Transfer Learning to Build a Scalable Model for the Declustering of Earthquake Catalogs

Florent Aden-Antoniow[1], William Benjamin Frank[1], and Leonard Seydoux[2]

[1]Massachusetts Institute of Technology
[2]Université Grenoble Alpes

November 21, 2022

Abstract

The rate of background seismicity, or the earthquakes not directly triggered by another earthquake, in active seismic regions is indicative of the stressing rate of fault systems. However, aftershock sequences often dominate the seismicity rate, masking this background seismicity. The identification of aftershocks in earthquake catalogs, also known as declustering, is thus an important problem in seismology. Most solutions involve spatio-temporal distances between successive events, such as the Nearest-Neighbor-Distance algorithm widely used in various contexts. This algorithm assumes that the space-time metric follows a bi-modal distribution with one peak related to the background seismicity and another peak representing the aftershocks. Constraining these two distributions is key to accurately identify the aftershocks from the background events. Recent work often uses a linear-splitting based on nearest-neighbor distance threshold, ignoring the overlap between the two populations and resulting in a mis-identification of background earthquakes and aftershock sequences. We revisit this problem here with both machine-learning classification and clustering algorithms. After testing several popular algorithms, we show that a random forest trained with various synthetic catalogs generated by an Epidemic Type Aftershock Sequence model outperforms approaches such as K-means, Gaussian-mixture models, and Support Vector Classifications. We evaluate different data features and discuss their importance in classifying aftershocks.

We then apply our model to two different actual earthquake catalogs, the relocated Southern California Earthquake Center catalog and the GeoNet catalog of New Zealand. Our model capably adapts to these two different tectonic contexts, highlighting the differences in aftershock productivity between crustal and intermediate depth seismicity.

Transfer Learning to Build a Scalable Model for the Declustering of Earthquake Catalogs

F. Aden-Antoniów[1], W. B. Frank[1], L. Seydoux[2]

[1]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Équipe Ondes et Structures, ISTerre, Université Grenoble-Alpes, UMR 5375, Gières, France

Corresponding author: Florent Aden-Antoniow, `faden@mit.edu`

Abstract

The rate of background seismicity, or the earthquakes not directly triggered by another earthquake, in active seismic regions is indicative of the stressing rate of fault systems. However, aftershock sequences often dominate the seismicity rate, masking this background seismicity. The identification of aftershocks in earthquake catalogs, also known as declustering, is thus an important problem in seismology. Most solutions involve spatio-temporal distances between successive events, such as the Nearest-Neighbor-Distance algorithm widely used in various contexts. This algorithm assumes that the space-time metric follows a bimodal distribution with one peak related to the background seismicity and another peak representing the aftershocks. Constraining these two distributions is key to accurately identify the aftershocks from the background events. Recent work often uses a linear-splitting based on nearest-neighbor distance threshold, ignoring the overlap between the two populations and resulting in a mis-identification of background earthquakes and aftershock sequences. We revisit this problem here with both machine-learning classification and clustering algorithms. After testing several popular algorithms, we show that a random forest trained with various synthetic catalogs generated by an Epidemic Type Aftershock Sequence model outperforms approaches such as K-means, Gaussian-mixture models, and Support Vector Classifications. We evaluate different data features and discuss their importance in classifying aftershocks. We then apply our model to two different actual earthquake catalogs, the relocated Southern California Earthquake Center catalog and the GeoNet catalog of New Zealand. Our model capably adapts to these two different tectonic contexts, highlighting the differences in aftershock productivity between crustal and intermediate depth seismicity.

Plain Language Summary

The seismic catalog of earthquakes that have occurred in a given fault zone is a window into the tectonic processes that occur at depth. These earthquakes rupture faults when the fault can no longer support the stress built-up by tectonic motion. When an earthquake occurs spontaneously from tectonic stresses, this mainshock will trigger aftershocks, which can themselves trigger even more aftershocks. Aftershock sequences often dominate catalogs due to the sudden increase in the number of earthquakes. Distinguishing between mainshocks and aftershocks requires an understanding of the connection between earthquakes. An accurate classification of mainshocks and aftershocks in a catalog allow notably a more precise investigation of the evolution of accumulated stress on a fault and is often a necessary tool to estimate the seismic hazard of a region. In this work, we develop a machine-learning algorithm to achieve such classification. We tested our model on large earthquake catalogs of Southern California and New Zealand to demonstrate the effectiveness of our approach. We show that our approach is generalizeable to any region of the world, independent of the style of seismicity or period of time.

1 Introduction

That earthquakes form clusters in time and space, regardless of the tectonic context, is a fundamental characteristic of earthquakes. The simplest classification of earthquakes breaks down the total seismicity rate into two categories: background events that are generated by long-term, large-scale tectonic forcings and aftershocks that are directly triggered by a background event or another aftershock. Isolating the background seismicity is crucial for a wide range of studies: from the monitoring of transient loading along faults (Marsan, Prono, & Helmstetter, 2013; Marsan, Reverso, et al., 2013; Reverso et al., 2015), fluid injections (Hainzl & Ogata, 2005; Bachmann et al., 2011; Kothari et al., 2020) for hazard assessments where declustering is essential to remove the spatial bias introduced by clustered seismic activity in non-declustered catalogs (Marzocchi & Taroni, 2014; Azak et al., 2018; Galina et al., 2019; Taroni & Akinci, 2021). Aftershocks can be modeled with empirical laws, such as Omori's law or the productivity law (Omori, 1894; Utsu, 1961), that predict the rate and magnitude of aftershocks following some background event. The classification of an earthquake as a mainshock or an aftershock then relies on the identification of the relationship between a given event and the events that preceded it. If an earthquake is likely to be triggered by a preceding earthquake, it can then be confidently labeled as an aftershock. Several algorithms have been proposed to tackle this classification problem, also known as declustering. Early declustering algorithms were mostly based on space-time windows (Gardner & Knopoff, 1974; Reasenberg, 1985). More recently, declustering approaches have made use of the Epidemic-Type Aftershock Sequence (ETAS) model (Ogata, 1988, 1998; Zhuang et al., 2002) or space-time-magnitude metric describing the link between event-pair (Baiesi & Paczuski, 2004; Zaliapin et al., 2008); we refer to (Van Stiphout et al., 2012) for an exhaustive list of declustering algorithms. Each method possess its own parameters but most of them rely on the estimation of the "reach" of each earthquakes to others.

Teng and Baker (2019) have recently compared several declustering methods: the space-time window based algorithms proposed by Gardner and Knopoff (1974) and by Reasenberg (1985), the Nearest-Neighbor-Distance (NND) Algorithm (Baiesi & Paczuski, 2004; Zaliapin et al., 2008) relying on a space-time-magnitude metric and stochastic declustering (Zhuang et al., 2002). Their study suggests that both Reasenberg and NND declustering methods are suitable for the declustering of the seismicity in particular to estimate the background seismic rate useful for seismic hazard analysis. The NND algorithm presents non-negligible advantages as it does not use tuning parameters other than characteristic features of the observed seismicity, such as the b value or the fractal dimension.

However, as pointed out by Bayliss et al. (2019), there is room to improve NND-based declustering. The frequency distribution of the NND metric is often observed as a bi-modal distribution characterizing the background seismicity and the aftershocks. A threshold is then used to split the distribution and classify the earthquakes into two categories, mainshocks or aftershocks, neglecting the overlap of both distributions. This implies that the declustered background seismicity will contain clustered seismicity and vice-versa. Bayliss et al. (2019) developed a probabilistic clustering framework using a Markov Chain Monte Carlo mixture modelling approach allowing the overlap of two Weibull function with the aim to quantify the uncertainties in event-pair linkage. Zaliapin and Ben-Zion (2020) recently introduced a modified version of the Nearest-Neighbor-Distance algorithm by introducing 3 additional steps. Their algorithm discriminates background events and aftershocks following a random thinning approach which aims to remove events according to a space-dependent threshold, which is estimated using randomized-reshuffled catalogs. In this work, we have decided to revisit the problem of the declustering of earthquake catalogs with a machine-learning based take on the NND algorithm but with the aim to keep the simplicity of the original algorithm. The declustering can be seen either as a clustering or a classification problem; both utilize data features to describe an earthquake. Machine learning techniques are now commonly used in seismology (Kong et al., 2019), especially for the detection and/or classification of seismic waveforms with neural networks (Perol et al., 2018; Li et al., 2018; Ross et al., 2018; Zhu & Beroza, 2019; Thomas et al., n.d.), driven by either supervised (Rubin et al., 2012; Lara-Cueva et al., 2016) or unsupervised learning algorithms (Seydoux et al., 2020; Shi et al., 2021; Steinmann et al., 2021). Beyond waveform classification, many other applications have benefited from machine learning algorithms, including but not limited to early warning systems (Kong et al., 2016), earthquake relocation (Trugman & Shearer, 2017), predicting fault slip cycles (Rouet-Leduc et al., 2017), and forecasting earthquake magnitudes (Asim et al., 2017; González et al., 2019; Hoque et al., 2020). However, only few studies have used machine learning to classify earthquakes within catalogs (Picozzi & Iaccarino, 2021), and they have not addressed the declustering problem. Here we implement and compare different machine learning algorithms apply to the the declustering of earthquake catalogs. Building on the nearest-neighbor algorithm, we perform a thoughtful comparison between clustering and classification approaches apply to synthetic catalogs with the aim to identify the most accurate algorithm. We apply our model to 2 real catalogs from Southern California and New Zealand and discuss the potential of our declustering model.

2 Method

2.1 Nearest-Neighbor Distance algorithm

The NND represents a good compromise between computational efficiency and stability, relying on a generalizeable metric of earthquake catalogs. The NND does not depend directly on the location and magnitude of earthquakes, but is rather based on the computation of a space-time metric taking into account both the Gutenberg-Richter law (Gutenberg & Richter, 1955) and the Omori-Utsu law (Omori, 1894; Utsu et al., 1995). It consists of the estimation of the metric η_{ij} between the jth event and a preceding event i. We call the nearest-neighbor event of j, the event i that minimizes this distance:

$$\eta_{ij} = t_{ij} \times (r_{ij})^{d_f} \times 10^{-bm_i} \tag{1}$$

where $t_{ij} = t_j - t_i$ is the inter-event time, $r_{ij} = |r_i - r_j|$ the inter-event physical distance (epicenter to epicenter; we ignore depth here), m_i is the magnitude of the event i and d_f is the fractal dimension. Zaliapin et al. (2008) introduced a re-scaled time-difference $T_{i,j}$ and spatial distance $R_{i,j}$ for discriminating clustered and non-clustered events in a 2D visualisation, thus allowing to account for both time and space distributions:

$$\begin{aligned} \eta_{ij} &= T_{ij} \times R_{ij} \\ T_{ij} &= t_{ij} \times 10^{-\frac{1}{2}bm_i} \\ R_{ij} &= (r_{ij})^{df} \times 10^{-\frac{1}{2}bm_i} \end{aligned} \tag{2}$$

For simplification, we refer to T_{ij} as the rescaled time and R_{ij} as the rescaled distance.

We measured the NND η for the relocated seismicity of Southern California, between 1981 and 2019 (Hauksson et al., 2012), and the GeoNet catalog of New Zealand, between 2010 and 2020 (Figure 1a and b). We identify two populations: aftershocks have systematically small rescaled times $T_{i,j}$ and distances $R_{i,j}$, while the mainshocks exhibit a wider distribution of rescaled time and distance (Figure 1c and d). The simplest way to discriminate between the two populations is by drawing a line at the local minimum of the η distribution (Figure 1e and f); this vertical line to split the NND distribution becomes a diagonal in the 2D representation of rescaled time and distance. Knowing that the background distribution follows a Weibull distribution (Zaliapin et al., 2008), by cutting the η distribution aggressively we neglect any overlap and thus a portion of mainshocks are considered as mainshocks and vice versa for aftershocks. To better constrain this problem, we consider the declustering of earthquake catalog as both a clustering problem and a classification problem to compare both approaches.

