Détection automatique et classification basée sur l'apprentissage machine des séismes de faible magnitude dans une région continentale stable

Alexandra Renouard

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE _ED 413_____

[UMR 7516]

THÈSE présentée par:

[Alexandra RENOUARD]

soutenue le : 04 novembre 2020

pour obtenir le grade de: **Docteur de l'université de Strasbourg**

Discipline/ Spécialité: Sismologie

Détection automatique et classification basée sur l'apprentissage machine des séismes de faible magnitude dans une région continentale stable

THÈSE dirigée par :

[Mme MAGGI Alessia]	Professeur, université de Strasbourg

RAPPORTEURS :

[Mr JOHNSON Paul]	Professeur, Los Alamos National Laboratory, New Mexico
[Mr DELOUIS Bertrand]	Professeur, Université de la Côte d'Azur

---

AUTRES MEMBRES DU JURY :

[Mr GAILLARD Pierre]	Ingénieur de recherche, CEA/DAM Bruyères-le-Châtel
[Mme HELMSTETTER Agnès]	Chargée de recherche CNRS, Université de Grenoble
[Mr VERGNE Jérôme]	Physicien, Université de Strasbourg
[Mme DOUBRE Cécile]	Physicienne adjointe, Université de Strasbourg
[Mr GRUNBERG Marc]	Ingénieur de recherche, Université de Strasbourg

UNIVERSITÉ DE STRASBOURG

École Doctorale des Sciences de la Terre et de l'Environnement ED 413

UMR 7516

Thèse présentée pour obtenir le grade universitaire de Docteur par

Alexandra RENOUARD

Détection automatique et classification basée sur l'apprentissage machine des séismes de faible magnitude dans une région continentale stable

Sous la direction de Pr. Alessia MAGGI, Université de Strasbourg

Soutenue le 04/11/2020 devant le jury composé de :

RAPPORTEURS	
Bertand DELOUIS	Professeur, Université de la Côte d'Azur
Paul JOHNSON	Professeur, Los Alamos National Laboratory
EXAMINATEURS	
Pierre GAILLARD	Ingénieur de Recherche, CEA/DAM Bruyères-le-Châtel
Agnès HELMSTETTER	Chargée de Recherche CNRS, Université de Grenoble
Jérôme VERGNE	Physicien, Université de Strasbourg
ENCADRANTS	
Alessia Maggi	Professeur, Université de Strasbourg
Cécile Doubre	Physicienne adjointe, Université de Strasbourg
Marc Grunberg	Ingénieur de recherche, Université de Strasbourg

Résumé

Les régions continentales stables sismiquement actives, comme celle du Graben du Rhin Supérieur, sont caractérisées par des taux de déformation très faibles et enregistrent une sismicité majoritairement faible à modérée. La compréhension des mécanismes qui gouvernent l'occurrence et la distribution de cette sismicité dans ces régions est fortement entravée par les capacités limitées des systèmes de détection à détecter les séismes de plus faible magnitude, dans des environnement qui sont souvent très anthropisés, et ce, malgré le déploiement intensif des réseaux de stations.

Afin d'améliorer la détection des séismes de faible magnitude dans notre zone d'étude, nous avons cherché à définir les facteurs qui limitent cette détection et avons développé une nouvelle procédure de détection automatique. Ce travail a mis en évidence deux principaux facteurs limitants : le niveau de bruit enregistré aux stations et le milieu de propagation des ondes sismiques. Si ces deux facteurs sont négligés dans les différentes étapes du processus de détection (pointé des temps d'arrivée des ondes sismiques, association des pointés pour inférer une origine, localisation de l'origine), des taux élevés de faux événements, associés à du bruit impulsif, et de vrais événements (tirs de carrière ou séismes), contaminés par du bruit, sont détectés.

En prenant en compte un nombre limité de paramètres qui gouvernent les différentes étapes du processus de détection des événements, nous avons été en mesure de réduire significativement la contamination par le bruit des vrais évènements détectés. Les paramètres ayant fourni les meilleurs résultats sont associés aux caractéristiques du bruit enregistré aux stations, à la géométrie du réseau de stations, ainsi qu'au milieu de propagation des ondes sismiques.

L'utilisation combinée de l'Homme et d'un algorithme d'apprentissage machine supervisé interprétable nous a permis de solidement classifier les différents types d'événements détectés : d'abord en vrais et faux évènements, puis en tirs de carrières et séismes. Cette approche hybride s'est avérée efficace pour classer les événements à travers une validation des règles de classification qui minimisent à la fois les effets liés au bruit et au milieu de propagation.

Les résultats de cette procédure de détection automatique sont prometteurs : 50% de séismes de magnitude inférieure à 1.2 sont détectés en plus. En outre, l'utilisation de l'apprentissage machine met à jour une variabilité spatiale dans l'efficacité des discriminants utilisés pour différencier les séismes des tirs de carrière, qui est à cartographier plus finement pour en comprendre l'origine. Aussi, ce travail de thèse promeut une plus large exploration de l'apprentissage machine au sein des observatoires sismologiques.

Mots-clefs : détection, discrimination, apprentissage machine supervisé, intelligence artificielle, tirs de carrière, bruit sismique, séismes de faible magnitude

Abstract

Seismically active stable continental regions, such as Upper Rhine Graben area, are subjected to very low strain rate conditions and mainly record low-to-moderate seismicity. In the highly anthropogenic context of Central Western Europe, and despite intense station deployments, small-magnitude earthquakes remain unevenly recorded, blurring our view of present-day earthquake behaviors. Under these conditions, earthquake occurrence and triggering mechanisms are difficult to explain.

In order to improve the detection capabilities of our station network, we studied the factors that affect earthquake detection performance and developed a new automatic detection procedure. We observe that the main limiting factors are related to station noise level and seismic wave propagation medium. If both factors are neglected during the detection process (P- and S-arrival picking, pick association to infer event origin locations, origin location), high rates of false events, related to impulsive noise, and real events (quarry blasts or earthquakes) contaminated by noise are detected.

By taking into account a limited number of parameters, we are able to significantly reduce the contamination of noise in the detection process of real events. The parameters that give the best results are associated to the space-time-varying noise characteristics of individual stations, the network geometry and the seismic wave propagation medium. By using a combination of both Human and interpretable supervised machine learning algorithm, we robustly classify the detected events in first, false vs real event, and second, quarry blast vs earthquake. This hybrid machine learning approach has proved to be efficient in event classification by validating classification rules that minimize noise and path effects.

Compared to the reference French National Catalog for the same time period, this detection procedure detects twice as many earthquakes with magnitudes less than 1.2. Furthermore, examination of the classification rules created in the earthquake-quarry blast classifier reveals a strong geographical variability in the effectiveness of signal discriminants, whose origin has to be investigated more deeply. This work also promotes a broader implication of hybrid intelligence monitoring within seismological observatories.

Key words : detection, discrimination, supervised machine learning, artificial intelligence, quarry blasts, seismic noise, small-magnitude earthquakes

Remerciements

« *Pies para qué los quiero si tengo alas para volar* »—Frida Khalo, 1953 La puissance des mots de Frida Khalo résonne encore depuis le quartier de Coyoacán de Mexico DF.

Je voulais remercier tout d'abord l'ensemble des membres du jury d'avoir accepté de participer à ma soutenance de thèse.

Je tiens à remercier en premier lieu mes encadrants de thèse, Alessia MAGGI, Cécile DOUBRE et Marc GRUNBERG, pour la confiance qu'il m'ont accordée en acceptant d'encadrer ce travail doctoral, pour leurs multiples conseils et pour toutes les heures qu'ils ont pu consacrées à diriger cette recherche. Je les remercie également chaleureusement de m'avoir soutenue dans tous mes choix et de m'avoir permis de m'exprimer librement, tout en me guidant scientifiquement. Je les remercie enfin de n'avoir jamais rechigné à m'envoyer dans des conférences diversifiées qui m'ont conduite, à travers les nombreuses discussions, à pousser ma réflexion toujours plus avant. J'ai également beaucoup appris à vos côtés et je vous adresse ma gratitude pour tout cela.

J'émettrai un remerciement particulier à Marc GRUNBERG pour son investissement dans le développement méthodologique de ce travail de thèse. Merci d'avoir toujours répondu positivement à mes sollicitations et d'avoir été force de propositions.

Je voulais remercier également Clément GRELLIER et Romain PESTOURIE pour leur aide. Merci pour votre gentillesse et votre disponibilité. J'ai également une pensée particulière pour Fabien ENGELS qui m'a sauvé la vie plusieurs fois dans mes batailles avec la machine... Je sais je te dois toujours une boîte de chocolats. Je me souviendrai de ce crash d'ordinateur pendant longtemps !

Un grand merci à Rémi DRETZEN également pour son travail colossal de discrimination manuelle quotidienne qui a permis d'apporter une sérieuse pierre à l'édifice. Merci d'avoir accepté de revoir manuellement certaines listes de séismes et de tirs de carrière.

Je tiens à profondément remercier aussi toutes les personnes impliquées dans le déploiement et la maintenance du réseau de stations permanentes et temporaires AlpArray qui ont fourni les données utilisées pour ce travail. Sans ces personnes ce travail n'aurait jamais pu voir le jour. Un remerciement particulier aux personnes impliquées dans le réseau de l'observatoire sismologique du Nord-Est de la France : Maxime BES DE BERC, Hélène JUND, Hervé WODLING, Céleste BROUCKE, Gauthier WEYLAND, et bien sûr Cécile DOUBRE qui dirige cet observatoire.

Je tiens particulièrement à remercier Antoine SCHLUPP pour m'avoir proposé de faire partie du comité d'organisation du colloque de l'AFPS2019. Ce fut une expérience enrichissante. Antoine, j'ai égalemennt appris beaucoup à tes côtés.

Je remercie également Anne Paul, membre du comité de suivi, pour la confiance qu'elle m'a accordée et l'apaisement qu'elle m'a fourni.

Une pensée émue pour Christophe Voisin qui m'a coachée à l'AGU de San Francisco

Ce travail n'aurait pas été le même sans les discussions élaborées au cours des différentes conférences, parfois furtives mais stimulantes, voire pour certaines fondamentalement marquantes : je pense notamment à Christian SUE, Oona SCOTTI, Simone CESCA, Andréas RIETBROCK, Claudio SATRIANO, Jean SCHMITTBUHL, Clément HIBERT, Jean François SEMBLAT, Huseyin Serdar KUYUK, Jean LETORT, Pierre GAILLARD, Arnaud MONTABERT, Catherine BERGE-THIERRY, Clara DUVERGER, etc... Mais il y a tant d'autres personnes à remercier car c'est aussi une vraie aventure humaine ! Et puis je garde en mémoire cette phrase lancée par Bertrand ROUET-LEDUC au sujet des publications :"c'est toujours le bon moment !"

