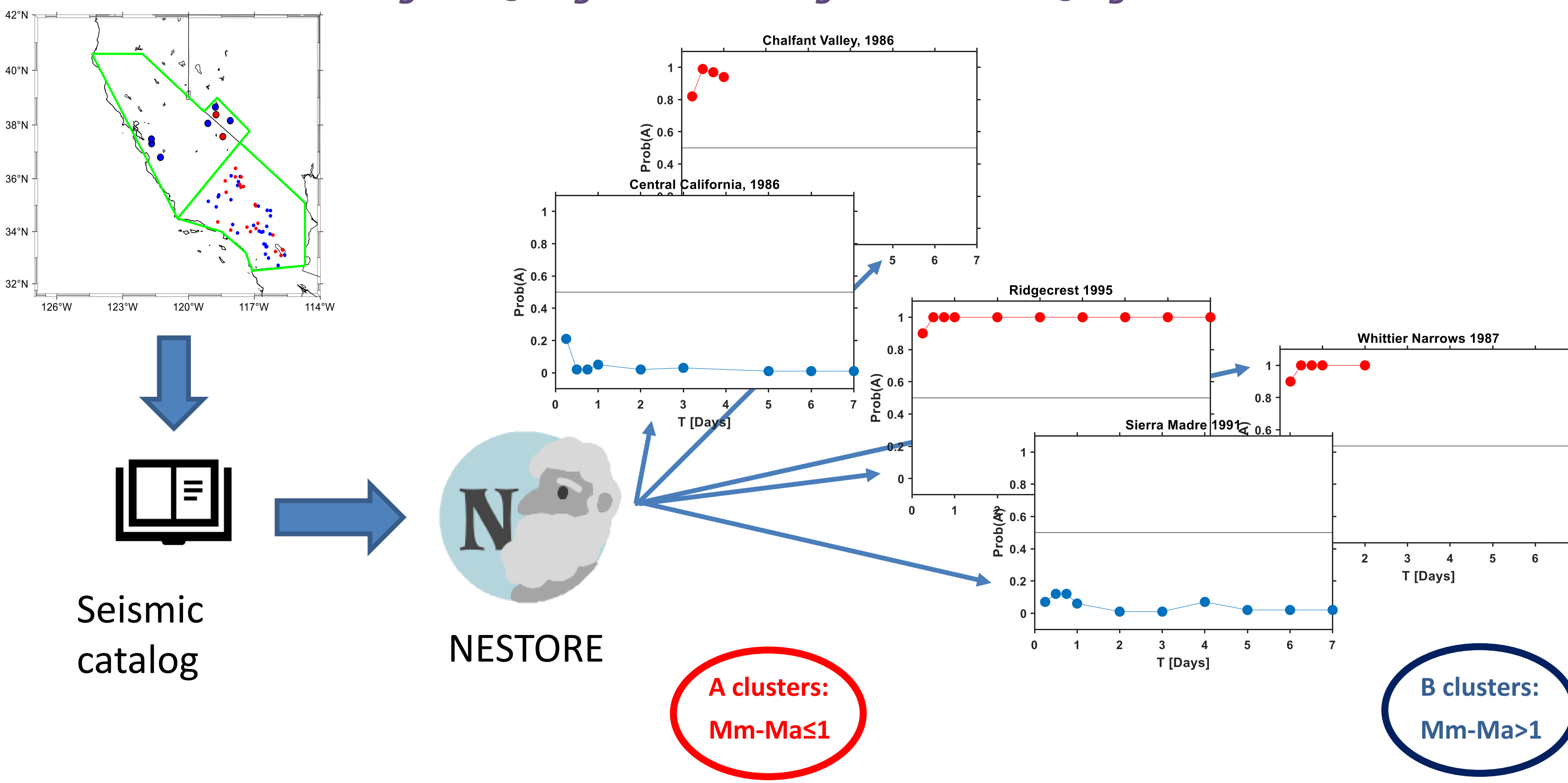
# Strong following earthquake forecasting by a pattern recognition approach in California

*S. Gentili and R. Di Giovambattista*
*sgentili@inogs.it and rita.digiovambattista@ingv.it*



# Motivation and adopted algorithm

- Large earthquakes that occur in a cluster after a strong event pose a significant threat to civil protection because they cause additional damage to already weakened buildings, raising the risk of building collapse.

- We divided the clusters into two classes based on the difference Dm between the o-mainshock (the first shock with M>4) and the strongest following earthquake magnitude

  (Type A: Dm≤1 Type B: Dm>1)

- **NESTORE** – (*NExt STrOng Related Earthquake*) is a software package based on machine learning approach for A clusters forecasting. It analyses the seismic data at increasing time intervals after the mainshock.

- One-node decision trees are used to train the algorithm using different features derived from seismic catalogues a short time after the o-mainshock. The output for each feature is a threshold.

