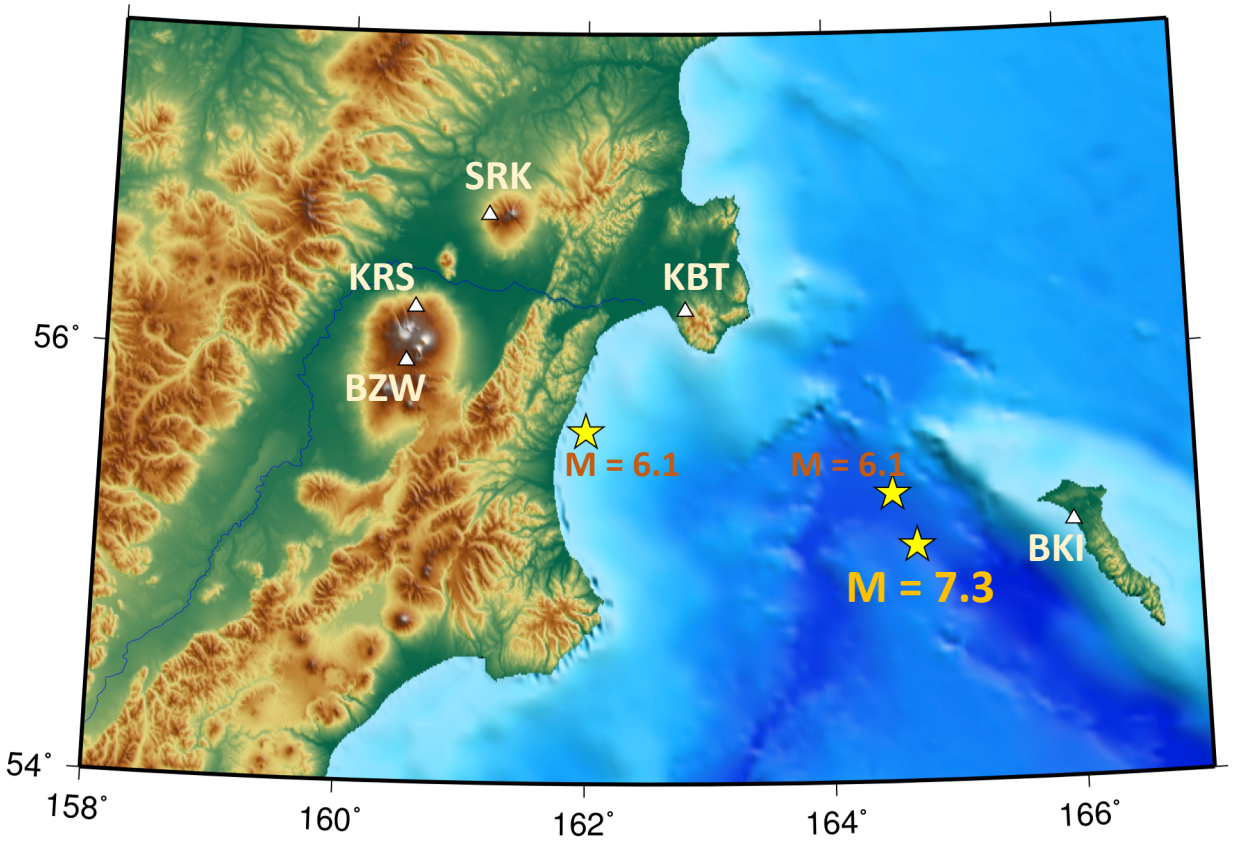
Shiveluch eruption : started on 2018-12-22 ???

Plots by Droznin D.



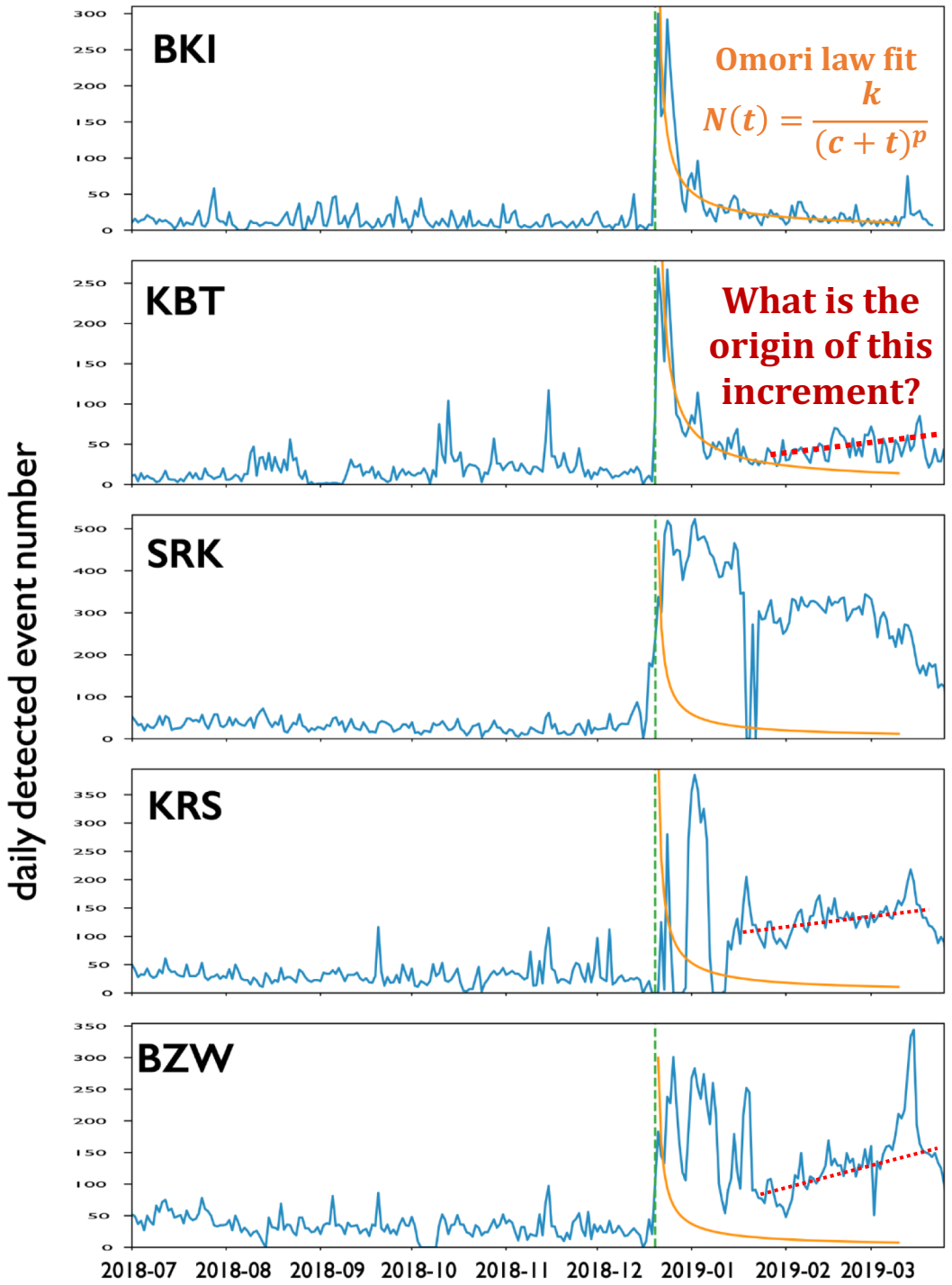
Detected earthquakes

July 2018 – April 2019



M=7.3 Earthquake : 2018-12-20 17:01:55 (UTC)

Shiveluch eruption : started on 2018-12-22 ???



Plots by Droznin D.

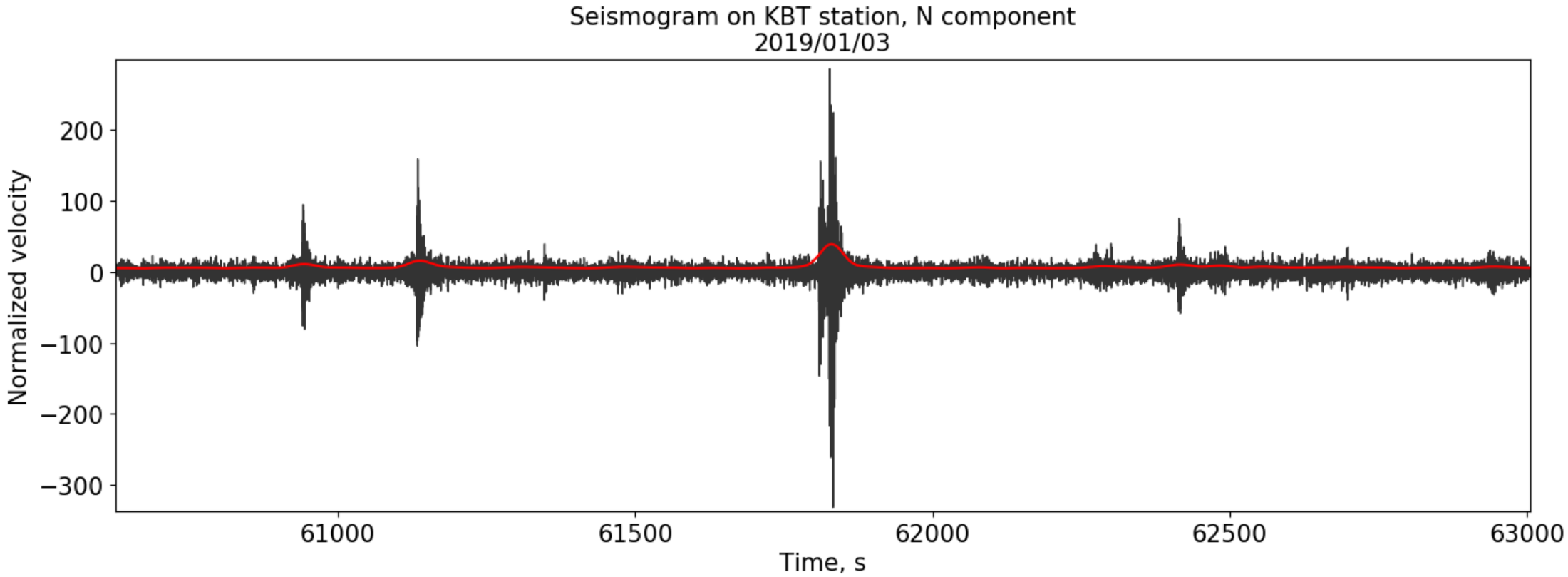
So, in this work we will study data from KBT station that recorded both tectonic and volcanic events and try to find a reason of increment in earthquakes number, i.e. was it connected to volcanic unrests or other tectonic activity

Outline

- Seismogram processing and earthquake detection
- Feature extraction
- Creating labeled set for supervised learning
- Spectra smoothing
- Results of clustering and classification
- Conclusions and further directions

Seismogram processing

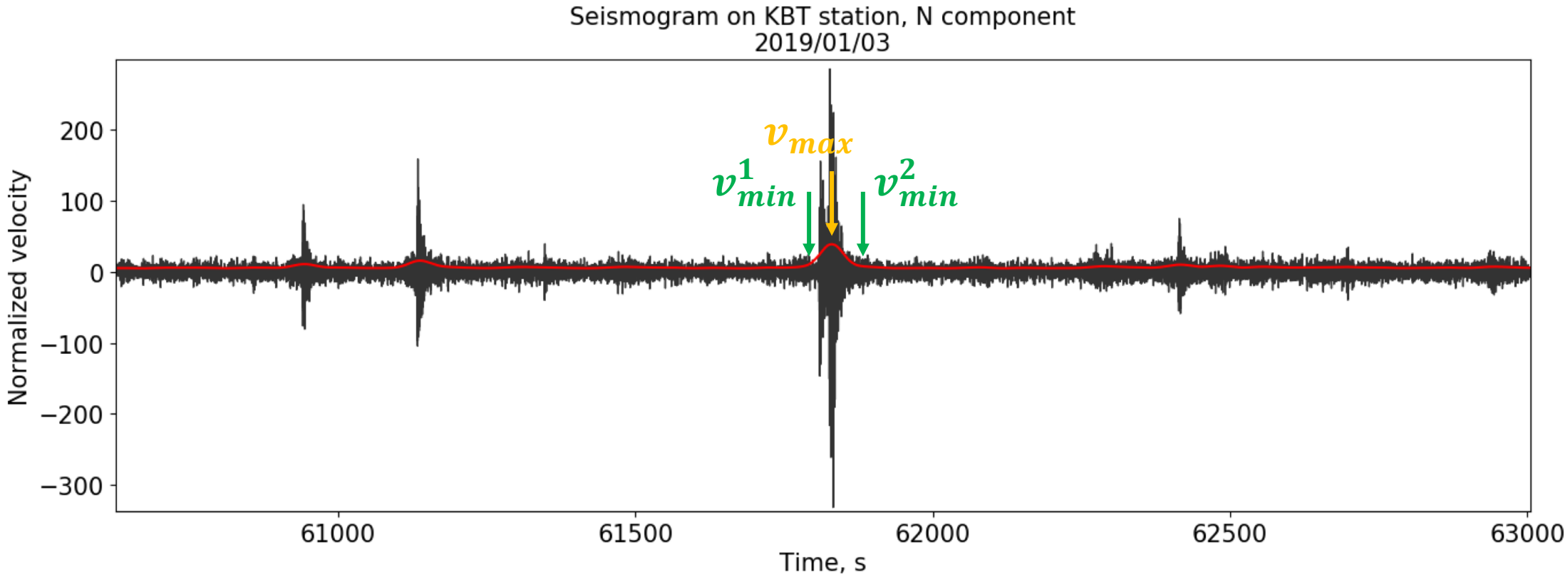
- Bandpass filtering (0.5 – 10 Hz) and decimating (Fs from 200 to 20 Hz)
- Seismogram smoothing with time window of 30 s



