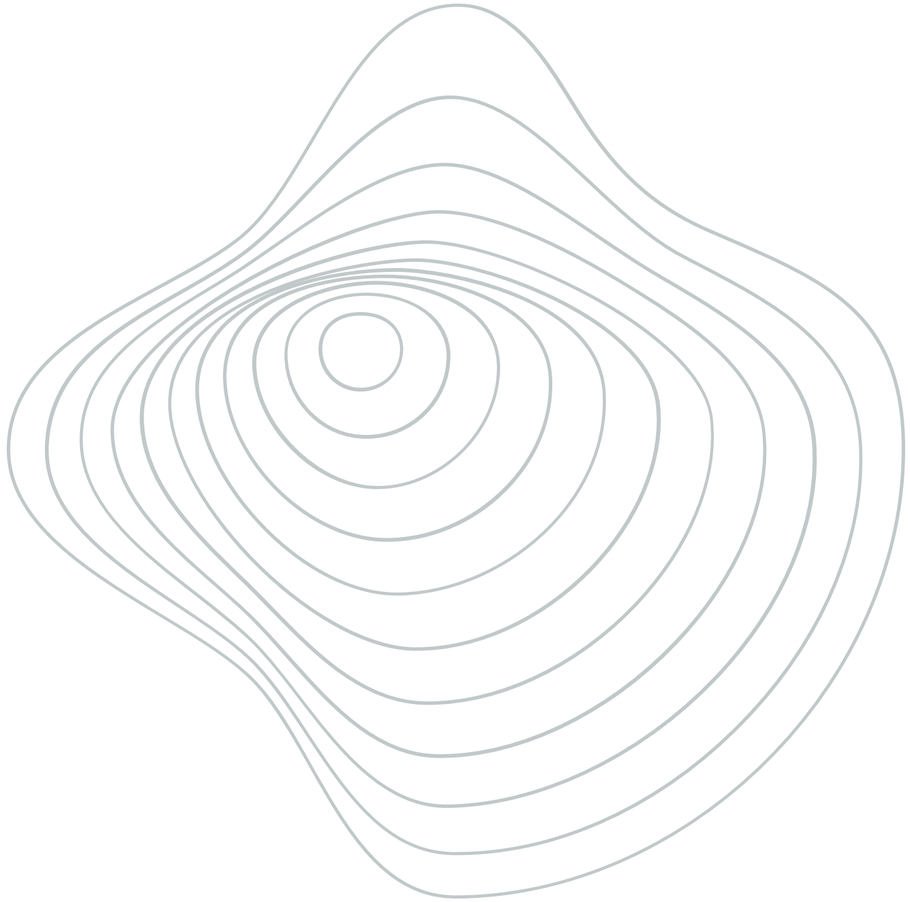
anomaly detection in time series: NASA bearing dataset

Andreoli Jacopo
12/12/2022

Motivations

OUTLIER DETECTION ANALYSIS ON THE NASA ACCELEROMETERS DATASET

The overall idea was that one of working on an outlier detection algorithm that deal with accelerometers data, so as to find characteristics and behaviors associated to this type of sensors

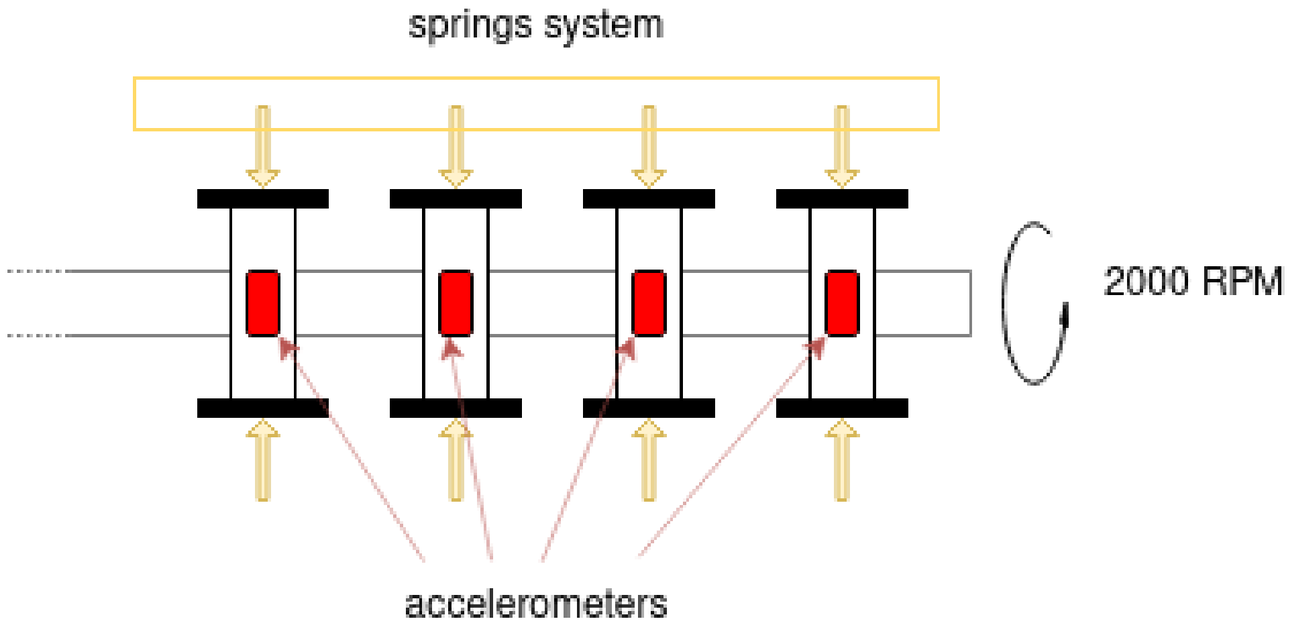


NASA bearing dataset

DISCLAIMER: only the 2nd test dataset is considered in the following slides

GENERAL DESCRIPTION

The dataset consists into measurements carried out from high-frequency accelerometers (20 KHz) that are installed on four different bearings. These bearings are installed on a shaft rotating constantly at 2000 RPM. A system consisting of a radial loads and springs is used for ensure the contact of the bearing with the shaft; lubrification is used for reduce friction.