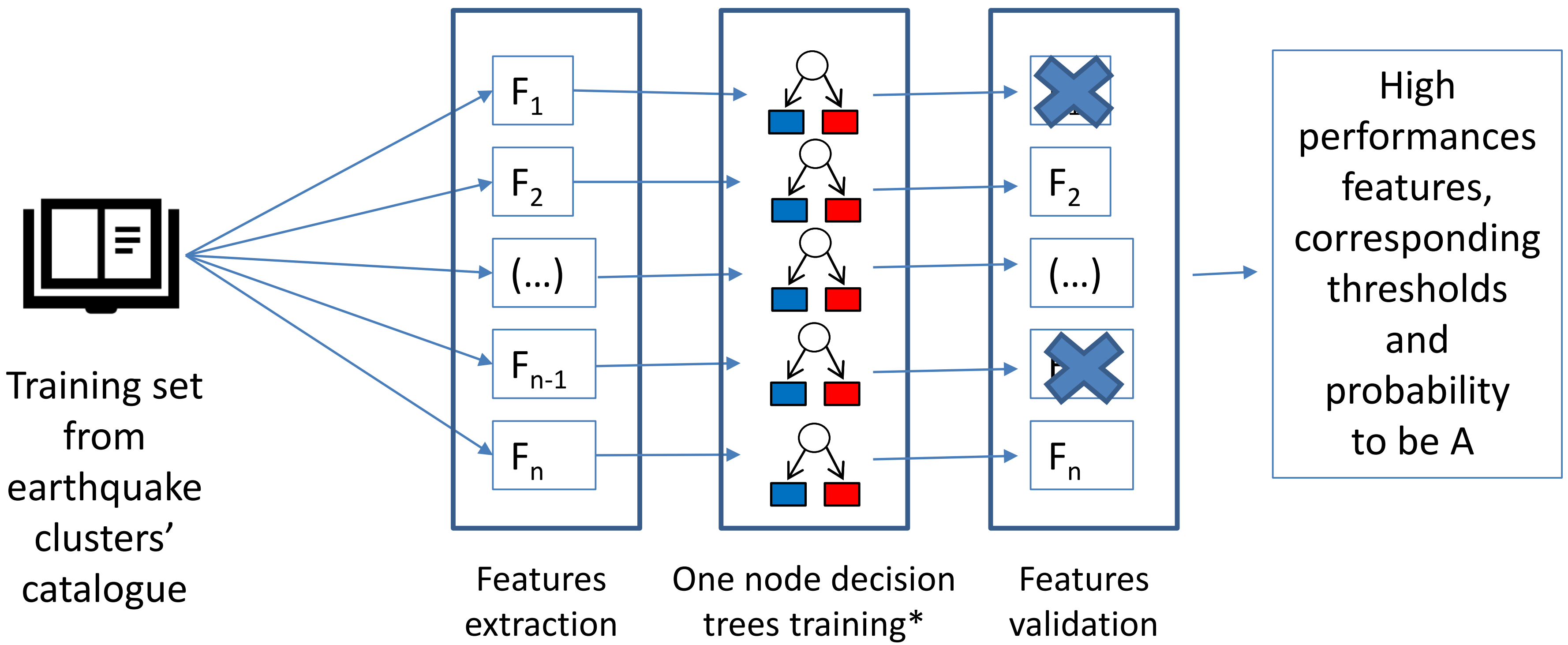
# Tested features

- N2=number of aftershocks (with magnitude Mm-2)
- S=total equivalent source area
- Q=cumulative radiated energy
- Vm=variation of magnitude from event to event
- Z=linear concentration of aftershock
- SLcum, SLcum2 =deviation of S from the long term trend (SLcum2 with sliding window)
- Qlcum, QLcum2=deviation of Q from the long term trend (QLcum2 with sliding window)