Earthquake detection

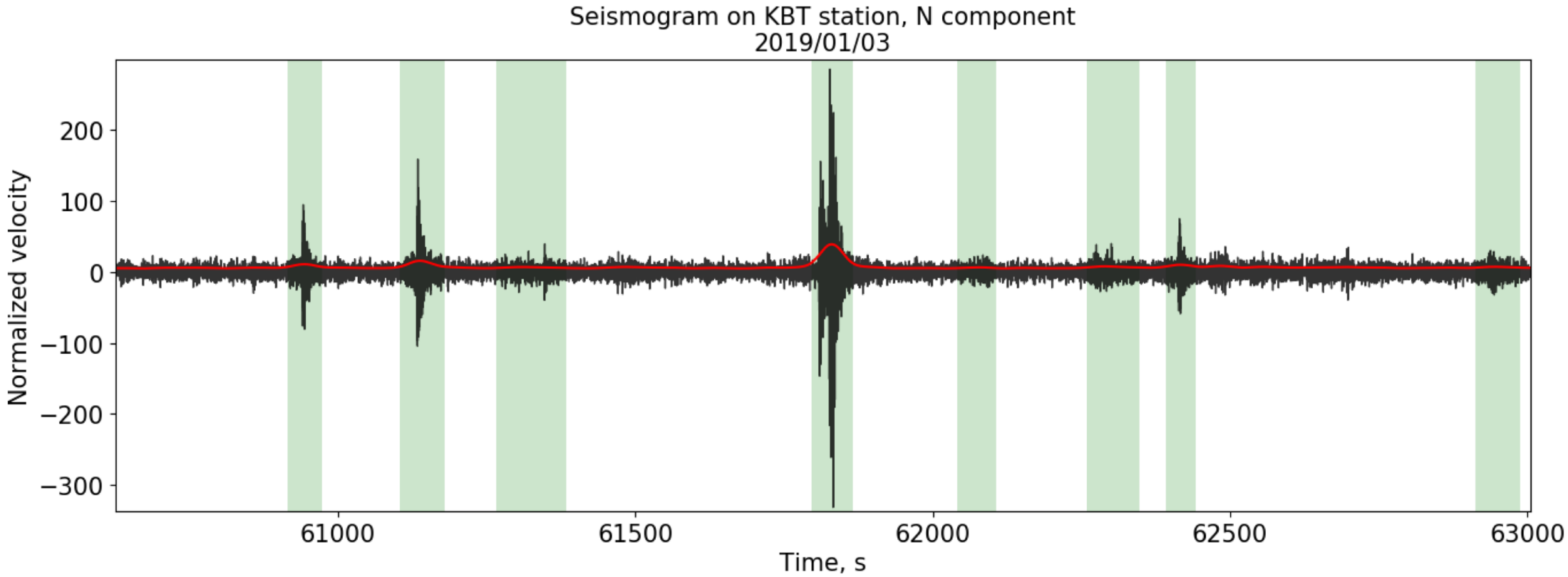
Signal-to-noise ratio: $SNR = v_{max}/v_{min}$

where $v_{min} = \max(v_{min}^1, v_{min}^2)$



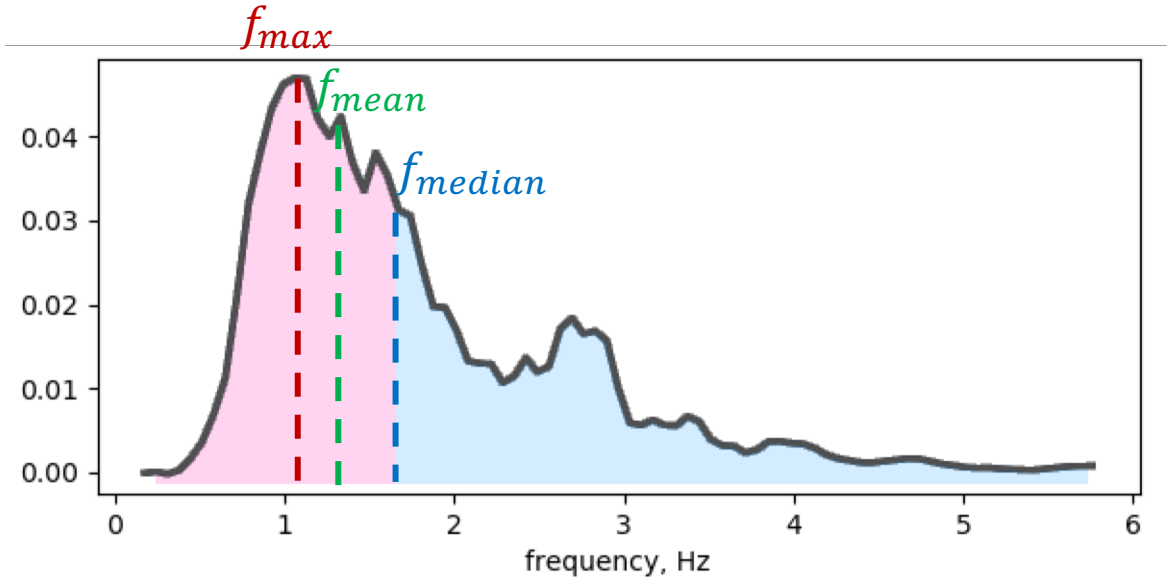
Earthquake detection

All detections with $\boldsymbol{SNR \geq 1.35}$ have been selected



Feature extraction

For each detection next parameters were estimated

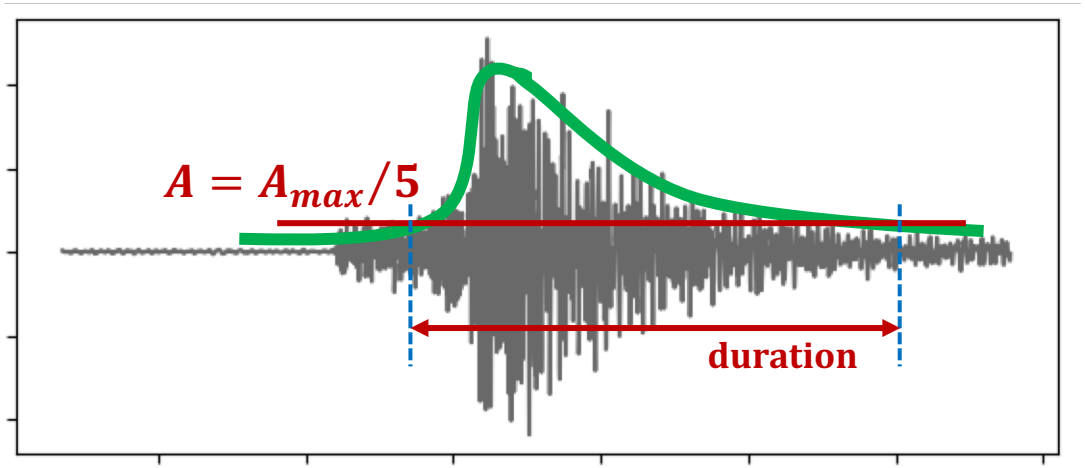


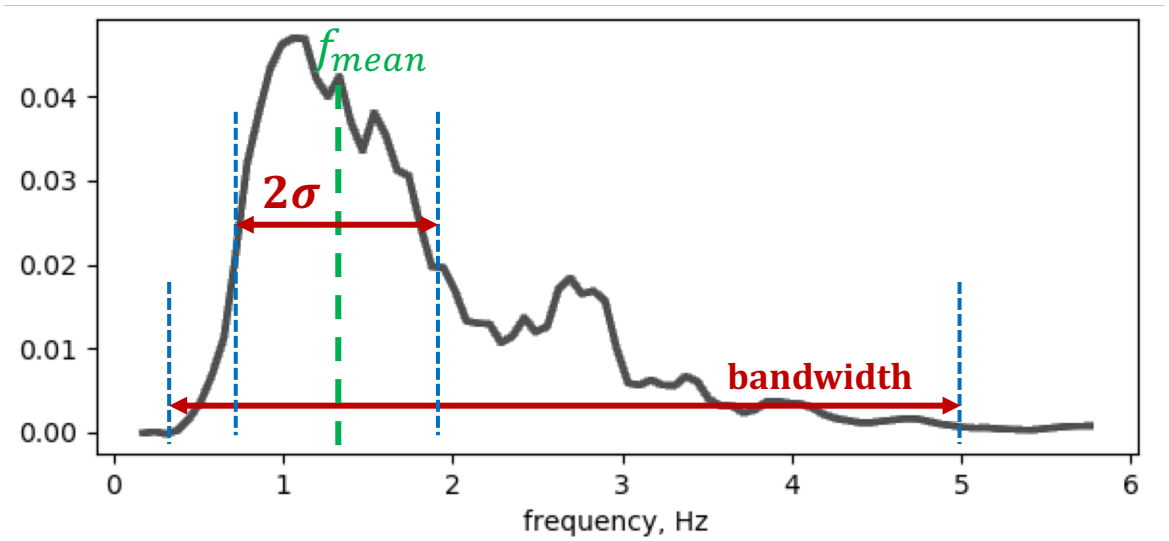
- Peak frequency f_{max}
- Mean frequency is defined as

$$f_{mean} = \frac{\sum_i P(f_i) f_i}{\sum_i P(f_i)}$$

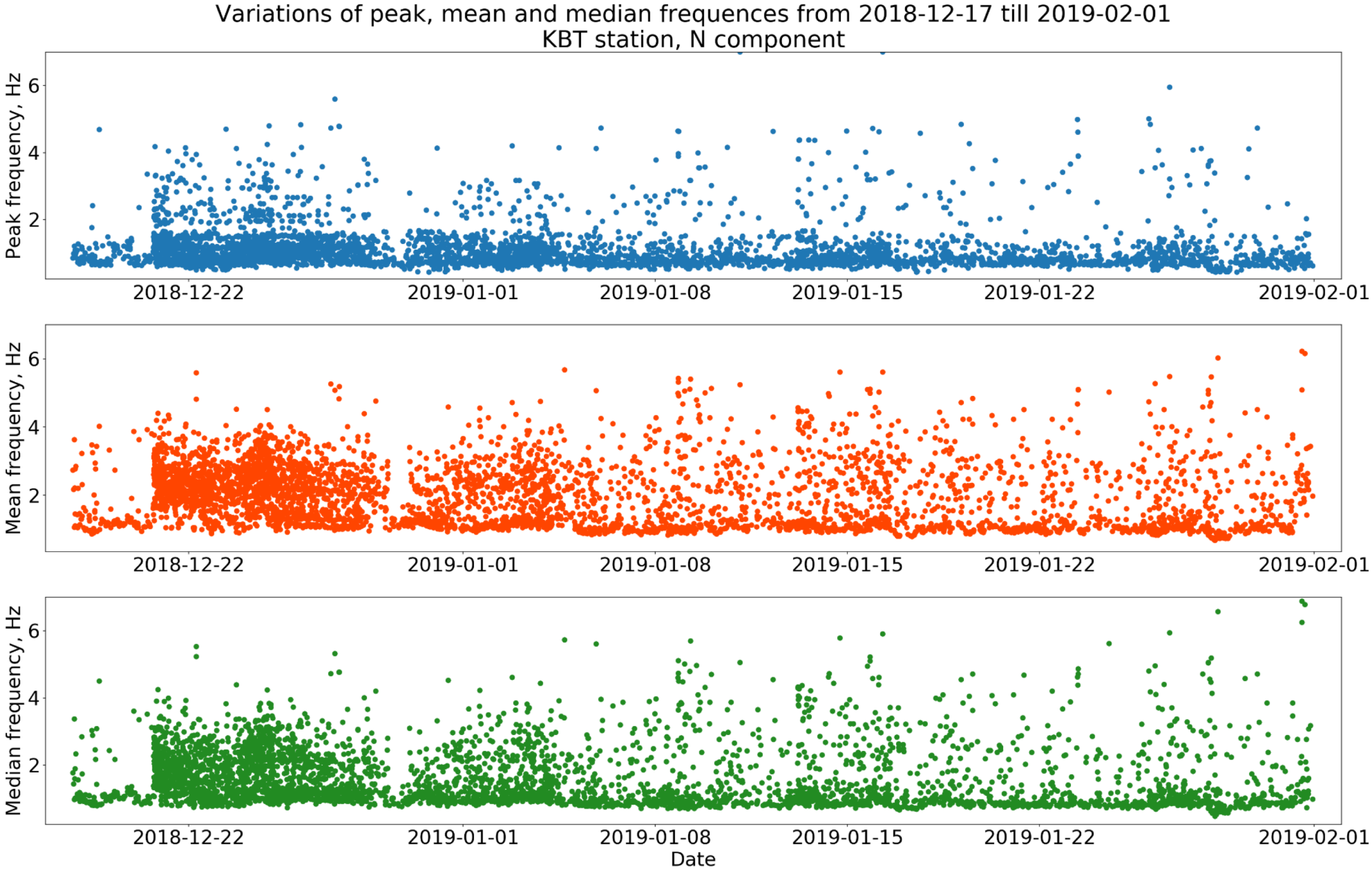
- Median frequency can be found from the condition

$$\sum_{j=1}^{j_{MDF}} P(f_j) = \sum_{j=j_{MDF}}^{M} P(f_j) = \frac{1}{2} \sum_{j=1}^{M} P(f_j)$$





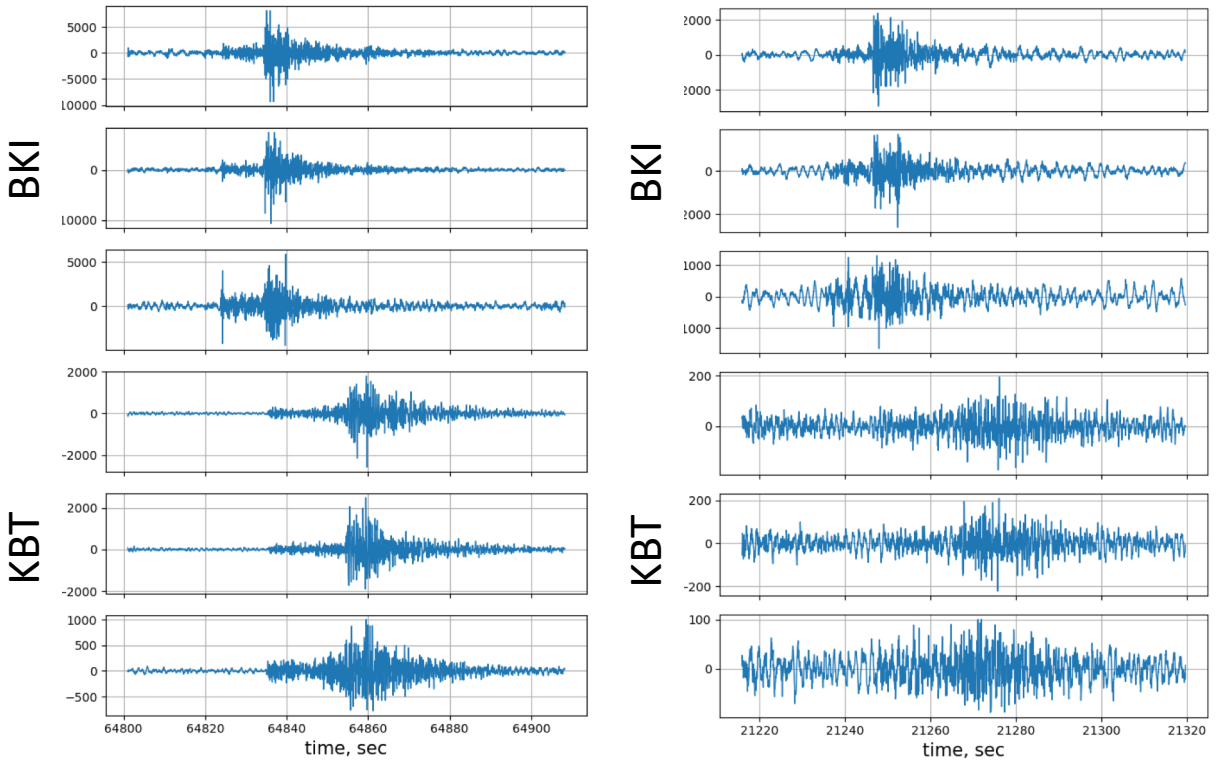
- Bandwidth is defined as difference between first and last frequencies where $P(f) \geq 0.01 P_{max}$
- The standard deviation from the mean frequency
- Signal amplitude
- Signal duration
- SNR of the detection


Variations of peak, mean and median frequences from 2018-12-17 till 2019-02-01
KBT station, N component
Peak frequency, Hz
Mean frequency, Hz
Median frequency, Hz
2
4
6
2018-12-22
2019-01-01
2019-01-08
2019-01-15
2019-01-22
2019-02-01
Date