There are 4 bearings

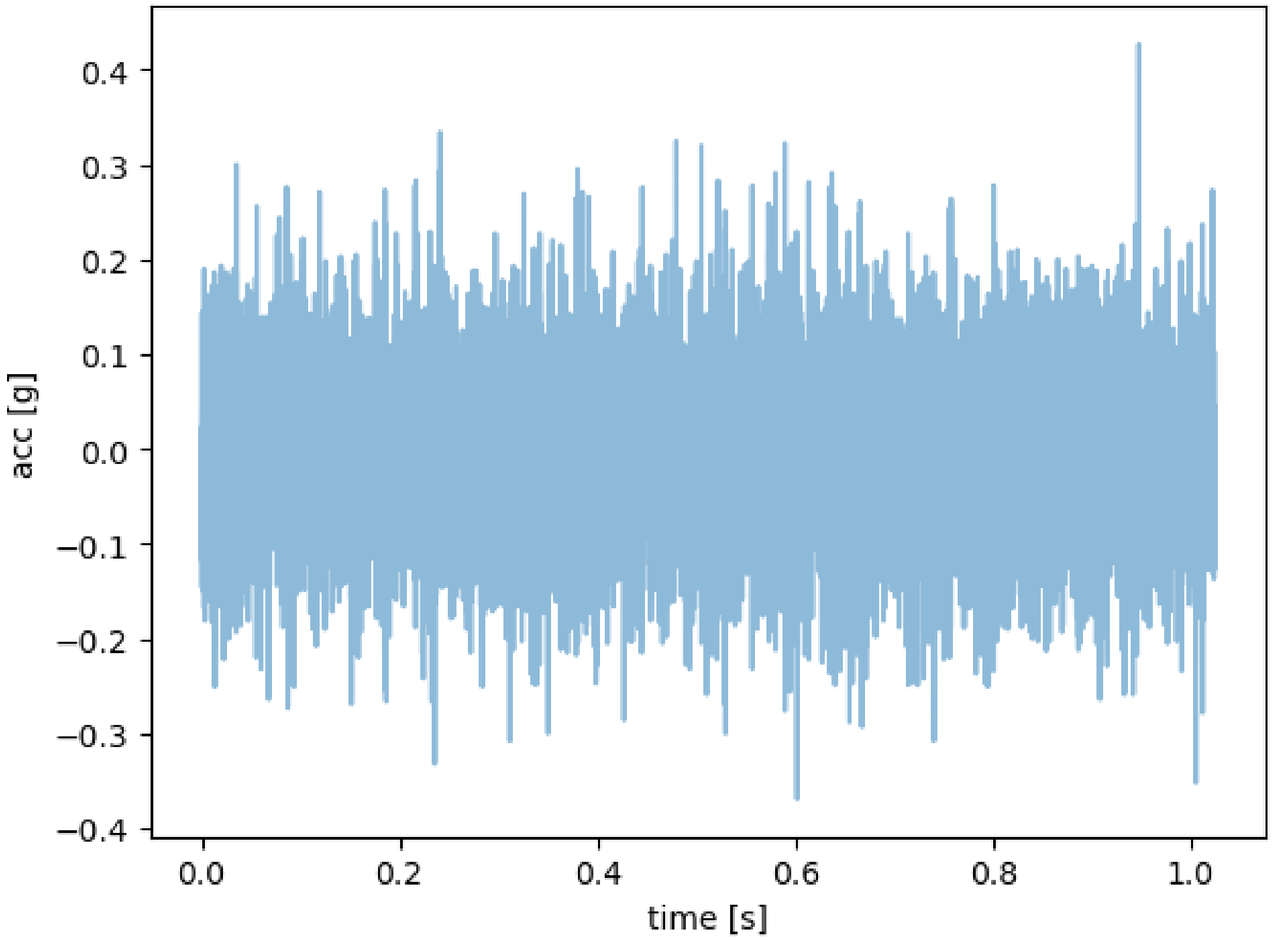
FINAL AIM

find an algorithm for automatic identification of bearing with defects analyzing the vibration measurements carried out at accelerometer level

Dataset organization

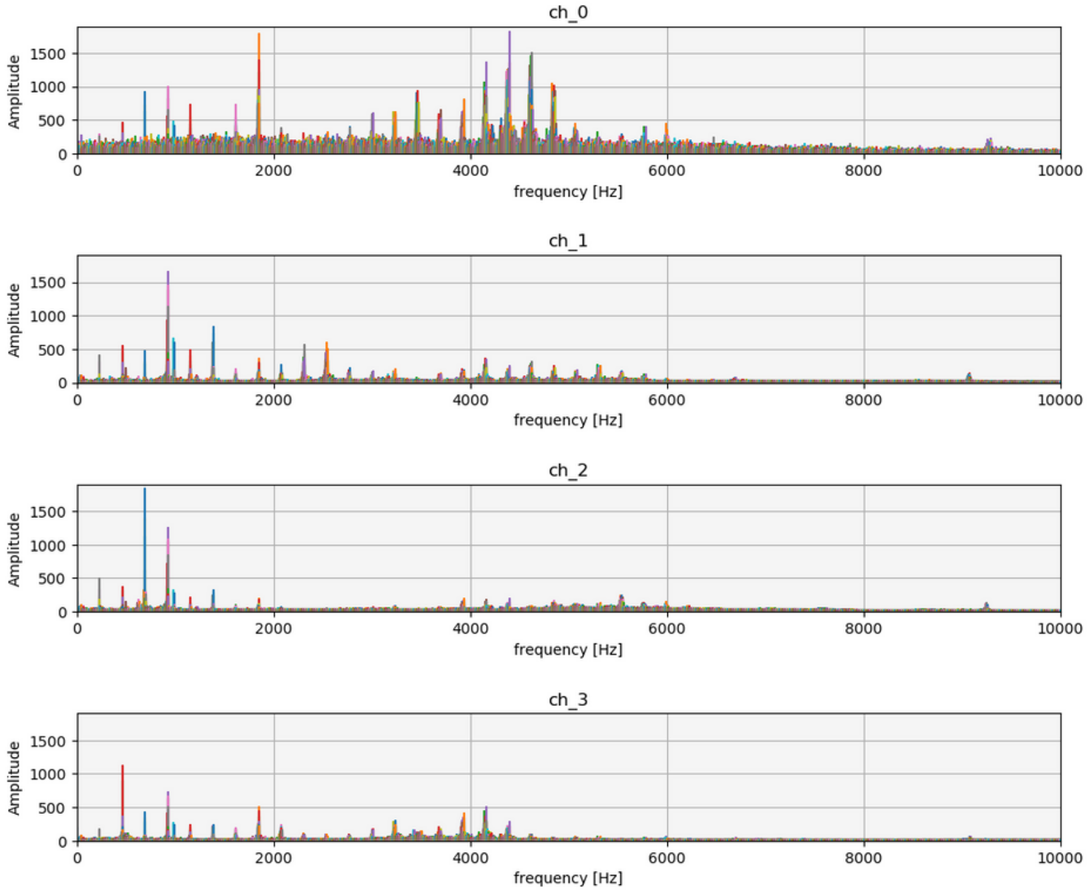
- 4 accelerometers, namely 1 for each bearing
- 984 ts
- each ts last for about 1s
- in the end, outer race failure occured in bearing 1

time series example
(2nd test, ch1, ts = 0)



Dataset Analysis

Since the dataset deal with vibration signal,it make sense to bring the signal analysis in the frequency domain, where it is possible to observe the different behaviors of the accelereometers



ts-fft 2nd test (plotted all 984 samples)

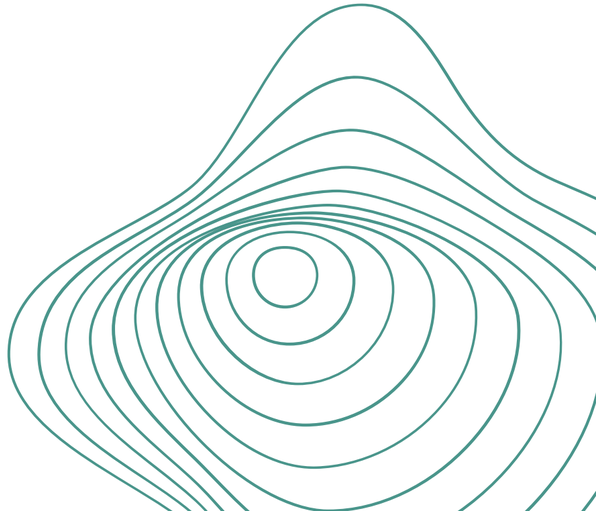
bearing 1 (ch_0) shows defect at the end