# NESTORE - TRAINING for each $T_J$

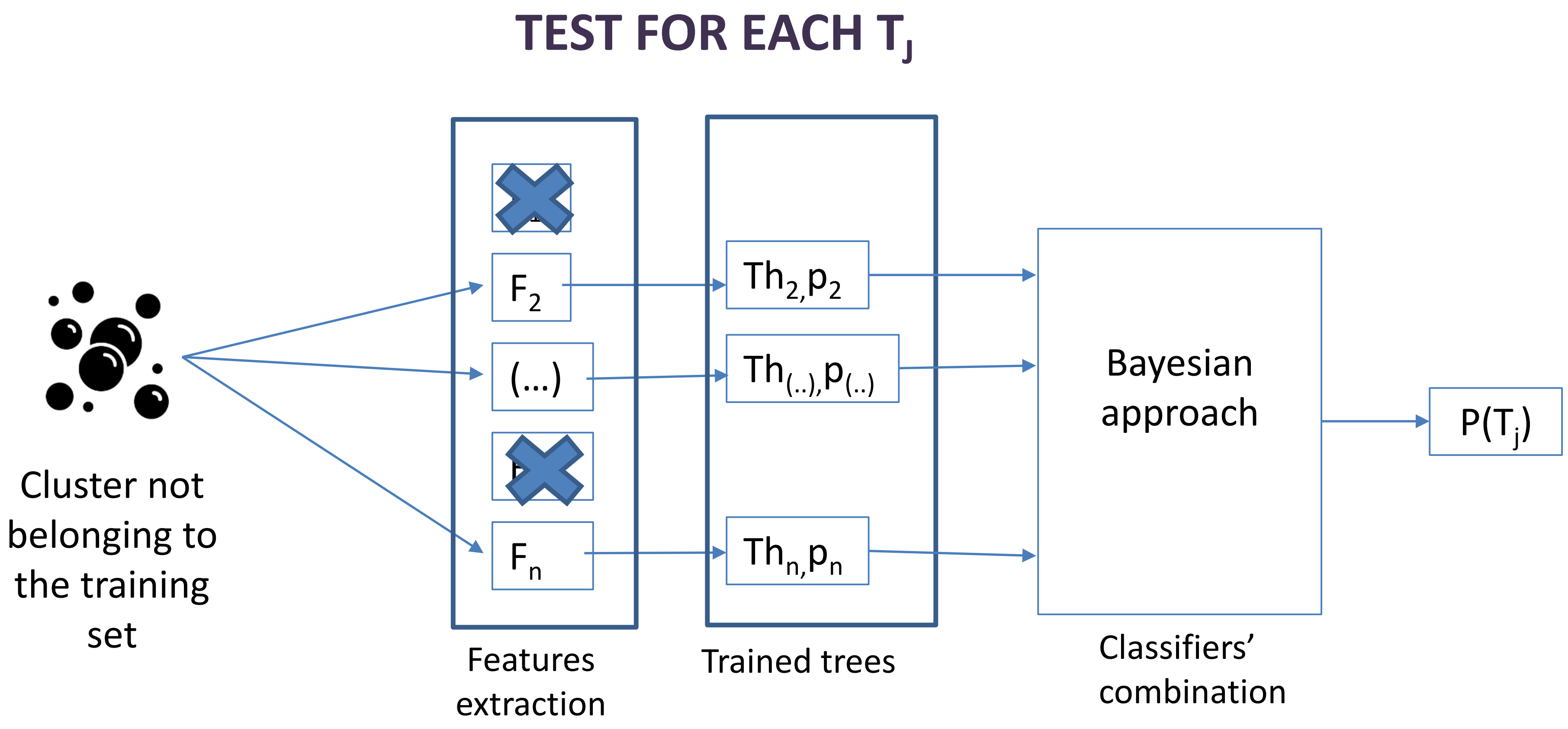
- NESTORE analyses the seismic data at increasing time intervals T=[0.25,0.5,0.75,1,2,…,7] days ($T_i$) after the mainshock.

- For relevant features at given $T_i$, it supplies a thresholds over which the cluster is classified as A and the probability to be an A.