J'ai également beaucoup apprécié les discussions avec Antoine SCHLUPP et Anne-Sophie MERIAUX (merci pour le Kombucha !) et également Christophe SIRA

Je remercie enfin Yves MAZABRAUD de m'avoir poussée à franchir le pas.

J'ai également une pensée émue pour mes voisins de bureau : Raphaël et Fatiah.

Je remercie mes proches et leur dirai que je suis assez admirative qu'ils m'aient soutenue sans jamais me lâcher.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Préface

> « *Often nature surprises us, such as when an earthquake, hurricane or flood is bigger or has greater effects than expected from hazard assessments. In other cases, nature outsmarts us, doing great damage despite expensive mitigation measures, or making us divert resources to address a minor hazard. We keep learning the hard way to maintain humility before the complexity of nature* »—Seth Stein, 2014

Le 13 octobre 2014 après-midi, île de Saint-Martin, Nord des Antilles. La sirène d'alerte retentit. A cet instant précis, tous les habitants de l'île savent que l'alerte rouge est désormais lancée : Gonzalo, un ouragan de catégorie 1 approche. Un peu chancelants, les saint-martinois pensent que ce sera comme d'habitude une "petite" tempête tropicale. La population se barricade mais elle va laisser nonchalemment des portails ouverts, des voitures non protégées, des vitres non consolidées, des bâteaux non mis à l'abri.

Nuit du 13 au 14 octobre 2014. Les portes claquent, les toits de tôle font un boucan du diable, les vitres tremblent. La nature rappelle à l'ordre, le réveil est lourd. L'eau et l'électricité sont coupées, les communications brouillées. Le paysage est apocalyptique : les panneaux signalétiques sont à terre, les arbres jonchent le sol, les câbles électriques trempent dans les flaques d'eau, des toits sont arrachés, des portails sont démantelés, les habitats de fortune ne sont plus que des amoncellements de débris, des cadavres de bateau habillent le bord de mer, des voitures ont été broyées par des morceaux de tôles projetés depuis le petit aéroport de Grand-Case à proximité. Les dégâts sont catastrophiques et démesurés, la population est abasourdie. Gonzalo a été un petit ouragan. Pourtant, l'expérience de Luis, Ouragan de catégorie 4 ayant ravagé l'île en 1995, avait laissé de lourdes cicatrices. "Luis" était en fait considéré comme un bon vieux copain. La mémoire collective en ce 13 octobre 2014 a été défaillante. Devoir de résilience oblige, la vie reprend rapidement son cours. Les habitats sont reconstruits identiquement, de bric et de broc, toujours plus proches de la

mer. Trois ans plus tard, le 6 septembre 2017, l'ouragan de catégorie 5, Irma, pulvérise l'île. Un ange passe...

Tout aléa naturel est donc indissociablement lié à la notion de risque. S'il est pourtant possible de prévoir à court terme la trajectoire des cyclones avec une assez bonne précision temporelle et donc d'anticiper, la vulnérabilité inhérente de l'île (fragilité des constructions, terres littorales facilement submergées, urbanisation intense dans les zones à risques), mai aussi le manque de préparation de ses habitants n'ont permis ni une minimisation des dégâts, ni une gestion de crise optimale. Un constat un peu plus lourd pourrait être également dressé en cas de séisme majeur. Un peu plus lourd car, contrairement aux cyclones, il est impossible de prévoir exactement où et quand un séisme aura lieu, et à quelle magnitude il sera.

Les séismes dévastateurs surprennent alors souvent. En Haïti, plusieurs décennies de constructions sans surveillance ont amené à un effondrement colossal des structures lors du séisme du 12 janvier 2010, provoquant la mort de plus de 230 000 personnes dans le district de Port-au-Prince. Ce séisme de magnitude 7.4 a été deux fois plus meurtrier que tout autre précédent séisme de magnitude équivalente (Bilham, 2010). Dans la région de Tohoku au Japon, le 11 mars 2011, un séisme de magnitude 9 déclenche un énorme tsunami, submergeant les digues de protection qui avaient été érigées jusqu'à 10 mètres de hauteur pour contrer les effets des séismes tsunamigéniques. Au bilan, l'alimentation nécessaire au maintien de la circulation d'eau pour refroidir les réacteurs de la centrale nucléaire de Fukushima est interrompue, plus de 19 000 morts et au moins 2 00 milliards de dollars de dégâts ont été recensés (Normile, 2012). Ce séisme a libéré environ 150 fois l'énergie du séisme de magnitude 7,5 qui était prévu par la cartographie des aléas (Stein, Geller et al., 2012).

Si Haïti et le Japon sont des exemples de régions reconnus historiquement comme étant naturellement actives, de nombreux séismes destructeurs complètement inattendus peuvent aussi être directement attribués aux activités humaines, comme l'exploitation d'énergie souterraine telle que le pétrole dans l'état d'Oklahoma aux Etats-Unis (X. Zhang et al., 2020), le gaz de shiste dans le Bassin de Sichuan en Chine (Lei, D. Huang et al., 2017) ou bien la géothermie dans la région de Pohang en Corée du Sud (Grigoli, Scarabello et al., 2018 ; K. W. Chang et al., 2020 ; E. J. Lee et al., 2020). Nucléés à des profondeurs très faibles (Klose, 2010), ces séismes ont des impacts socio-économiques disproportionnellement élevés car ils sont généralement situés proche de zones urbanisées, dans des régions continentales stables, qui concentrent 90 % d'une population mondiale souvent peu préparée à endiguer un séisme destructeur (Kraft et al., 2009).

« *Earthquakes are a collective experience* »—Richard M. Allen, 2013

Chaque société est donc confrontée à un double-défi. Le premier est de comprendre les risques sismiques auxquels elle est confrontée pour en atténuer les effets, et donc de décider du niveau de sécurité à atteindre. Le deuxième est d'évaluer la capacité à se remettre d'événements extrêmes (R. Eyre et al., 2020 ; Markhvida et al., 1999).

De problèmes sanitaires post-sismiques qui peuvent se poser tels que la recrudescence de maladies infectieuses (Chin, 2011 ; Reina Ortiz et al., 2017), ou la perturbation du fonctionnement des hôpitaux (Almeida et al., 2020), de problèmes sociaux aussi comme l'aide nécessaire pour cultiver la résilience psychologique chez les survivants de séismes tragiques (Sasaki et al., 2019), et de problèmes économiques évidemment comme la répartition des budgets alloués pour la consolidation de bâtiments avec des structures parasismiques (Stein, Liu et al., 2017), chaque membre de la société est acteur du double-défi à relever. Et le scientifique en fait parti.

A travers l'étude de l'aléa sismique, le scientifique a un rôle central à jouer dans la compréhension des risques sismiques et leur évolution. Ainsi, si les séismes destructeurs peuvent balayer en un court instant les constructions humaines, ils peuvent balayer aussi des concepts scientifiques tenaces qui gouvernaient la compréhension humaine de ces derniers. La science des séismes tire donc les leçons de son propre objet d'étude : « Seismology : Shaking up earthquake theory » (Chui, 2009), « The lessons of Tohoku-Oki » (Avouac, 2011), « Why giant earthquakes keep catching us out » (Lay, 2012), « Beware of slowly slipping faults » (P. Z. Zhang, 2013). Ainsi, par exemple, le bien installé modèle de rebond élastique, introduit à la suite du séisme de San Francisco de 1906 (Reid, 1910), bien avant l'avènement de la tectonique des plaques, a fini par lâcher. Face à la variabilité de la récurrence des séismes, force est de constater que la périodicité ou la quasi-périodicité de cette récurrence est finalement un phénomène plutôt rare dans la nature (Matthews et al., 2002 ; Kagan et al., 2012 ; Y. Chen et al., 2020). Seulement, des nouveaux paradigmes scientifiques ont pu être mis à jour parce que l'émergence des nouvelles technologies d'observation, comme le suivi temporel de la position du sol par les capteurs GNSS, a permis d'apporter un regard neuf sur les phénomènes sismiques observés.

Cependant, ces changements de paradigmes scientifiques sont en fait difficiles à installer. Avec seulement un siècle d'histoire détaillée des séismes, une instrumentation sismique moderne pour enregistrer les mouvements du sol et des méthodes analytiques pour en extraire l'information développés que très récemment (1970-2000), la rareté des données d'observation a cristallisé pendant longtemps nos connaissances sur l'occurrence des séismes (interaction entre séismes, physique de la rupture et facteurs déclencheurs par exemple).

En conséquence, les cartes d'aléas sont souvent encore construites sur la base de postulats, comme le modèle du cycle sismique, qui ne reflètent pas le comportement non linéaire de l'occurrence des séismes (Stein, Geller et al., 2012). Il n'est alors pas très étonnant que des séismes dévastateurs inattendus aient révélé des zones qui abritaient un potentiel sismique auparavant sous-estimé (Lay et Kanamori, 2011 ; Lay, 2012).

Si la rareté des données d'observation fut un frein à la compréhension des phénomènes physiques qui sous-tendent l'occurrence des séismes, la technologie de la détection des séismes est désormais en pleine révolution (E. Z. Cochran et al., 2018 ; Kong et al., 2019 ; Bergen et al., 2019). Cette révolution prépare de nouvelles observations sans précédent sur les séismes et leurs impacts. En effet, des réseaux sismologiques denses à taux d'échantillonage élevé sont désormais aisément déployés (Li et al., 2018 ; Meng et al., 2018), le développement et l'installation de nouvelles générations de sismomètres portatifs bon marché sont en expansion (Clayton et al., 2015 ; Christensen et al., 2017), et la détection acoustique distribuée (DAS), une technologie émergente qui convertit la fibre optique en capteurs sismiques, est en plein essor (Lindsey et al., 2017 ; Williams et al., 2019). De grands volumes de données sont alors produits et disponibles (Yoon, Bergen et al., 2019). Par exemple, le centre de gestion de données des Institutions de Recherche Incorporée de Sismologie (IRIS) a actuellement archivé plus de 600 To de données sismologiques (IRIS-DMC Archive, 2020). Les progrès de la technologie informatique, avec l'augmentation de la puissance de calcul et de la mémoire de stockage, le traitement parallèle et distribué, le développement de nouveaux algorithmes d'exploration de données et de l'intelligence artificielle, rendent possible le traitement massif de toutes ces données (Yoon, Bergen et al., 2019).