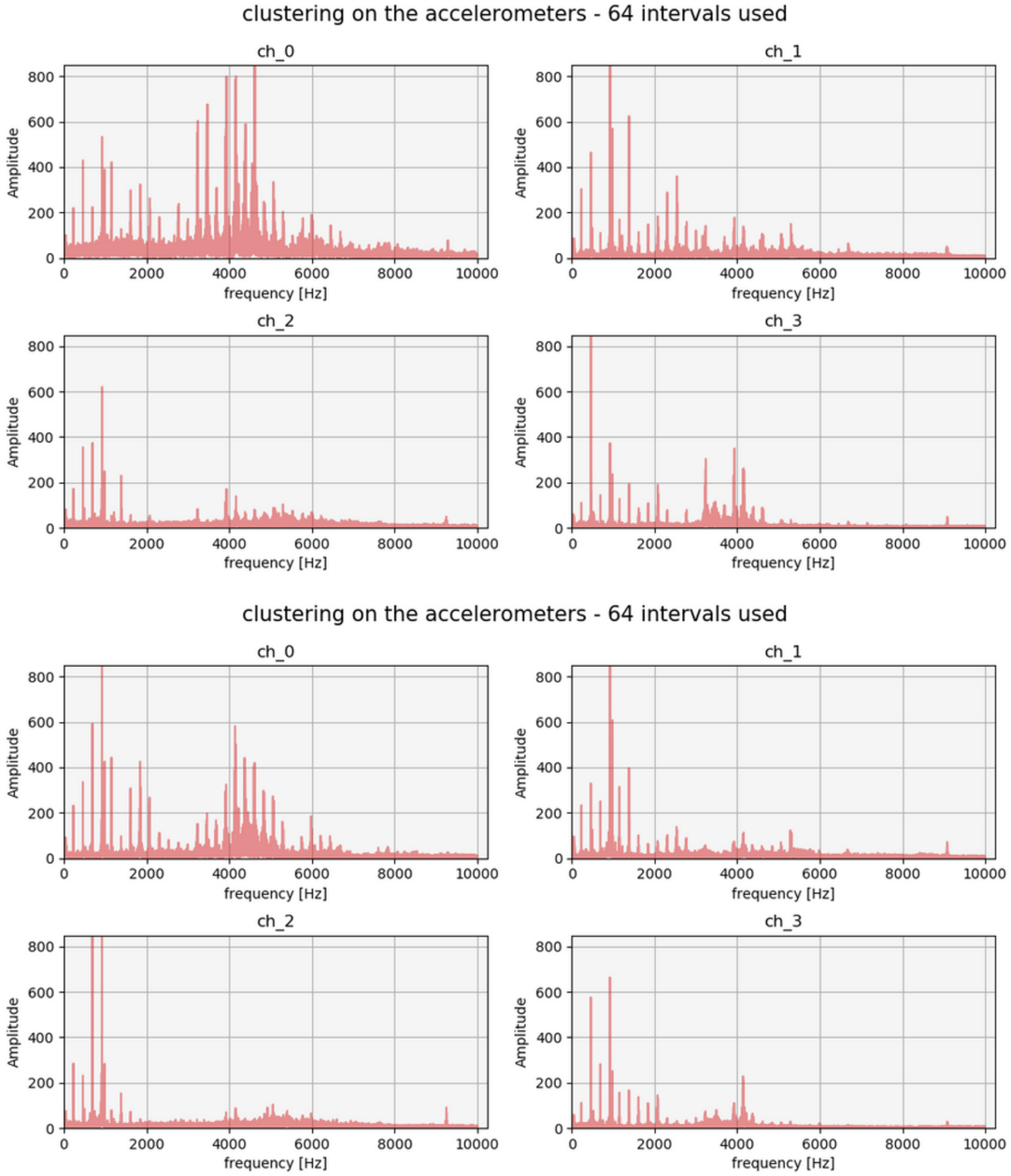
1st approach

given the fourier transform in the frequency domain, the 10000 frequencies (given the Nyquist frequency 20000/2) are divided in 64 intervals and the following characteristics are extracted:

- max peak value
- mean
- std
- kurtosis
- skew

then a simple k-means algorithm is used for the classification of the signal. The results was good, however the algorithm was not able to discretize cases in which all accelerometers face something different from the past, that is related to the working condition and not in presence of defect