2.2 A Clustering or a Classification problem

To decluster an earthquake catalog is to distinguish the mainshocks from the aftershocks. In a clustering (unsupervised learning) approach, we would expect both populations to separate without any *a priori* on the data set or on any sort of model that would have generated the earthquakes in the catalog. This idealized approach ignores the difficulty of evaluating the quality of the prediction because there is no explicit loss function that quantifies the quality of a given clustering. In a classification (supervised learning) approach, one can evaluate how well a model performs during training and testing phases by comparing the results with the synthetic labels. In our particular case of earthquake declustering, training such a model can be achieved by using a Epidemic-Type Aftershock Sequence (ETAS) model to generate synthetic catalogs in an exercise of transfer learning.

In this work, we have compared the results of four well-known machine learning models. The unsupervised learning algorithms we tested here are the K-means algorithm (linear and deterministic) and a Gaussian-Mixture model (non-linear and probabilistic). The supervised learning models we tested rely on either a Support Vector Machine Classifier (linear and deterministic) or a Random Forest classifier (non-linear and probabilistic). We describe each of the four models below.

The K-means algorithm will separate the samples into k groups, here $k = 2$ to account for mainshocks and aftershocks, by minimizing the cluster variances, also called inertia, $\sum_{i=0}^{N_k}(||x_i - \mu_k||^2)$.

A Gaussian Mixture model (GMM) is a probabilistic model and considers the sample distribution as a mixture of a finite number K of Gaussian distributions with unknown parameters; each Gaussian is defined by a mean μ_k and a variance σ_k. GMM is initialized by applying the k-means algorithm in order to assign a label to each data point and a first set of model parameters. Then a two-step technique called Expectation Maximization is performed to estimate the mixture model parameters. First, the membership weights (or probabilities) ϕ_{ik} for each sample with the given set of model parameters are computed. For each data point, the membership weight is defined as $\sum_{k=1}^{K} \phi_{ik} = 1$. In a second step, it infers the new set of model parameters for the given membership weights by maximizing the log-likelihood function. This is an iterative process and eventually converges to a final set of parameters.

A Support Vector Machine Classifier (SVC) is an unsupervised learning model that can be used to solve classification or regression problems. We decided to use a linear kernel for the SVC to have a supervised, linear and deterministic model to compare with K-means. With SVC, a sample is considered as a k-dimensional vector, with k the number of features. The linear SVC can be trained to separate the data set with $k-1$ dimensional hyperplane(s). If the solution of this problem is not unique, the best hyperplane is the one that create the largest separation or margin between the two populations; the "best" hyperplane maximizes the distance from it to the nearest sample(s) on each side, also called the support vectors.

A Random Forest (RF) is an ensemble classifier represented by a forest of decision trees. One main advantage of a RF is to prevent over-fitting of the training data set. The ensemble of uncorrelated trees is generated following a method of Bootstrap Aggregating (also called bagging): each tree is trained with a randomly selected (with replacement) sub-set of the data. Until now, RFs have been most often used to classify events such as earthquakes, landslides, geyser or period of volcanic activation based on time-frequency features (Rubin et al., 2012; Hibert et al., 2017; Maggi et al., 2017; Provost et al., 2017; Yuan et al., 2019; Dempsey et al., 2020); an RF model has not yet been applied to seismic catalogs. During the training of a tree, at each node, the best split is found either from all input features or more generally from a random subset of it, thus contributing to the randomness which reduces potential overfitting. From this set of features, the idea is to find the feature and the associated threshold that allows the best splitting of the remaining data-set into two groups, based on the provided data labels. The selected feature and the given threshold minimize the class impurity of the two populations according to the Gini impurity index computed from the known class of the samples. The Gini impurity at a split is the probability that a randomly chosen sample would be incorrectly labeled. It is expressed for the node n:

$$I_G(n) = 1 - \sum_{i=1}^{K} p_i^2 \tag{3}$$

This is done recursively until one sample remains in the leaves or when the trees reach a maximum depth. When all trees are built, the classification of a sample is done by feeding it to the ensemble of trees and comparing its features to the selected feature and its respective threshold at each node. At the end of each tree, in the leaf where the sample falls, a prediction is made. The class prediction is then obtained by a vote of all trees. The final probability to belong to each class is derived by the percentage of trees in the forest that voted for a given classification.

We also considered using Gradient Boosted Trees as a supervised learning model but knowing that the approach relies on the fit of the prediction residuals in a series of weak learners, i.e. shallow decision trees, we believe it would be difficult to use for the declustering of seismic catalogs without over-fitting the training data set. We also considered a neural network, but we judged the training and the estimation of the model's architecture parameters too demanding computationally and likely not adequate for this particular problem with a potential small number of features compared to a large number of samples.

2.3 Synthetic catalogs for training

To train and later test the supervised learning models, the Support Vector Machine Classifier and the Random Forest, we need to build a labelled data set, i.e. synthetic catalogs where we know which earthquakes is a mainshock or an aftershock. To do so, we have followed the model of Epidemic-Type Aftershock Sequence (ETAS) (Ogata, 1988; Marsan, Reverso, et al., 2013; Marsan et al., 2017). The spatio-temporal distribution of earthquakes $\lambda(x, y, t)$ is defined as the sum of the background seismicity $\mu(x, y, t)$ and the aftershocks $\nu(x, y, t)$:

$$\lambda(x, y, t) = \mu(x, y, t) + \nu(x, y, t) \tag{4}$$

The mainshock distribution $\mu(x, y, t)$ is characterized by a Poisson point-process: the events are independent from each other and the inter-event time follows the Poisson probability density function as:

$$f(\tau) = \frac{1}{T_0} \exp\left(-\frac{\tau}{T_0}\right) \quad \text{for } \tau \geq 0 \tag{5}$$

We have drawn coordinates from a 2D probability density function (PDF) to give a location for each mainshocks. This PDF can be uniform for a randomly distributed background seismicity or can be generated from an existing seismicity.= by

Following Marsan, Reverso, et al. (2013), the aftershock distribution $\nu(x, y, t)$ can be expressed as a spatio-temporal distribution, which results from the product of the Omori-Utsu law, a productivity law and a power spatial density:

$$\nu_(x, y, t) = \sum_{i}^{N_{after}} \frac{K 10^{a(m_i - m_c)}}{(t - t_i + c)^p} \times \frac{(\gamma - 1) L_i^{\gamma - 1}}{2\pi \left((x - x_i)^2 + (y - y_i)^2 + L_i^2\right)^{(\gamma-1)/2}} \tag{6}$$

K and a are constant and part of the aftershock productivity of the i-th event which follow a Poisson law with the average, with K and a constants (Gu et al., 2013). t_i is the occurrence time of the mainshock and L_i is the characteristic length in kilometers (Utsu & Seki, 1955) such as $L_i = L_0 \times 10^{(m_i - m_c)/2}$ with $L_0 = 0.1$km (Marsan, Reverso, et al., 2013; Reverso et al., 2016).

2.4 Training Data Set, Feature Space and Model Optimization

We create the mainshock seismicity in three steps:

(1) To account for variability in the inter-event time, we have randomly selected an averaged number of events per day for each catalog ranging from 1 to 3. The total average number

of mainshocks $< N >$ is obtained by multiplying the average number of events per day by the duration of the catalog desired. Here the duration is set to 8000 days ($\sim$ 22 years) to account for long-term seismic interactions, which we will discuss later on. The actual number of mainshocks N is drawn from a Poisson-law with a mean equal to $< N >$. The averaged inter-event time T_0, which is simply N divided by the actual duration of the catalog, is injected in equation 5 to draw N inter-event times. The origin times of the N mainshocks are simply obtained by computing the cumulative sum of the inter-event times.

(2) To account for variability in the inter-event distance, we have used a 2D-PDF obtained from the declustered interface seismicity for Northern-Chile from Aden-Antoniów et al. (2020). From this catalog, we built a earthquake density map, smoothed it with a Gaussian filter and normalized it by the total event count. Using a 2D PDF that is not uniform and generated from a subduction zone seismicity is allowing us to simulate a broader distribution of distance between the events.

(3) To obtain the magnitude of each mainshock, we have randomly selected a b-value, uniformly distributed between 0.8 and 1, and did the same for the completeness magnitude, uniformly distributed from 2 to 3. The magnitudes are then drawn from the probability density function following the Gutenberg-Richter law (Gutenberg & Richter, 1955) with a maximum magnitude set to 7.5. We selected this maximum magnitude on the basis that for larger magnitude, some catalogs with low b value presented aftershock sequences lasting for decades continuously triggering M6+ events resulting in catalogs constituted of a disproportionate number of aftershocks in comparison of mainshocks.

We have generated the aftershocks and the aftershocks of the aftershocks and so on following the method presented in section 2.3. To generate each catalog, we have drawn a set of parameters of equation 6 such as the p and c, corresponding to the parameters of the Omori-Utsu law (Omori, 1894; Utsu, 1957; Utsu et al., 1995), the parameter γ for the power spatial density function, K and a for the productivity law. The magnitude of each aftershocks is drawn from the same PDF as the mainshocks of the a given catalog. An example of these synthetic catalogs is shown Figure S1 and the range of each parameters of the ETAS model are shown Table 1. Finally, we have labeled each events of the catalogs with 0 if it is a mainshock or 1 if it is an aftershock. We have generated 200 synthetic catalogs representing a total of 5,461,475 earthquakes, 68% of which are aftershocks. This corresponds to an average of 27,307 events per catalog with a minimum of 9,382 and a maximum of 386,054.

For each event j in each synthetic catalog, we conducted the search of its nearest-neighbor i which is not necessarily it's mainshock if j is an aftershock. We estimated the fractal dimension d_f necessary to compute the nearest-neighbor-distance metric for each catalog following the Minkowski-Bouligand approach, also called the box counting method. For different size boxes that divide the area covered by the synthetic catalogs, we counted the number of boxes actually containing earthquakes. The relation between the size of the boxes and the number of non-empty boxes can be described by a power law with the fractal dimension d_f as exponent. The d_f of the synthetic catalogs are on average 1.7 with minimal variations.

For both supervised and unsupervised approaches, we need to identify the relevant features that are useful to distinguish background events from aftershocks. To keep our model generalizable, the features necessary to classify an earthquake should not rely on absolute locations or origin times. The NND framework (equation 3) provides several relative metrics well suited to use as data features. To predict the label of event j, we describe the link with its nearest neighbor i with five features: (1) the rescaled time T_{ij} and (2) the rescaled distance R_{ij} both described in equation 3; (3) the difference in magnitude $\Delta m_{ij} = m_i - m_j$ between the event j and its nearest-neighbor i, which we would expect to be high if j is an aftershock of a larger event; (4) N_p the number of siblings or events that share the same nearest neighbor as event j, which if relatively high would suggest that event j is one of many aftershocks following a background event; and (5) N_c the number of offspring of the event j, which if high would suggest that event j has generated many aftershocks. Associated together, these features provide information on the linkages between earthquakes so that we may distinguish between mainshocks and aftershocks.

Of the 200 synthetic catalogs we generated (see section 2.3), we used 100 catalogs to train the supervised learning models; the remaining 100 serve to test and compare all the different models on a known data set. Besides training a supervised learning model, one should also estimate the best hyper-parameters of the model, i.e. the unique parameters defining the architecture of the model. To do so we have conducted a stratified k-Fold cross-validation (Mosteller & Tukey, 1968). This consists of first splitting the training data set into k groups or folds, in a *stratified* fashion. In this context, stratified means that in each fold, a similar percentage of samples of each label as the original data set is preserved, this is particularly useful and recommended when the population to classify are not equally represented in the data set. A new model is then trained k times and at each time a different group is left out to be used later for validation purposes. The whole procedure is repeated for each subset of potential hyperparameters, allowing a better estimation of the model performance and a more robust selection of the hyperparameters. The best model hyperparameters are chosen by computing the average validation accuracy of the k folds. One chose the number of folds with the goal that each fold should remain statistically representative of the whole training data set; usually 5 or 10 folds are sufficient.