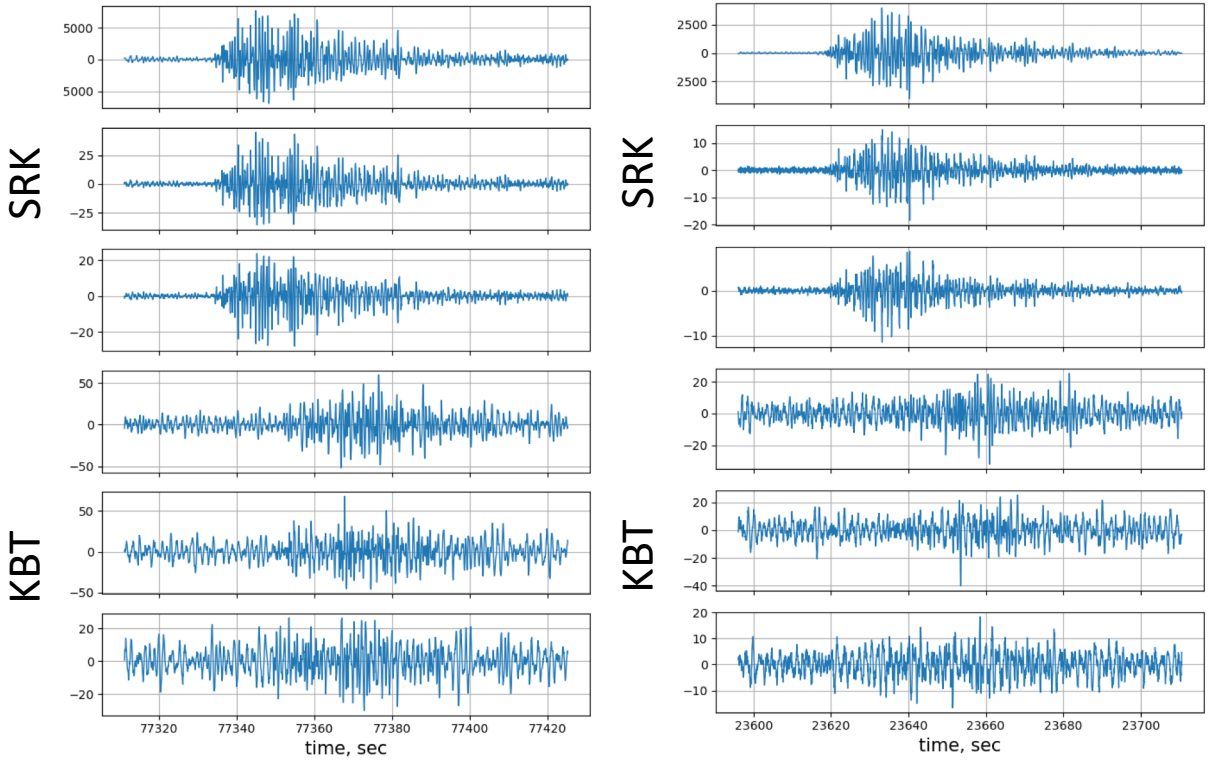
Creating labeled set

To create the labeled set, i.e. set of events which class is known (tectonic or volcanic), we took two reference stations to be more confident about detections on KBT station

BKI ⟶ “pure” tectonic events



SRK ⟶ “pure” volcanic events

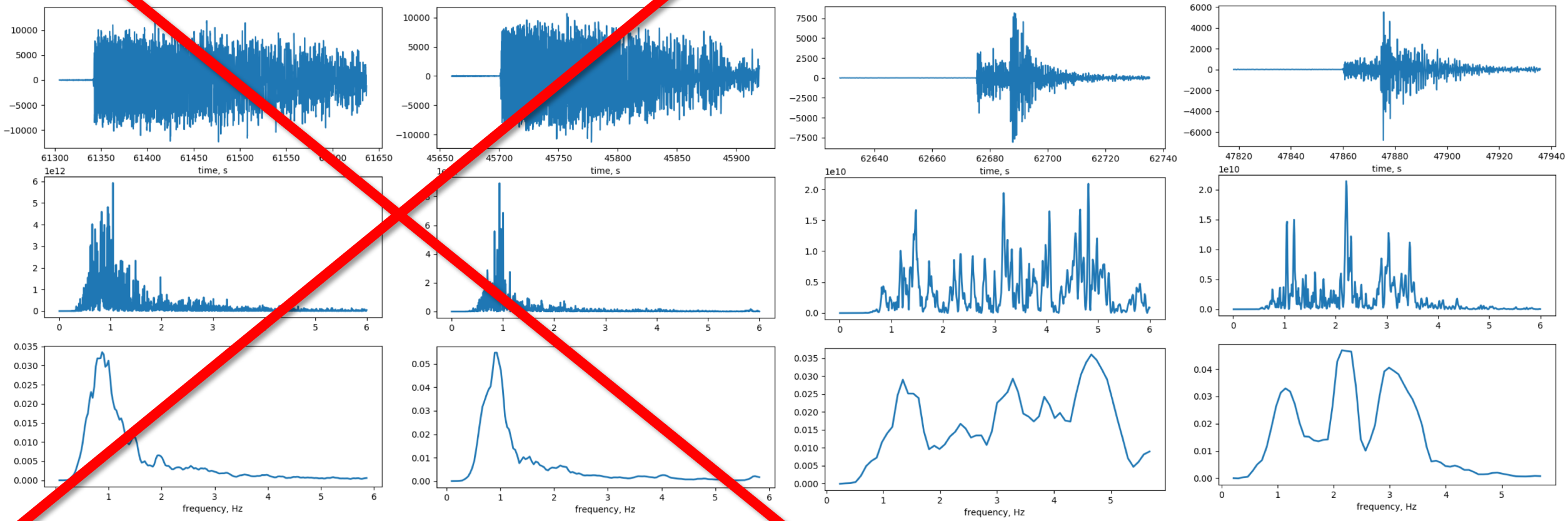


We had to remove strong tectonic events from the labeled set, because later it was found that they are accompanied with saturation effect on the stations that lead to frequency content similar to volcanic events

M=7.3

M=6.1

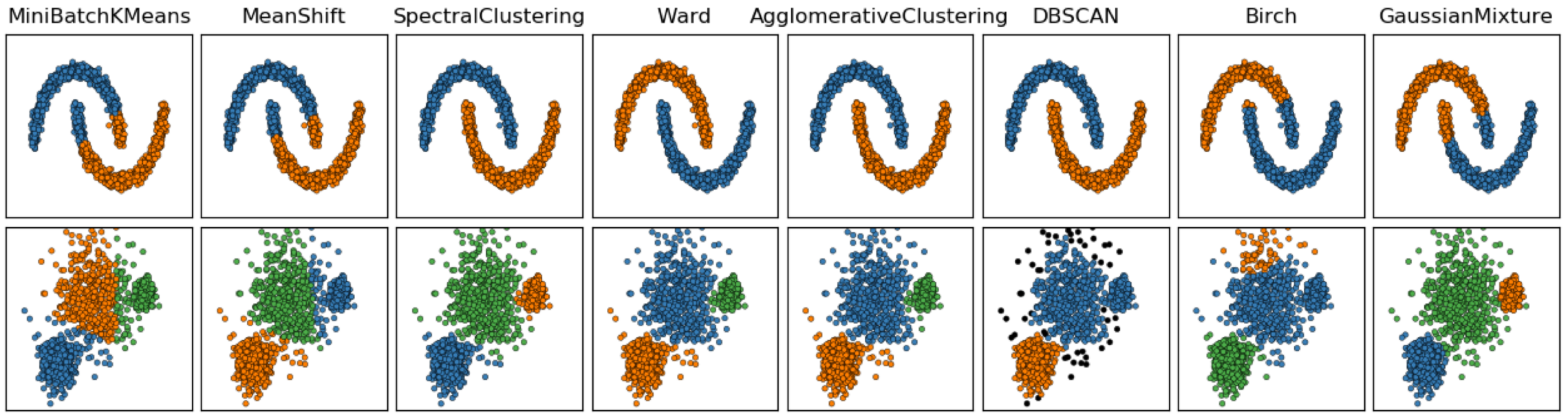
moderate events



Clustering

is automatic grouping of similar objects into sets
and is the class of unsupervised machine learning methods

One can see that it is an ambiguous problem, and the result varies with the chosen method



from scikit-learn.org

Clustering

We chose a couple of quite simple methods

K-means

- set of N samples $X \longrightarrow K$ disjoint clusters C
- each cluster is described by the mean of μ_j of the samples in the cluster – **centroids**
- algorithm aims to choose centroids that minimize the inertia, or within-cluster sum-of-squares criterion

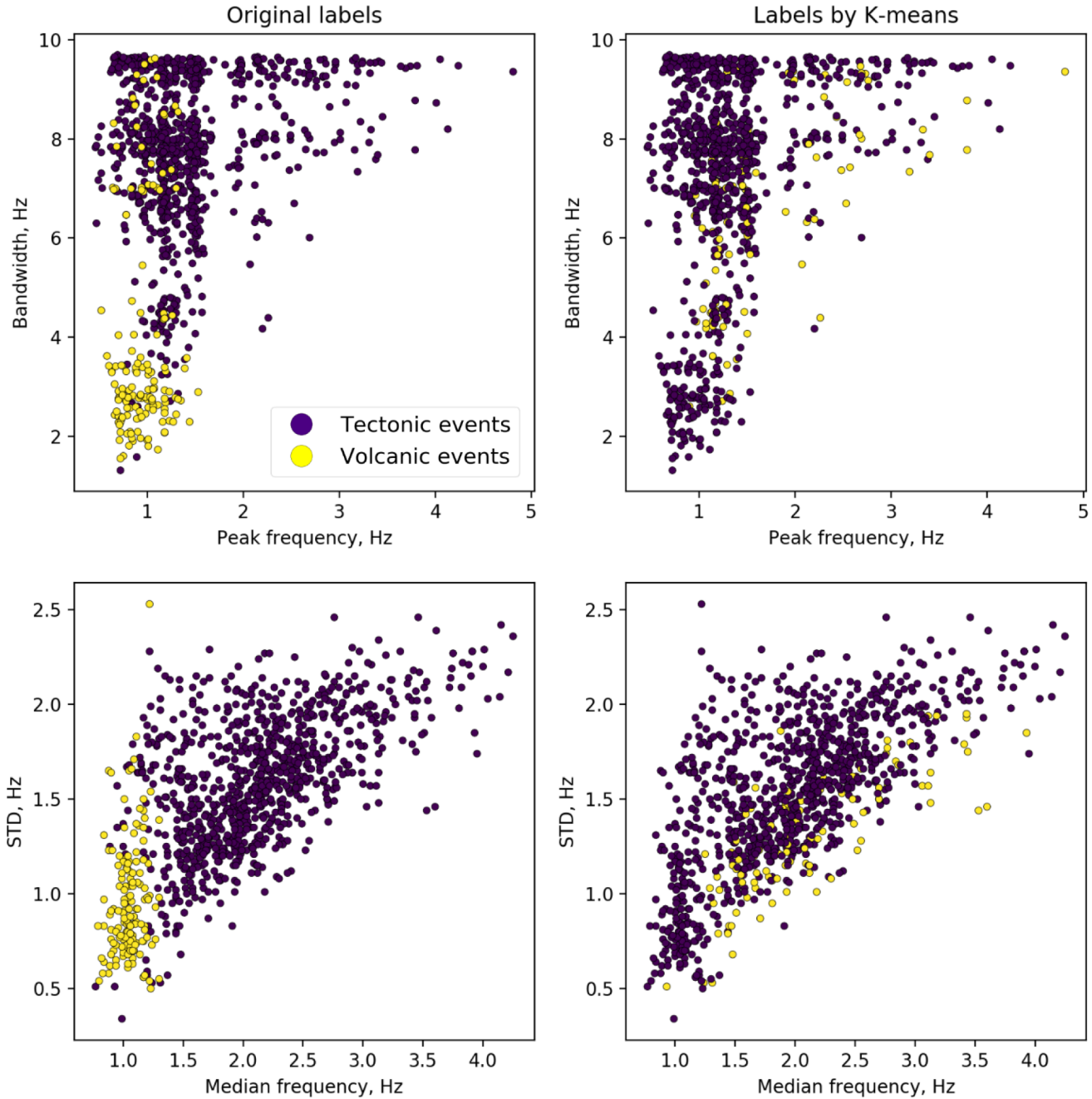
$$\sum_{i=1}^{N} \min_{\mu_j \in C} \left(\|x_i - \mu_j\|^2 \right)$$

Agglomerative clustering

- is a hierarchical clustering using a bottom up approach: each observation starts in its own cluster, and clusters are successively merged together
- linkage criteria determines the metric used for the merge strategy

K-means clustering using features

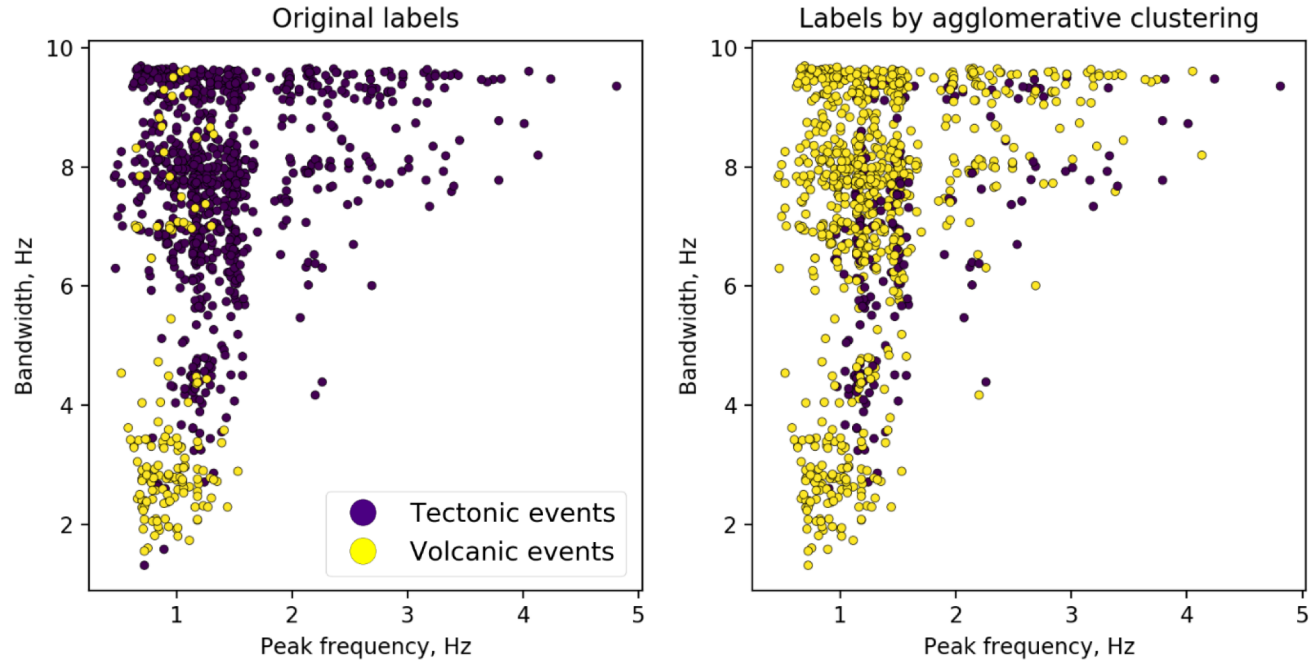
Performance of clustering algorithms on labeled set is poor and the result is inadequate