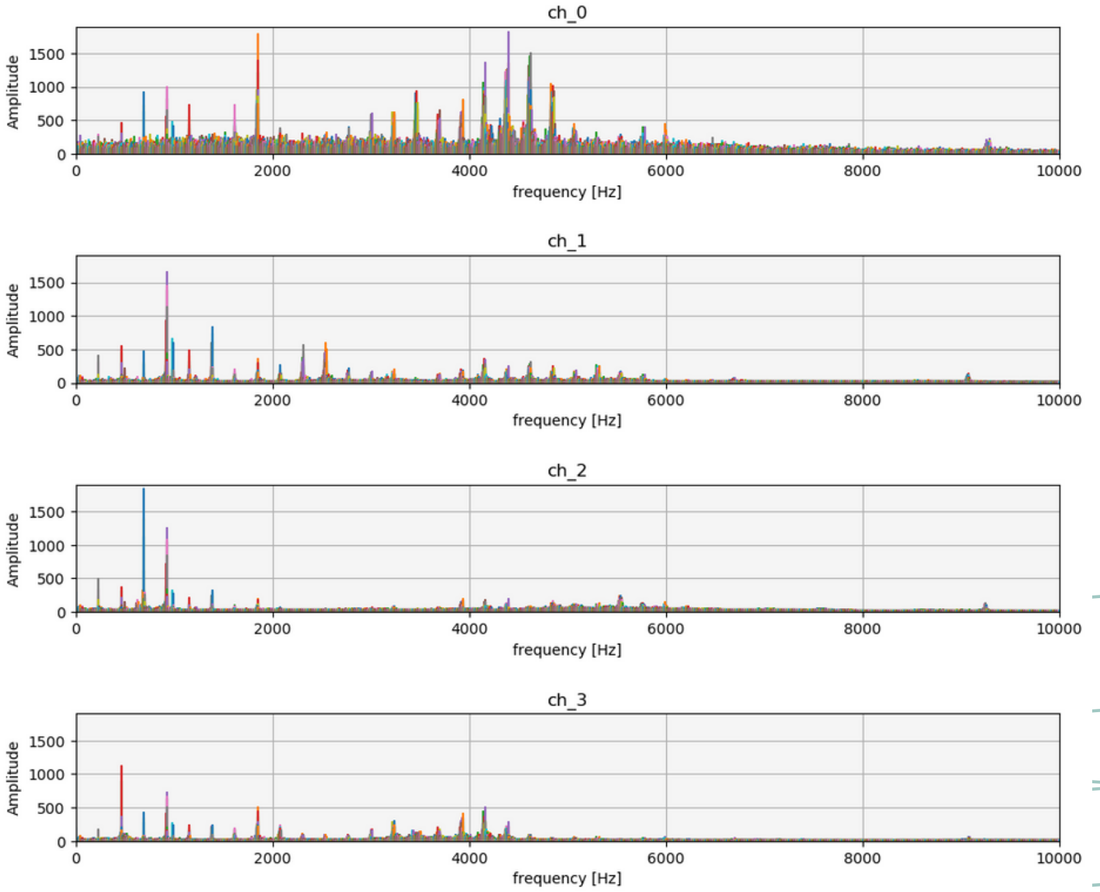




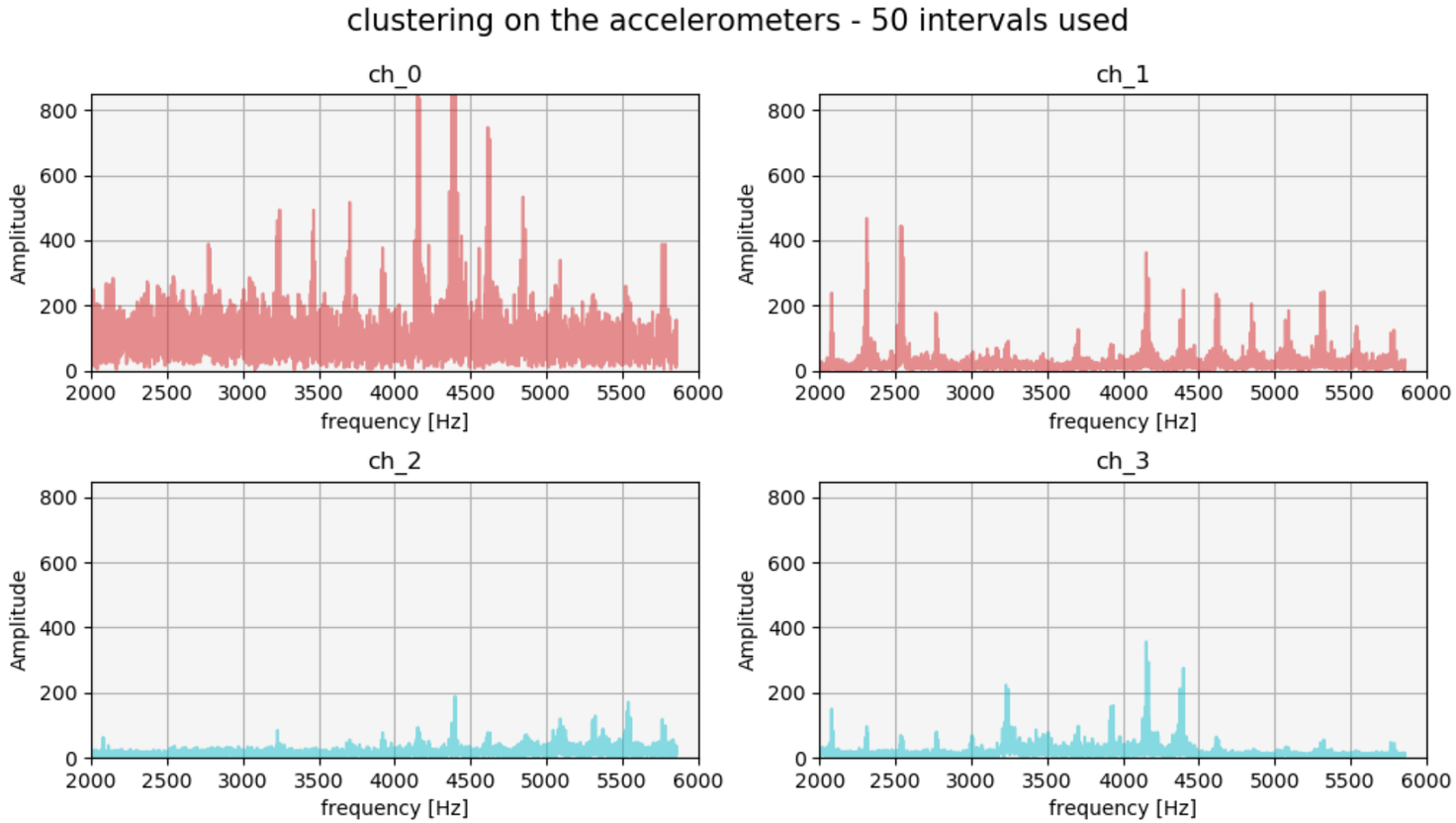
Some examples are reported here; the working condition bring the accellerometer to new behavior that differ from most of other samples, but dose not identify a defect. The broken accellerometer is not distinguished from the others

1st approach - rivisited

A cause of the results obtained in the previous clustering could deal with the intrinsic parameters of the accelerometers, like presence of shift, difference placing position over the bearing, different friction at the section of contact, ... hence, there could be some difference at lower frequencies. Then, from the plot show in the previous slide, reported here for simplicity, at frequencies higher then 6 KHz, there is no resonance, also for the broken bearing (in all three tests). Hence the idea is to restrict the interval of study from 2KHz -> 6KHz



Only with this simple assumption, the results obtained with the trivial k-means algorithm was good, considering that at the end, all the samples belonging to the outlier set belong to the channel 0, with the exception of only one sample, reported here



Hence 1/984 sample seems to not be correctly clustered. As before, it seems that the clustering method is not able to differentiate the actual working condition with respect to the normal behavior of the accellerometer in time

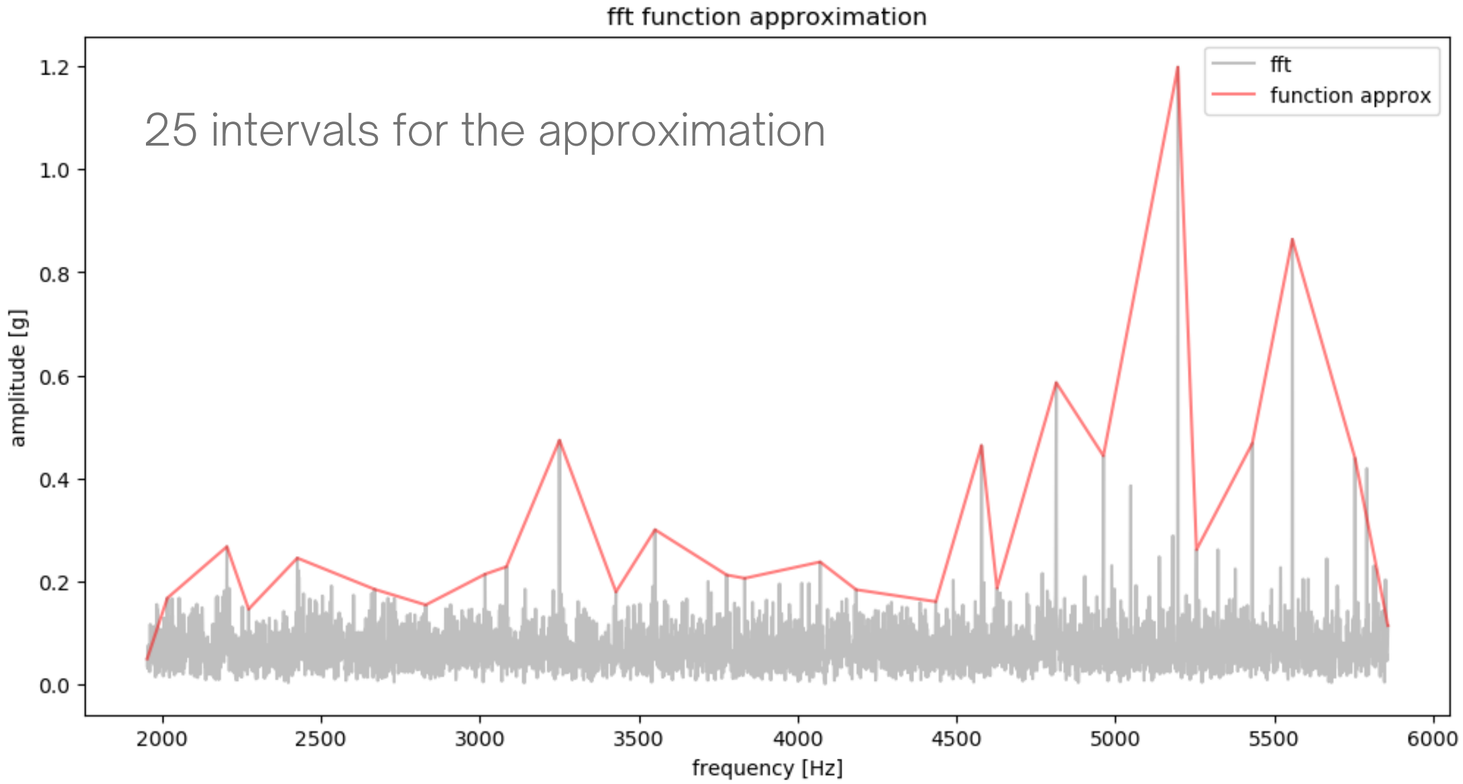
the idea is to find a different path to solve the problem, observing if it is able to find a better classification