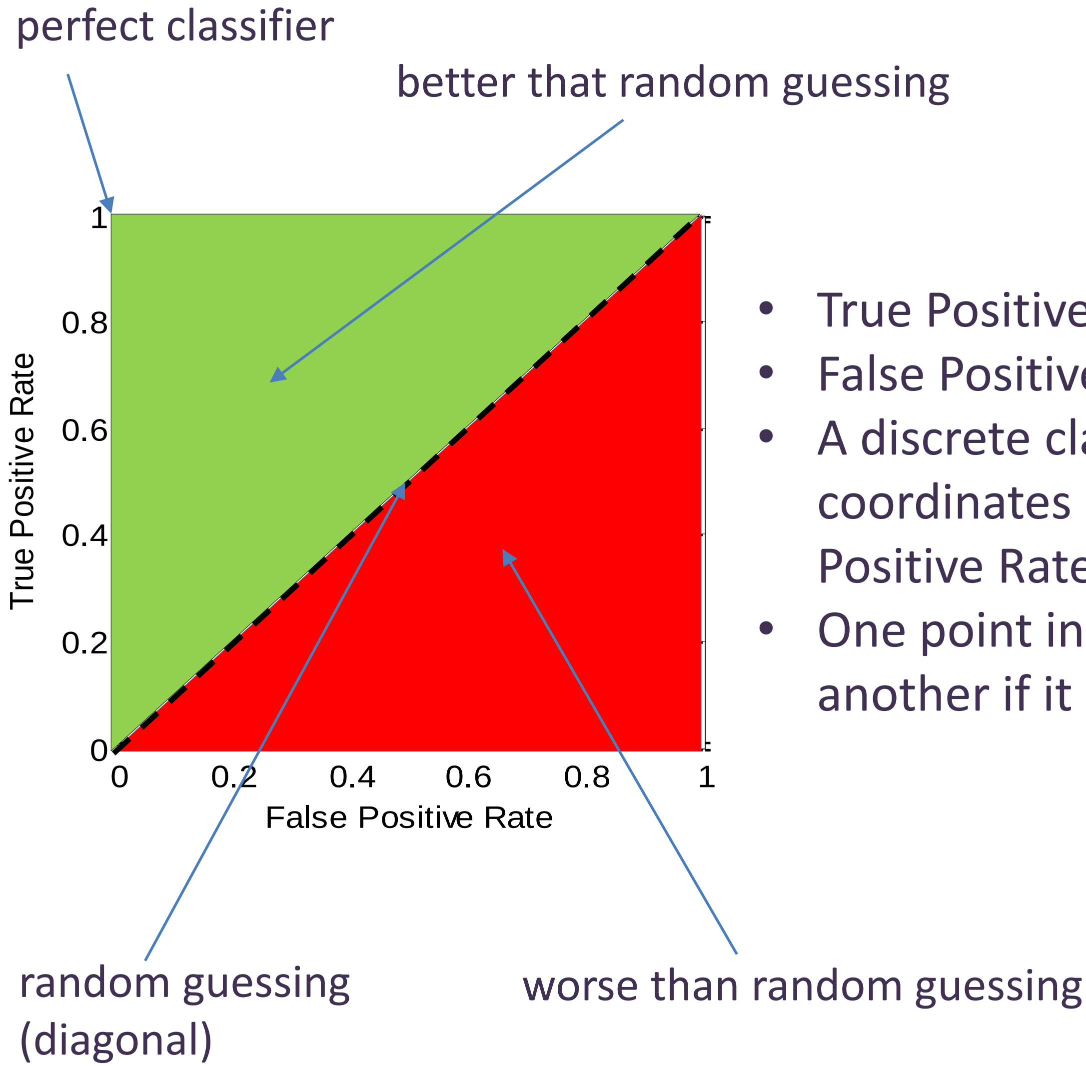


# NESTORE classification

After the training NESTORE can supply the probability to have an A cluster at time intervals $T_i$

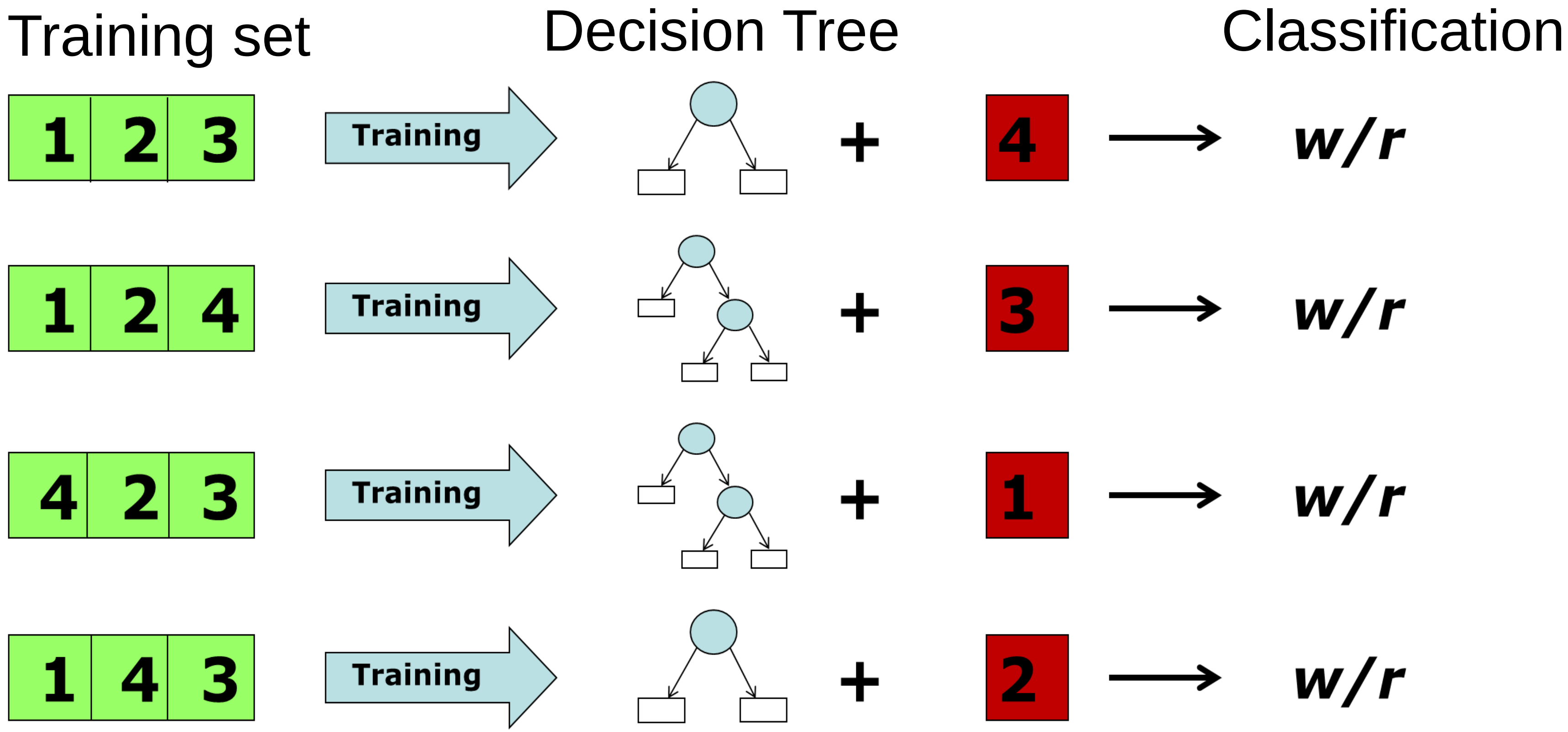


# ROC graph




- True Positive Rate= True(A) / All(A)
- False Positive Rate= False(B)/ All(B)
- A discrete classifier produces a point of coordinates (False Positive Rate, True Positive Rate) in the ROC graph.
- One point in ROC space is better than another if it is closer to the point (0,1)