La sismologie encaisse donc des changements rapides, radicaux et à multiples facettes, et doit se ré-inventer. Elle doit se réinventer car si la sismologie a toujours été gouvernée par la donnée, elle y croule dorénavant dessous, et beaucoup plus que ce que les chercheurs ne peuvent analyser en utilisant des méthodes conventionnelles. Le nouvel enjeu est donc de donner du sens à toute cette donnée, et les nouvelles technologies en sont le catalyseur principal. Il y a là la nécessité d'accepter que ce ne soit pas la physique qui, dans un premier temps, gouverne la réflexion scientifique mais plutôt la donnée... C'est à la fois un vrai challenge mais aussi une grande opportunité, une voie ouverte vers la créativité.

La sismologie doit aussi se réinventer parce-qu'avec l'essor des nouveaux sismomètres portatifs à faible coût, chaque citoyen peut désormais en installer un chez lui (E. Z. Cochran et al., 2018). Cela signifie alors que l'accès à la connaissance scientifique devient universelle, et n'est plus cantonnée à une seule élite. Et les attentes sociales en matière de science, de technologie et d'innovation n'ont jamais été aussi élevées (Sato et al., 2016).

Chaque citoyen peut donc être pleinement acteur de la découverte scientifique mais aussi contestataire de la science institutionnelle. A l'heure de la « science ouverte » et la « science citoyenne participative », pour une imprégnation radicale et durable des changements de paradigmes qui viennent rythmer le champ de la connaissance scientifique, et une diffusion positive et unanime de cette connaissance sur l'ensemble des membres de la société, il apparaît indispensable de cultiver un écosystème scientifique où acteurs de la gestion opérationnelle des données collectées, de la recherche fondamentale, ainsi que de l'ingénierie de l'aléa et du risque sismique forment un réseau intimement connecté. D'une catastrophe à l'autre, d'un bout à l'autre de la planète, les séismes sont une expérience collective.

Ce travail de thèse s'inscrit donc dans ce panorama, comme une petite brique posée. S'il ne traite pas directement des grands séismes, il se focalise plutôt sur les plus petits. Un non-sens ? Peut-être pas. Le chapitre 1 est alors consacré plus spécifiquement au développement de la problématique de recherche ainsi que les questions de recherche scientifiques qui y en découlent. Le chapitre 2 définit en quoi l'objet d'étude qui est choisi est un objet intéressant pour répondre aux questions de recherche soulevées dans le chapitre 1. Le chapitre 3 et 4 exposent la méthodologie de recherche mise en oeuvre pour répondre aux questions de recherche. Le chapitre 6 offre une conclusion partielle à ce travail et présente les nouvelles perspectives de recherche qui s'ouvrent.

Chapitre 2

Introduction

> « *How small is small enough ?* »—J.E. Ebel, 2008

Sommaire

2.1 Motivation générale : pourquoi détecter les petits séismes ?

2.1.1 Les petits séismes dominent les catalogues

La distribution de la taille des séismes pour une région donnée est généralement décrite par la relation empirique de Gutenberg-Richter (Ishimoto et al., 1939 ; Gutenberg et al., 1944). Cette relation indique que la fréquence des magnitudes des séismes suit une distribution exponentielle (Figure 2.1) :

$$\log N(M) = a - bM \quad \text{for } M \geq M_c \tag{2.1}$$

N(M) représente le nombre cumulé de séismes de magnitude égale ou supérieure à M ; M_c est la magnitude de complétude : tous les événements de magnitude $M \geq M_c$ sont supposés être enregistrés dans un catalogue donné. Le paramètre a décrit le niveau de sismicité global ou le niveau de productivité des séismes, qui peut varier largement d'une région à l'autre. La valeur b décrit la relation entre le nombre de petits et de grands séismes.

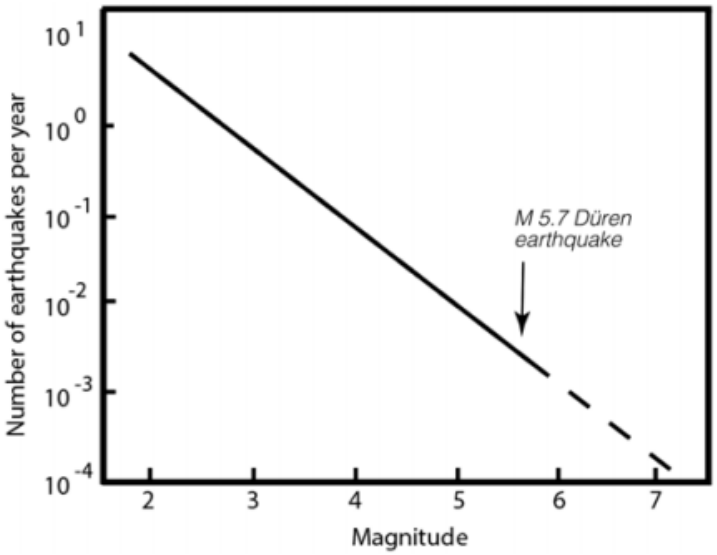


Figure 2.1: Relation empirique de Gutenberg-Richter établie pour le Nord-Ouest de l'Europe, dans la zone inférieure du Graben du Rhin (D'après Vanneste et al., 2013)

Une valeur b égale à 1 signifie que la fréquence des événements de magnitude M = 2 est dix fois celle des événements de magnitude M = 3 (Fiedler et al., 2018 ; Brodsky, 2019b). Si les jeux de données globaux et régionaux suivent souvent une distribution fréquence-magnitude avec b = 1, des variations locales de la valeur b entre 0.4 et 2.0 sont aussi observées (Wiemer et al., 2002).

Certains auteurs attribuent ces fluctuations de la valeur b à des artefacts résultant du sous-échantillonnage des événements, des erreurs de calcul des magnitudes et des capacités de détection non homogènes (Shi et al., 1982 ; Frohlich et al., 1993 ; Kagan, 1999 ; Kagan, 2002 ; Kagan, 2010 ; Amorese et al., 2010). D'autres auteurs considèrent ces variations spatio-temporelles comme une approximation de divers régimes tectoniques, de la contrainte de cisaillement ou bien de la pression interstitielle (Scholz, 1968 ; Wiemer et al., 1997 ; Wiemer et al., 2002 ; Schorlemmer et al., 2005 ; Bachmann et al., 2012 ; Tormann et al., 2014 ; Gulia, Tormann et al., 2016).

Déterminer statistiquement la distribution des magnitudes des séismes n'est en fait pas si simple que cela. En effet, de nombreux *a priori*, tels que la taille de la zone de mesure de la valeur b ou bien le choix de la valeur b habituelle pour une région donnée, sont nécessaires pour mener à bien une étude statistique sur la taille et la distribution des séismes. Ces différents postulats de départ représentent le talon d'Achille de ces études statistiques, d'autant plus qu'un catalogue de séismes est intrinsèquement incomplet (Brodsky, 2019b).

De ce fait, la valeur b extraite de la relation de Gutenberg-Richter est fortement dépendante du seuil inférieur de magnitude détectée (Mignan et Woessner, 2012 ; Godano et al., 2014). Certes, les petits séismes dominent systématiquement les catalogues de séismes du fait de la courte période d'enregistrement couverte par ces derniers (Hanks, 1992 ; Pacheco et al., 1992 ; Ross, Trugman et al., 2019). Cependant, ces catalogues étant très peu exhaustifs pour les gammes de petite magnitude, beaucoup d'autres petits séismes en sont inexorablement absents. Les raisons pour cela sont notamment l'hétérogénéité spatio-temporelle des réseaux de stations sismologiques et les seuils limites de détection (Helmstetter, 2005 ; Gulia et Wiemer, 2019).

L'amélioration des capacités de détection et de localisation d'un réseau sismologique d'un ou de deux ordre(s) de magnitude (par exemple de magnitude M 3.0 à M 2.0 puis M 1.0) peut augmenter d'un facteur 10 à 100 le nombre de séismes détectés par an. Seulement, abaisser le seuil de détectabilité d'un réseau implique un fort surcoût (coût de calcul ou charge manuelle de travail supplémentaire). Est-ce que ce coût en vaut le bénéfice ? **Autrement dit, est-ce qu'il est si important d'enregistrer et de traiter des sismogrammes de tous les séismes jusqu'à la magnitude 2.0 ? magnitude 1.0 ? Magnitude 0 ou en-dessous ?**

2.1.2 Les petits séismes cartographient plus finement le comportement sismique d'une région

•**Comportements sismiques long-terme**

Depuis le séisme de San Francisco de 1906, un des paradigmes dominants de la sismologie a été le cycle sismique décrit simplement par le modèle du rebond élastique : lors de la période intersismique, les contraintes s'accumulent peu à peu sur une faille verrouillée du fait du mouvement relatif des plaques ou des blocs qu'elle sépare ; lors de la phase cosismique, les contraintes sont relâchées par le glissement sur le plan de faille associé au séisme (Reid, 1910) (Figure 2.2a).

Ce modèle implique l'occurrence de séismes périodiques donnant lieu à une accumulation régulière de déplacements cumulés (Figure 2.2). Or, les longues séquences d'enregistrement des séismes, maintenant accessibles pour les régions à fort taux de déformation, montrent un comportement beaucoup plus complexe. En effet, dans la plupart des systèmes tectoniques actifs, on observe que les grands séismes ont lieu plus souvent en clusters regroupés dans le temps, alternant avec des intervalles de quiescence longs et variables (Wallace, 1987 ; Sieh et al., 1989 ; Agnon, 2014 ; D. Clark, McPherson et Van Dissen, 2012 ; D. Clark, McPherson, T. Allen et al., 2014 ; Ratzov et al., 2015 ; Salditch et al., 2020 ; Y. Chen et al., 2020). (Figure 2.2b et Figure 2.3).

Cette distribution des séismes en Escalier du Diable (Mandelbrot, 1982 ; Turcotte, 1997) est une caractéristique des systèmes dynamiques complexes. Cela suggère donc que ces motifs particuliers d'occurrence sismique manifestent le comportement de systèmes non linéaires élaborés à partir de multiples composants (i.e. les failles et/ou les segments de faille) qui interagissent entre eux (Liu et al., 2016). Ces interactions entre failles se traduisent par des transferts de contrainte statique, dynamique, et/ou viscoélastique ou bien des perturbations des conditions de chargement régional par des ruptures de failles locales.