2nd approach

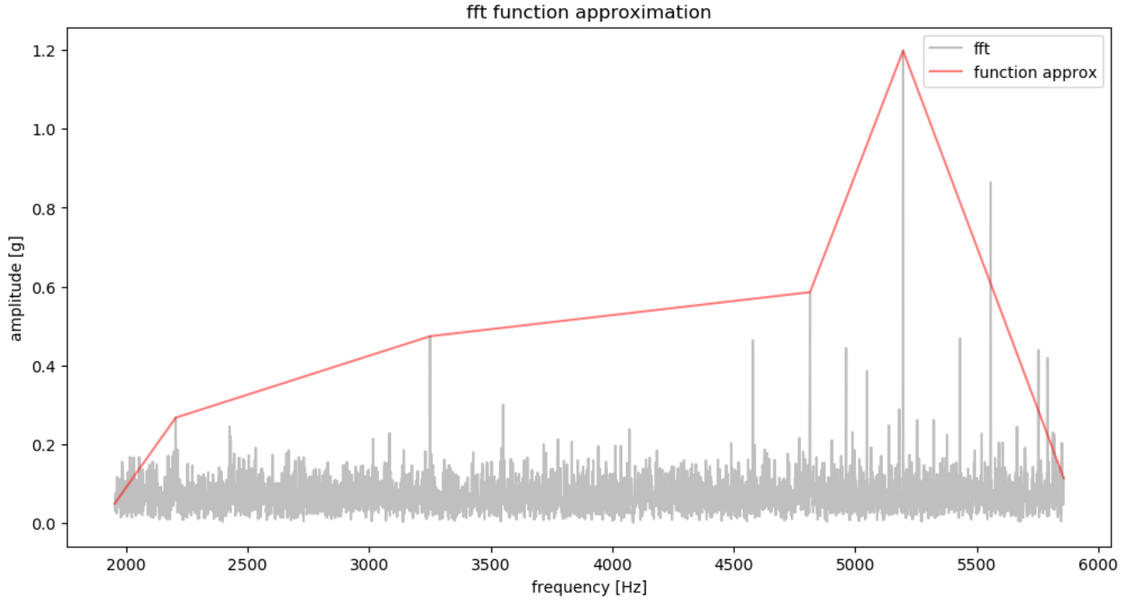
The idea in the first approach is to differentiate the functions underlying the Fourier transform of the different channels exploiting some parameters that try to encode the function description. Then the clustering method output the classification based on the variation of such parameters.

In this approach, so as to enhance the accuracy, the idea is to directly find a characteristics directly on the Fourier transform function instead of passing through the parameters evaluation

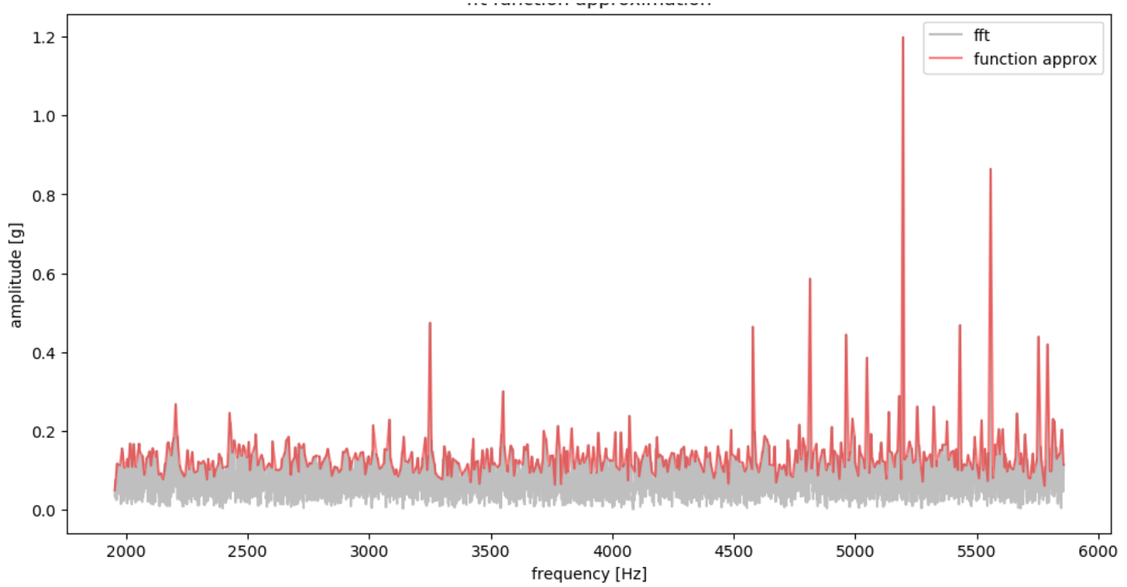
The first step considered is to find a description of the Fourier transform that is able to encode its variation wrt different channels considered. The idea is to exploit a function description based on maximum peaks found in different intervals in which the frequency is divided (as 1st approach did)



The function is built taking maxium peak at each interval; the correct number of intervals needs to be chose given the variation in frequencies observed