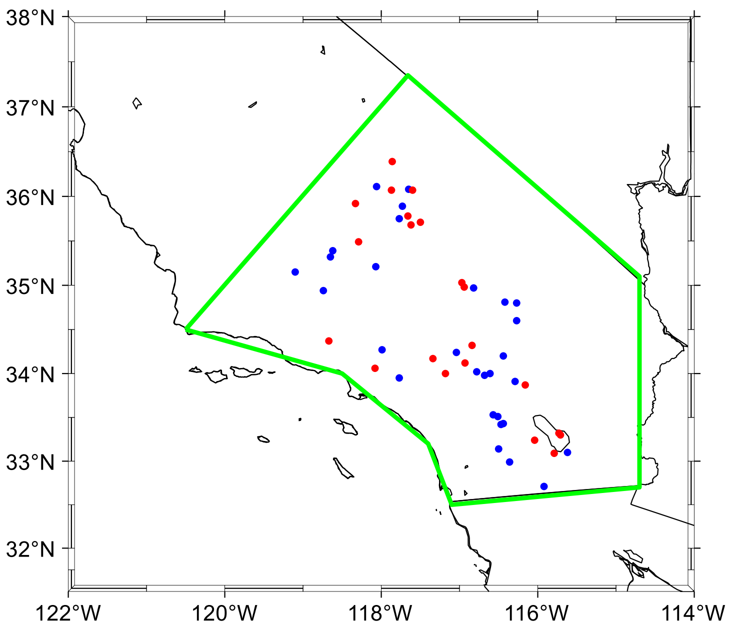
# LOO method




LOO method is a testing procedure in which the learning algorithm is applied once for each instance, using all other instances as a training set and using the selected instance as a test set

# Database and clusters identification: Southern California

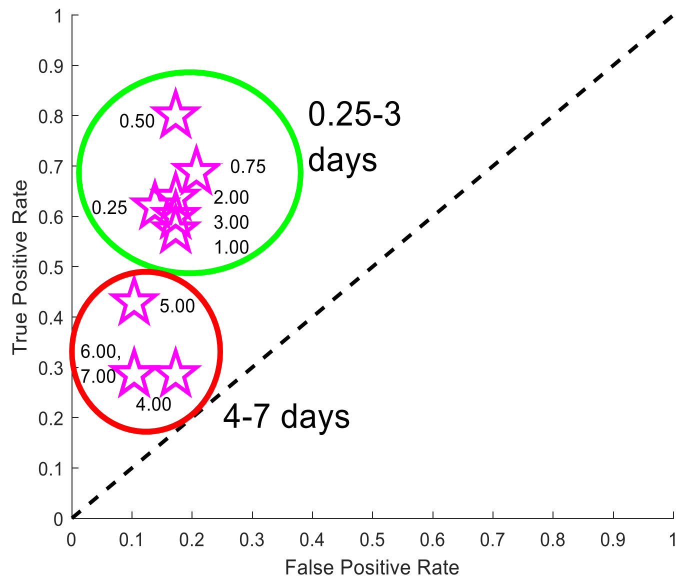
https://service.scedc.caltech.edu/eq-catalogs/date_mag_loc.php




- Database: Southern California earthquake catalog [1981, 2020]
- Clusters identified by a windowing algorithm (Kagan 2002 for space and Gardner and Knopoff 1974 for time extensions)

Dots: clusters mainshock location (50 clusters); red: type A clusters; blue: type B clusters.
Region of analysis: green region