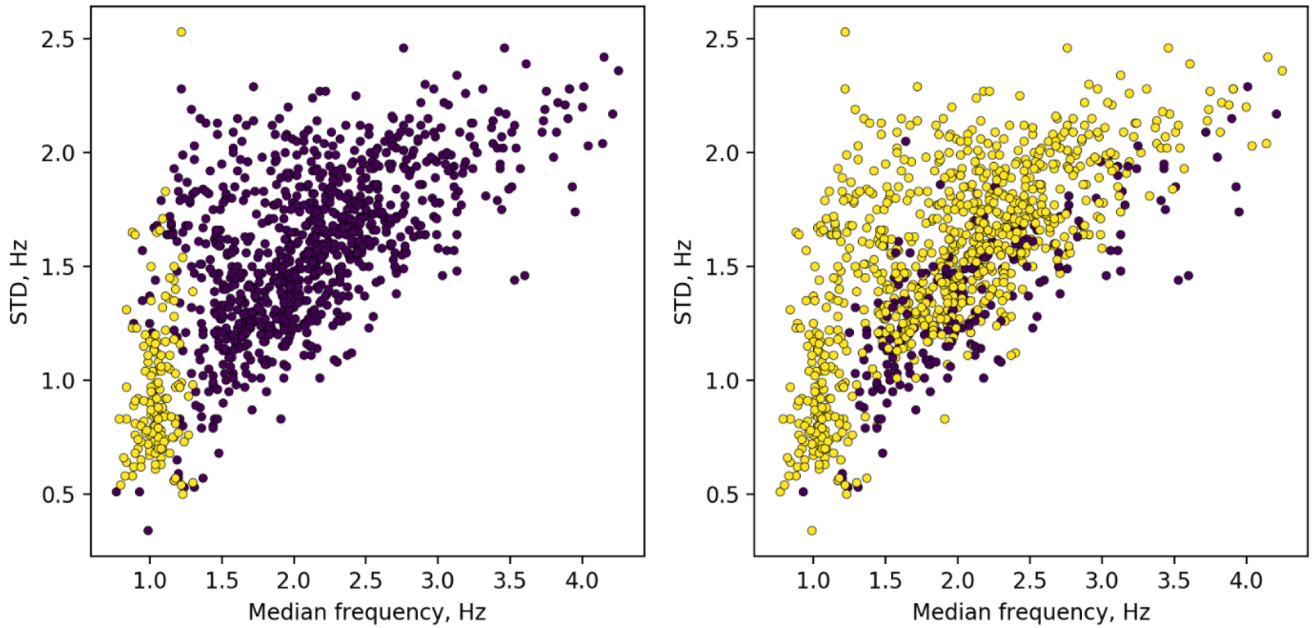


Agglomerative clustering using features

Performance of clustering algorithms on labeled set is poor and the result is inadequate

So we decided to use spectral representations of signals instead of features

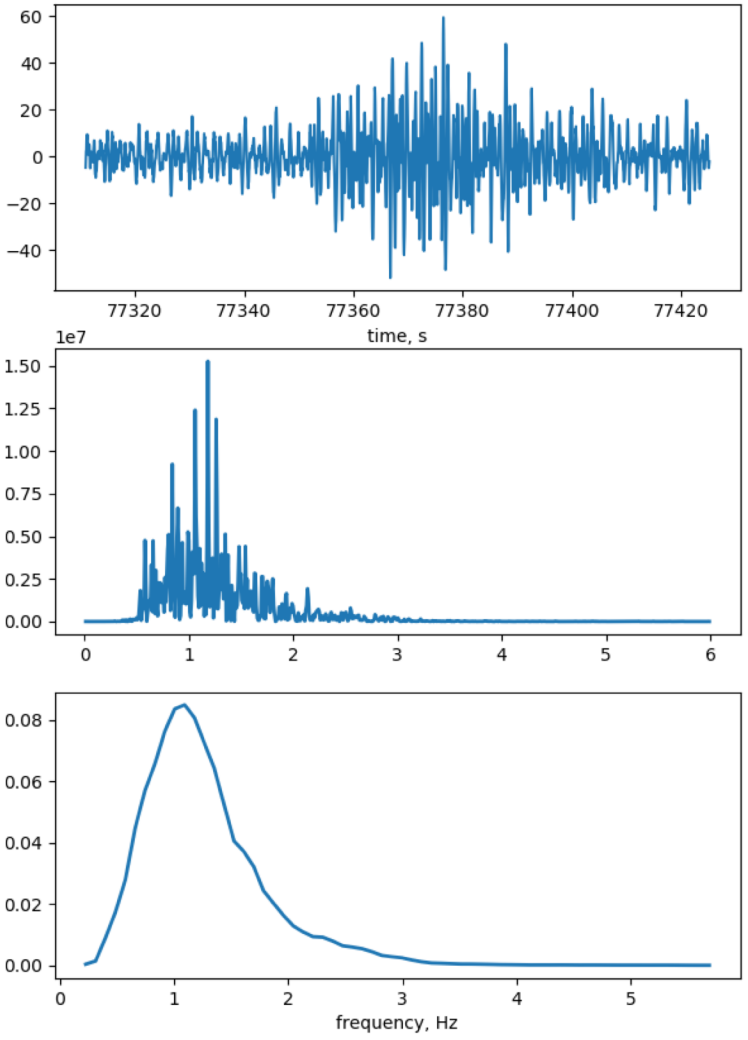




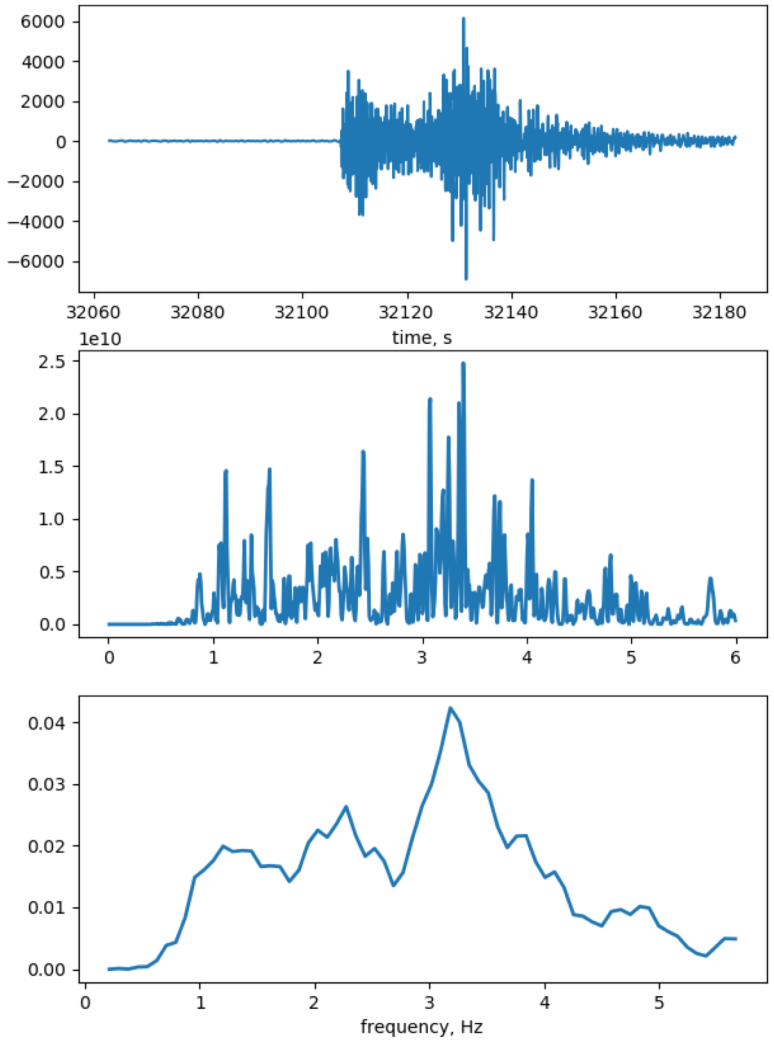
Clustering using smoothed spectra

Sliding window of 50 points, high-frequencies are cut off
8 features ⟶ set of 85 frequencies

Volcanic event



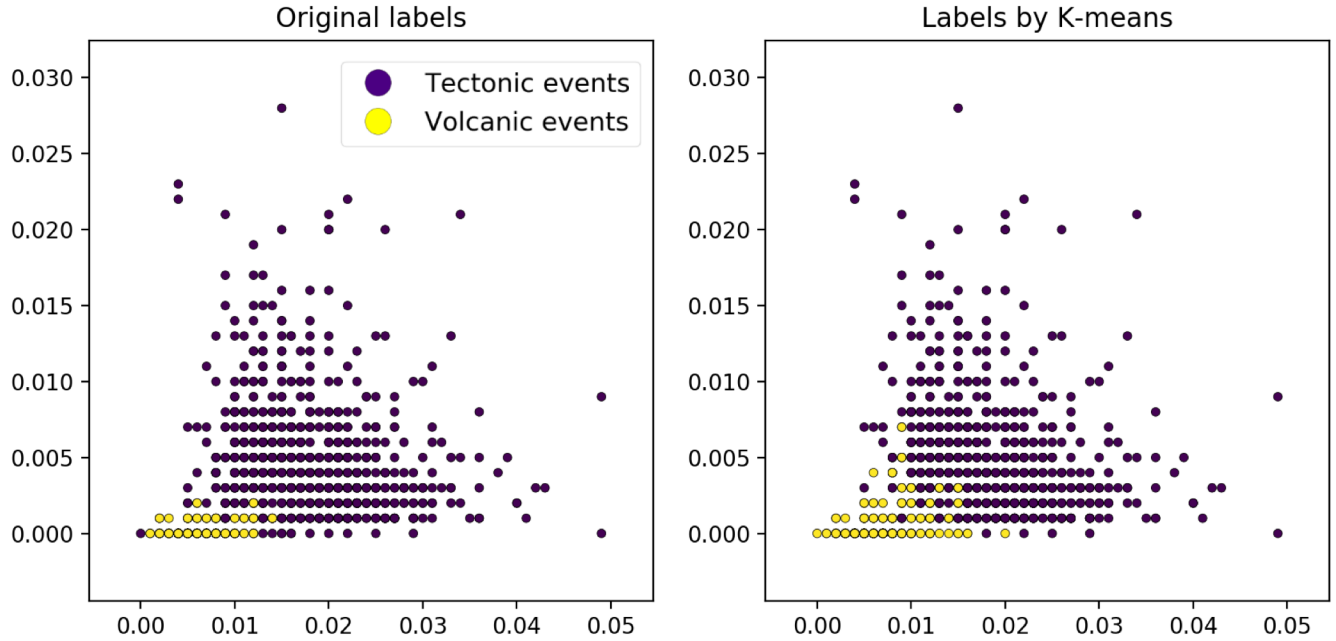
Tectonic event

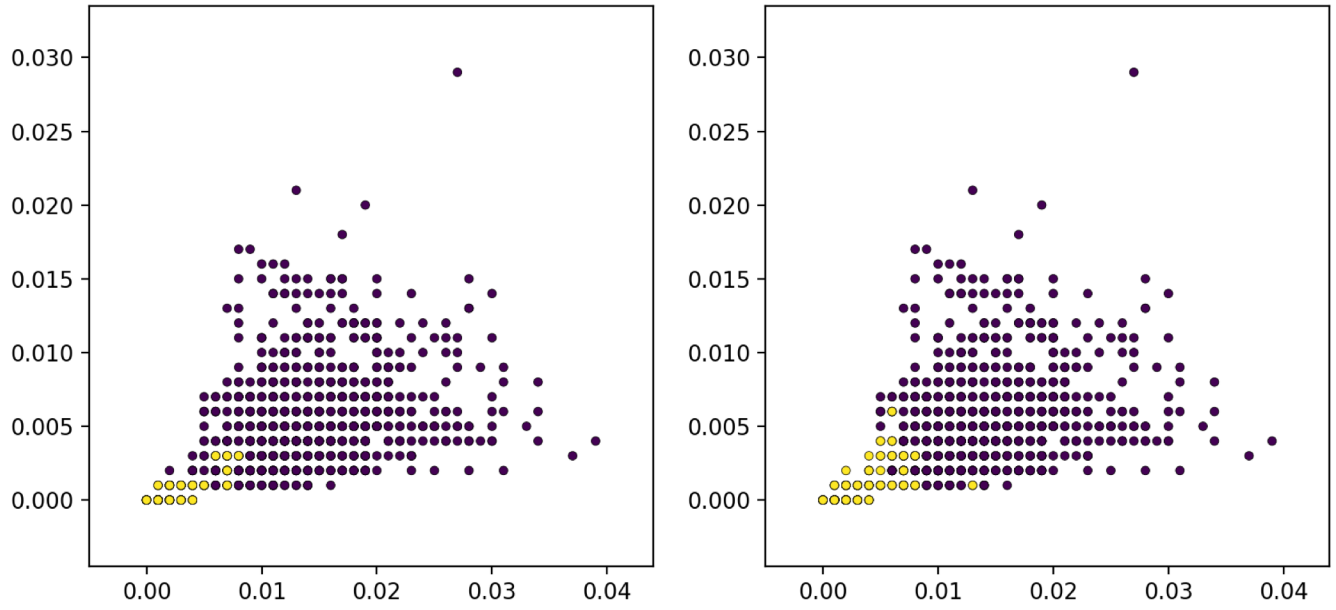


K-means clustering using smoothed spectra

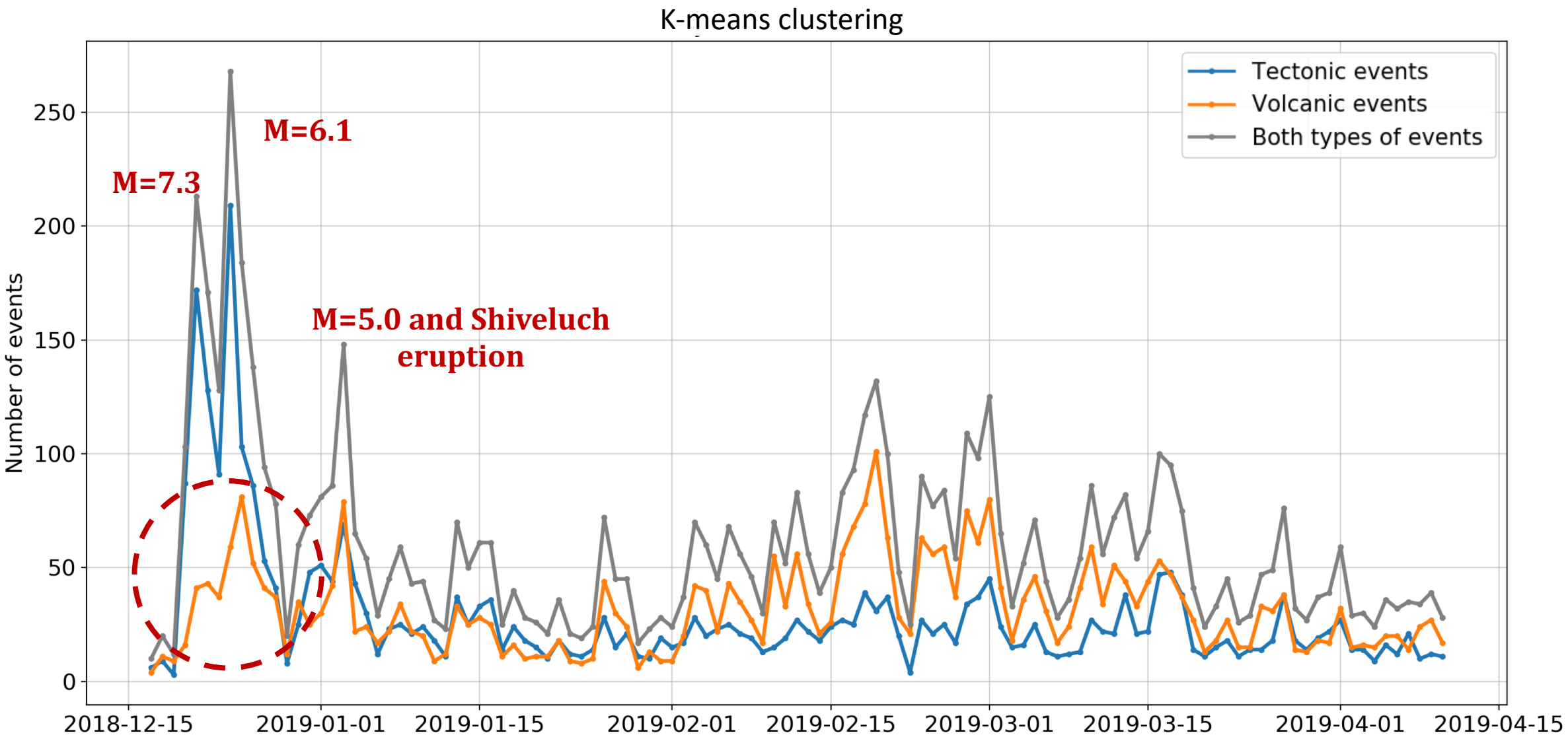
It can be seen that even one of the simplest clustering algorithms have not so bad performance on the labeled set

What if we present the entire data set to the algorithm?