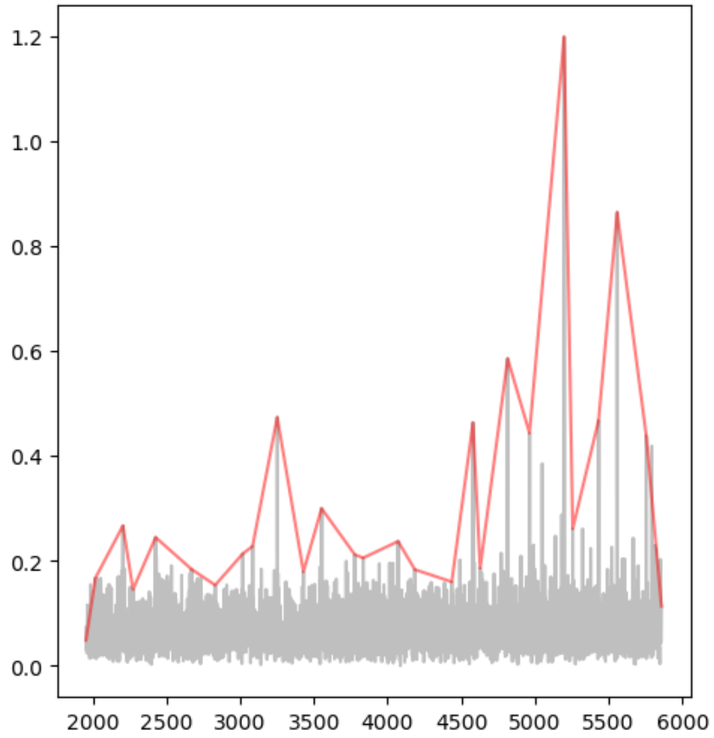


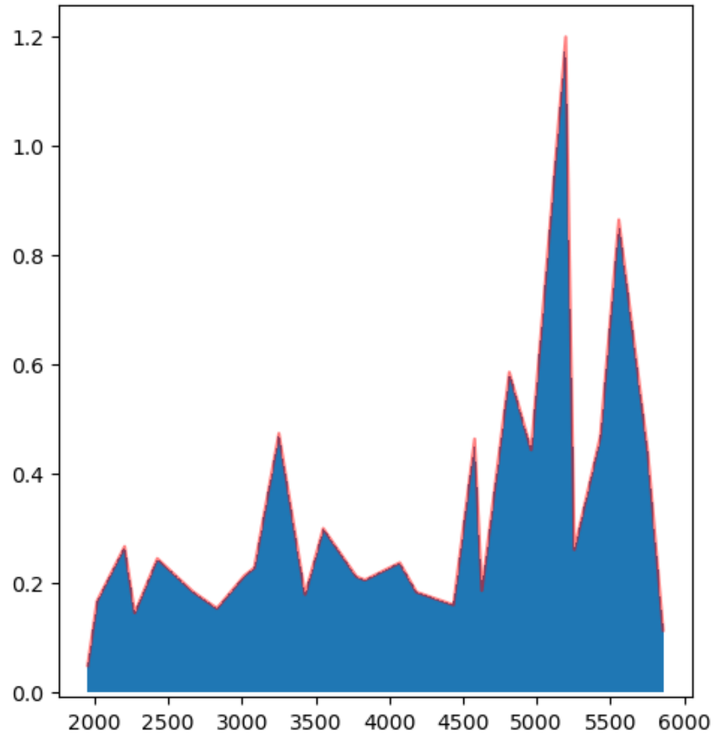
4 approximation intervals



500 approximation intervals

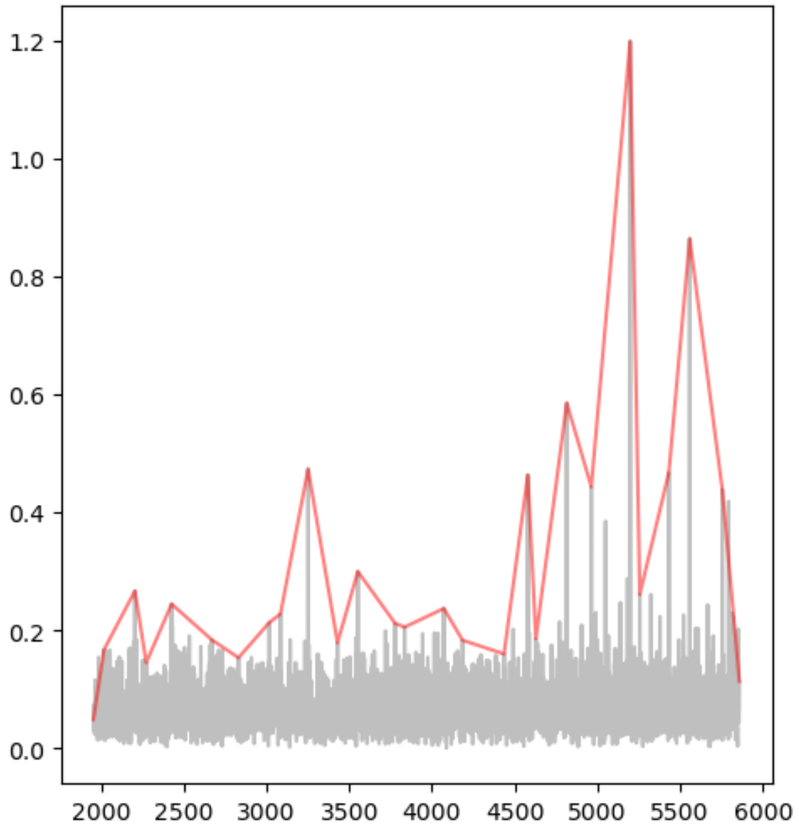
The idea is that: each accelerometer does not provide a perfect description of the frequency in which peaks are detected. Hence, it is not possible to compare the different functions in a points to points manner. This is due to the nature of the signals, that will not be the same and also for the always present noise in the measurements. Inestead, the approximation give the possibility to find a points to points comparison. However, in order to make it more robust, it is not considered the difference between peaks, but instead between the different integrals computed in the intervals

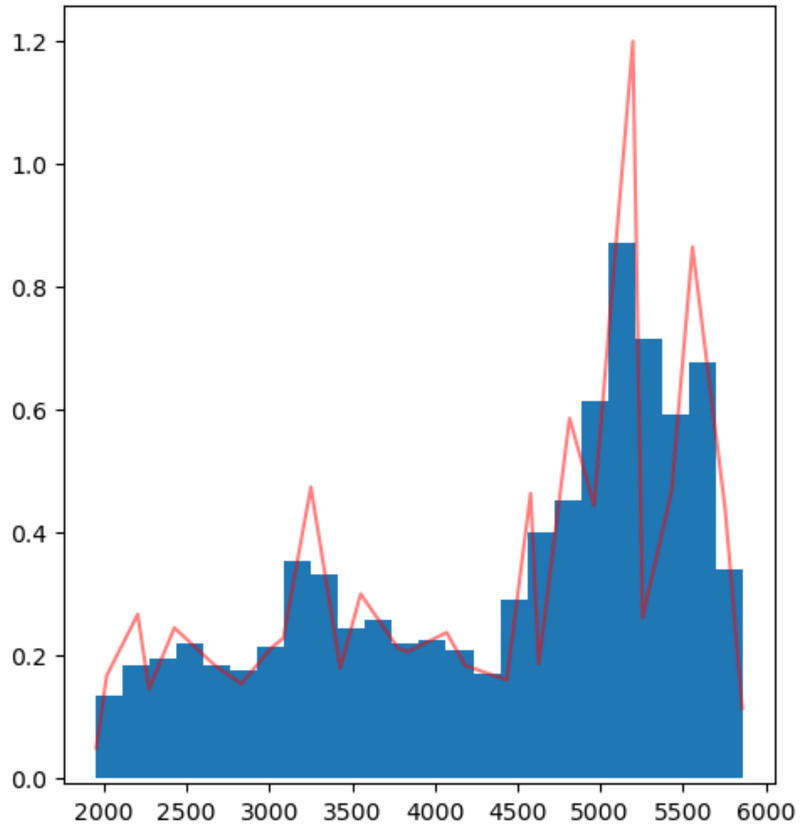




The integral is computed in a Rienmannian (finite) fashion: the frequency interval is divided in 500 intervals

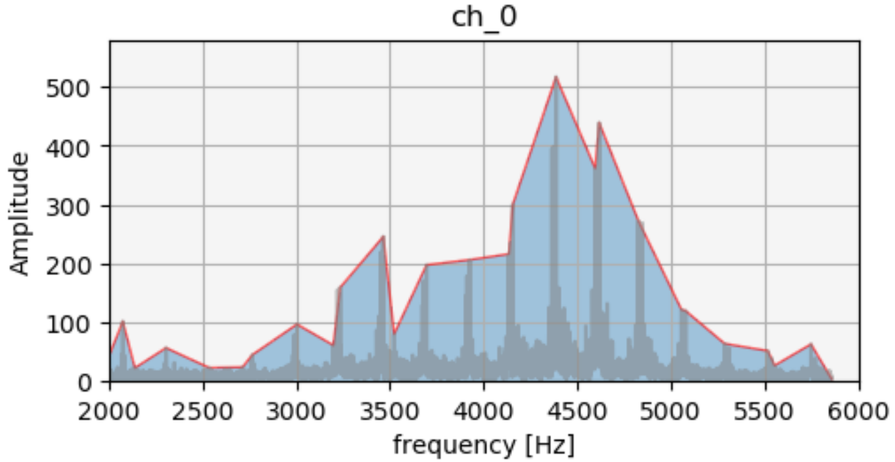
In the case of integral approximation, as the intervals enhance, as the integral calculation is more accurate. However, from the computation viewpoint, it does not make sense to use too high values. Some problem arise with low value for the number of intervals, since the integral approximation become poor