# Performance estimation in Southern California

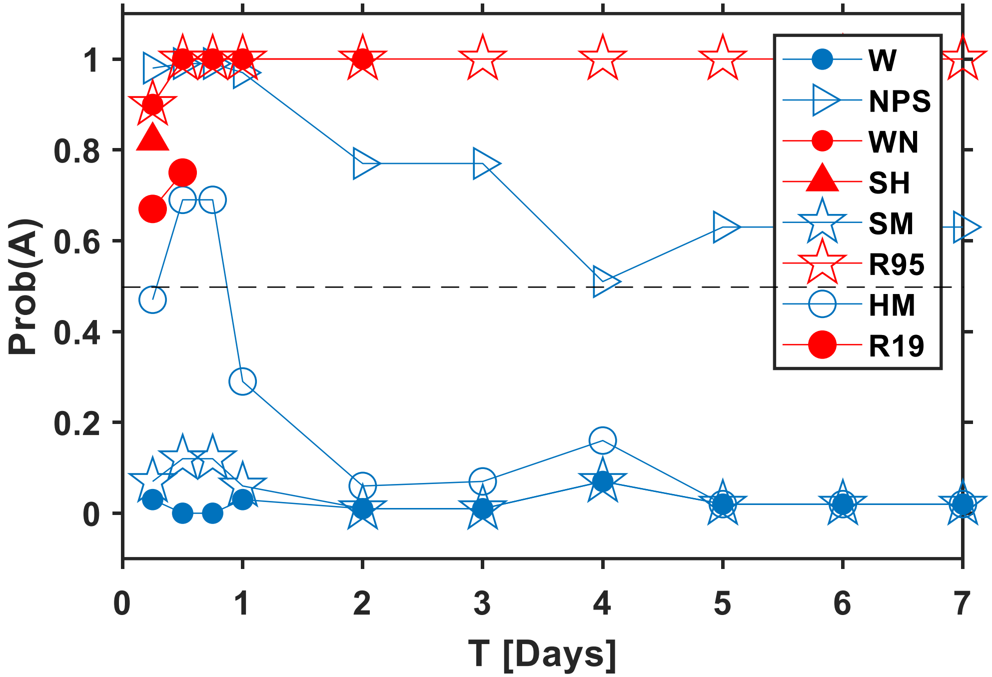



NESTORE performances

Estimate using the Leave One Out (LOO) method:

- For all time periods, the NESTORE is always in the upper left triangle, corresponding to reliable performances.
- The TPR decreases for T>3 days.

# Case studies: eight Mm≥5.8 clusters



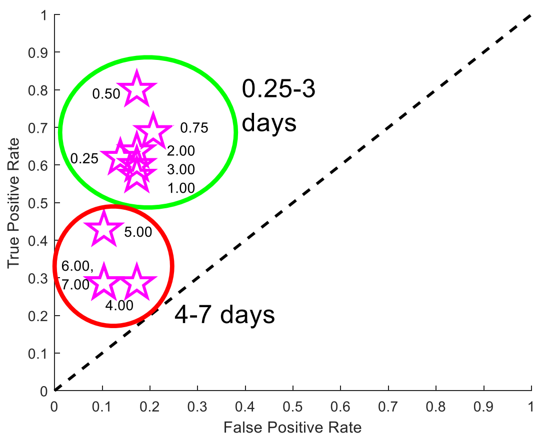

Red symbols: A clusters Blue ones: B clusters

| Cluster | Abbr. |
|---|---|
| Westmorland | W |
| North Palm Springs | NPS |
| Whittier Narrows | WN |
| Superstition Hill | SH |
| Sierra Madre | SM |
| Ridgecrest | R95 |
| Hector Mine | HM |
| Ridgecrest | R19 |

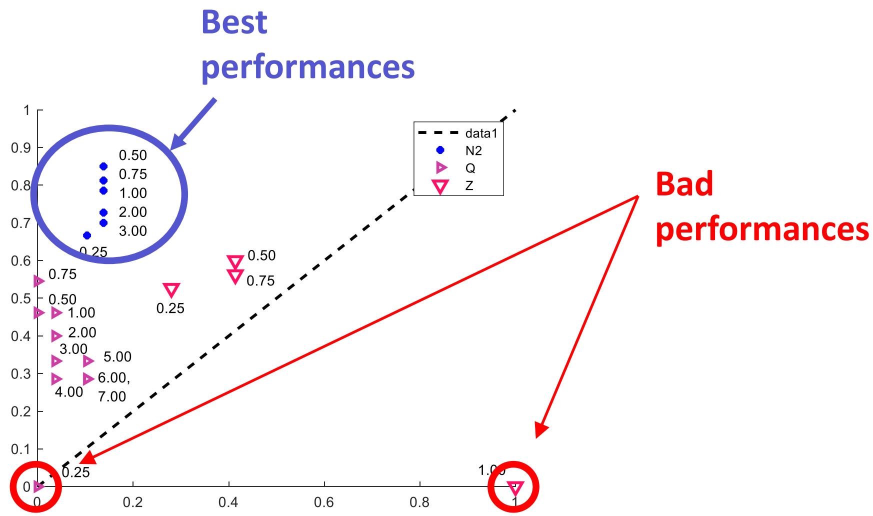
- Most clusters are correctly classified from six hours after the mainshock.
- For one cluster (HM) the classification becomes correct after one day, one cluster (NPS) is an outlier

# What can be enhanced?

- The TPR decreases for T>3 days.
- The feature performances may be poor for some $T_i$
- We compared the results obtained by different training set in LOO analysis
- Bad performances are related to an overfitting for some training sets
- It is related with the small databases and to the unbalanced classes (for large Ti)