There are several hyper-parameters that control the architecture of the Random Forest, including: (1) the number of trees; (2) the number of features that are randomly selected at each split to minimize the Gini impurity, from 1 to total the number of features; (3) the minimum number of training samples in a leaf. We do not consider here other parameters such as the maximum depth of a tree or the minimum number of sample to allow a split because they would be in competition with parameter (3). Our most accurate RF model, with hyperparameters chosen after a Stratified 5-Fold Cross-Validation (approximately 620,000 samples per fold), is composed of 100 trees, 2 features considered at each split and a minimum number of sample in each leaf of 1.

The linear SVC is mainly controlled by the soft margin parameter C. It is a regularization parameter: a larger C will result in smaller acceptance margins, where the model will allow fewer aftershocks to be on the mainshocks side of the hyper-plane; a lower C will allow larger acceptance margins which in return will generate a simpler and less precise hyper-plane, implying a lower training accuracy. Similar to the RF hyperparameters estimation, we performed a 5-Fold cross-validation to look for the best parameter C with the same training and testing data set and found that the value $C = 0.1$ gives the best validation accuracy.

3 Results

3.1 Model Comparison

Fernández-Delgado et al. (2014) found that the RF is the best classification model among many other available algorithms, including Support Vector Machine or Logistic Regression, but we prefer to conduct our own comparison tailored to the data set at hand. A rather implicit hyper-parameter of this study is the number of feature to use. To investigate its impact on the accuracy of the different models, we have predicted the label of the earthquakes in 100 test catalogs with a different number of features n_f for each model. If $n_f = 1$, only η_{ij}, the nearest-neighbor distance metric is used. For $n_f = 2$, R_{ij} and T_{ij} are used. Then if $n_f = 3$, the magnitude difference d_m is added to the list of features, then N_p for $n_f = 4$ and finally N_c for $n_f = 5$. We have trained both supervised learning models, the Random Forest and the Support Vector Machine Classifier, beforehand on the training data set with the corresponding number of features and each time their hyperparameters have been determined using a 5-Fold Cross-Validation.

We also compared the four models to the classical way of declustering with the NND, which is to use a threshold η_0 to split the nearest-neighbor distance distribution η_{ij} in two. We estimated this threshold by rounding the membership probabilities given by the GMM only using η_{ij} distribution. η_0 is often defined as the local minimum or in its vicinity between the two lobes of the η_{ij} distribution if the two populations are easily discernible (e.g. Zaliapin & Ben-Zion, 2013; Gu et al., 2013; Davidsen et al., 2015; Shebalin & Narteau, 2017; Peresan & Gentili, 2018; Aden-Antoniów et al., 2020). To account for the overlap of the two populations using the GMM, we drew a random number between 0 and 1 for each event to test and if the probability to be an aftershocks is greater than this random number, the event is labelled as aftershock; if not we label it as a mainshock.

An important consideration is most machine learning models benefit from a particular pre-processing of the data. For most algorithms dealing with several features, such as the K-means, GMM or the SVC, the different features need to be scaled or standardized, for example by removing the mean and dividing by the standard deviation. Because most models treat the features as an ensemble, they tend to weigh higher numerically large features, regardless of the unit of the features, which is clearly problematic. Accordingly, we have considered the logarithm of η_{ij}, R_{ij} and T_{ij} because they span over a large range; for example η_{ij} distribution is comprised between 10^{-15} and 10^3 (Figure 1c and d); this is sometimes called a kernel trick. We emphasize though that we did not have to do this for the RF model. Because the RF randomly selects a single feature at each split it spares us any extra step of scaling because the features are not directly compared to each other. The comparison of the different models according to the number of features n_f on the testing catalogs is shown in Figure 2a. For each number of features n_f, each testing catalog has been declustered separately and we estimated the accuracy of the model's prediction. We can then look at the testing accuracy confidence interval (32%-68%) for each model with respect to the median (50%) and the mean (Figure 2a).

The classical approach consisting of splitting the η_{ij} into two groups with the threshold η_0 performs well, with an average testing accuracy of 88%, in comparison to the unsupervised learning models, except when 2 features are considered where the GMM performs slightly better. We note that this is not too surprising as the large majority of events are easily classified, such as aftershocks with small values of η_{ij}. The "tricky" part that is key is to distinguish between aftershocks and mainshocks at the border between the two populations. This is where we observe that the supervised learning models have an advantage over the other approaches. Introducing this non-linear transformation of the nearest-neighbor distance metrics clearly benefits the Linear Support Vector Machine Classifier, as it is the best model when using one or two features with a maximum testing accuracy greater than 90%. The additional features d_m, N_p and N_c seem to only benefit the RF model which reaches a maximum testing accuracy around 92% using all five features. We look into the feature importance in the RF model as it is possible to measure which feature as been used the most to do a split. We can have access to the uncertainty of the importance by considering each tree or estimator separately. We can see Figure 2b that R_{ij} and T_{ij} constitute more than 70% of the model. Each feature is important in this model as even N_c, which as a very limited importance, is essential to explain more than 95% of the model.

Figure 3 shows the difference between the different models for the declustering of a testing catalog as an example which can give us a lot of information about the differences between the models (See other examples in Figures S2, S3 and S4). We selected the number of features that gave the best testing accuracy for each model (Figure 2a): the best model of GMM uses only T_{ij} and R_{ij} while the KM, the SVC and the RF show better results using all the features. In Figure 3a, we can identify where the misclassified events are located in the 2D rescaled-feature space. Not surprisingly, the classical approach shows a clear boundary between the two populations of earthquakes and the resulting misclassified events distribution correspond directly to the overlap between the two distributions. The simple threshold on η_{ij} still was able to correctly classify almost 90% of the catalog that was away from this boundary. For this catalog, the KM is the least accurate model with around 82% of accuracy; it has clearly missed large chunks of both aftershocks and mainshocks. The GMM, on the contrary, identifies the region of overlap between the aftershock and the mainshock distributions, however the accuracy is similar to that of the classical approach. This means that if the GMM can handle the overlap as expected, it does not improve significantly the quality of the overall prediction of the catalog. Both supervised learning models show an improvement over the classical method and unsupervised learning models with accuracies of around 93% and 98% respectively for the SVC and the RF. Because we used a linear kernel for the SVC, it is not surprising to see we have a clear boundary between the mainshock and aftershock distributions. It is clearly visible in the rescaled-feature space, especially because the additional features don't seem to improve the model prediction capability as seen in Figure 2a. The RF, while greatly improving the accuracy of this catalog, reduces the overlap uncertainties range and does not show a clear boundary between the two populations. It in fact draws a region where it is difficult to access the true nature of an earthquake, whether it is a long-term aftershock not directly "linked" to its mainshock or a mainshock linked to a spatially distant earthquake but with a very small separation in time. If we compare the η_{ij} distributions with the Weibull function of the true mainshock distribution, we can also see where there are differences between the models (Figure 3b). The confusion matrices (Figure 3c) allow us to see that the RF model is as good at labelling mainshocks as it is at labelling aftershocks, while the other models have a tendency of favoring one population over the other. Only based on synthetic catalogs built using an ETAS model, the RF seems to be a promising model to use for a real case application as it is flexible, does not require a scaling of the features and is able to reproduce with high fidelity the mainshock distribution.

3.2 Application to Southern-California and New Zealand of the Random Forest Model

We have shown in the previous section that the Random Forest is our ideal machine learning model for the declustering of seismic catalog, whose hyperparameters have been optimized with a k-Fold Cross-Validation. We have also shown that the accuracy of the model's predictions increases as we include additional data features: the rescaled distance R_{ij}, the rescaled time T_{ij}, the difference in magnitude d_m between the event considered and its nearest neighbor, N_p and N_c. Here, we present the declustering with our RF model of two actual, extensive seismic catalogs shown in Figure 1: the relocated Southern Californian earthquake catalog (Hauksson et al., 2012) and the GeoNet catalog of New Zealand.

For Southern California, we have used the SCEDC catalog from 1981 to 2019 (Hauksson et al., 2012), relocated with GrowClust (Trugman & Shearer, 2017). We estimated the catalog completeness at magnitude 2.3 with a b-value of 1.03 following the maximum curvature method (Wiemer & Wyss, 2000) and maximum-likelihood method (Aki, 1965). The total number of events whose magnitude is greater than 2.3 is 72,911. We estimated the fractal dimension of this catalog to be approximately 1.6 using the box counting method, similar to our synthetic catalogs and to what was used by Corral (2003), Zaliapin and Ben-Zion (2013) and Moradpour et al. (2014). The result of declustering with the Random Forest model is shown Figure 4. As expected from our model testing, the RF model is capable of resolving an overlap of the aftershock and the mainshock distributions (Figure 4a and b). Both the 2D distribution Rij-T_{ij} and the distribution of the NND η (Figure 4a and b) show that the separation of the aftershocks from the background seismicity is similar to what we observed with our synthetic catalogs. We minimized the L2-norm to fit both the aftershocks and mainshocks distributions to two Weibull functions as suggested by Bayliss et al. (2019). When we look at the prediction of the model and how the probability of being an aftershock is distributed, we observe how the model struggled to predict the label of the events. We show in Figure 4b the distribution of the probability to be an aftershock.

We note that it is more difficult for the model to label an event as a mainshock, because the probability distribution is relatively flat from 0 to 50%; the model is more "decisive" in determining an aftershock, given that more aftershocks are clearly labeled as aftershocks (probabilities greater than ∼75%). By looking at the classification of the catalog through time, Figure 4c, we clearly identify the largest aftershock sequences associated with major regional earthquakes (vertical black lines). We can also see how the mainshock seismicity rate is changing through the years, although these variations could be related to changes in network configuration or detection methods. Overall though, this background rate remains remarkably constant. This suggests that our model is able to effectively distinguish between real mainshocks and aftershocks.

We took a closer look at the aftershock sequences of four major earthquakes that occurred in Southern California: the M7.3 Landers earthquake in 1992 (Figure 4e), the M7.1 Hector Mine earthquake of 1999 (Figure 4f), the M7.2 Baja California in 2010 (Figure 4g), and finally the recent 2019 M7.1 Ridgecrest earthquake (Figure 4h). We analyzed the seismicity 180 days prior to and following these earthquakes, separating the seismicity into inside and outside groups, depending on whether each earthquake was further or closer than 350 km from the mainshock epicenter. We observe that we are able to well isolate the aftershock sequence that resembles a characteristic Omori-type sequence, leaving a fairly constant background sequence. We remark that the background or the mainshock seismicity rate slightly decreases in the days following the Landers, the Hector Mine and Ridgecrest earthquakes (Figure 4f and h) only to recover after a week or two. We speculate that this is the case is due to the fact that network analysts focus on the aftershock sequence for several days and weeks following the mainshock, creating this artificial rate decrease in background seismicity. Inside the 350km perimeter, the decrease of the background seismicity rate could also be explained by the fact that there are few mainshocks distinguishable (with the data features we have used) from the ongoing an aftershock sequence.

We also applied our random forest model to the GeoNet New Zealand catalog; we used the seismicity recorded by GeoNet from 2010 to 2020. Similar to Southern California, we estimated the completeness magnitude to be 2.6 with the maximum curvature method (Wiemer & Wyss, 2000), associated with a b value of 0.89. The resulting catalog is composed of 64,200 earthquakes. This catalog comprises a various type of tectonic and volcanic settings originating from New Zealand volcanoes, the Hikurangi subduction zone where the Pacific plate subducts beneath the North Island (Australian plate), and more shallow activity related to crustal intraplate faulting in the South Island along the Alpine fault and the transpressional transition from the Alpine fault to the Hikurangi subduction zone. We estimated the fractal dimension of the catalog at around 1.9 from the box counting method; this is slightly higher than our estimate for the Southern California catalog, but we suggest it is reasonable given the presence of intermediate depth seismicity related to the Hikurangi subduction zone and the larger study region considered. The results of our declustering are shown Figure 5. Comparably to Southern California, the declustering of the GeoNet catalog is in agreement with our expectations as we were able to fit two Weibull functions to the mainshock and the aftershock η_{ij} distributions (Figure 5b). We also observe a similar distribution of probability to be an aftershock (Figure 5c) with Southern California, where aftershocks are more decisively classified. When looking at the seismicity rates for both mainshocks and aftershocks, we clearly identified the aftershock sequences of large earthquakes which occurred in the last 10 years over New Zealand (Figure 5d). We can see that the seismicity rate was slightly higher between 2010 and 2012, the background rate remains constant afterwards. We focus again on four major aftershock sequences to evaluate the quality of the declustering: the M7.1 Darfield earthquake of 2010 (Figure 5e), the M6.5 Cook Strait earthquake which occurred in 2013 (Figure 5f), the M7.1 East Coast event of 2016 (Figure 5g) preceding the complex M7.8 Kaikoura earthquake of 2016 (Figure 4h). For all of these sequences, we observe that these major events did not affect the seismicity outside a radius of 350km from the mainshock; away from these sequences the seismic activity was calm as shown by the constant rate of mainshocks and aftershocks. We can also observe that these major events had no incidence on the background seismicity inside their area as it does not show significant deviations from a constant rate.