Clustering using smoothed spectra

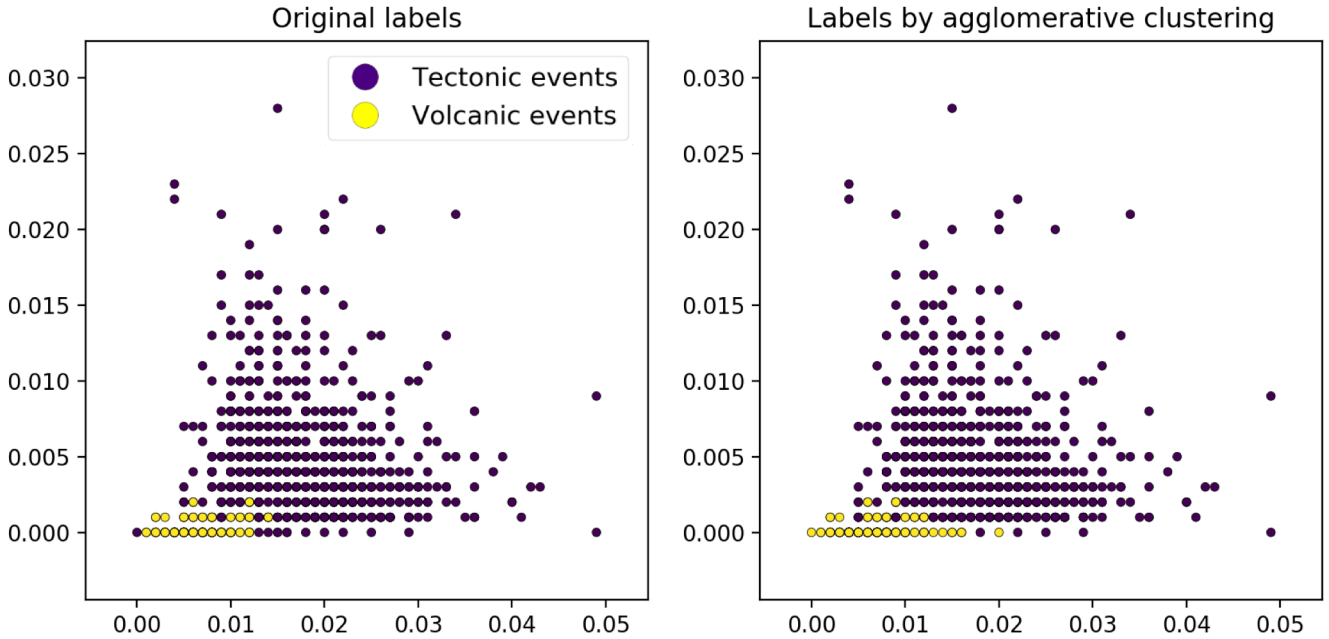


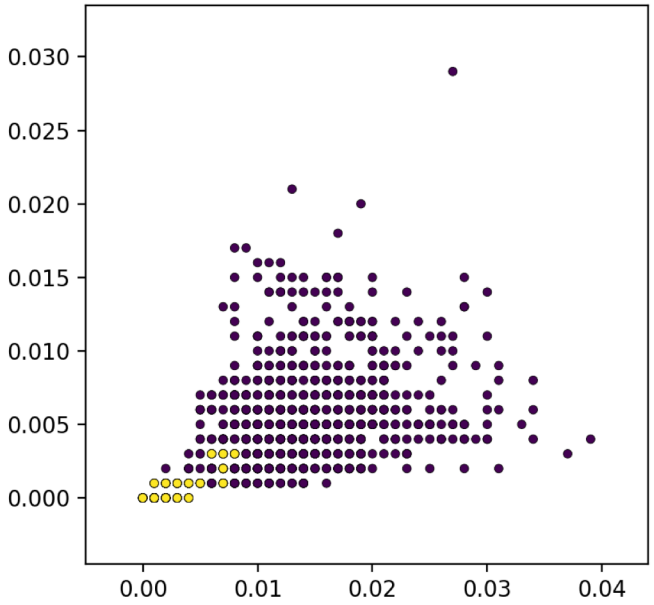
Shiveluch activation

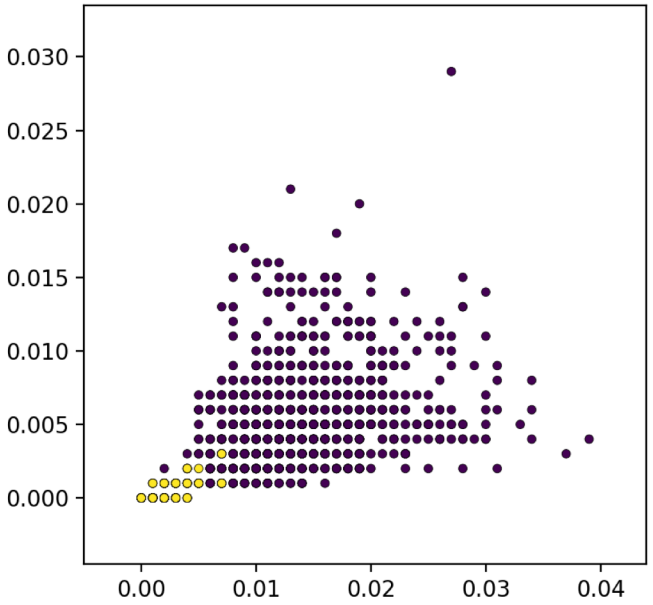
December	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Шивелуч	Ж	Ж	Ж	О	О	О	О	О	О	О	О	О	О	О	О	Ж	О	О	О	О	О	К	О	О	К	К	К	К	К	К	О

Data on the volcanic activity from emsd.ru

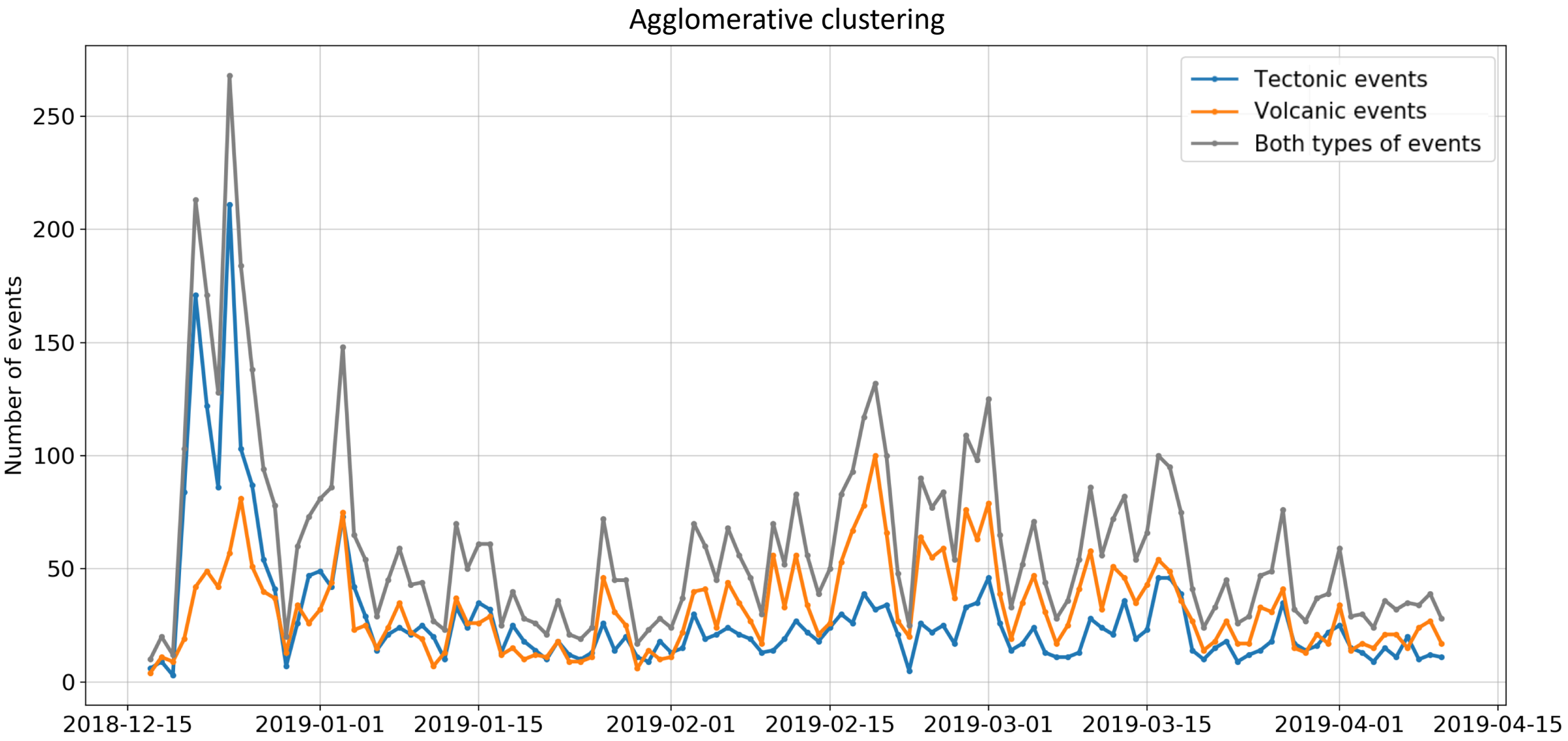
Agglomerative clustering using smoothed spectra







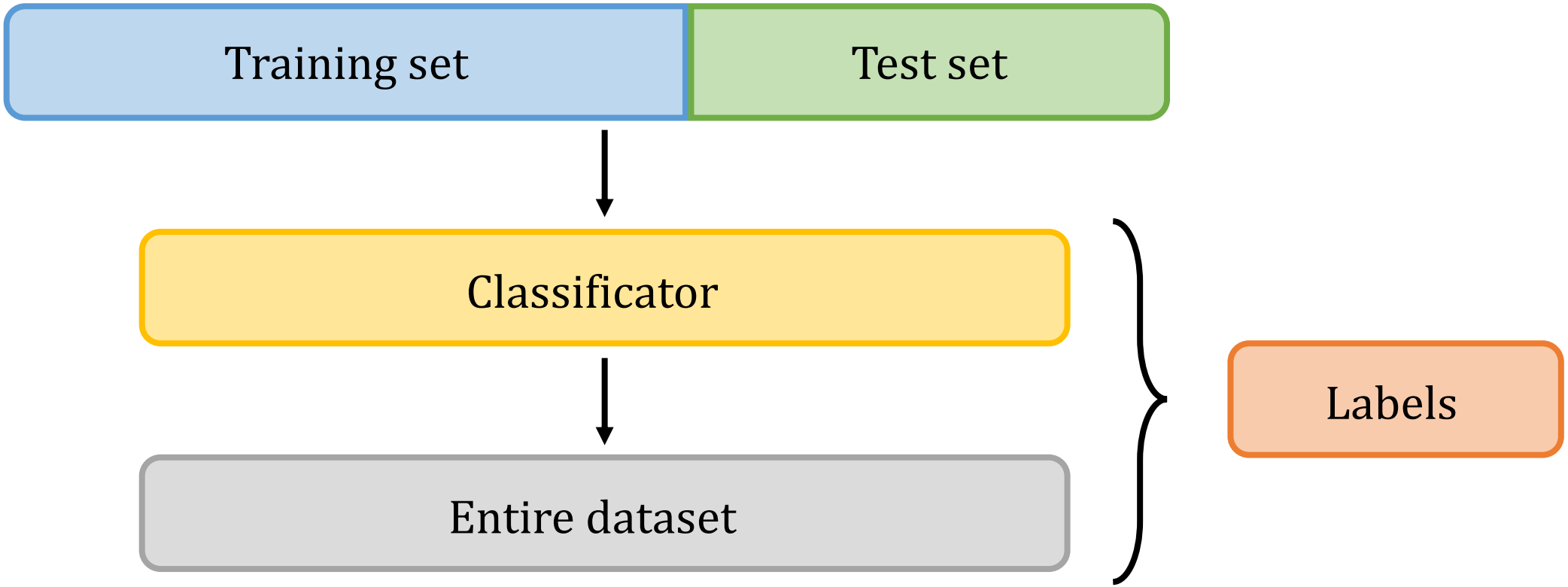
Clustering using smoothed spectra



Classification

is identifying to which category an object belongs to
and is the class of supervised machine learning methods

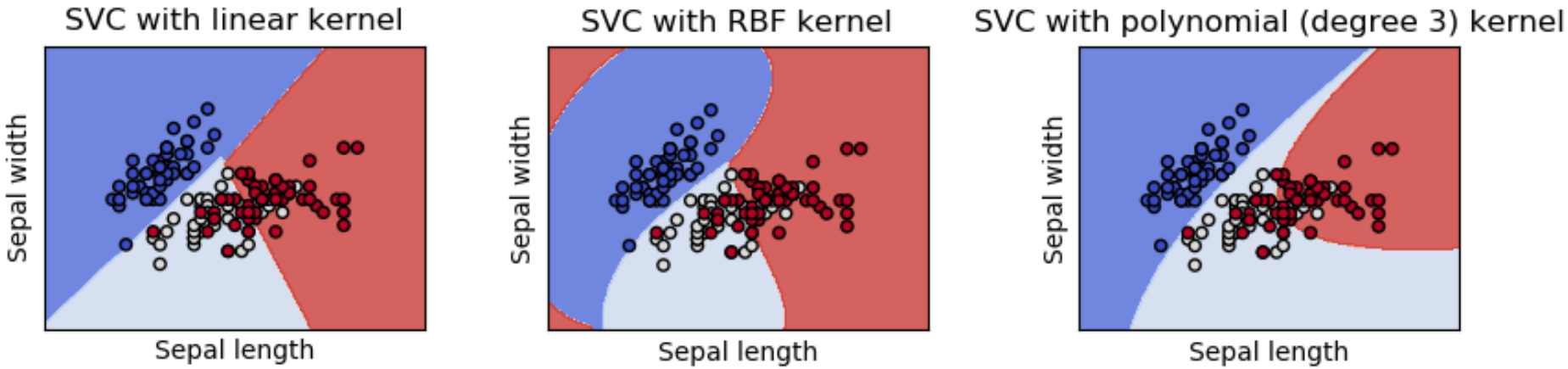
Labeled data: 902 tectonic and 273 volcanic events