De ce fait, chaque grand séisme ayant eu lieu sur une faille (ou segment(s) de faille) du système peut affecter les contraintes et les taux de chargement sur les autres failles (Dolan et al., 2007 ; Luo et al., 2012). Il n'est donc pas rare que des grands séismes rompent des segments de failles multiples comme cela a été le cas pour les séismes de Kunlun en 2001 (Mw 7.8, Chine), d'El Mayor-Cucapah en 2010 (Mw 7.2, Mexique) et de Kaikōura en 2016 (Mw 7.8, Nouvelle-Zélande) (Fletcher et al., 2017 ; Xu et al., 2018 ; Improta et al., 2019).

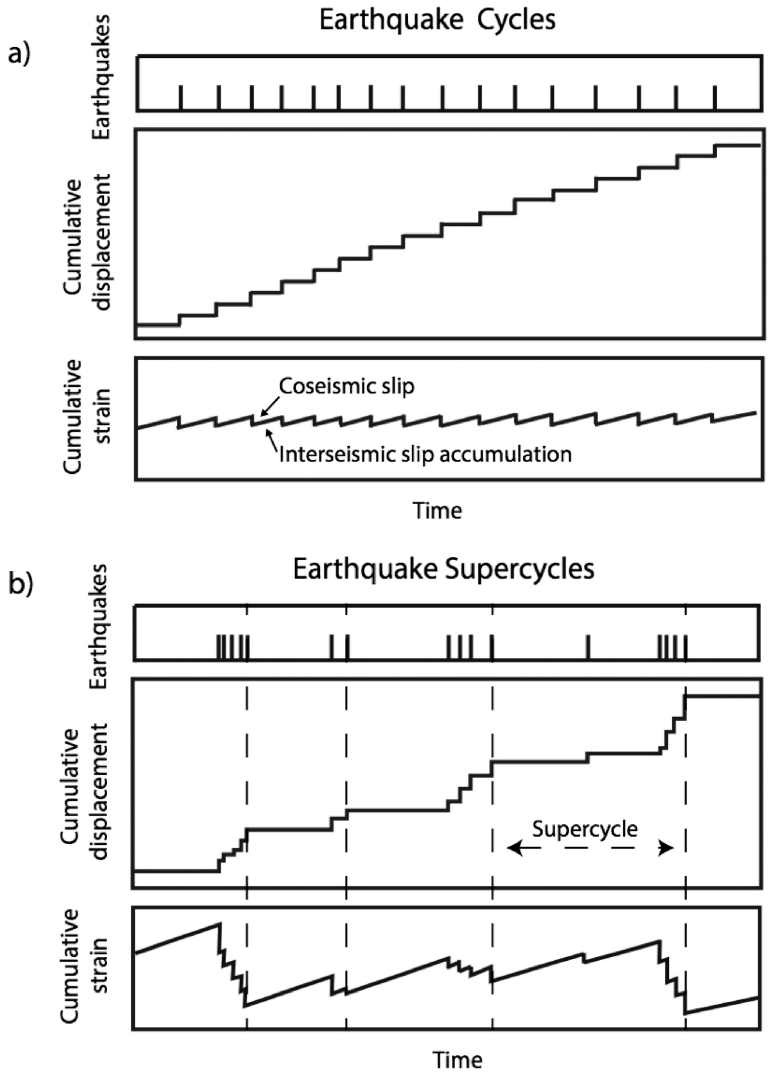


FIGURE 2.2: Comparaison globale des schémas d'occurrence des séismes, des déplacements et de la déformation cumulés sur une faille selon l'hypothèse classique du cycle sismique (a) et selon l'hypothèse des supercycles (b) (D'après SALDITCH et al., 2020)

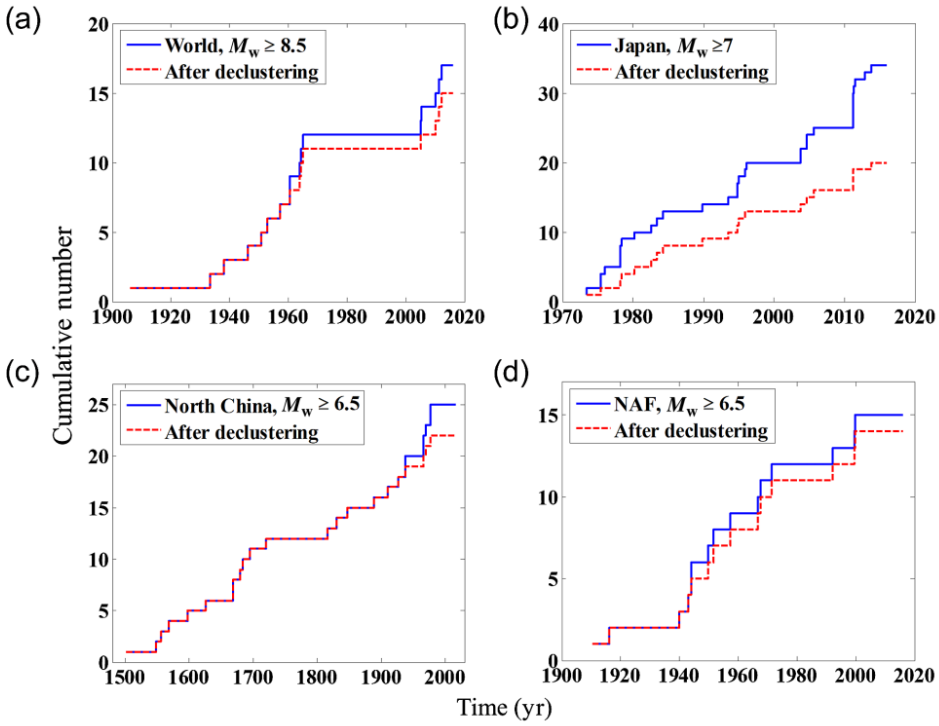


FIGURE 2.3: Motifs temporels des grands séismes (a) dans le monde, (b) au Japon, (c) dans le Nord de la Chine et (d) sur la faille Nord Anatolienne en Turquie (NAF). Les lignes continues représentent les catalogues globaux, et les lignes discontinues représentent les résultats après déclusterisation. (D'après Y. CHEN et al., 2020).

Chaque faille individuelle présente alors un taux de chargement spécifique et variable, qui est affecté par les séismes précédents ayant eu lieu soit sur cette même faille, soit sur d'autres failles du système. A cela s'ajoutent les perturbations de contraintes locales telles que l'érosion (Calais et al., 2010), le réajustement isostatique post-glaciaire (Ding et al., 2019), les variations de densité lithosphérique (Levandowski et al., 2017 ; Murphy et al., 2019) ou bien les fluides (Kumar et al., 2017). Ainsi, inscrit dans un système entier de failles en interaction, le comportement d'une faille court-terme est plus difficilement appréhendable que ce que le modèle de rebond élastique préludait.

De plus, les grands séismes sont finalement très peu fréquents, et les catalogues instrumentaux ont des périodes d'enregistrement courtes et sont intrinsèquement incomplets. Par conséquent, l'analyse de tels catalogues montre inéluctablement une vue biaisée de la sismicité long-terme (Salditch et al., 2020). Il est alors difficile d'estimer par exemple un temps de récurrence moyen des séismes ou bien d'identifier si les quelques événements enregistrés d'un catalogue se sont produits au sein d'un même cluster de séismes ou s'ils s'étendent à la fois sur la période d'activité du cluster et la période de quiescence (Stein, Liu et al., 2017 ; Y. Chen et al., 2020). La plupart des enregistrements actuels contiennent donc des séquences de séismes raccourcies qui sous-estiment le degré d'apériodicité des séismes. Ce qui rend d'autant plus difficile l'estimation de l'aléa sismique (Ellsworth et al., 1999 ; M, 2007).

Cette difficulté est amplifiée pour les **régions continentales intraplaques** où les grands séismes sont des phénomènes encore plus rares, voire inexistants, et où les intervalles de quiescence sont beaucoup plus longs que dans les zones interplaques. Ces régions se déformant très lentement, les données historiques et paléosismologiques, combinées aux données néoctectoniques, sont indispensables pour comprendre la chronologie des déformations sismiques, et estimer l'aléa sismique (T. I. Allen, 2020 ; D. J. Clark et al., 2020).

Des indices de déformation de surface associés à l'occurrence de séismes passés peuvent être effectivement préservés sur des milliers d'années. Seulement, si l'intervalle entre deux séismes engendrant des ruptures en surface est beaucoup plus long que les processus d'érosion et de sédimentation qui viennent modeler le paysage actuel, les traces d'activité de faille datant d'avant la fin du Quaternaire sont alors perdues (Walker et al., 2015 ; Abdrakhmatov et al., 2016).

Par exemple, entre 1968 et 2018, même si 90% des séismes enregistrés dans le craton australien ont engendré des déformations de surface, ces derniers ne peuvent être associés à aucune évidence néotectonique d'activité de faille. Seulement, on ne peut pas exclure la possibilité que des marqueurs de rupture précédente aient été supprimés (King et al., 2019). L'absence d'indicateurs de déformations de surface n'est pas donc un révélateur formel d'inactivité d'une faille sur le long terme.

Par conséquent, assigner un label « actif/inactif» à une faille (ou segment(s) de faille) dans ces zones continentales intraplaques, basé sur l'occurrence (ou la non occurrence) d'un séisme dans les quelques derniers milliers d'années n'est pas un indicateur robuste de futur potentiel sismogénique (D. Clark et Mcpherson, 2011 ; D. Clark, McPherson et Van Dissen, 2012 ; Boncio et al., 2018).

Sans apport conséquent d'études paléosismologiques et néotectoniques approfondies, il est alors difficile d'estimer l'aléa sismique dans ces régions continentales stables où les témoins de déformation de surface se font rares et où les mesures géodésiques le long des failles ne décèlent pas d'indices forts de déformation cumulée (Grutzner et al., 2017 ; Vallage et al., 2020).

Par ailleurs, la grande incertitude associée aux localisations des séismes des catalogues instrumentaux ne permet pas non plus d'établir clairement une relation entre les hypocentres et les failles projetées en surface, ni d'accéder finement à la géométrie 3D de ces failles, à supposer que ces séismes soient inscrits sur le même plan (D. J. Clark et al., 2020 ; Ross, E. S. Cochran et al., 2020). De plus, l'hétérogénéité des catalogues sismiques (calcul des magnitudes, réseaux sismiques évoluant) ainsi que leur courte période d'enregistrement révèlent une cartographie de la sismicité largement incomplète.