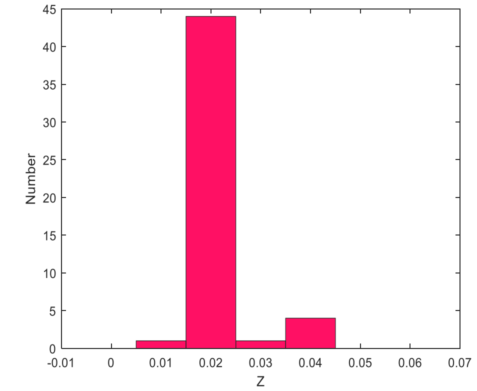

NESTORE performances

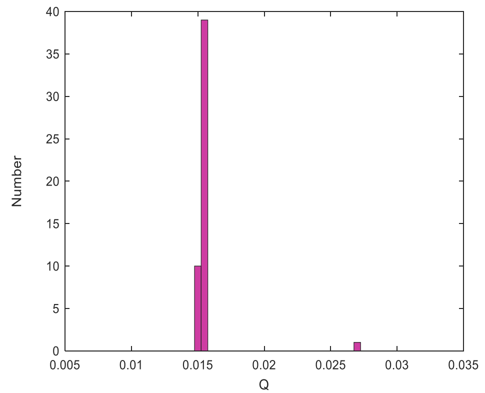



Some features' performances

# A robust NESTORE (rNESTORE)

| T [days] | S | Z | SLcum | QLcum | SLcum2 | QLcum2 | Q | Vm | N2 |
|---|---|---|---|---|---|---|---|---|---|
| 0.25 | 0 | 88 | 0 | 0 | 0 | 0 | 62 | 78 | 100 |
| 0.50 | 0 | 100 | 0 | 100 | 0 | 0 | 76 | 100 | 100 |
| 0.75 | 0 | 100 | 0 | 67 | 0 | 0 | 78 | 100 | 100 |
| 1 | 0 | 28 | 14 | 0 | 14 | 0 | 93 | 100 | 100 |
| 2 | 0 | 3 | 13 | 0 | 100 | 0 | 93 | 100 | 100 |
| 3 | 0 | 0 | 13 | 0 | 100 | 0 | 92 | 97 | 100 |
| 4 | 0 | 0 | 14 | 0 | 72 | 0 | 94 | 0 | 11 |
| 5 | 100 | 0 | 11 | 0 | 8 | 0 | 97 | 0 | 0 |
| 6 | 100 | 0 | 11 | 0 | 8 | 0 | 100 | 0 | 0 |
| 7 | 100 | 0 | 11 | 0 | 8 | 0 | 100 | 0 | 0 |

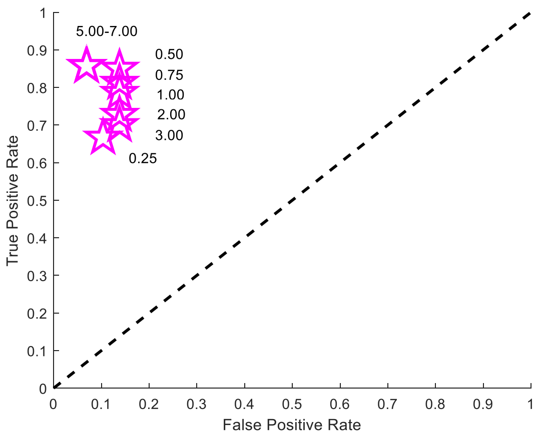


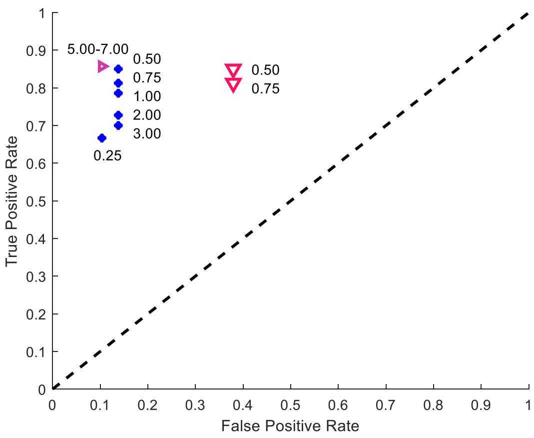

- We selected only time intervals for which at least 95% of the training sets NESTORE finds a reliable threshold
- We choose as threshold the mean of the thresholds of different training sets

# Coherence test for Southern California for rNESTORE

- The training set and the test set are the same; this is an internal coherence test not a evaluation of the classifier
- However, we can see a great improvement in Q and Z feature performances, also because they are no more considered reliable for the Ti for which they failed
- NESTORE performances at 5-7 days are improved due to the more stable features