Classification

Support vector machine (SVM)

- constructs a hyper-plane or set of hyper-planes in a high or infinite dimensional space
- good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called **margin**)
- different Kernel functions can be specified for the decision function



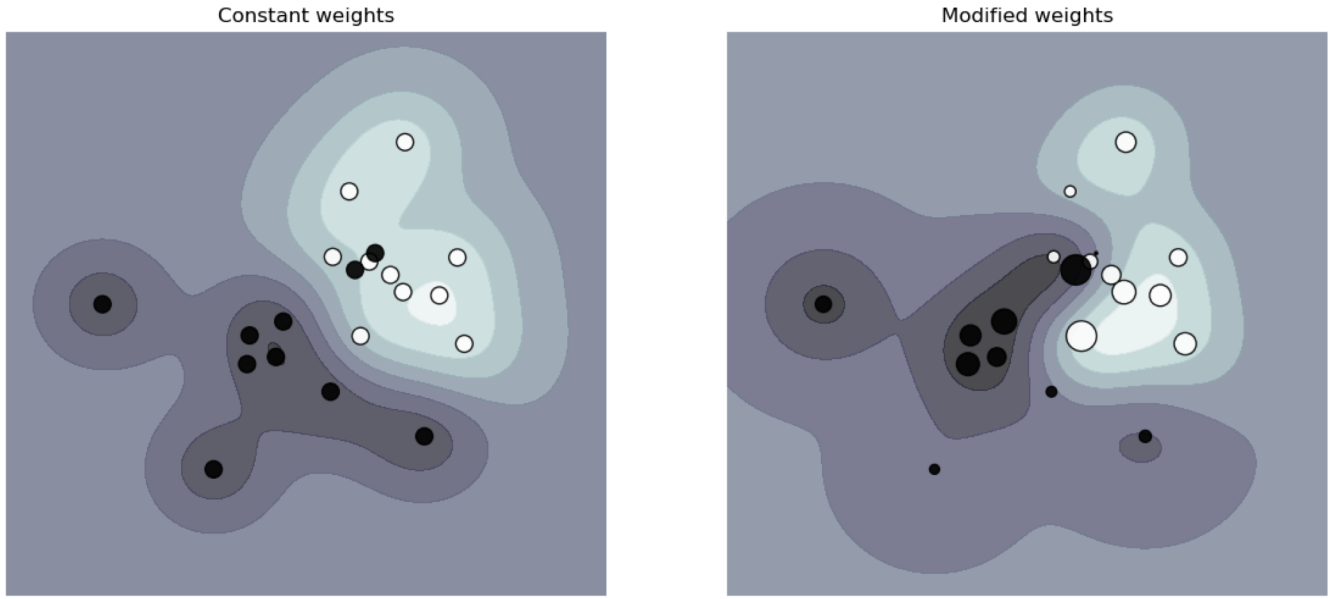
modified from scikit-learn.org

Classification: SVM

Accuracy of labeling on the test set with different kernels using different signal representations

	features			**spectra**		
	linear	polynomial	rbf	linear	polynomial	rbf
weighted*	0.974	0.974	0.982	0.971	0.977	0.971
unweighted	0.982	0.987	0.982	0.977	0.971	0.977

* The weighting makes classifier puts more emphasis on getting these points right. So we tried to put weight to volcanic class

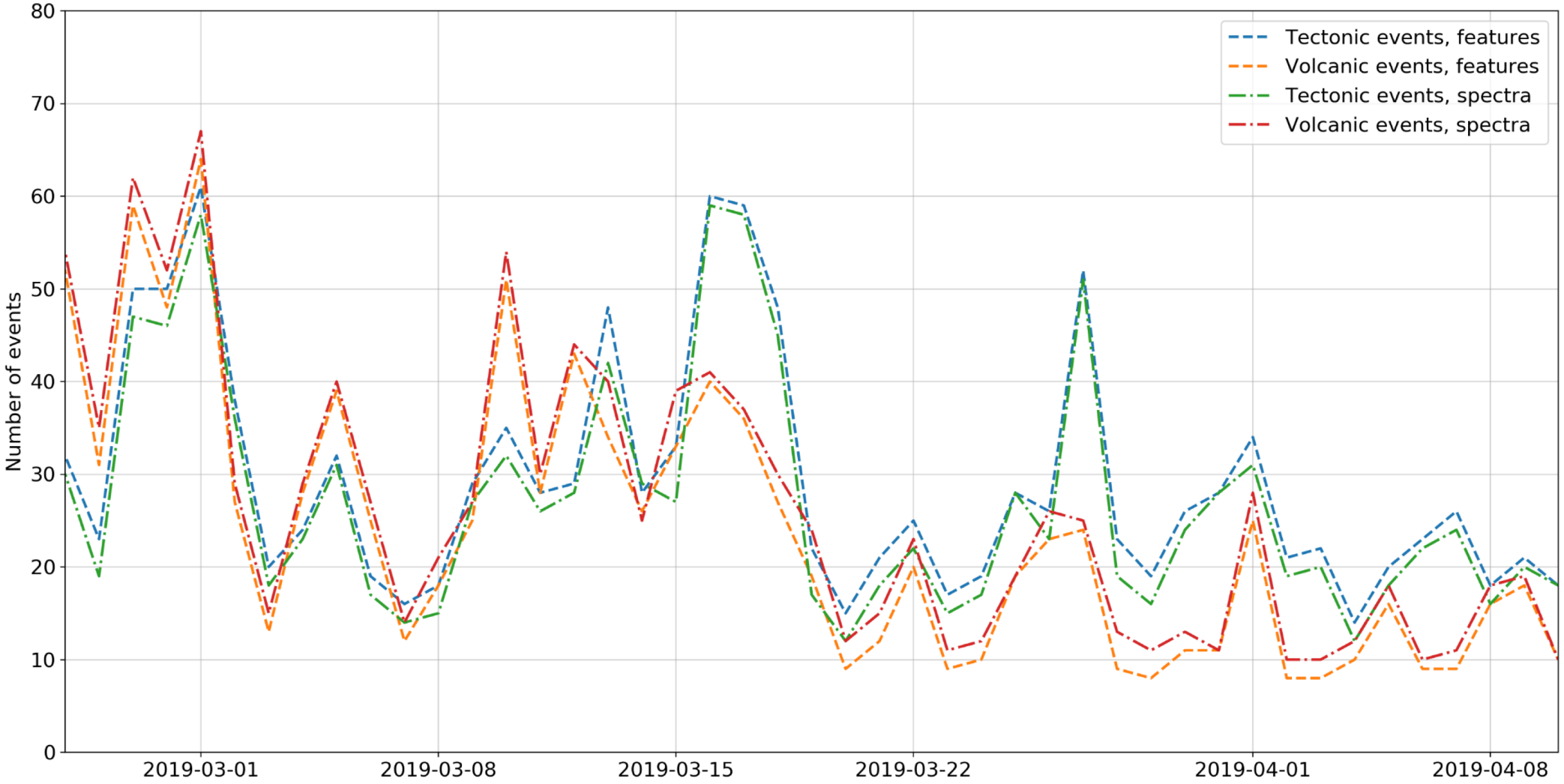


from scikit-learn.org

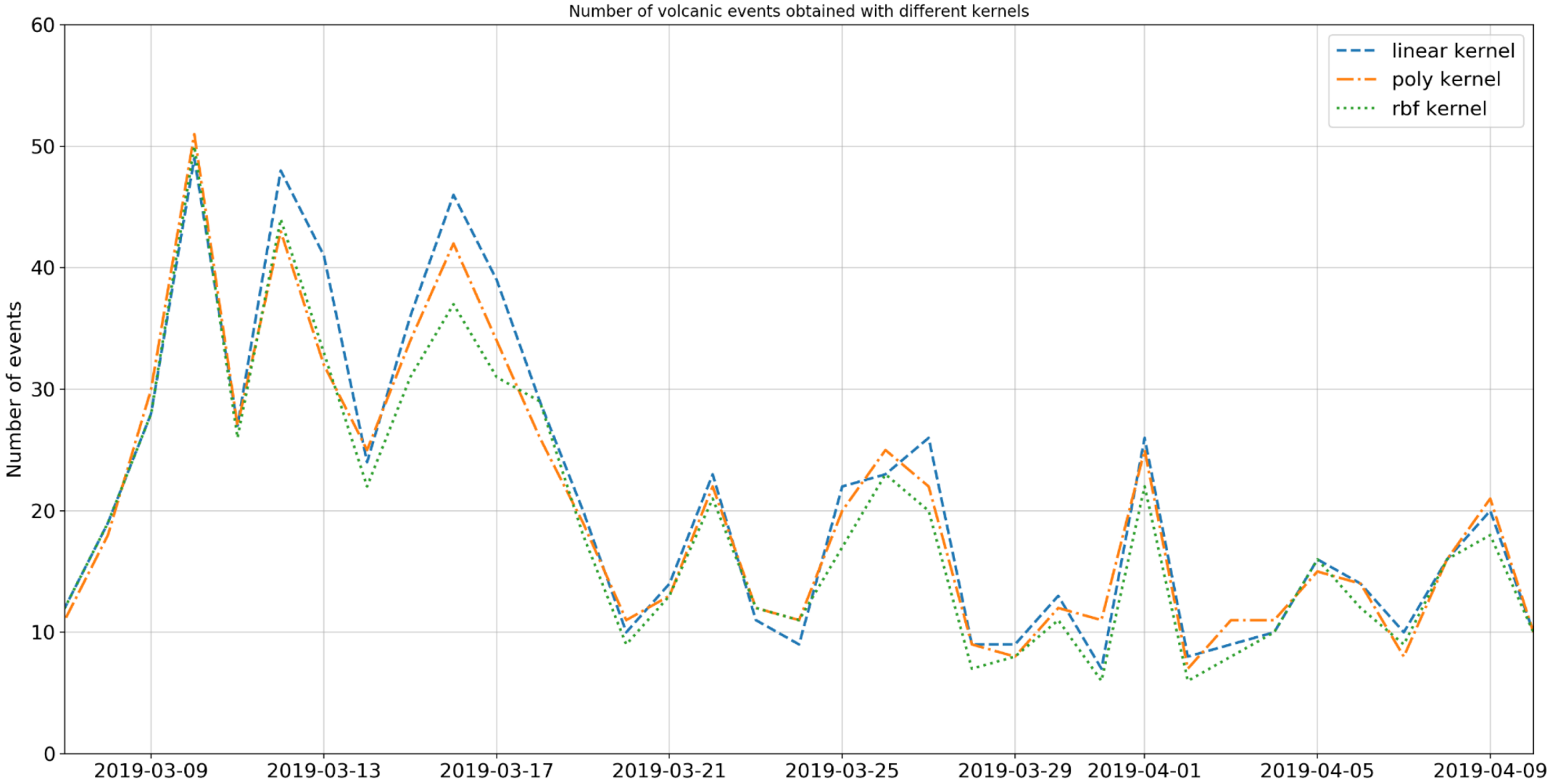
It can be seen that the performance of the SVM on the test set is stable to the parameters choice