We can speculate that our results could be different depending on how a catalog was built. If events are detected visually and picked manually by an operator, the magnitude completeness of a catalog could vary significantly in space and time during major aftershock sequences as more events would be added to the catalog close to these sequences. By presenting these two applications of our model to real seismicity in very distinct tectonic contexts, we have demonstrated the model robustness and its increased capability of classifying mainshocks and aftershocks compared to the classical approach of the Nearest-Neighbor-Distance approach, especially at the critical boundary between aftershocks and mainshocks in the 2D NND space.

4 Discussion and conclusions

In this work we present a new transfer learning approach to solve one of the classical problems in modern seismology, the declustering of aftershocks. As each seismicity catalog is different, we built a generalizeable model that is capable of declustering regardless of tectonic context. We based our model on one of the most used algorithms to decluster earthquake catalogs, the nearest-neighbor distance (NND) approach. This approach relies on the computation of a metric η_{ij} between one event and the events that precede it to identify its nearest neighbor, interpreted here as the mainshock that triggered the event in question. Recent work has focused on thresholding the NND η_{ij} to define the boundary between mainshocks and aftershocks, but we have shown that this neglects in most catalogs a significant overlap of the two populations. We tackled this problem with a data-science and machine learning approach, testing different supervised or unsupervised learning models to identify which model is best suited to decluster. We use features issued from the NND algorithm, differences in magnitude, the number of times the nearest neighbor has been selected as such, and the number of potential offspring. Our model of choice is a Random Forest trained on synthetic catalogs built following an ETAS model with a different set of parameters. The RF is the model with the best testing accuracy on synthetic catalogs. The RF model does not require any preprocessing, standardization or normalization of the features and it is scalable; regardless of the size of the catalog to be declustered, the model classifies each earthquake one by one rather than the distribution of all earthquakes. The hyperparameters of the model have been selected following a 5-Fold Cross-Validation. We applied our declustering model to two real catalogs: the relocated Southern California catalog from 1981 to 2019 and New Zealand GeoNet catalog from 2010 to 2020. We have shown that our model significantly improves the quality of the declustering in comparison of the classical use the NND metrics. Additionally, our approach is simple and fast to train, especially in comparison to Monte Carlo Markov Chains inversions that can be computationally demanding (Bayliss et al., 2019) or stochastic declustering (e.g. Zhuang et al., 2002) that does not make use of relative NND metrics and relies on many more parameters.

One has to keep in mind that these results reflect the performance of the model trained on a set of synthetic catalogs with a certain duration and fixed ETAS parameters for each catalog and stationary background seismic rates. Hainzl et al. (2016) found that aftershock sequences only last for about 100 days for moderate mainshocks (M = 6) and to a few years for larger events ($M \geq 7$). The estimated duration of an aftershock sequence is impacted by the background activity rate, which can potentially hide the trailing edge of a postseismic sequence as the process rate slows down with time. We have tested our models on shorter catalogs (approximately 500 days long) and noticed that the relative performance remains similar (see Figure S5 and S6). However, models trained on shorter catalogs do not reach similar accuracy when applied to longer catalogs. This is not surprising as long-term aftershocks could eventually be considered as mainshocks which means a larger nearest-neighbor-distance metric with their real mainshocks than what the model has been trained on. Indeed, another earthquake could have occurred closer in time and/or space but eventually not sufficiently close for these long-term aftershocks to be labelled as aftershocks anymore. If this problem is difficult to evaluate, better prediction performances are actually achieved if we don't force the nearest neighbor of the aftershocks to be their real mainshock, thus somewhat challenging the model.

Another aspect to keep in mind is that we used only five features to build the model presented here. It is of course possible to add more features by considering not only the nearest-neighbor but the N-nearest-neighbors, providing information to the model to classify an earthquake based on its relationship with many associated events, not just its single nearest neighbor. We can also consider using the nearest neighbor of the nearest neighbor, creating a family (Zaliapin & Ben-Zion, 2013). This would be a likely extension of our approach, as one aftershock can trigger another, and so on. Finally we can also imagine using N Random Forests to classify N neighbors based on a single set of features. Prelimary tests with our training data set, however, did not improve the accuracy of our model significantly, but this could represent a potential improvement for future work.

Regarding the comparison between the Southern California and the New Zealand catalogs, we could observe a significant difference between the shapes of their η_{ij} distributions (Figure 1e and f). The GeoNet catalog exhibits a higher number of mainshocks than aftershocks, respectively 57% against 43% (Figure 5), while the Southern California catalog is 70% aftershocks (Figure 4a). This difference can be explained by comparing their Nearest-Neighbor Distance distributions (see Figures 4b and 5b). Southern California's η_{ij} distribution shows the two expected lobes associated with mainshocks and aftershocks; the mainshock lobe of the η_{ij} distribution is less pronounced for New Zealand. The NZ catalog, while being shorter, has a higher background seismic rate according to our declustering model (see Figures 5d and 5d). We investigated this significant difference by comparing the η_{ij} distributions of the mainshock and aftershock seismicity as a function of depth (Figure 6a). We observe a rapid decrease in seismicity with depth of the Southern California seismicity that reaches approximately a maximum of 30-40 km depth, while NZ earthquakes can still be observed until 300 km depth. This reflects the difference in tectonics between the two regions and the type of seismicity observed in both catalogs: crustal seismicity in Southern California and a seismicity dominated by subduction earthquakes in New Zealand, whether located on the upper plates, the plate interface or at intermediate depths. We see for both regions that there are more aftershocks than mainshocks at shallow depth, above 25km. To understand how much deep mainshocks account for the observed NND distribution for NZ, we segmented the η_{ij} distribution for different depth ranges (Figure 6b). The seismicity comprised between 0 and 25 km exhibits a bimodal NND distribution and includes most of the aftershocks within the entire catalog. We see that the seismicity deeper than 50 km corresponds mostly to mainshocks. A similar behavior has been observed for Northern Chile by Aden-Antoniów et al. (2020), where no aftershocks were observed at intermediate depth ($z \geq 70$km); this results in a single-lobe NND distribution. The limited aftershock productivity at intermediate depths has been documented for different subduction zones in the past (e.g. Frohlich, 1987; Wiens et al., 1994). Only large intermediate-depths earthquakes seem to be able to trigger significant aftershock sequences but it still depends on the subduction zone (Frohlich, 1989). Temperature, geometry and hydration of the subducting slab seem to be the main parameters to control the aftershock productivity at these depths (e.g. Wiens, 2001; Ye et al., 2017; Cabrera et al., 2021).

To conclude, it is important to keep in mind that this model has been trained on synthetic data in an exercise of transfer learning. The synthetic data is controlled by range of reasonable ETAS parameters (see Table 1), but neither do these parameters or even the ETAS model reflect the entire spectrum of possibilities within earthquake catalogs. If this model might not be applicable to very particular seismicity i.e. mining related seismicity or seismicity with high earthquake rate variations, we suggest that it will be useful in a large majority of study regions. This work also provides a route to go beyond the classical use of the NND as the framework would be easy to reproduce.

Acknowledgments

F.A-A would like to thank Sid Kothari from the University of Western Ontario and Eric Beaucé from the Massachusetts Institute of Technology for interesting and constructive discussions and comments. We also thank everyone who has contributed to the SCEDC and GeoNet catalogs. We would like to thank the editor X and two anonymous reviewers for their comments and suggestions that greatly improve the manuscript. The codes allowing the use of our best Random Forest model for the declustering of seismic catalog created for this study are available here.

References

Aden-Antoniów, F., Satriano, C., Bernard, P., Poiata, N., Aissaoui, E.-M., Vilotte, J.-P., & Frank, W. (2020). Statistical analysis of the preparatory phase of the m w 8.1 iquique earthquake, chile. *Journal of Geophysical Research: Solid Earth*, *125*(6), e2019JB019337.

Aki, K. (1965). Maximum likelihood estimate of b in the formula log n= a-bm and its confidence limits. *Bull. Earthq. Res. Inst., Tokyo Univ.*, *43*, 237–239.

Asim, K., Martínez-Álvarez, F., Basit, A., & Iqbal, T. (2017). Earthquake magnitude prediction in hindukush region using machine learning techniques. *Natural Hazards*, *85*(1), 471–486.

Azak, T. E., Kalafat, D., Şeşetyan, K., & Demircioğlu, M. (2018). Effects of seismic declustering on seismic hazard assessment: a sensitivity study using the turkish earthquake catalogue. *Bulletin of Earthquake Engineering*, *16*(8), 3339–3366.

Bachmann, C. E., Wiemer, S., Woessner, J., & Hainzl, S. (2011). Statistical analysis of the induced basel 2006 earthquake sequence: introducing a probability-based monitoring approach for enhanced geothermal systems. *Geophysical Journal International*, *186*(2), 793–807.

Baiesi, M., & Paczuski, M. (2004). Scale-free networks of earthquakes and aftershocks. *Physical review E*, *69*(6), 066106.

Bayliss, K., Naylor, M., & Main, I. G. (2019). Probabilistic identification of earthquake clusters using rescaled nearest neighbour distance networks. *Geophysical Journal International*, *217*(1), 487–503.

Cabrera, L., Ruiz, S., Poli, P., Contreras-Reyes, E., Osses, A., & Mancini, R. (2021). Northern chile intermediate-depth earthquakes controlled by plate hydration. *Geophysical Journal International*, *226*(1), 78–90.

Corral, A. (2003). Local distributions and rate fluctuations in a unified scaling law for earthquakes. *Physical Review E*, *68*(3), 035102.

Davidsen, J., Gu, C., & Baiesi, M. (2015). Generalized omori–utsu law for aftershock sequences in southern california. *Geophysical Journal International*, *201*(2), 965–978.

Dempsey, D., Cronin, S. J., Mei, S., & Kempa-Liehr, A. W. (2020). Automatic precursor recognition and real-time forecasting of sudden explosive volcanic eruptions at whakaari, new zealand. *Nature communications*, *11*(1), 1–8.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, *15*(1), 3133–3181.

Frohlich, C. (1987). Aftershocks and temporal clustering of deep earthquakes. *Journal of Geophysical Research: Solid Earth*, *92*(B13), 13944–13956.

Frohlich, C. (1989). The nature of deep-focus earthquakes. *Annual Review of Earth and Planetary Sciences*, *17*(1), 227–254.

Galina, N., Bykova, V., Vakarchuk, R., & Tatevosian, R. (2019). Effect of earthquake catalog declustering on seismic hazard assessment. *Seismic Instruments*, *55*(1), 59–69.

Gardner, J., & Knopoff, L. (1974). Is the sequence of earthquakes in southern california, with aftershocks removed, poissonian? *Bulletin of the Seismological Society of America*, *64*(5), 1363–1367.

González, J., Yu, W., & Telesca, L. (2019). Earthquake magnitude prediction using recurrent neural networks. In *Multidisciplinary digital publishing institute proceedings* (Vol. 24, p. 22).

Gu, C., Schumann, A. Y., Baiesi, M., & Davidsen, J. (2013). Triggering cascades and statistical properties of aftershocks. *Journal of Geophysical Research: Solid Earth*, *118*(8), 4278–4295.