Outre la faible représentativité des données sismologiques, l'absence de marqueurs morphotectoniques et géodésiques forts de déformation ainsi que la complexité des systèmes de failles qui accommodent cette faible déformation (Matos et al., 2018), l'aléa sismique sera d'autant plus difficile à estimer que le comportement des séismes s'éloigne du comportement poissonnien (Y. Chen et al., 2020 ; Vallage et al., 2020), que l'évaluation de la magnitude maximale associée au futur plus grand séisme est plus que spéculative (Neely et al., 2018), et que la présence encore active de répliques associées à des chocs principaux historiques ou préhistoriques n'est pas encore robustement établie (Toda et al., 2018).

Enfin, aucun mécanisme n'est universellement accepté pour expliquer le déclenchement et le comportement des séismes dans ces régions continentales stables (Soto-Cordero et al., 2018 ; Gallen et al., 2018 ; Bezada et al., 2019 ; Leclere et al., 2019. De ce fait, en l'absence de théorie robuste sur l'occurrence des séismes dans ces régions, l'incertitude épistémique dans la caractérisation des sources sismiques pour l'évaluation de l'aléa dans ces zones restera élevée (Griffi et al., 2020).

C'est donc sur ce dernier point crucial que va intervenir la détection des petits séismes. Comme cela a été soulevé par Brodsky, **2019a**, nous sommes souvent conduits en sismologie à analyser de près les quelques exemples que la nature nous fournit à intervalles irréguliers. Par conséquent, nous devons manier avec prudence ces quelques données rares avant d'en tirer des conclusions générales hâtives (Brodsky, **2019a**). Les grands séismes sont effectivement extrêmement rares par rapport à l'abondante sismicité de magnitude faible qui est enregistrée à travers le globe. De part la haute fréquence d'occurrence de ces séismes de faible magnitude, mieux les détecter permettrait d'apporter plus de robustesse statistique aux comportements sismiques court-terme observés, afin de mieux comprendre leur origine. Ceci est particulièrement important pour les zones continentales intraplaques, où la sismicité de magnitude faible à modérée reste difficile à expliquer puisque les taux de déformation y sont très faibles. Cependant, comprendre l'origine de cette microsismicité et son rôle dans la description du comportement sismique d'une région présente un réel enjeu du point de vue du risque sismique, étant donné que ces régions concentrent 90 % de la population mondiale (Hirose et al., **2014**).

•Comportements sismiques court-terme

Une détection plus fine des petits séismes permet de rendre visible entièrement des séquences de séismes qui étaient auparavant fragmentées (Brodsky, 2019b). Cette détection plus continue un niveau supérieur d'interactions entre les séismes, à la fois au sein même d'une séquence individuelle (essaims ou répliques par exemple), mais également entre les différentes séquences enregistrées (par exemple : précurseurs/chocs principaux, chocs principaux/répliques, chocs principaux/post-séismes déclenchés de façon dynamique ou statique) (Figure 2.4).

Ces interactions plus complexes sont non seulement capables de connecter des séquences ou des séismes qui étaient à première vue isolé(e)s, mais peuvent surtout renseigner sur les détails des processus physiques qui sous-tendent l'initiation, le déclenchement ou bien la migration spatio-temporelle de l'ensemble des séismes repérés (Ross, Trugman et al., 2019).

La détection plus fine des petits séismes rend donc possible la description plus précise de l'évolution spatio-temporelle de séquences de séismes spécifiques, y compris par la prise en compte des événements précoces voire précurseurs. Dans les paragraphes suivants, je m'intéresse à quelques séquences individuelles de séismes, qui sont intensément étudiées dans la littérature.

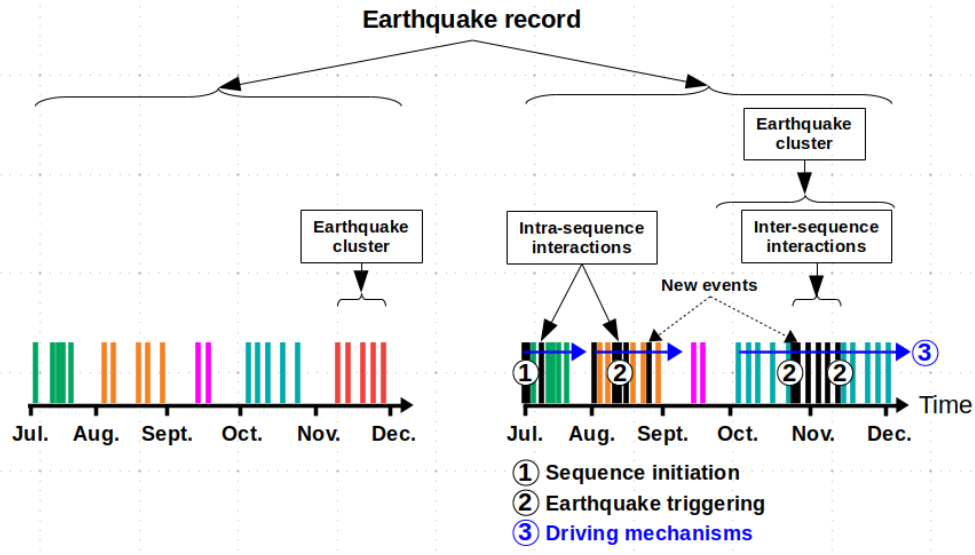


FIGURE 2.4: Enregistrement sismique théorique sur 6 mois (de juillet à décembre) avant l'amélioration de la détection des petits séismes (à gauche) et après (à droite). Chaque amas de séismes est représenté par une couleur. Les traits noirs correspondent aux nouveaux séismes détectés. Cette détection supplémentaire permettrait par exemple de : (1) mieux résoudre l'initiation d'une séquence individuelle de séismes et d'en définir son mécanisme déclencheur ; (2) compléter une séquence individuelle de séismes et comprendre leur degré d'interaction ainsi que leur mécanisme de déclenchement, (3) relier des séquences isolées entre elles et investiguer plus sur les mécanismes qui contrôlent l'évolution spatio-temporelle de l'ensemble de ces séismes.

Les répliques. Une quantité importante des répliques qui ont lieu après un choc principal est souvent absente des catalogues de séismes existants, surtout si elles sont de faibles magnitudes ou trop nombreuses (superposition des signaux associés empêchant leur utilisation, Kagan, 2004 ; Peng, Vidale et al., 2006). Pourtant, réussir à les detecter puis les localiser en plus grand nombre amènerait des contraintes importantes sur la géométrie du plan de faille sur lequel le choc principal a eu lieu (Bulut et al., 2007 ; Yang et al., 2009 ; Peng et Zhao, 2009) et sur l'extension latérale et en profondeur du segment qui a rompu (C. H. Chang et al., 2007 ; Peng et Zhao, 2009 ; Yang et al., 2009 ; Yin et al., 2018).

Une meilleure description de l'évolution spatio-temporelle des répliques est donc essentielle pour d'une part comprendre les mécanismes physiques qui contrôlent leur déclenchement (Enescu et al., 2007), et pour d'autre part suivre la déformation post-sismique autour de la zone de rupture associée au choc principal, pouvant impliquer d'autres segments de failles voisins (Hsu et al., 2006 ; C. H. Chang et al., 2007). Les mécanismes physiques à l'origine des répliques font encore l'objet de nombreux débats (Lippiello et al., 2015). En utilisant des événements de plus faible magnitude, une meilleure robustesse statistique peut être atteinte, saisissant plus finement les conditions qui définissent l'état de contrainte crustale comme par exemple celles de la pression interstitielle lithostatique (Shebalin et al., 2017).

La plupart des séquences de répliques sont relativement transitoires, le taux d'occurrence décroissant au fil des jours, des mois ou des années avant d'atteindre les niveaux de fond, en suivant globalement la loi d'Omori (Utsu et al., 1995) et la loi empirique de Bath (Shearer, 2012). Cependant, dans certaines zones intraplaques, des groupes d'événements persistants peuvent se produire sur des échelles de temps beaucoup plus grandes, comme la séquence en cours dans la zone de New Madrid dans l'est des États-Unis (J. Wang, Main et al., 2017). Seulement dans ce cas, l'étude court-terme de séquences de petits séismes ne permettra pas d'apporter plus de contraintes sur l'étude de ces séquences persistantes si particulières.

Les essaims sismiques. Les essaims sismiques sont des séquences de séismes concentrées dans le temps et l'espace sans aucun choc principal évident (Vidale et al., 2006). Ces essaims peuvent se produire dans des régions volcaniques (De Barros, Bean et al., 2013 ; McNutt, 2005), des régions à faible taux de déformation (Hainzl, 2004), le long de failles glissant asismiquement (Llenos et al., 2009 ; Roland et al., 2009), dans les zones de subduction (Vallee et al., 2013), ou lors des stimulations hydrauliques anthropiques des réservoirs (Keranen et al., 2018 ; Wei et al., 2015).

La question de savoir pourquoi la sismicité se développe comme un essaim, plutôt que comme une séquence de choc principal/répliques, est fondamentale. Grâce à une meilleure détection des petits séismes, la résolution spatio-temporelle plus fine des essaims de séismes aiderait alors à mieux trancher sur les facteurs réels qui les déclenchent (Hatch et al., 2020). En effet, des facteurs comme la pression des fluides (Vidale et al., 2006 ; Hainzl et al., 2012 ; Shelly et al., 2016) ou bien le glissement asismique (Delahaye et al., 2009 ; Hirose et al., 2014) sont souvent évoqués. Seulement, des études récentes ont montré que l'activité des essaims serait en fait contrôlée par les deux facteurs à la fois : des phases d'accumulation de pression de fluide déclencheraient un glissement asismique, qui lui-même induirait des séquences de sismicité à migration rapide (Bourouis et al., 2007 ; Guglielmi et al., 2015 ; De Barros, Guglielmi et al., 2018 ; Cappa et al., 2019 ; De Barros, Cappa et al., 2020).

De plus, une étude plus détaillée des essaims sismiques permettrait de mieux comprendre leur rôle dans les mécanismes précurseurs de futurs grands séismes (Brodsky et Lay, 2014 ; Rhoades, 2010). Indicateurs de glissement lent, une image détaillée de ces derniers pourrait également contribuer à suivre finement la naissance et l'évolution d'un glissement lent (Nadeau et al., 1998 ; A. Kato et al., 2014 ; Reverso et al., 2016 ; Nishikawa et al., 2017).