Next slides will show the comparison of the results on the entire dataset

Classification: features vs. spectra



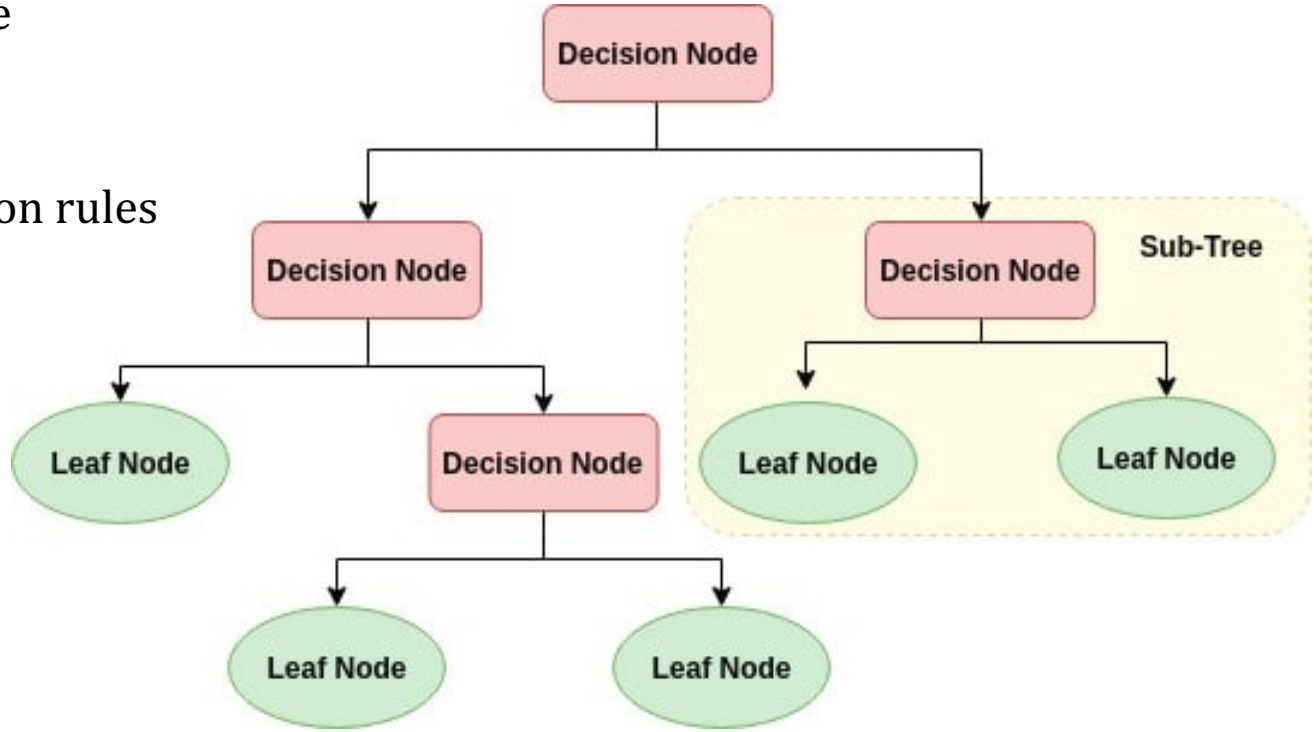
Classification: kernel choice



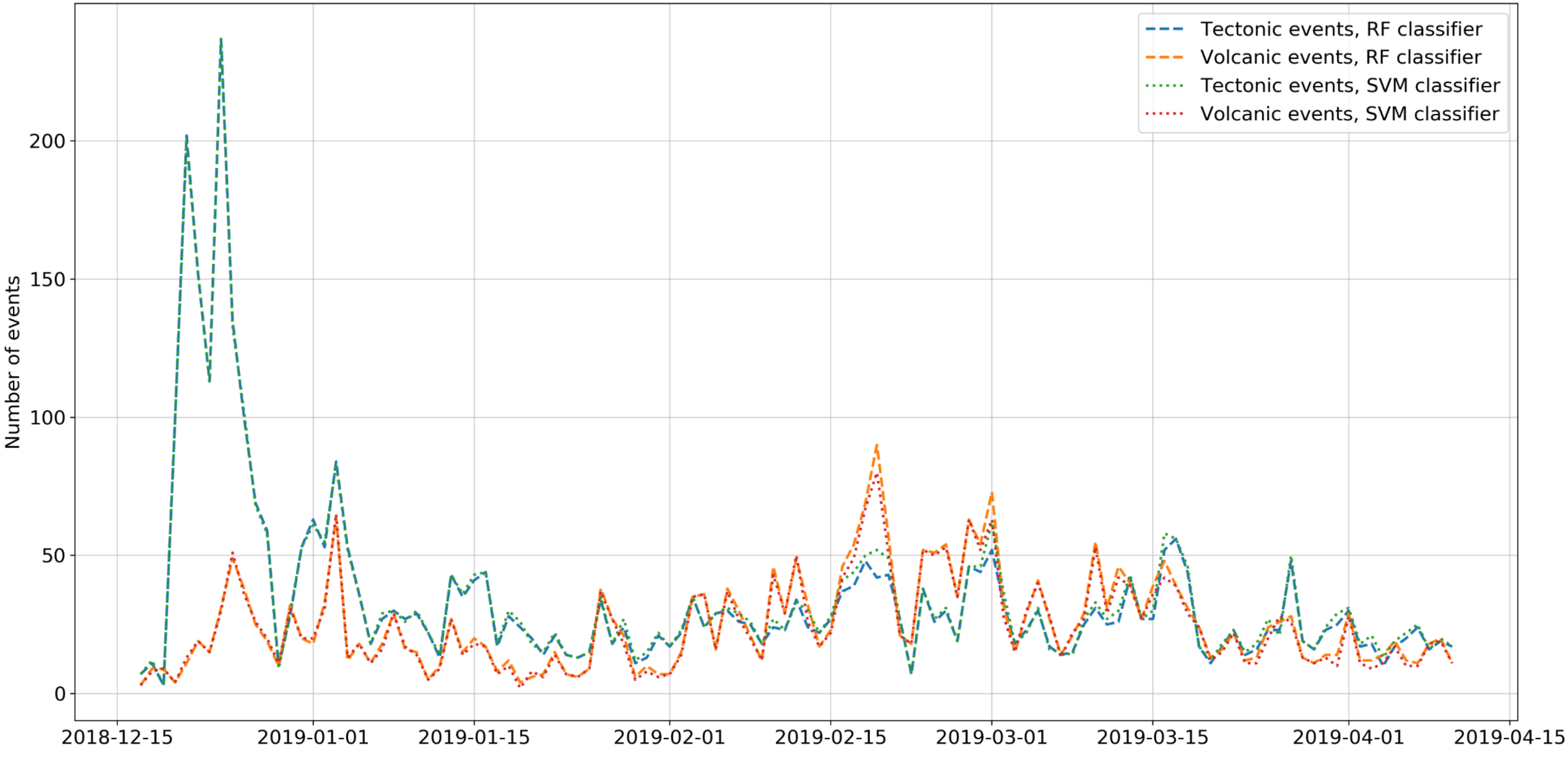
Classification

Random forest

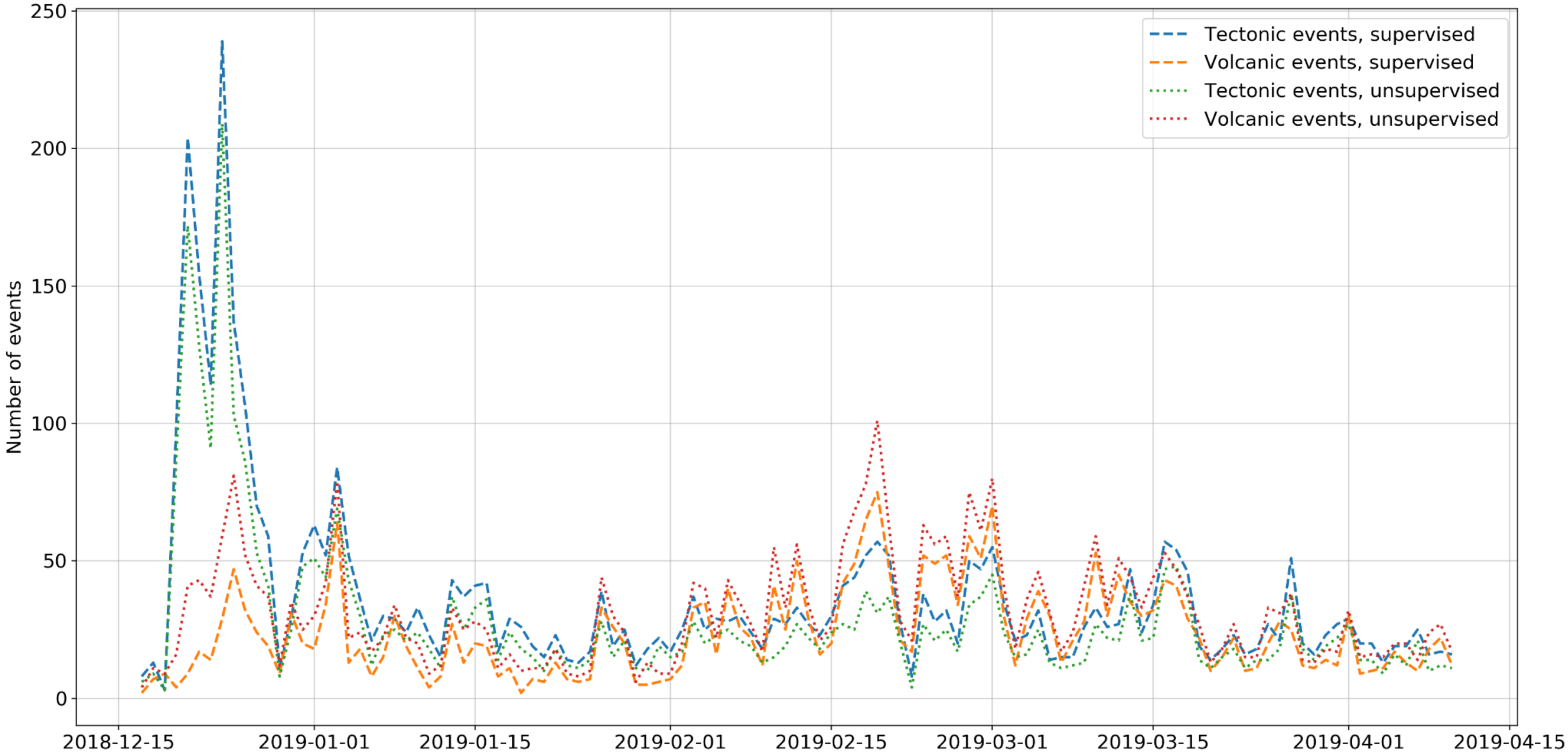
- is an ensemble learning method that constructs a multitude of decision trees at training time and outputs the class that is the mode of the classes of the individual trees
- a decision tree is a flowchart-like structure:
  - an internal node is a "test" on an attribute
  - a branch is the outcome of the test
  - a leaf node is a label
  - the paths from root to leaf are classification rules



Classification: RF vs. SVM



Comparison of unsupervised and supervised learning



Scattering coefficients

Well, we have seen that regular Fourier Transform (FT) worked well with seismic signals and algorithms performance was quite good.

But it is known that the FT is applicable when frequency content is stationary , and most of the signals in nature are non-stationary

So, FT has high resolution in frequency domain, but zero resolution in time domain

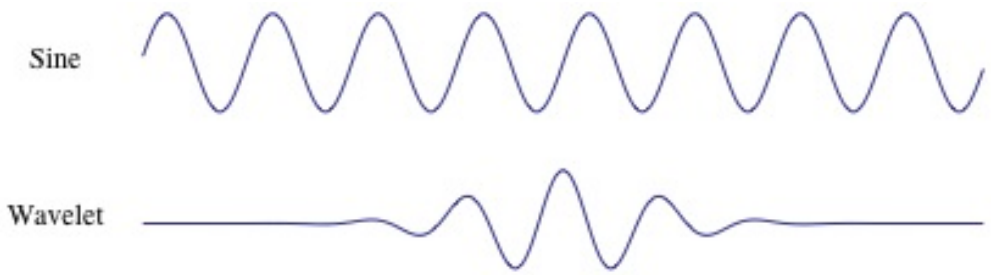
Scattering coefficients

A better approach for analyzing signals with a dynamical frequency spectrum is the Wavelet Transform (WT).

The WT has a high resolution in both the frequency- and the time-domain.

How does the Wavelet Transform work?

The FT uses a series of sine-waves with different frequencies to analyze a signal. In fact, a signal is represented through a linear combination of sine-waves.



The WT uses a series of functions called wavelets, each with a different scale.

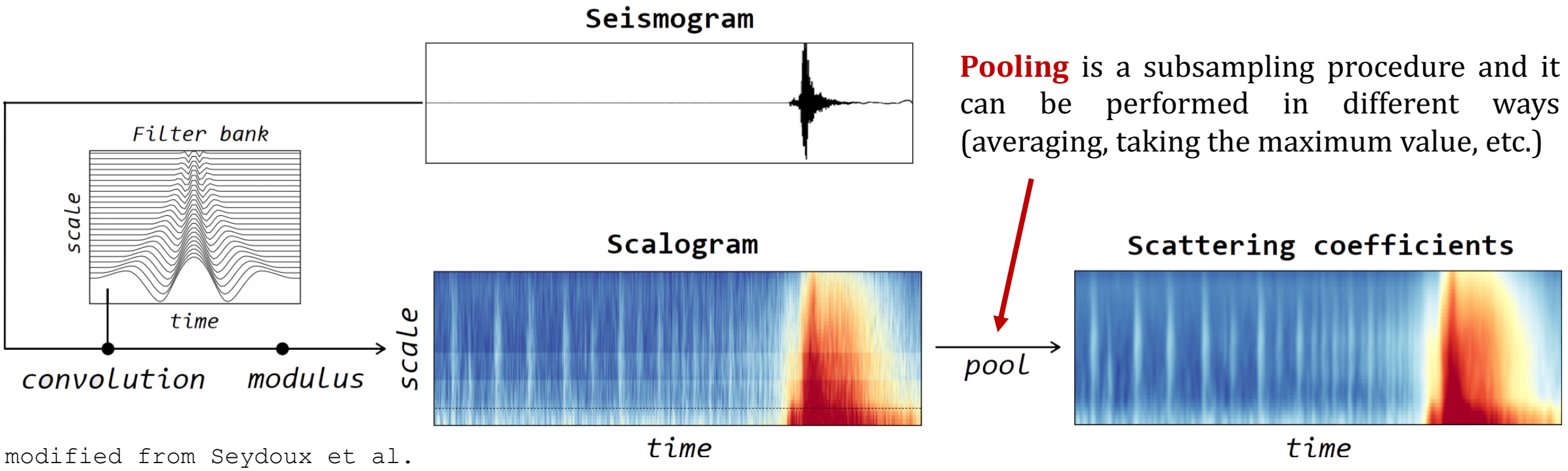
from ataspinar.com

Scattering coefficients

Original (mother) wavelet moves along the signal from its beginning to the end and at each point the **convolution** of the wavelet with the signal is calculated. After that the wavelet is scaled that it becomes larger and the procedure repeats.

So, the **scalogram** is the time-scale representation of a signal and the output of the WT

Scattering coefficients



modified from Seydoux et al.

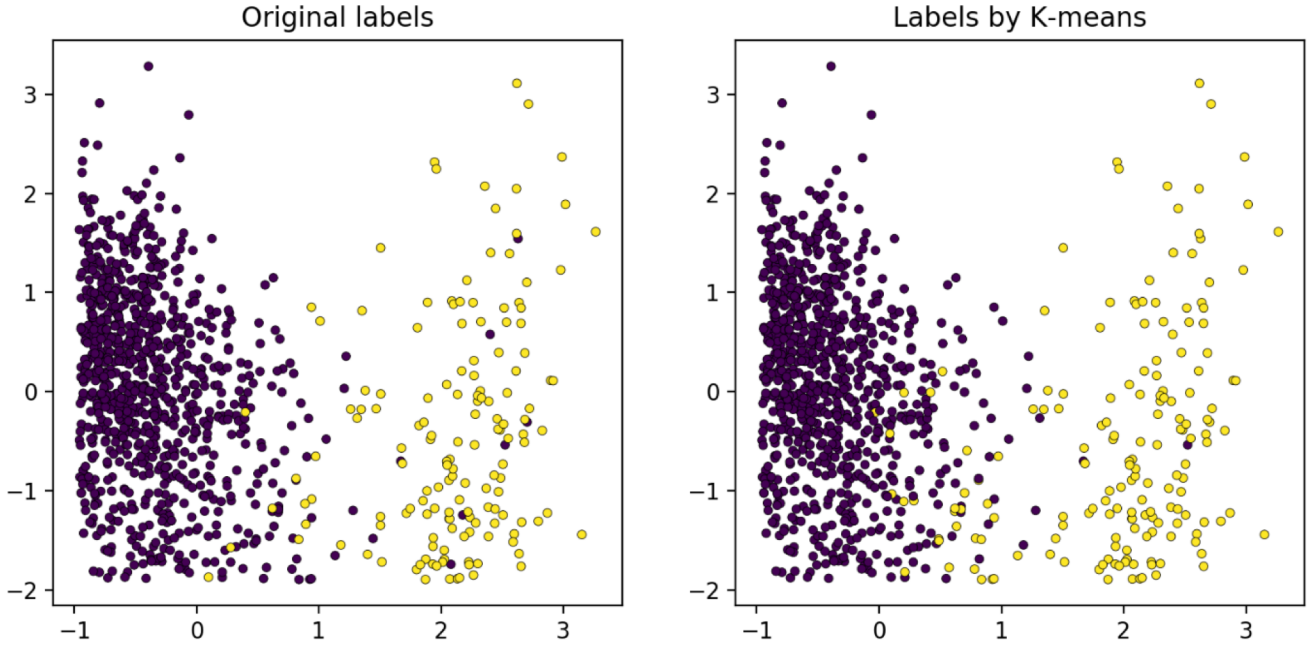
Pooling is a subsampling procedure and it can be performed in different ways (averaging, taking the maximum value, etc.)