Gutenberg, B., & Richter, C. (1955). Magnitude and energy of earthquakes. *Nature*, *176*(4486), 795–795.

Hainzl, S., Christophersen, A., Rhoades, D., & Harte, D. (2016). Statistical estimation of the duration of aftershock sequences. *Geophysical Journal International*, *205*(2), 1180–1189.

Hainzl, S., & Ogata, Y. (2005). Detecting fluid signals in seismicity data through statistical earthquake modeling. *Journal of Geophysical Research: Solid Earth*, *110*(B5).

Hauksson, E., Yang, W., & Shearer, P. M. (2012). Waveform relocated earthquake catalog for southern california (1981 to june 2011). *Bulletin of the Seismological Society of America*, *102*(5), 2239–2244.

Hibert, C., Provost, F., Malet, J.-P., Maggi, A., Stumpf, A., & Ferrazzini, V. (2017). Automatic identification of rockfalls and volcano-tectonic earthquakes at the piton de la fournaise volcano using a random forest algorithm. *Journal of Volcanology and Geothermal Research*, *340*, 130–142.

Hoque, A., Raj, J., Saha, A., & Bhattacharya, P. (2020). Earthquake magnitude prediction using machine learning technique. In *International conference on computational intelligence, security and internet of things* (pp. 37–53).

Kong, Q., Allen, R. M., & Schreier, L. (2016). Myshake: Initial observations from a global smartphone seismic network. *Geophysical Research Letters*, *43*(18), 9588–9594.

Kong, Q., Trugman, D. T., Ross, Z. E., Bianco, M. J., Meade, B. J., & Gerstoft, P. (2019). Machine learning in seismology: Turning data into insights. *Seismological Research Letters*, *90*(1), 3–14.

Kothari, S., Shcherbakov, R., & Atkinson, G. (2020). Statistical modeling and characterization of induced seismicity within the western canada sedimentary basin. *Journal of Geophysical Research: Solid Earth*, *125*(12), e2020JB020606.

Lara-Cueva, R., Carrera, E. V., Morejon, J. F., & Benitez, D. (2016). Comparative analysis of automated classifiers applied to volcano event identification. In *2016 ieee colombian conference on communications and computing (colcom)* (pp. 1–6).

Li, Z., Meier, M.-A., Hauksson, E., Zhan, Z., & Andrews, J. (2018). Machine learning seismic wave discrimination: Application to earthquake early warning. *Geophysical Research Letters*, *45*(10), 4773–4779.

Maggi, A., Ferrazzini, V., Hibert, C., Beauducel, F., Boissier, P., & Amemoutou, A. (2017). Implementation of a multistation approach for automated event classification at piton de la fournaise volcano. *Seismological Research Letters*, *88*(3), 878–891.

Marsan, D., Bouchon, M., Gardonio, B., Perfettini, H., Socquet, A., & Enescu, B. (2017). Change in seismicity along the japan trench, 1990–2011, and its relationship with seismic coupling. *Journal of Geophysical Research: Solid Earth*, *122*(6), 4645–4659.

Marsan, D., Prono, E., & Helmstetter, A. (2013). Monitoring aseismic forcing in fault zones using earthquake time series. *Bull. Seismol. Soc. Am.*. doi: 10.1785/0120110304

Marsan, D., Reverso, T., Helmstetter, A., & Enescu, B. (2013). Slow slip and aseismic deformation episodes associated with the subducting Pacific plate offshore Japan, revealed by changes in seismicity. *J. Geophys. Res. E Planets*. doi: 10.1002/jgrb.50323

Marzocchi, W., & Taroni, M. (2014). Some thoughts on declustering in probabilistic seismic-hazard analysis. *Bulletin of the Seismological Society of America*, *104*(4), 1838–1845.

Moradpour, J., Hainzl, S., & Davidsen, J. (2014). Nontrivial decay of aftershock density with distance in southern california. *Journal of Geophysical Research: Solid Earth*, *119*(7), 5518–5535.

Mosteller, F., & Tukey, J. W. (1968). Data analysis, including statistics. *Handbook of social psychology*, *2*, 80–203.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, *83*(401), 9–27.

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, *50*(2), 379–402.

Omori, F. (1894). Investigation of aftershocks. *Rep. Earthquake Inv. Comm*, *2*, 103–139.

Peresan, A., & Gentili, S. (2018). Seismic clusters analysis in northeastern italy by the nearest-neighbor approach. *Physics of the Earth and Planetary Interiors*, *274*, 87–104.

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

Picozzi, M., & Iaccarino, A. G. (2021). Forecasting the preparatory phase of induced earthquakes by recurrent neural network. *Forecasting*, *3*(1), 17–36.

Provost, F., Hibert, C., & Malet, J.-P. (2017). Automatic classification of endogenous landslide seismicity using the random forest supervised classifier. *Geophysical Research Letters*, *44*(1), 113–120.

Reasenberg, P. (1985). Second-order moment of central california seismicity, 1969–1982. *Journal of Geophysical Research: Solid Earth*, *90*(B7), 5479–5495.

Reverso, T., Marsan, D., & Helmstetter, A. (2015). Detection and characterization of transient forcing episodes affecting earthquake activity in the aleutian arc system. *Earth and Planetary Science Letters*, *412*, 25–34.

Reverso, T., Marsan, D., Helmstetter, A., & Enescu, B. (2016). Background seismicity in boso peninsula, japan: Long-term acceleration, and relationship with slow slip events. *Geophysical Research Letters*, *43*(11), 5671–5679.

Ross, Z. E., Meier, M.-A., Hauksson, E., & Heaton, T. H. (2018). Generalized seismic phase detection with deep learning. *Bulletin of the Seismological Society of America*, *108*(5A), 2894–2901.

Rouet-Leduc, B., Hulbert, C., Lubbers, N., Barros, K., Humphreys, C. J., & Johnson, P. A. (2017). Machine learning predicts laboratory earthquakes. *Geophysical Research Letters*, *44*(18), 9276–9282.

Rubin, M. J., Camp, T., Van Herwijnen, A., & Schweizer, J. (2012). Automatically detecting avalanche events in passive seismic data. In *2012 11th international conference on machine learning and applications* (Vol. 1, pp. 13–20).

Seydoux, L., Balestriero, R., Poli, P., De Hoop, M., Campillo, M., & Baraniuk, R. (2020). Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning. *Nature communications*, *11*(1), 1–12.

Shebalin, P., & Narteau, C. (2017). Depth dependent stress revealed by aftershocks. *Nature communications*, *8*(1), 1–8.

Shi, P., Seydoux, L., & Poli, P. (2021). Unsupervised learning of seismic wavefield features: clustering continuous array seismic data during the 2009 l'aquila earthquake. *Journal of Geophysical Research: Solid Earth*, *126*(1), e2020JB020506.

Steinmann, R., Seydoux, L., Beaucé, E., & Campillo, M. (2021). Hierarchical exploration of continuous seismograms with unsupervised learning. *Earth and Space Science Open Archive ESSOAr*.

Taroni, M., & Akinci, A. (2021). Good practices in psha: declustering, b-value estimation, foreshocks and aftershocks inclusion; a case study in italy. *Geophysical Journal International*, *224*(2), 1174–1187.

Teng, G., & Baker, J. W. (2019). Seismicity declustering and hazard analysis of the oklahoma–kansas region. *Bulletin of the Seismological Society of America*, *109*(6), 2356–2366.

Thomas, A., Inbal, A., Searcy, J., Shelly, D., & Bürgmann, R. (n.d.). Identification of low-frequency earthquakes on the san andreas fault with deep learning. *Geophysical Research Letters*, e2021GL093157.

Trugman, D. T., & Shearer, P. M. (2017). Growclust: A hierarchical clustering algorithm for relative earthquake relocation, with application to the spanish springs and sheldon, nevada, earthquake sequences. *Seismological Research Letters*, *88*(2A), 379–391.

Table 1. Parameter ranges for the modeling of synthetic catalogs. These values were used as boundaries to draw, for each catalog, the ETAS parameter from a uniform distribution.

parameter	min	max
m_c	2.0	3.0
b	0.8	1.0
$< \tau >$	1	3
c	10^{-8}	1
p	1.0	1.3
K	0.12	0.18
a	0.8	1.05
γ	1.5	2.5

Utsu, T. (1957). Magnitudes of earthquakes and occurrence of their aftershocks. *Zisin, Ser. 2*, *10*, 35–45.

Utsu, T. (1961). A statistical study on the occurrence of aftershocks. *Geophys. Mag.*, *30*, 521–605.

Utsu, T., Ogata, Y., et al. (1995). The centenary of the omori formula for a decay law of aftershock activity. *Journal of Physics of the Earth*, *43*(1), 1–33.

Utsu, T., & Seki, A. (1955). A relation between the area of the aftershock region and the radius of the sensibility circle. *Zisin*, *3*(34), 1955.

Van Stiphout, T., Zhuang, J., Marsan, D., Stiphout, V., Zhuang, J., & Marsan, D. (2012). Theme V-Models and Techniques for Analyzing Seismicity Seismicity Declustering. *CORSSA*. doi: 10.5078/corssa

Wiemer, S., & Wyss, M. (2000). Minimum magnitude of completeness in earthquake catalogs: Examples from alaska, the western united states, and japan. *Bulletin of the Seismological Society of America*, *90*(4), 859–869.

Wiens, D. A. (2001). Seismological constraints on the mechanism of deep earthquakes: Temperature dependence of deep earthquake source properties. *Physics of the Earth and Planetary Interiors*, *127*(1-4), 145–163.

Wiens, D. A., McGuire, J. J., Shore, P. J., Bevis, M. G., Draunidalo, K., Prasad, G., & Helu, S. P. (1994). A deep earthquake aftershock sequence and implications for the rupture mechanism of deep earthquakes. *Nature*, *372*(6506), 540–543.

Ye, L., Lay, T., Bai, Y., Cheung, K. F., & Kanamori, H. (2017). The 2017 mw 8.2 chiapas, mexico, earthquake: Energetic slab detachment. *Geophysical Research Letters*, *44*(23).

Yuan, B., Tan, Y. J., Mudunuru, M. K., Marcillo, O. E., Delorey, A. A., Roberts, P. M., . . . others (2019). Using machine learning to discern eruption in noisy environments: A case study using co2-driven cold-water geyser in chimayó, new mexico. *Seismological Research Letters*, *90*(2A), 591–603.

Zaliapin, I., & Ben-Zion, Y. (2013). Earthquake clusters in southern California I: Identification and stability. *J. Geophys. Res. Solid Earth*. doi: 10.1002/jgrb.50179

Zaliapin, I., & Ben-Zion, Y. (2020). Earthquake declustering using the nearest-neighbor approach in space-time-magnitude domain. *Journal of Geophysical Research: Solid Earth*, *125*(4), e2018JB017120.

Zaliapin, I., Gabrielov, A., Keilis-Borok, V., & Wong, H. (2008). Clustering analysis of seismicity and aftershock identification. *Physical review letters*, *101*(1), 018501.

Zhu, W., & Beroza, G. C. (2019). Phasenet: a deep-neural-network-based seismic arrival-time picking method. *Geophysical Journal International*, *216*(1), 261–273.

Zhuang, J., Ogata, Y., & Vere-Jones, D. (2002). Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, *97*(458), 369–380.