Les précurseurs. La détection accrue des petits séismes pourrait apporter là encore une robustesse statistique quant à la présence d'activité sismique précurseur (Goebel et al., 2013 ; Malin et al., 2018). Actuellement, la valeur pronostique des précurseurs est fortement débattue : des précurseurs seraient observés pour seulement 10 à 50% des chocs principaux étudiés (Mori et al., 1997 ; X. Chen et al., 2016 ; Marsan et al., 2014). De ce fait, établir des catalogues de séismes de haute résolution (meilleure précision des paramètres hypocentraux et magnitude de complétude plus faible) représente un double-enjeu. Le premier est la possibilité de rechercher systématiquement une activité précurseur de façon à en estimer la fréquence réelle dans la nature (Martinez-Garzon et al., 2019 ; Trugman et al., 2019 ; Ende et al., 2020). Le deuxième enjeu concerne l'approfondissement des connaissances relatives aux mécanismes physiques qui participent à l'occurrence des précurseurs et leur lien avec les chocs principaux. En effet, pour l'instant, deux écoles de pensées s'affrontent (Mignan, 2014) : l'école déterministe qui affirme que les précurseurs constituent une réponse à un glissement précurseur sur une faille (ou segment(s) de faille) comme par exemple un glissement lent (Bouchon, Durand et al., 2013 ; Tokuda et al., 2019 ; Yao et al., 2020), et l'école stochastique qui postule que les précurseurs font partie d'un processus naturel de déclenchement des séismes en cascade par transfert de contraintes inter-séismes (Gulia et Wiemer, 2019 ; Pino et al., 2019).

Les séismes déclenchés dynamiquement. De grands séismes peuvent déclencher dynamiquement, par propagation des ondes résultant du choc principal, d'autres séismes plus distants. Une détermination plus poussée des petits séismes pourrait révéler une sismicité de plus faible magnitude reliée indirectement à un plus grand séisme comme cela a été le cas dans le Sud de la Californie après le séisme d'El Mayor-Cucapah en 2010 (Ross, Trugman et al., 2019).

...Et toutes les autres séquences de séismes identifiées. Ainsi, révéler des séquences spatio-temporelles de séismes plus complexes pourraient mettre en évidence plus systématiquement plusieurs mécanismes moteurs de génération des séismes comme l'association d'un glissement asismique avec une diffusion de pression de fluide (Ross, Rollins et al., 2017). La complexification de ces séquences spatio-temporelles identifiées par une détection plus accrue des petits séismes constitue donc le terrain idéal pour une caractérisation plus aboutie des facteurs qui déclenchent les séismes en général, que ce soit par des contraintes différentielles transitoirement élevées (Jamtveit et al., 2018 ; Levandowki et al., 2018) ou bien des mécanismes locaux d'affaiblissement (par exemple, une pression élevée du fluide interstitiel, Gardonio et al., 2018).

La clusterisation des séismes est donc une des caractéristiques dominantes de la sismicité naturelle et anthropique (Ross, Trugman et al., 2019). Les types les plus étudiés de clusterisation incluent les répliques, les précurseurs, les essaims (Zaliapin et al., 2008). A travers l'étude plus approfondie des petits séismes, l'analyse plus précise d'un cluster de séismes ou une combinaison de plusieurs d'entre eux constitue une des perspectives majeures pour comprendre la redistribution et/ou le transfert des contraintes sismiques, ainsi que leur origine, la genèse des séismes et la dynamique globale de la lithosphère en somme (Romanowicz et al., 1993).

2.1.3 Avantages des bases de données actuelles pour la détection des petits séismes

Ces dix dernières années, des réseaux sismologiques se sont densifiés et produisent des données de meilleure qualité avec un contenu fréquentiel beaucoup plus large (Jousset et al., 2018). Les projets japonnais Hi-net, américains USArray ou européens AlpArray sont des exemples marquants d'instrumentation de vastes régions par des réseaux sismologiques très denses (Hetényi et al., 2018 ; Fuchs et al., 2019 ; T. Zhou et al., 2020). Initialement destinés à l'imagerie des structures profondes, ces réseaux représentent également une bonne opportunité de réduire la magnitude de complétude d'une région donnée et de construire des catalogues sismiques de qualité. Il en est de même pour les récents réseaux denses de capteurs type nodes ou de capteurs bas-coût, qui se multiplient partout dans le monde.

En conséquence, de hauts volumes de données de qualité sont désormais disponibles, et offrent une opportunité unique d'obtenir une image de la sismicité plus haute résolution (Bouchon, Karabulut et al., 2011 ; H. Kato et al., 2012 ; Schaumberg et al., 2020). Par exemple, au 1er juin 2020, le Centre de Gestion de Données des Institutions de Recherche Incorporée pour la Sismologie (IRIS DMC) a stocké près de 650 Terabytes de données sismologiques (Figure 2.5).

Cependant, alors que la quantité de données acquises par les réseaux de stations toujours plus denses augmente continuellement, la qualité des données reste en fait toujours entravée par la présence de bruit systématique enregistré aux stations et un échantillonnage spatial souvent biaisé (sources et stations sismiques inégalement réparties, P.-F. Chen et al., 2019).

Même si des réseaux plus denses proches des sources cibles ont des capacités de détection plus élevées, un ensemble de paramètres tels que la qualité du réseau (niveau de bruit, distribution spatiale), les caractéristiques de la source, les effets de propagation des ondes et le système d'acquisition vont venir limiter les capacités réelles de détection (Kwiatek et al., 2016).

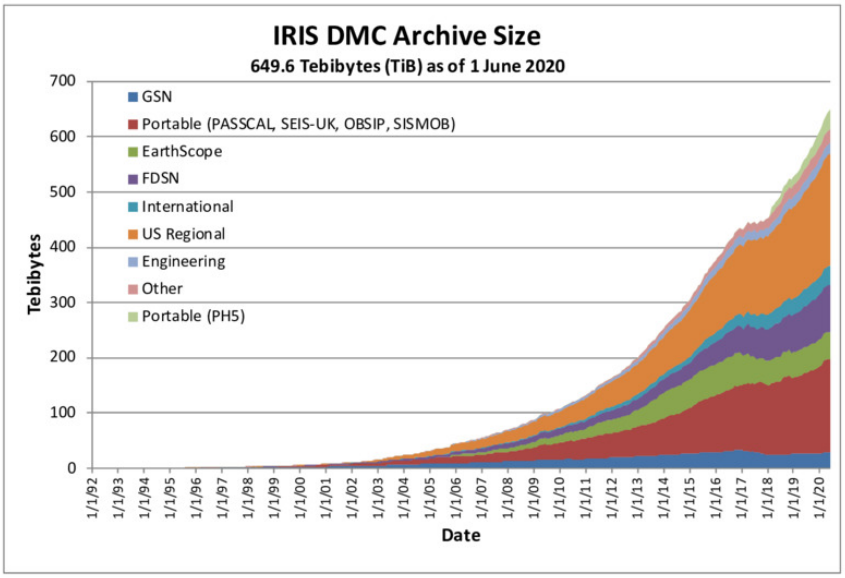


FIGURE 2.5: Evolution temporelle de l'archivage des données sismologiques au sein du Centre de Management des Données des Institutions de Recherche Incorporée pour la Sismologie (IRIS DMC) depuis 1992. (D'après http://ds.iris.edu/files/stats/data/archive/Archive_Growth.jpg)

De cette façon, si de hauts volumes de données disponibles sont capables d'offrir une mine d'or d'informations sur les petits séismes, il semblerait que cette masse de donnée ne garantisse pas une récupération optimale de cette information.

2.2 Limitations à la détection des petits séismes

2.2.1 Présentation générale du système de détection utilisé dans les observatoires sismologiques

Les deux systèmes de détection les plus utilisés dans les observatoires sismologiques sont Earth Worm (EW ; C. Johnson et al., 1995 ; http://www.earthwormcentral.org) et SeisComP3 (SC3 ; http://www.seiscomp3.org). Conçus pour recevoir les flux de données en temps réel, ces deux systèmes utilisent des algorithmes de détection basés sur l'amplitude du signal enregistré, archivent les données de manière continue, et déterminent automatiquement une localisation ainsi qu'une magnitude pour chaque événement détecté (Utheim et al., 2014). Cette localisation est généralement revue par un analyste.

L'algorithme principal de détection utilisé pour déceler les évènements sismiques est basé sur le calcul des valeurs moyennes de l'amplitude absolue d'un signal sismique sur deux fenêtres temporelles mobiles consécutives. La fenêtre temporelle courte (STA) est sensible aux événements sismiques tandis que la fenêtre temporelle longue (LTA) fournit des informations sur l'amplitude temporelle du bruit sismique à une station donnée (R. Allen, 1978 ; R. Allen, 1982). Un rapport des deux valeurs moyennes estimées sur ces deux fenêtres, le rapport STA/LTA, est calculé. Ce rapport est comparé en continue à une valeur-seuil définie par l'utilisateur, le niveau seuil de déclenchement STA/LTA (Trnkoczy, 1999). Si le rapport excède ce seuil à un nombre de stations données, un déclenchement est déclaré et un pointé, qui correspond au temps d'arrivée des ondes sismiques (principalement ondes de volume P et S) est créé (Figure 2.6A).

Un ensemble de temps d'arrivée des différentes phases sismiques (P et S) pour chaque station est donc généré. Un algorithme d'association va par la suite nucléer et localiser les événements sismiques (Yeck et al., 2019). L'association d'événements consiste à rassembler les pointés de différentes stations dans une certaine fenêtre temporelle. Si le nombre de pointés dans cette fenêtre temporelle est supérieur à un seuil prédéfini, l'algorithme d'association va relier les différents pointés (les premiers temps d'arrivée des ondes P et S) à une localisation hypocentrale approximative (Grigoli, Scarabello et al., 2018). Si cette procédure réussit, un événement sismique est déclaré (Figure 2.6B et C).

Une fois que cet événement est déclaré, sa localisation peut être affinée grâce à des méthodes de localisation plus avancées (par exemple NonLinLoc, Lomax, Virieux et al., 2000) et des modèles de vitesse plus détaillés.

Les méthodes de localisation basées sur le pointé du temps d'arrivées des ondes sismiques reposent sur la minimisation des résidus entre les temps d'arrivée théoriques et observés des ondes de volume (ondes P et S), et utilisent des algorithmes d'inversion itératifs (Thurber, 1985) ou globaux (Lomax, Virieux et al., 2000).

Ces trois étapes essentielles (pointés, association, localisation) constituent le coeur du système de détection standard classiquement utilisé dans la plupart des organismes en charge de la surveillance sismique.