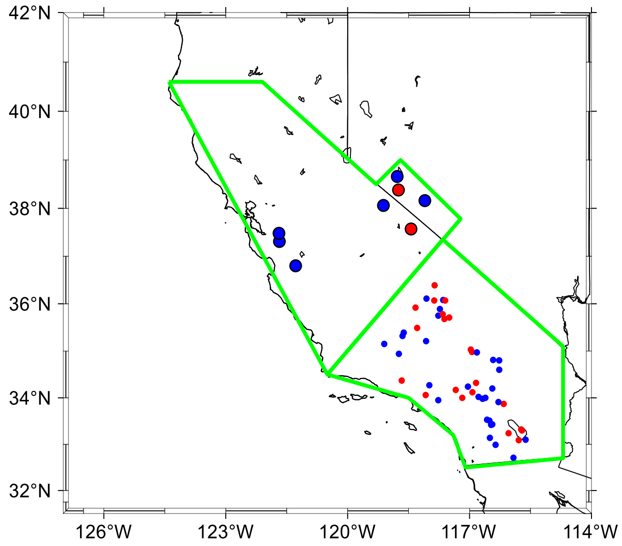

NESTORE performances




Some features' performances

# Independent test set: Northern California Nevada border

https://earthquake.usgs.gov/earthquakes/search/



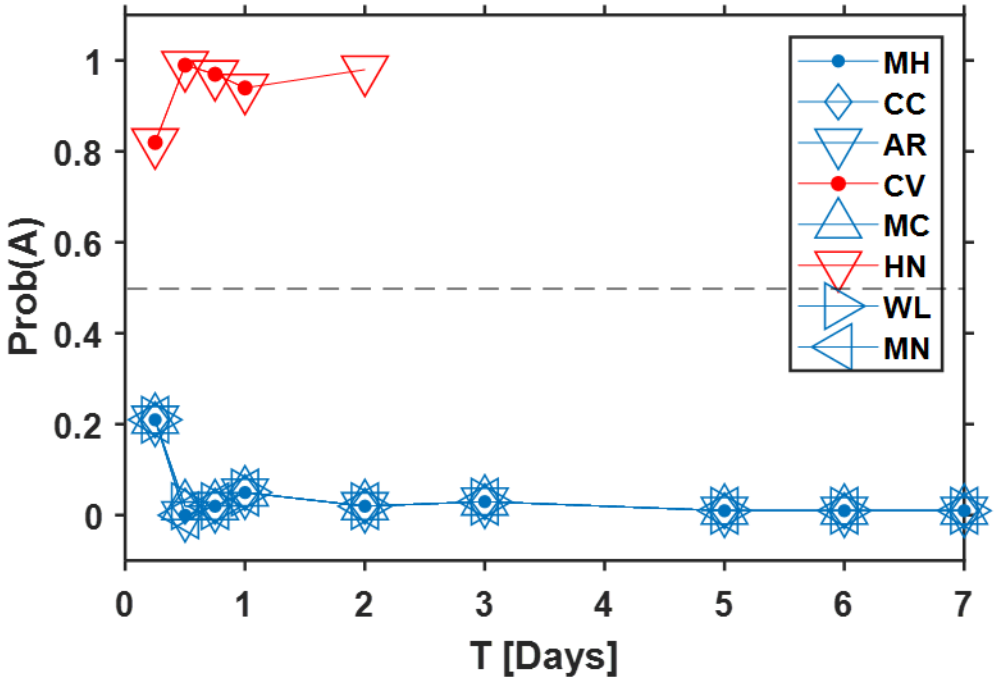

- Database: Comprehensive Earthquake Catalog (ComCat) [1981, 2021]
- Clusters identified by a windowing algorithm (Kagan 2002 for space and Gardner and Knopoff 1974 for time extensions)

Circles: test set clusters mainshock location (8 clusters); red: type A clusters; blue: type B clusters.

Dots: training set clusters mainshock location (50 clusters); red: type A clusters; blue: type B clusters.

Region of analysis: upper green region

# Test of rNESTORE on an independent test set




Red symbols: A clusters Blue ones: B clusters

| Cluster | Abbr. |
|---|---|
| Morgan Hill | MH |
| Central California | CC |
| Alum Rock | AR |
| Chalfant Valley | CV |
| Mono County | MC |
| Hawthorne Nevada | HN |
| Walker Lake | WL |
| Mina Navada | MN |

- All the cluster's classifications are correct and the classifications in time are coincident or almost coincident
- The poorer performances for T=6 hours are because only one feature (N2) is used

# Conclusions

- One of the main problems in statistical seismology is the small number of data available
- In Southern California, e.g., we selected 50 clusters in 40 years with the desired completeness magnitude
- We present a pattern recognition method NESTORE, optimized for small datasets
- We applied the method to Southern California seismicity where it shows a false alarm rate < 0.2 and a hit rate up to 0.80
- We proposed a robust version of the algorithm, **rNESTORE**, based on the comparison of the trainings for different datasets
- rNESTORE correctly classifies all the clusters of a small independent database in Northern California/Nevada

# For further details


- Gentili S. and Di Giovambattista R. (2017). Pattern recognition approach to the subsequent event of damaging earthquakes in Italy. Physics of the Earth and Planetary Interiors, 266, 1-17. https://www.researchgate.net/publication/314126252_Pattern_recognition_approach_to_the_subsequent_event_of_damaging_earthquakes_in_Italy
- Gentili S. and Di Giovambattista R. (2020). Forecasting strong aftershocks in earthquake clusters from northeastern Italy and western Slovenia. Physics of the Earth and Planetary Interiors, 303, 106483. https://arxiv.org/ftp/arxiv/papers/2005/2005.02779.pdf
- Gentili S. and Di Giovambattista R. (2020). How strong will be the following earthquake? DOI: 10.5194/egusphere-egu2020-8184 EGU General Assembly 2020 Online | 4–8 May 2020. https://meetingorganizer.copernicus.org/EGU2020/EGU2020-8184.html