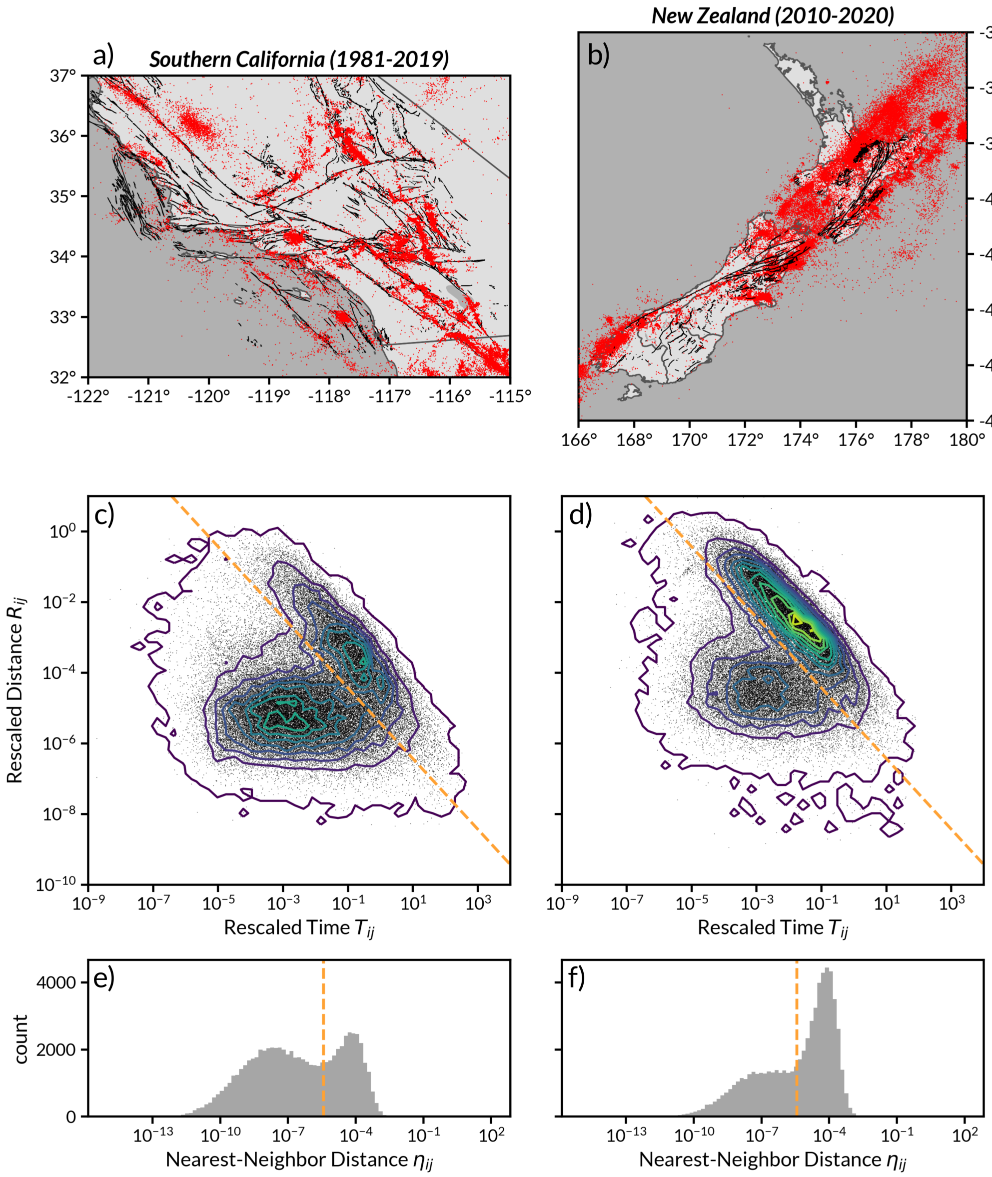

Figure 1. Examples of real catalogs. The seismic catalogs used in this study are the Southern California catalog (a) from 1981 to 2019 (Hauksson et al., 2012) and New Zealand (b) from 2010 to 2020 from GeoNet. (c) and (d) show the respective 2D distribution of the Nearest-Neighbor distance (NND) as a function of the rescaled time T_{ij} and distance R_{ij} while (e) and (f) show their 1-D NND metric distribution. The dashed orange line represents the threshold that would be used for a classical approach of the NND for the declustering of earthquake catalogs.

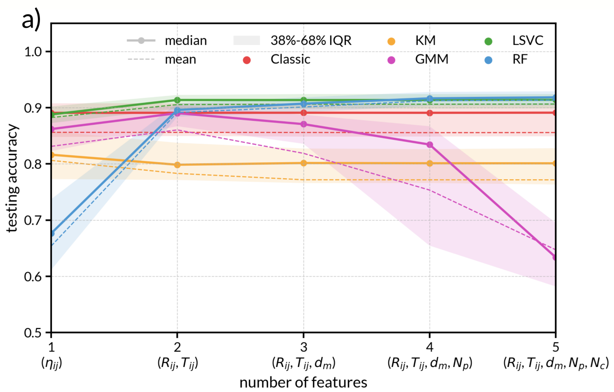

Figure 2. Machine Learning algorithms comparison and feature importance for the Random Forest model. (a) We declustered 100 synthetic catalogs using different number of features with different models and estimated the testing accuracies. The shaded area represent the 38% - 68% inter-quartiles range of the accuracy distributions, the solid line represent the median of this distributions and the dashed line is their mean. (b) We estimated the features importance in the Random Forest using all 5 features. The black lines show the relative importance of these features in a descending order for the model while the black line is the cumulative. The vertical bars shows the standard variation of the features importance. This was computed from the distribution of each feature importance across all trees of the forest.

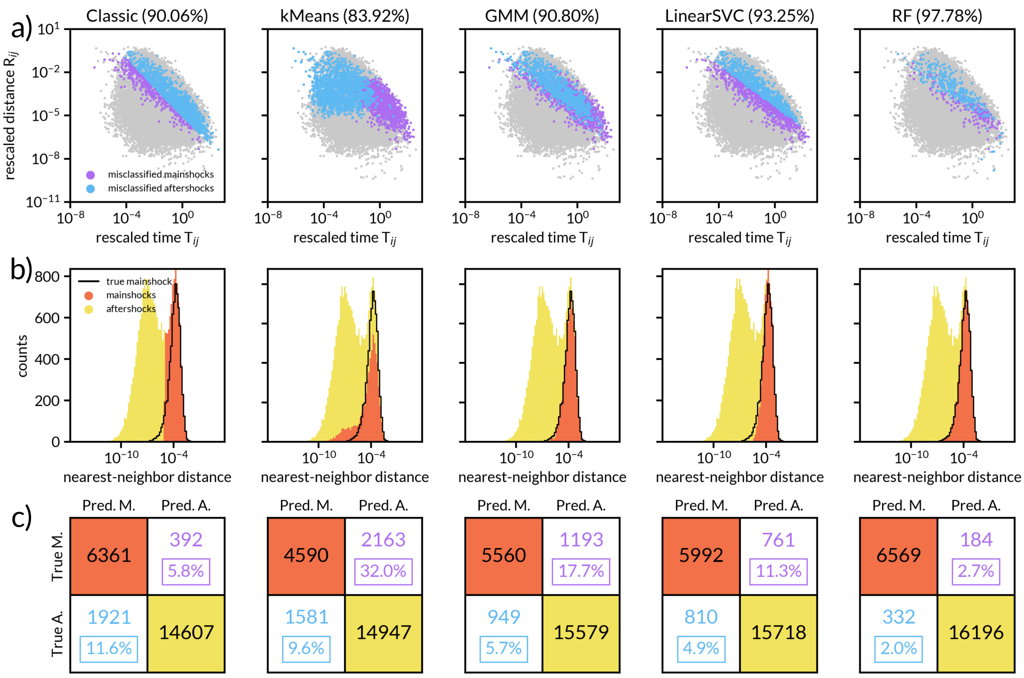

Figure 3. Example of the declustering of a synthetic catalog with different machine learning algorithms. (a) For the different algorithms, we show the distribution of the nearest neighbor rescaled distance R_{ij} and time T_{ij} where the color corresponds to the mis-labelled events whether there are originally mainshocks (purple) or aftershocks (blue). The accuracy of the model is displayed in the title of each subplot. (b) η_{ij} distribution obtained by applying each model. The orange corresponds to the inferred mainshock population while yellow is for the aftershock population. The black curve shows the real background distribution. (c) shows the different confusion matrices where we show the number of correctly identified samples in the colored cells and mis-labelled events with their corresponding proportion regarding their population in the white cells.

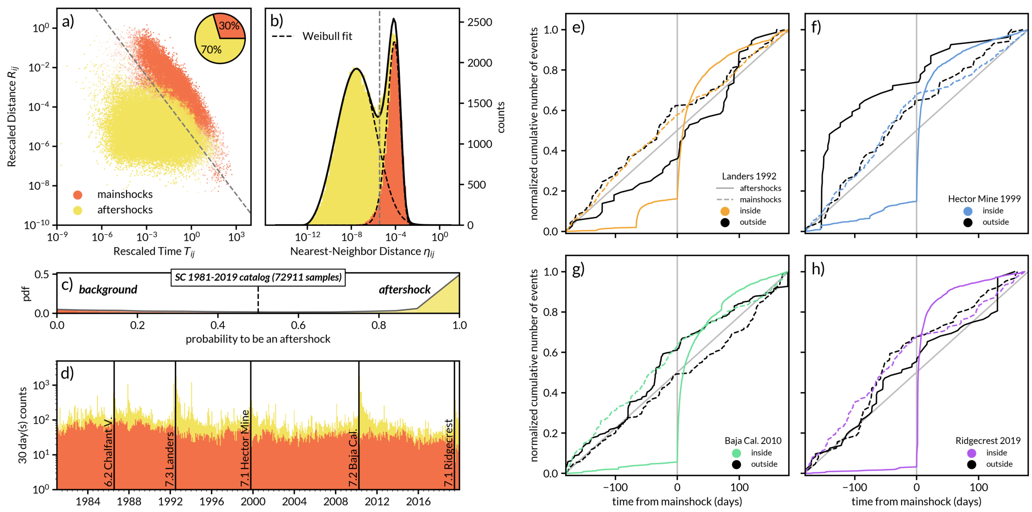

Figure 4. Our best Random Forest model for the declustering of earthquake catalog applied to Southern California seismicity between 1981 and 2019. (a) shows the nearest-neighbor rescaled distance R_{ij} and time T_{ij} distribution. The color indicate if the events have been classified as mainshocks (orange) or aftershocks (yellow). The dashed grey line shows where the classical approach of the NND whould have separated the two populations. (b) shows the obtained the two stacked η_{ij} distributions. The dashed black lines corresponds to the fit of a Weibull function to both distributions while the black line shows the resulting sum and fit to the overall η_{ij} distribution. (c) displays the probability density function of the probability to be an aftershock, obtained from the data. (d) exhibits the stacked histograms on the earthquake count per window of 30 days. The black vertical lines mark the date of large earthquake that occurred during the period covered by the catalog. (e) - (h) show the normalized cumulative number of mainshocks (dashed curves) and aftershocks (plain curves) for 4 remarkable sequences of aftershocks. The color indicates whether these events are located at distance larger or lower than 350km.

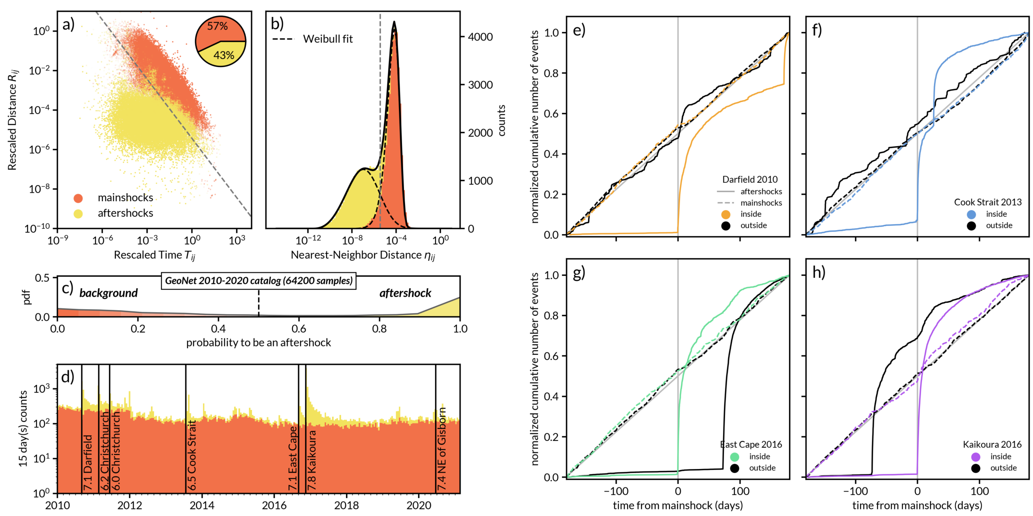

Figure 5. Our best Random Forest model for the declustering of earthquake catalog applied to New Zealand seismicity between 2010 and 2020. Similar to Figure 4.