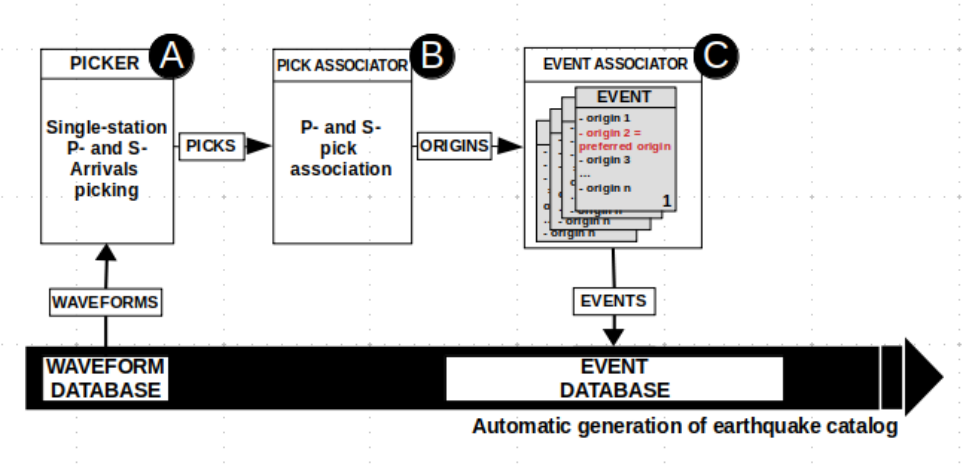


Figure 2.6: Procédure de détection sous SeisComP3. (A) Pointés automatiques des temps d'arrivées des ondes P et S avec le module *Scautopick*. (B) Association des pointés automatiques P et S avec le module *Scautoloc* et/ou *Scanloc*. Chaque association réussie engendre la création d'une origine qui est localisée avec l'algorithme LOCSAT. (C) Association des origines produites à un événement spécifique avec le module *Scevent*. Chaque événement présente une origine préférentielle fixée. La procédure SeisComP3 génère automatiquement un catalogue multi-origine. Les événements peuvent être localisés plus finement avec le module *Screloc* par exemple en utilisant un autre algorithme de localication (comme NonLinLoc) et/ou un autre modèle de vitesse.

Même si les méthodes de détection et de localisation des séismes basées sur la forme d'onde sont également largement utilisées en sismologie et très efficaces pour détecter les petits séismes (Kao et al., 2004 ; Grigoli, Cesca et al., 2013 ; Grigoli, Scarabello et al., 2018 ; Pesicek et al., 2014 ; Yoon, O'Reilly et al., 2015 ; Yoon, Y. Huang et al., 2017 ; M. Zhang et al., 2015 ; Wei et al., 2015 ; Tong et al., 2016 ; Perol et al., 2018), elles restent cependant très coûteuses en calcul (Z. Zhang et al., 2019). Par conséquent, les méthodes basées sur le pointé des temps d'arrivées des ondes sismiques restent toujours dominantes pour les opérations routinières de surveillance des séismes en temps réel (Grigoli, Scarabello et al., 2018 ; Z. Zhang et al., 2019).

2.2.2 Première limitation : le processus d'association, goulot d'étranglement des systèmes de détection

Une des principales lacunes des systèmes de détection standard est la possibilité de détecter tout type de signal transitoire impulsif, autre que ceux associés aux séismes (Ross, Trugman et al., 2019). En effet, les algorithmes d'association standards sont principalement basés sur l'information apportée par les temps d'arrivées (Figure 2.7). Par comparaison entre les temps d'arrivées observés et les temps théoriques calculés à partir d'un modèle de vitesse aux différentes stations, ces derniers associent les arrivées dans une fenêtre temporelle qui semblent compatibles avec une source réaliste (McBrearty et al., 2019).

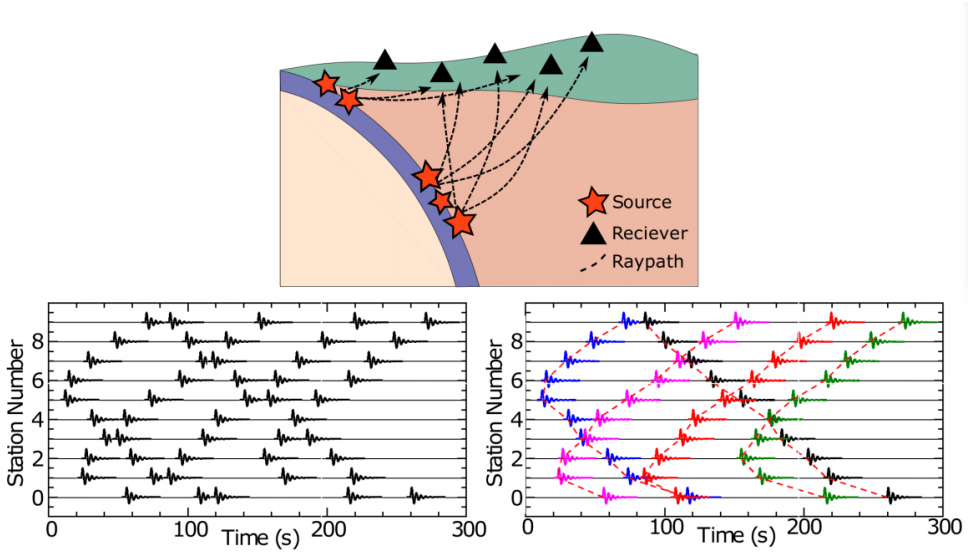


Figure 2.7: Exemple schématique du processus d'association basée sur le temps d'arrivée des ondes sismiques. En haut : Plusieurs sources sismiques, localisées sur une interface de subduction, produisent des ondes impulsives qui se propagent aux cinq stations sismologiques. En bas à gauche : ensemble des arrivées observées à travers le réseau sismique avant association (le signal est ici monophasé). En bas à droite : ensemble des arrivées correctement associées, colorées pour les cinq sources distinctes et reliées par une courbe (ligne pointillée rouge). D'après McBrearty et al., 2019.

Seulement, ce processus d'association perd progressivement de l'efficacité et de la précision à mesure que les algorithmes de détection deviennent plus sensibles (C. E. Johnson et al., 1997 ; McBrearty et al., 2019). Quand les seuils de détection sont effectivement abaissés pour détecter les événements de plus faible magnitude, les algorithmes de détection sont confrontés à de plus faibles rapports signal sur bruit. De cette façon, ils deviennent sensibles à la moindre irrégularité, impulsivité ou hausse d'amplitude véhiculée par le bruit sismique ambiant enregistré.

Par conséquent, en plus d'une augmentation de la quantité de pointés correspondant aux phases P et S détectés, de nombreux pointés correspondant à du bruit transitoire impulsif sont aussi produits. Ces faux pointés sont alors traités comme s'ils correspondaient aux arrivées d'ondes sismiques propagées depuis une source sismique (Figure 2.8).

En plus de la diminution du seuil de détection, le taux de pointés automatiques sera également amplifié de par l'augmentation des volumes de flux de données à traiter provenant des réseaux sismologiques plus denses. Ce taux sera aussi exacerbé en contexte urbain, là où le niveau de bruit d'origine anthropique enregistré est très intense (Díaz et al., 2017 ; Poli et al., 2020). Dans ces environnements, une grande quantité de signaux d'origine anthropique est générée, créant alors de nombreux pointés supplémentaires.

De ce fait, alors qu'aucune forme d'onde n'est utilisée pour affiner le processus d'association, des jeux de pointés d'une grande variabilité, créés sur des fenêtres temporelles très courtes, sont très facilement associés. La proportion de fausses associations produites à partir du bruit transitoire est alors fortement augmentée, de même que la proportion d'associations provoquées par des événements autres que les séismes (tirs de carrière, activités géothermiques, glissements de terrain par exemple).

De plus, sur la base uniquement de temps d'arrivée compatibles sur une fenêtre temporelle donnée, la probabilité d'associer des arrivées reliées à des phases sismiques avec des arrivées reliées à du bruit est largement accrue (Figure 2.8).

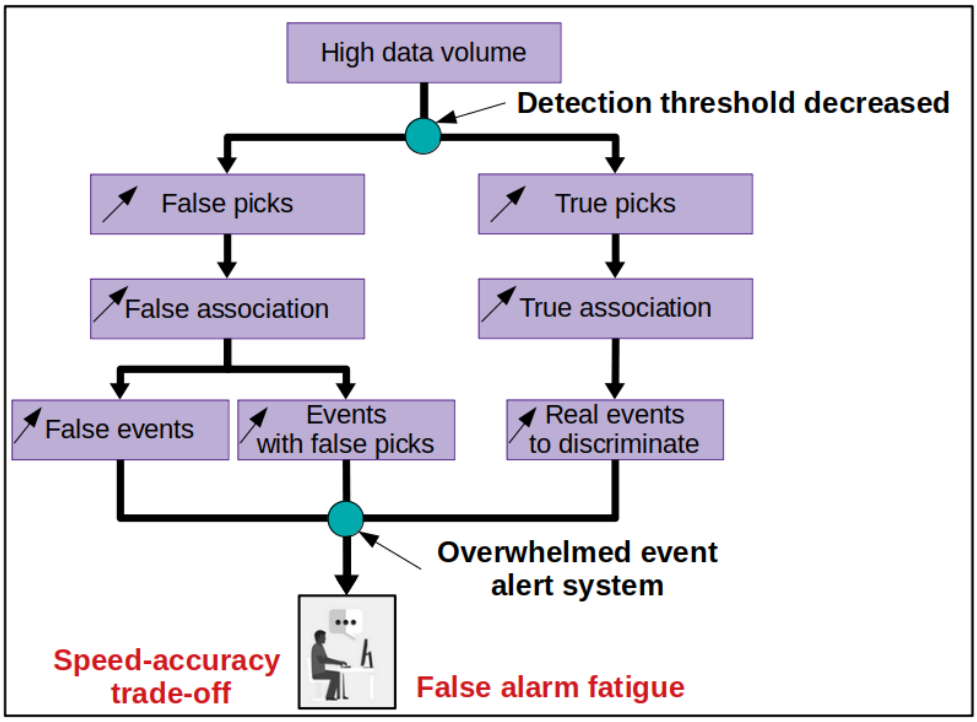


FIGURE 2.8: Hauts volumes de données et seuils de détection diminués dans les observatoires sismologiques : un encombrement rapide des systèmes d'alerte des événements.