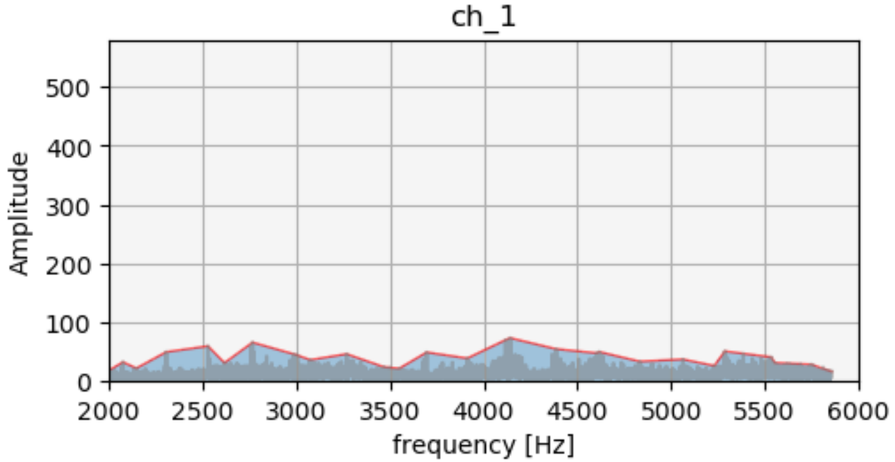


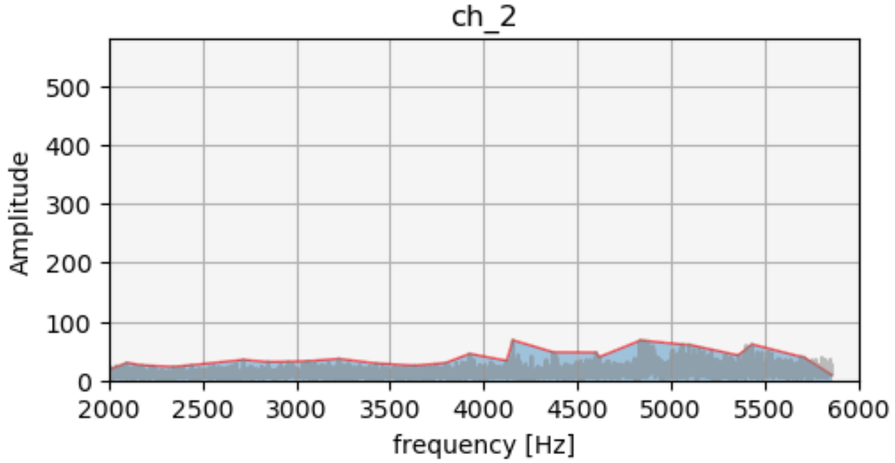


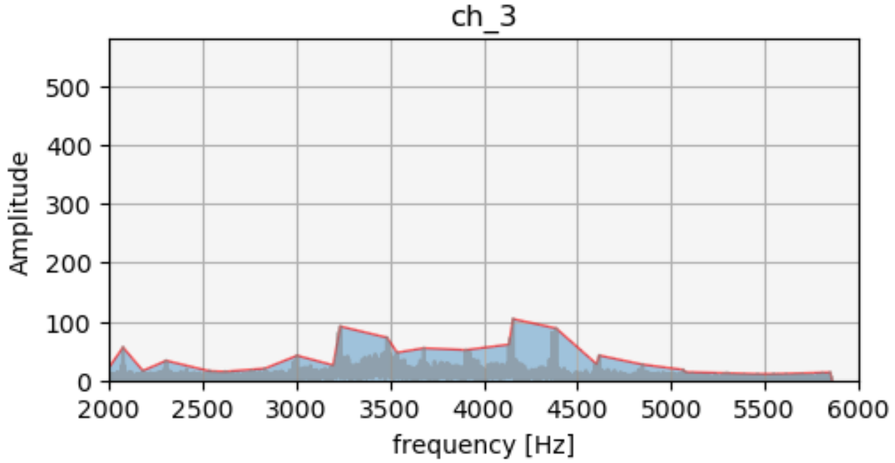
Arriving at this point, the question is: how to find a metrics that differentiate the good accelerometers wrt to broken one considering also the timestamp in which the measurements are taken?

- make the difference of the integral between the channel so as to find how a channel is "different" wrt to the other
- average wrt the integral difference of the other channel
- normalize for the channel more "similar" wrt the others