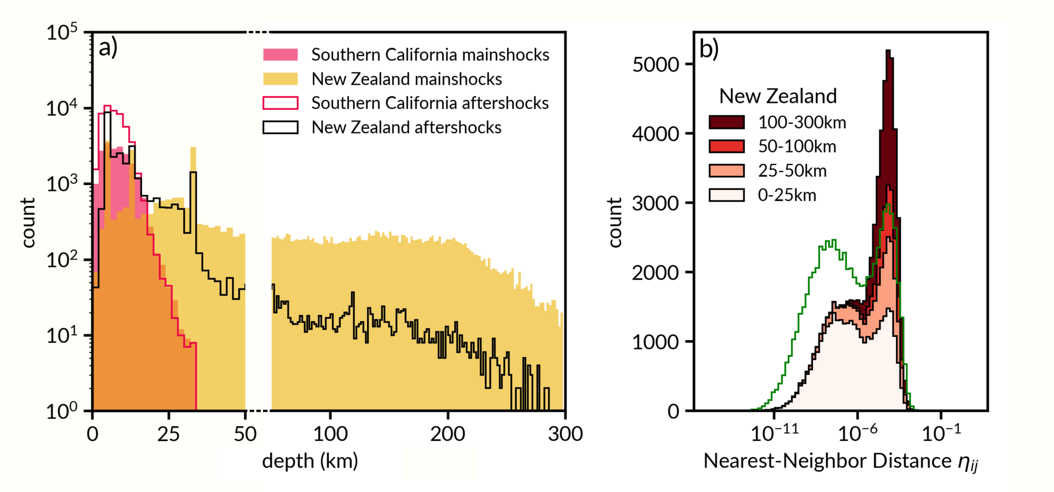

Figure 6. Mainshock/Aftershock depth distributions in relation to the Nearest-Neighbor Distance distribution for New Zealand. a) The histograms represent the distribution of the depth of the background seismicity and aftershocks for the Southern California and New Zealand. b) The NND η_{ij} distribution for the New Zealand catalog is represented by the colored histograms for different depth range. The green histogram represents the complete NND η_{ij} distribution for Southern California for comparison purposes.

Supporting Information for "Transfer Learning to Build a Scalable Model for the Declustering of Earthquake Catalogs"

F. Aden-Antoniów[1], W. B. Frank[1], L. Seydoux[2]

[1]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

[2]Équipe Ondes et Structures, ISTerre, Université Grenoble-Alpes, UMR 5375, Gières, France

Contents of this file

Figures complementing information reported in the main text

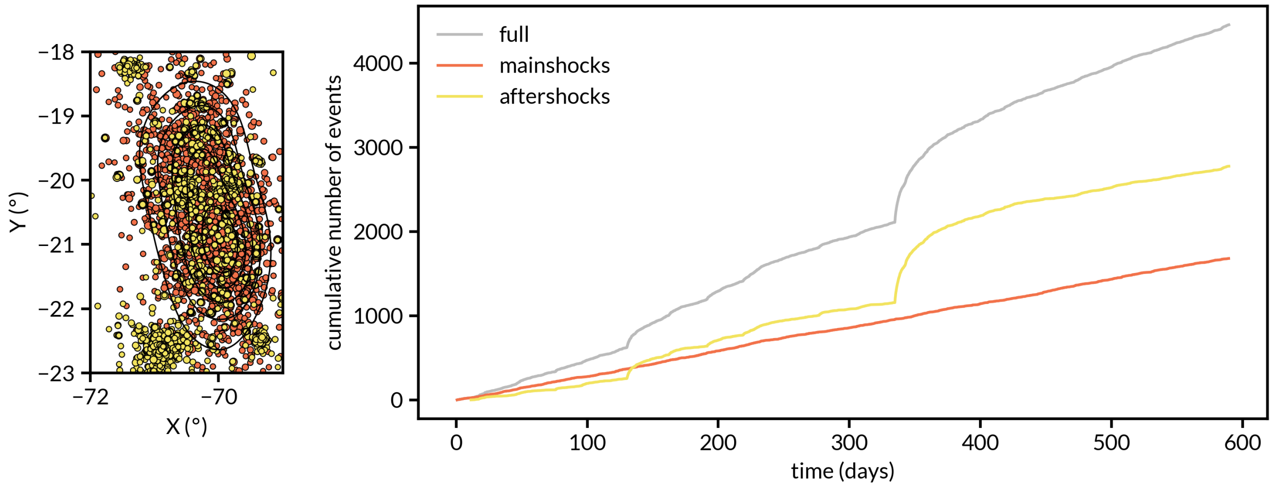

Figure S1. Example of a short synthetic catalog generated for this study. (left) Spatial distribution of the synthetic mainshocks (orange) and aftershocks (yellow). The black contours represent the 2D probability function use to generate the location of these earthquakes. (right) Cumulative number of the mainshocks and the aftershocks.

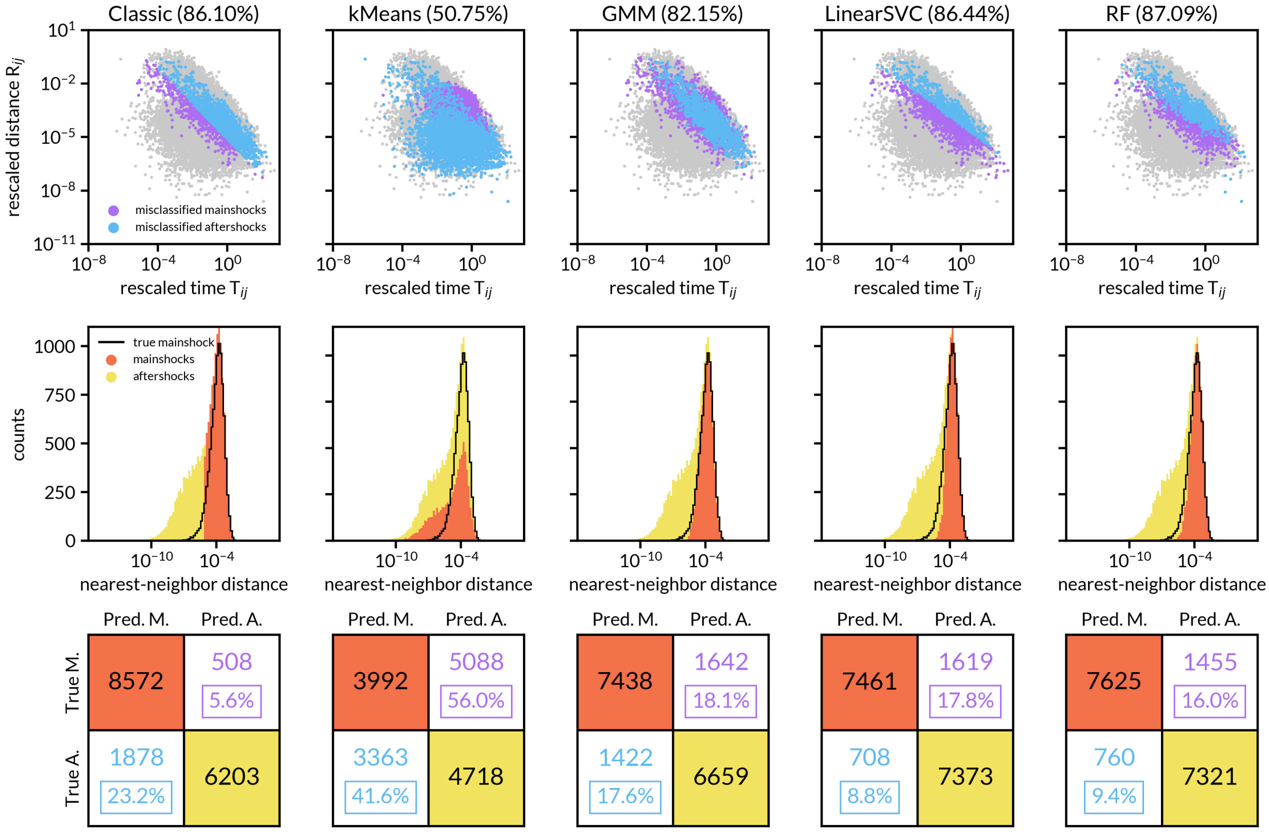

Figure S2. Example of the declustering of a synthetic catalog with different machine learning algorithms. (a) For the different algorithms, we show the distribution of the nearest neighbor rescaled distance R_{ij} and time T_{ij} where the color corresponds to the mis-labelled events whether there are originally mainshocks (purple) or aftershocks (blue). The accuracy of the model is displayed in the title of each subplot. (b) η_{ij} distribution obtained by applying each model. The orange corresponds to the inferred mainshock population while yellow is for the aftershock population. The black curve shows the real background distribution. (c) shows the different confusion matrices where we show the number of correctly identified samples in the colored cells and mis-labelled events with their corresponding proportion regarding their population in the white cells.

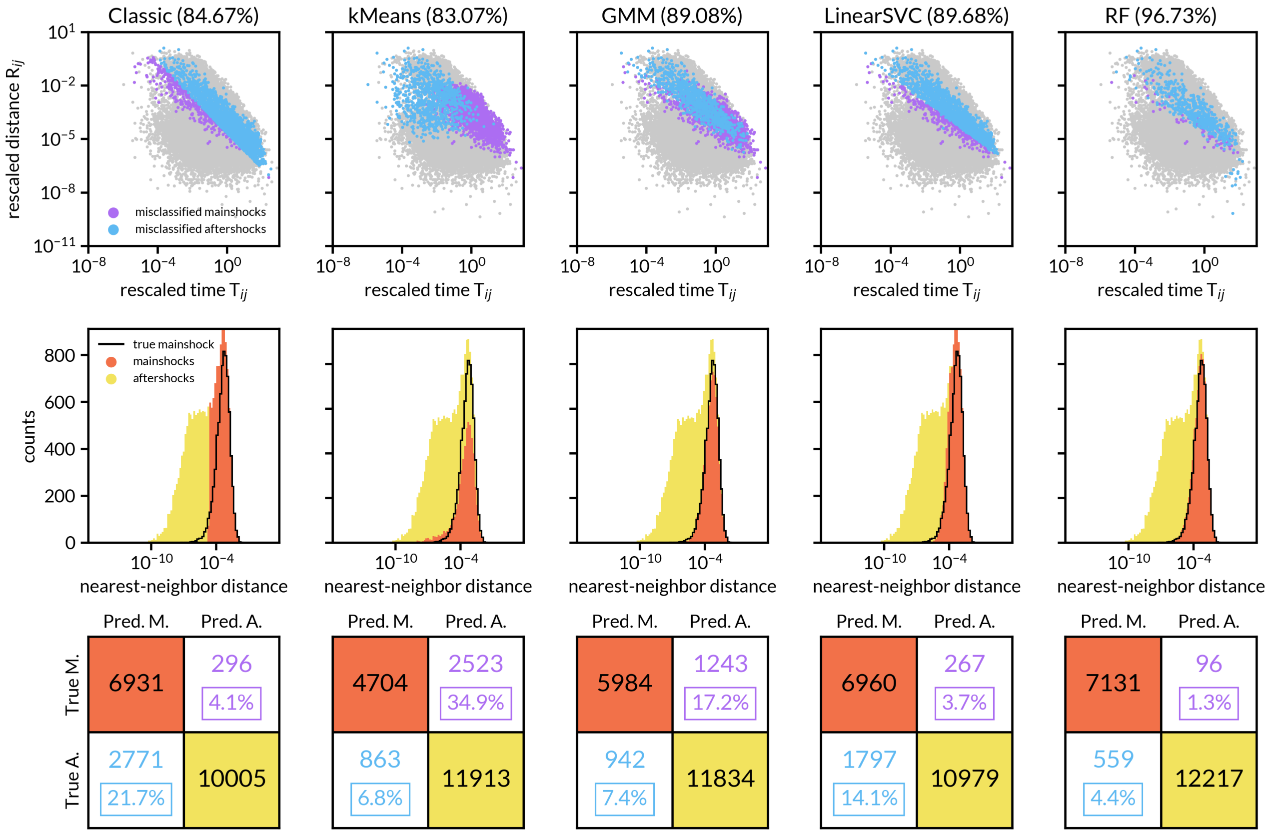

Figure S3. Other example of the declustering of a synthetic catalog with different machine learning algorithms. Same as S2.