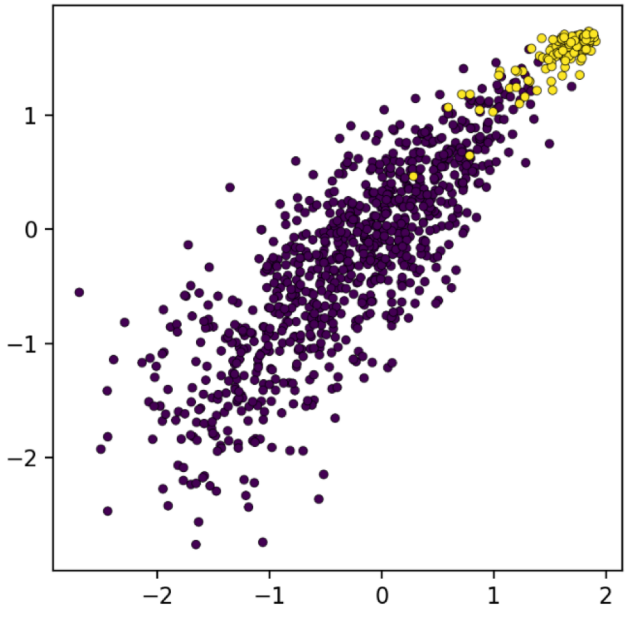
Now one event is described by 3458 coefficients
(vs. 8 features and 85 values of a smoothed spectrum)

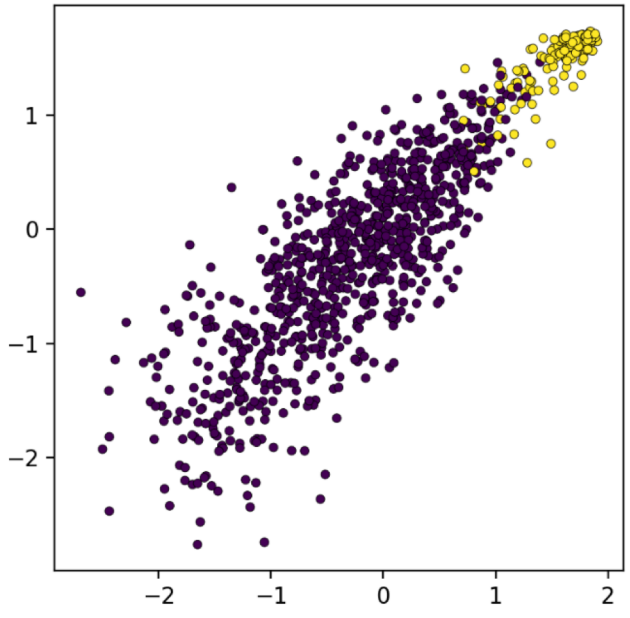
Clustering using scattering coefficients

Performance of clustering algorithms on the labeled set is very good:

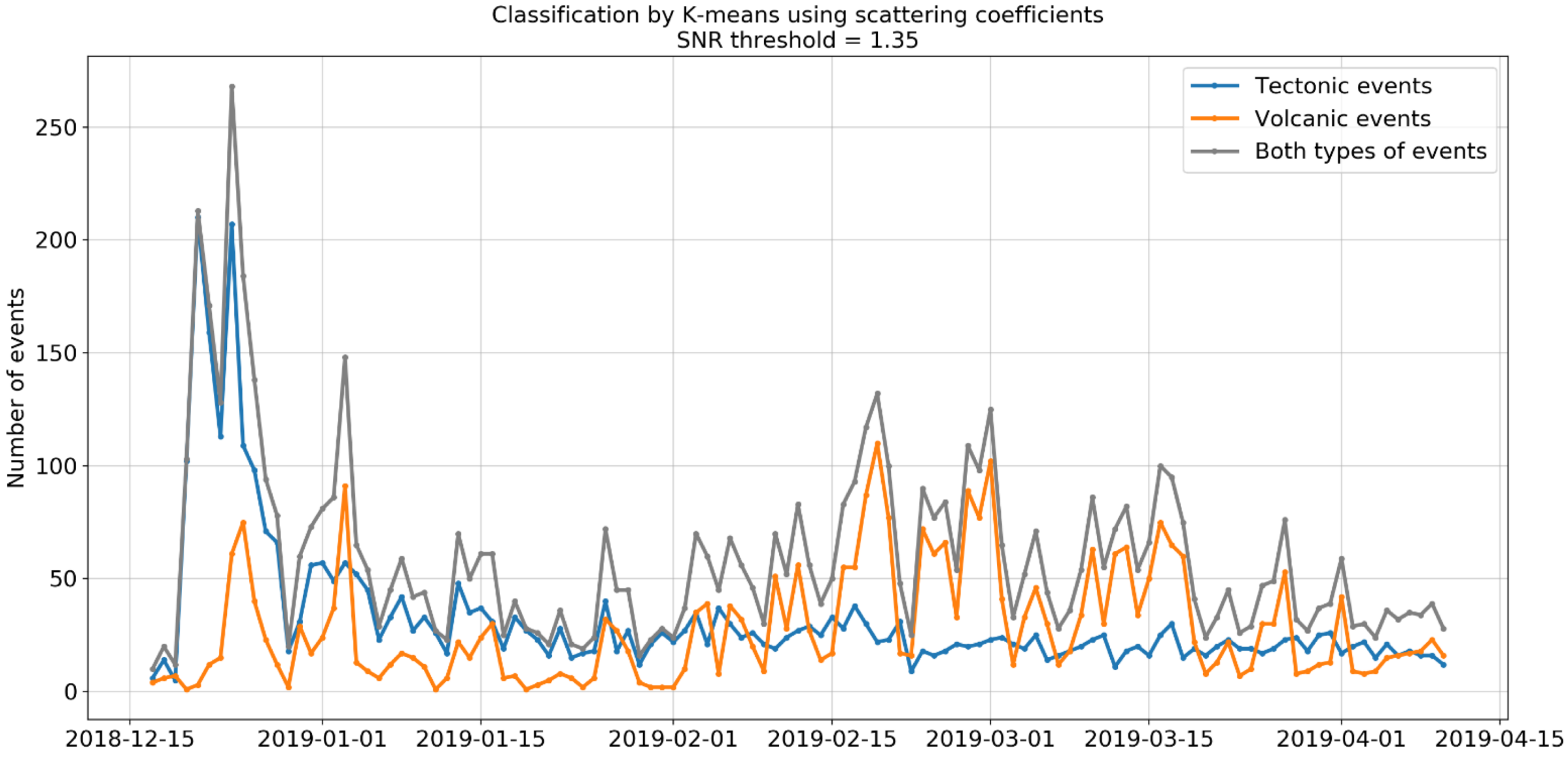
1000/1041 correct predictions both for K-means and agglomerative clustering algorithm







If we present the entire dataset for clustering algorithms the main difference from the previous results is the higher volcanic activity according to the algorithm



In fact, all activity plots have similar form and features that slightly depend on parameters chosen for this or that algorithm (kernel type for SVM, number of trees in Random Forest, etc.)

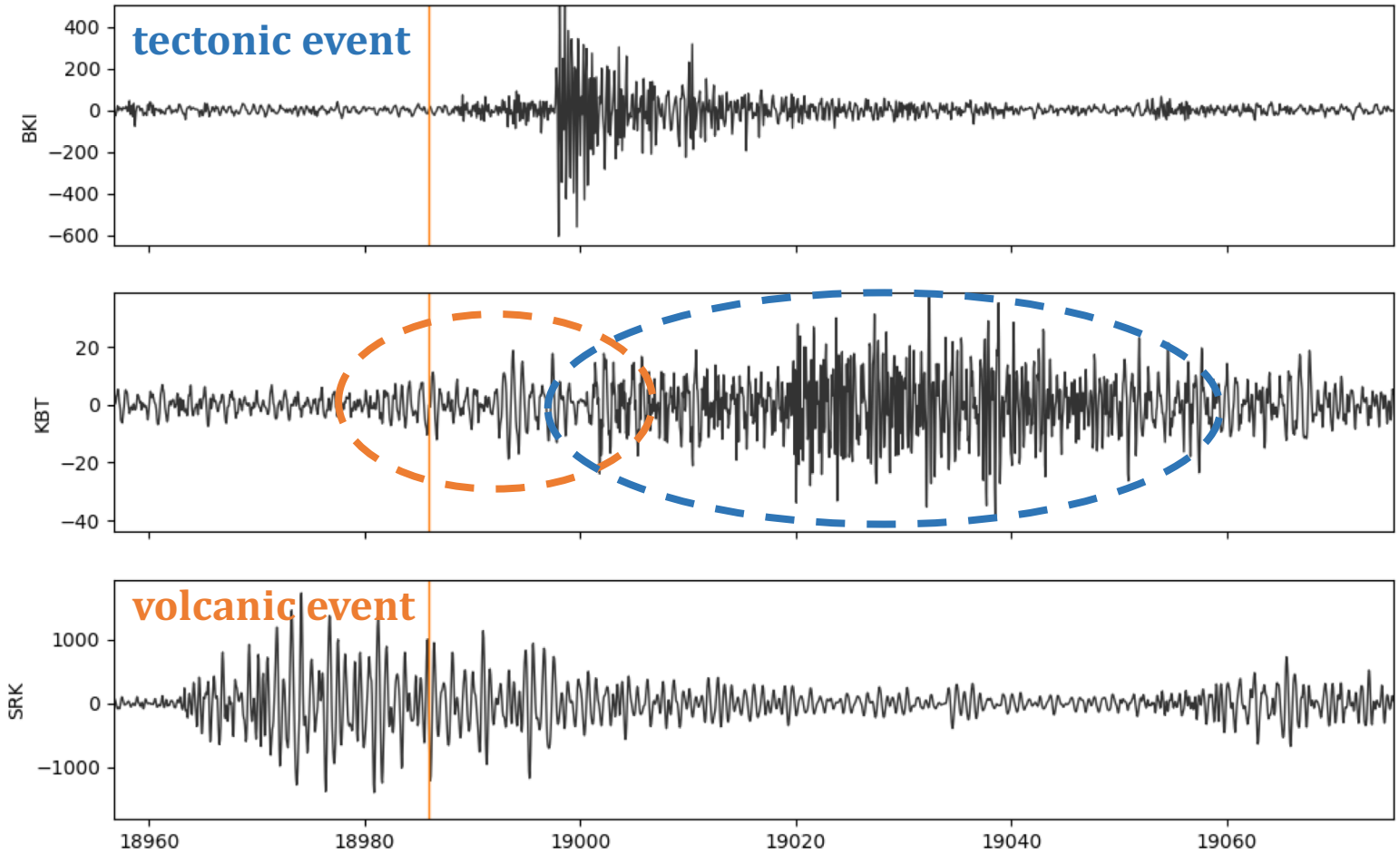
So, we will not show this plots further and consider main types of algorithms to assess their perfomance better

Manual seismograms processing

The horizontal (N) components of BKI, KBT and SRK stations were used.

Classification algorithms returned two classes of events:
0 for tectonic and 1 for volcanic.

During manual processing, the following types of events were introduced:
2 for noise/calibration signal, 3 for **overlap**, i.e. the case when both a tectonic earthquake and volcanic one occurred at the same time and at KBT station they formed a complex signal



Manual seismograms processing

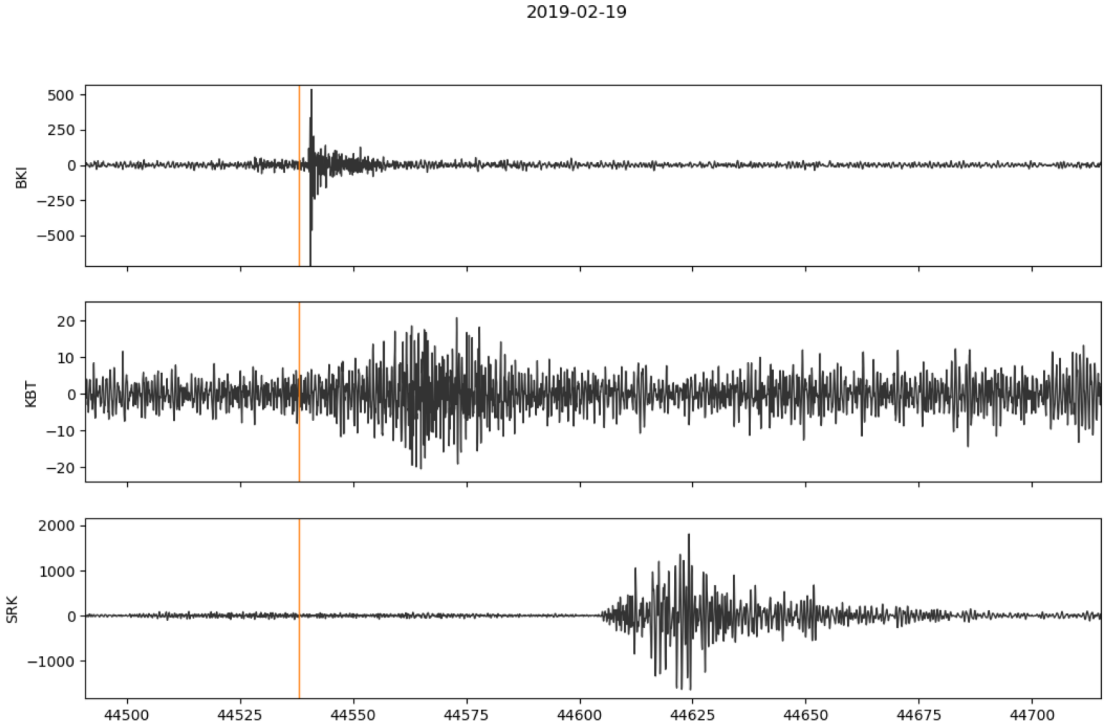
Due to high activity levels the following days were considered: December 22-25, 30, 2018, January 3 and February 19, 2019 with total number of checked detections of **1091**

41 of them were defined as **calibration signals/noise** and **67** of them are labeled as **overlaps** of tectonic and volcanic events

Manual seismograms processing

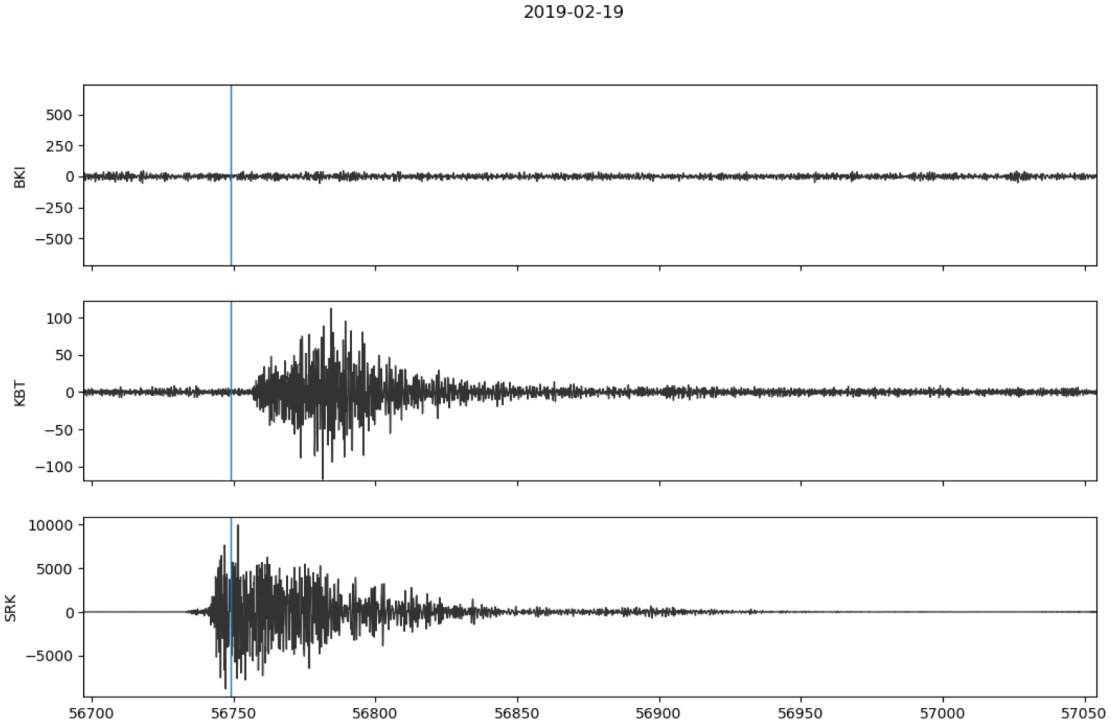
The main type of mistakes is the false classification of a weak tectonic event as a volcanic one

false volcanic event



Very rarely very strong volcanic earthquakes were classified as tectonic ones

false tectonic event



The least number (60) of the first type errors are made by the Random Forest algorithm when using smoothed spectra, but at the same time it makes relatively many errors of the second type (9)

Conclusions

- The results of manual processing showed that, regardless of the signal representations used, supervised algorithms provide better results: tectonic earthquakes are less often classified as volcanic.
- Results are quite stable relatively different classifiers and their main parameters
- Deviation of the aftershocks distribution from the Omori law cannot be explained only by volcanic activity