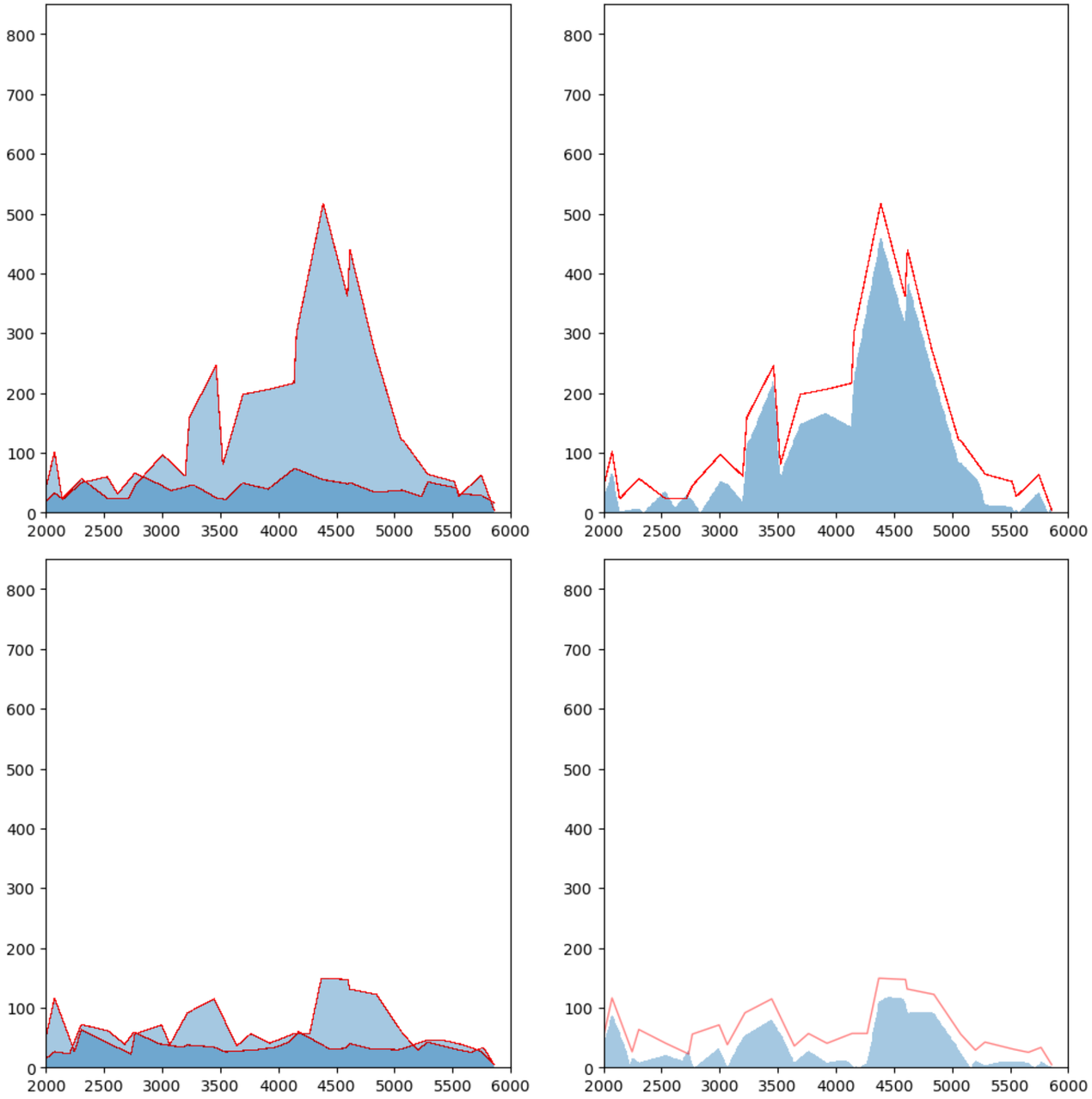








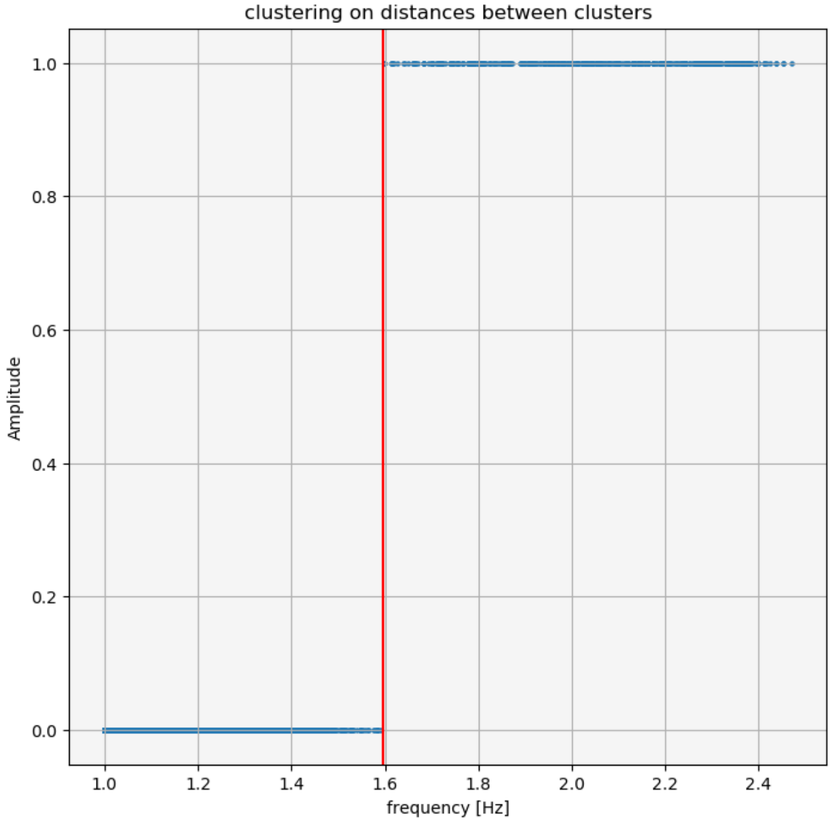
Difference of integral examples

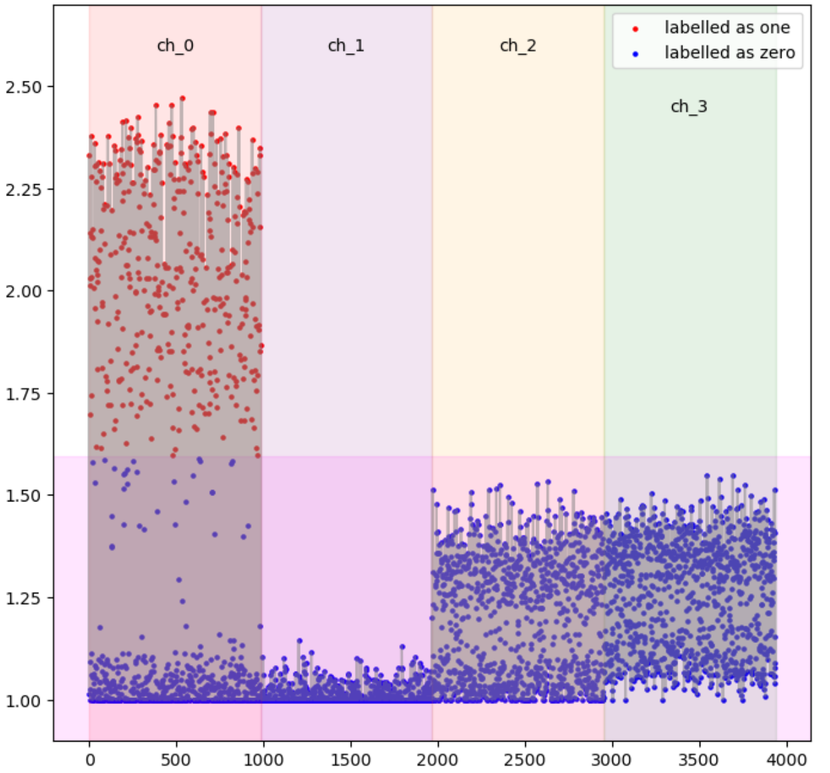


Broken Vs not broken

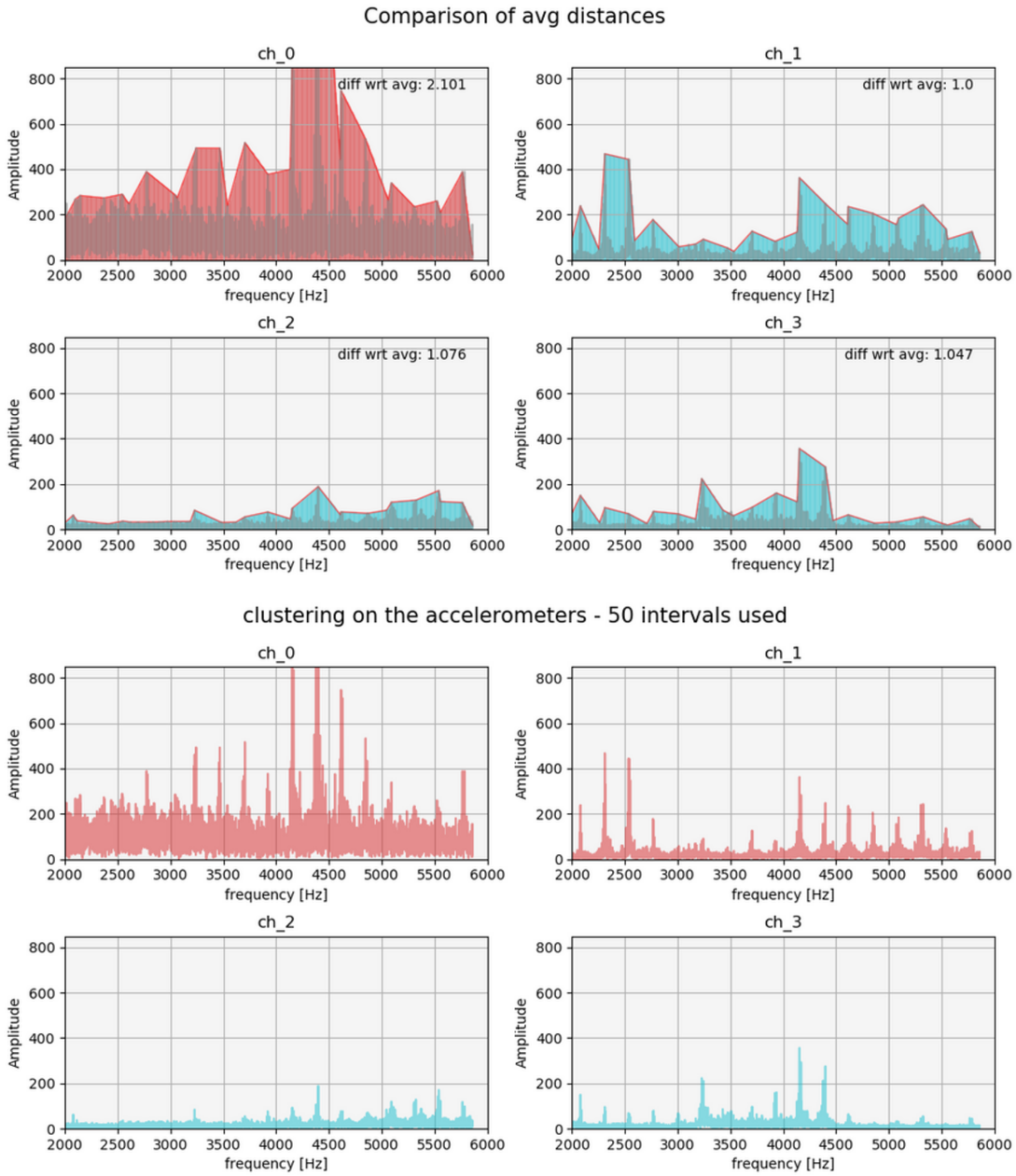
Not broken Vs not broken

At this point, each channel of each ts has only one value. In this case a 1D k-means is applied. The results show that the algorithm is able to distinguish all the samples correctly.





Comparing the sample wrong labelled



The results for the clustering with 1st-2nd method is reported in the linked folder. What is important to remark is that NO LABELS are present; it is considered something abnormal the signals that have different component at high frequency. For further comparison, one can refer to the two Jupiter notebook attached in this repository