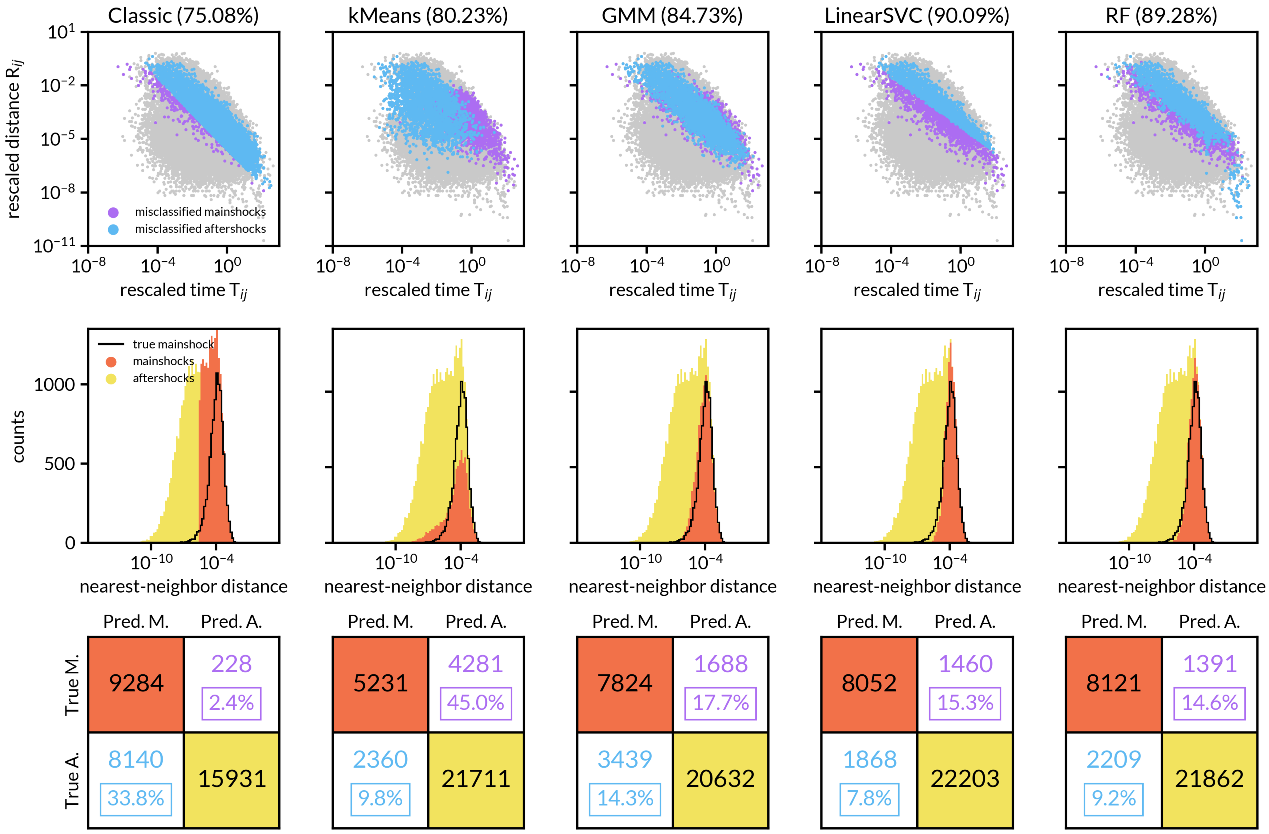

Figure S4. Other example of the declustering of a synthetic catalog with different machine learning algorithms. Same as S2.

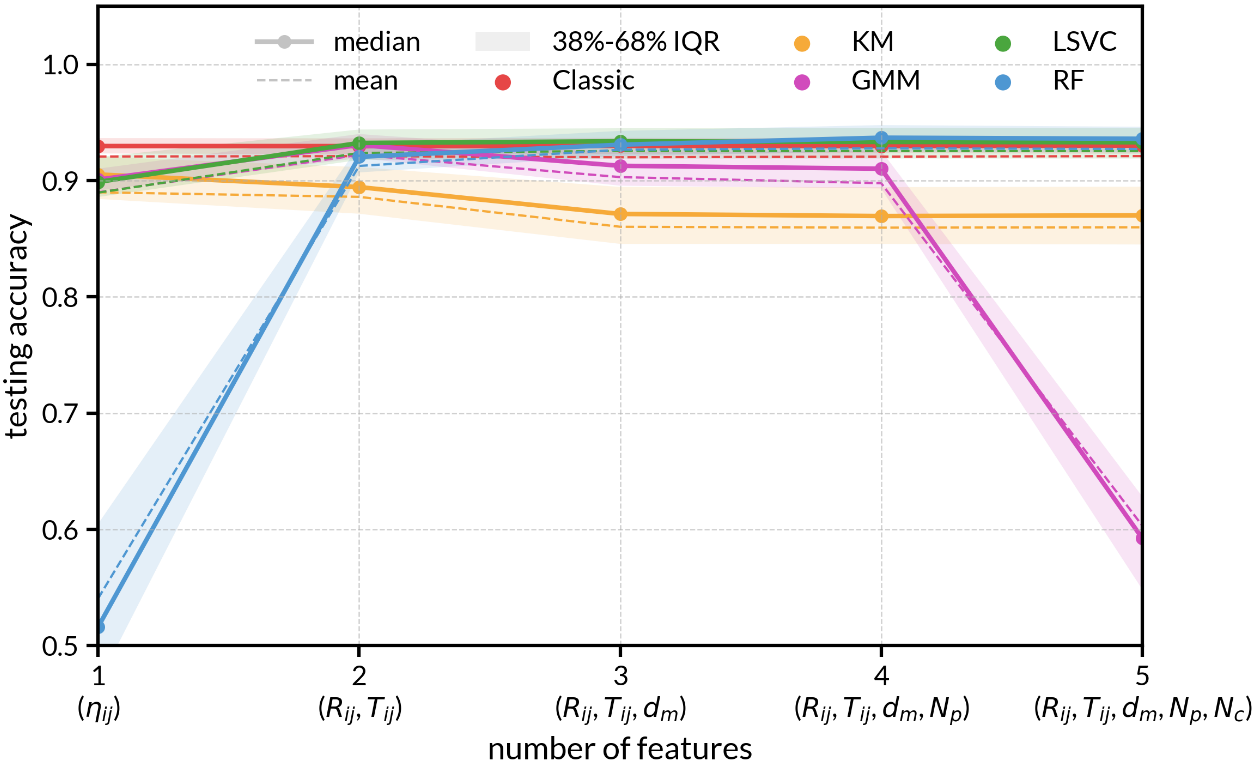

Figure S5. Machine Learning algorithms comparison with different catalog duration for training and testing phases. (a) We declustered 100 synthetic catalogs of an approximated duration of 1.5 years using different number of features with different machine learning models (KM: K-Means, GMM: Gaussian Mixture Model, LSVC: Linear Support Vector Classifier, RF: Random Forest) and estimated the testing accuracies. The Linear Support Vector Classifier and the Random Forest have been trained on synthetic catalogs with a duration of approximately 10 years. The shaded area represent the 38% - 68% inter-quartiles range of the accuracy distributions, the solid line represent the median of this distributions and the dashed line is their mean.

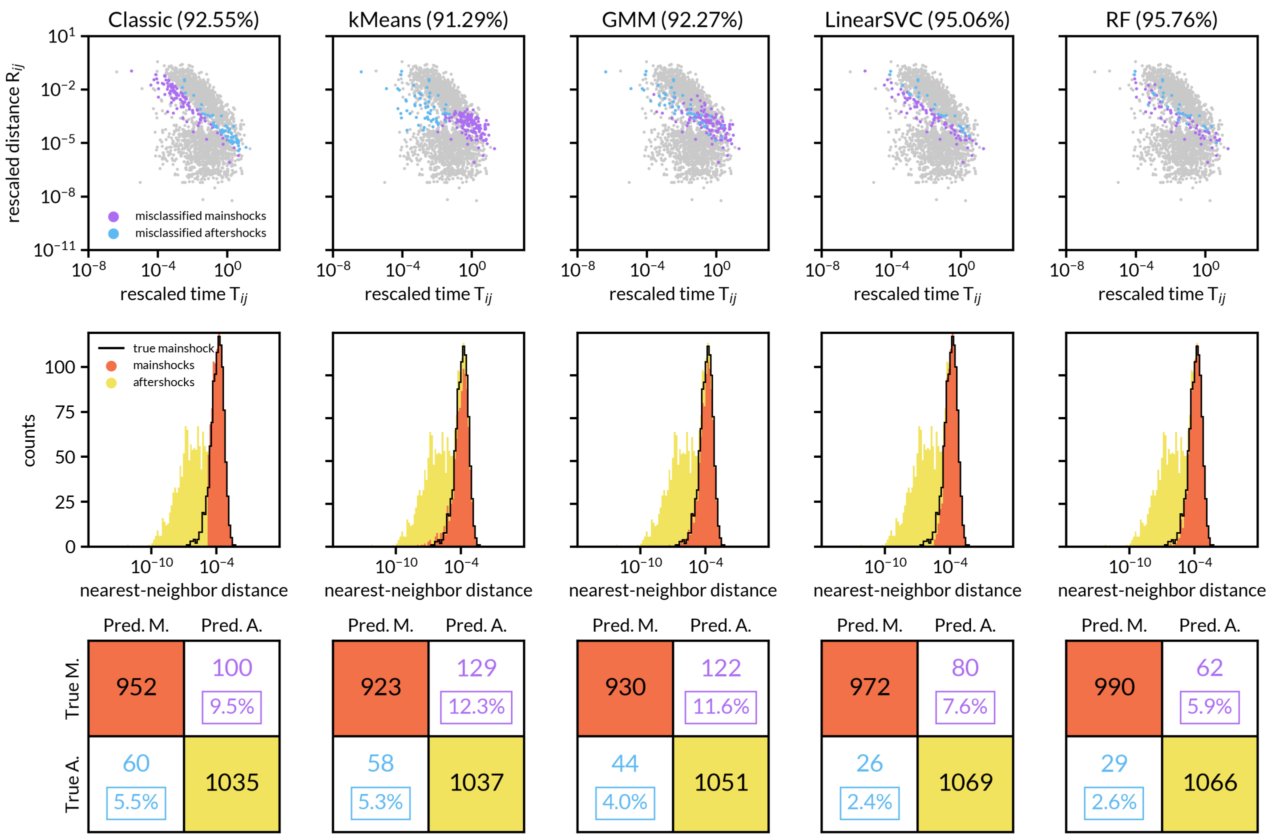

Figure S6. Example of the declustering of a short synthetic catalog (1.5 year) with different machine learning algorithms. (a) For the different algorithms, we show the distribution of the nearest neighbor rescaled distance R_{ij} and time T_{ij} where the color corresponds to the mis-labelled events whether there are originally mainshocks (purple) or aftershocks (blue). The accuracy of the model is displayed in the title of each subplot. (b) η_{ij} distribution obtained by applying each model. The orange corresponds to the inferred mainshock population while yellow is for the aftershock population. The black curve shows the real background distribution. (c) shows the different confusion matrices where we show the number of correctly identified samples in the colored cells and mis-labelled events with their corresponding proportion regarding their population in the white cells.