2.2.3 Deuxième limitation : une hausse des détections parasitaires

Avec la sensibilité croissante des algorithmes de détection et la hausse des volumes de données à traiter, la détection standard des petits séismes engendre un taux considérable de détections d'événements autres que ces séismes (Figure 2.9). Et le processus d'association en est une des principales causes.

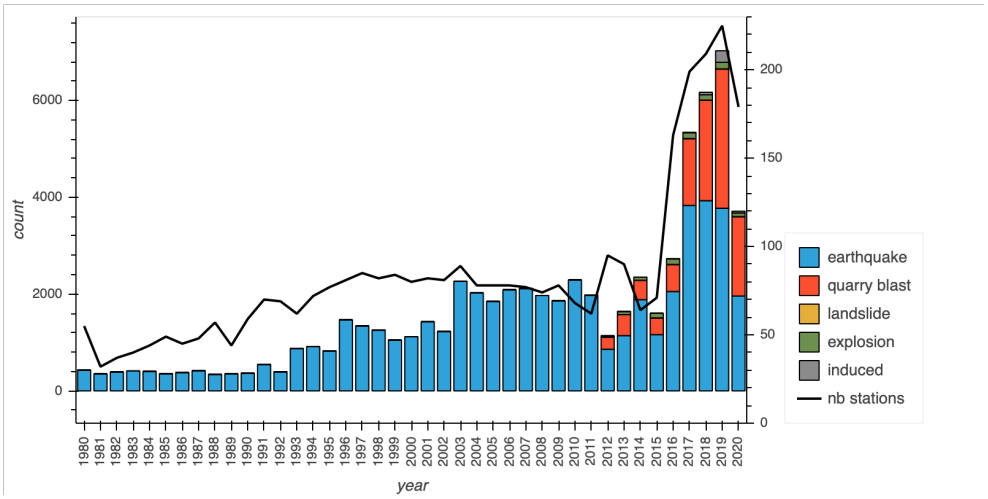


FIGURE 2.9: Détection des événements par le réseau de surveillance sismique BCSF-RéNaSS depuis 1980. L'année 2012 marque le début de l'intégration des tirs de carrière dans le catalogue après discrimination manuelle. La densification du réseau de stations dont les données sont intégrées au système de localisation depuis 2014 (courbe noire) a conduit à une détection d'environ 2 fois plus de séismes (en bleu) et 10 fois plus de tirs de carrière (en rouge), si l'on prend comme référence l'année 2012. Un peu plus anecdotique en termes de nombre, un nombre croissant d'explosions (en vert) ainsi que de sismicité induite (en gris) par l'activité géothermique (plus particulièrement de la région Grand-Est) est aussi détecté. La période 2016-2019 correspond à la période de déploiement du réseau temporaire AlpArray dont les stations françaises, allemandes, belges, italiennes et suisses ont été utilisées pour la localisation manuelle des événements.

Des milliers de faux événements provoqués par un pointé quasi-systématique de bruit transitoire impulsif sont d'abord aisément détectés (Figure 2.10).

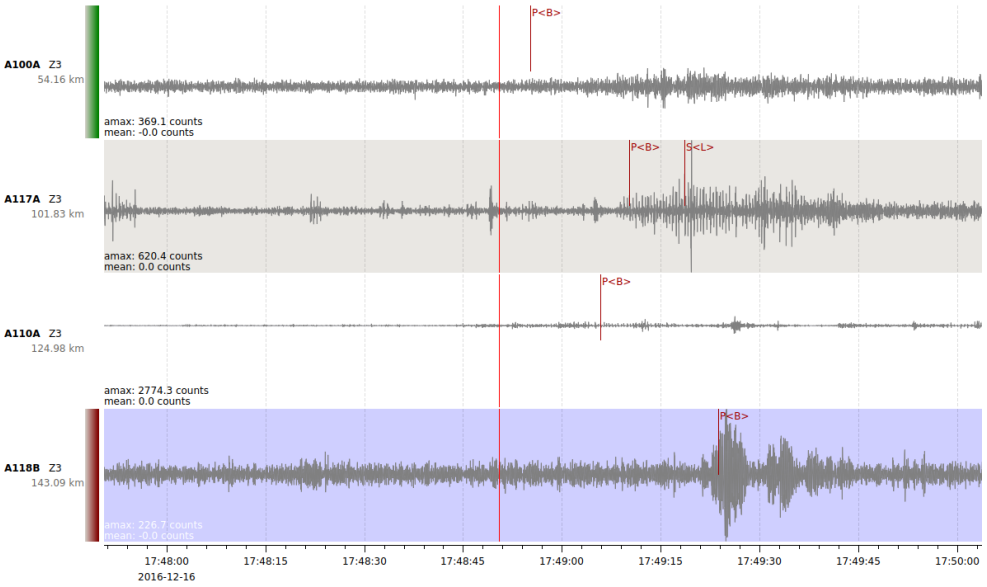


FIGURE 2.10: Exemple de faux événement détecté par l'association de bruit pointé pour quatre stations sismiques dans une fenêtre temporelle compatible pour engendrer une détection.

En conséquence, des catalogues massifs d'événements contaminés par l'existence de faux événements liés au bruit sont générés. De plus, parmi ce taux considérable de faux événements à traiter, des quantités non négligeables de vrais événements - comme les séismes naturels ou d'origine anthropique - se trouvent facilement pollués par du bruit. Et cette pollution est d'autant plus fréquente si le paramétrage associé à la qualité des pointés automatiques et du processus d'association n'est pas affiné spécifiquement en fonction du niveau de bruit à chaque station et des vitesses de propagation des ondes sismiques dans le milieu (Figure 2.11).

Dans cette configuration de réseaux denses, revoir manuellement tous les événements issus des catalogues automatiques devient une tâche nettement plus difficile à accomplir.

En outre, en plus de ces derniers événements, d'autres événements, de même ordre de magnitude que les séismes, s'ajoutent à la liste déjà trop longue d'événements à discriminer manuellement. Dans les environnements très urbanisés, ces événements supplémentaires sont principalement d'origine anthropique, et particulièrement des tirs de carrière. La détection opérationnelle des petits séismes dans ces contextes implique donc de discriminer manuellement les séismes (Figure 2.12) d'autres événements comme les tirs de carrière (Figure 2.13).

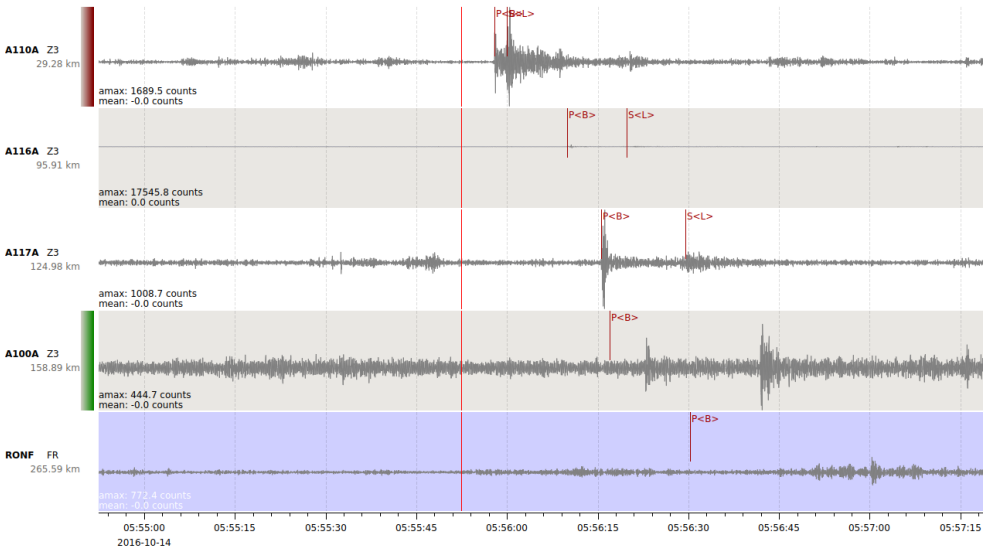


FIGURE 2.11: Exemple de fausse association ayant généré un vrai événement contaminé par du bruit pointé à la station RONF (en bleu). Si cet événement n'est pas nettoyé manuellement par un analyste, ce dernier sera conservé en tant qu'événement dans le catalogue, même si les incertitudes des paramètres hypocentraux risquent d'être significatifs.

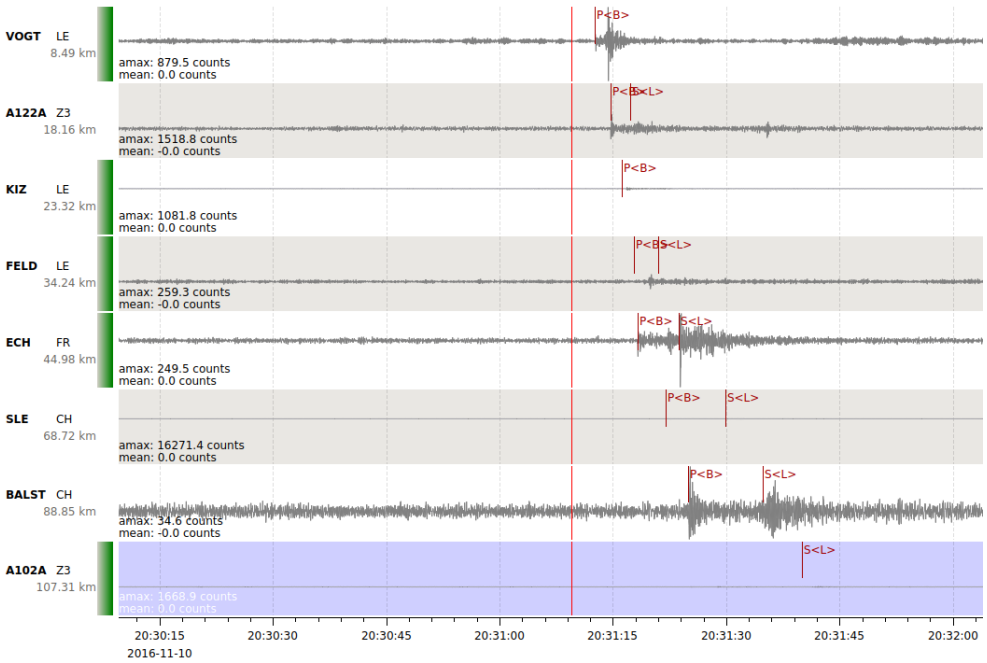


FIGURE 2.12: Exemple de séisme enregistré dans la plaine d'Alsace, près de la ville de Colmar (Magnitude Locale composante verticale MLv 1.0)

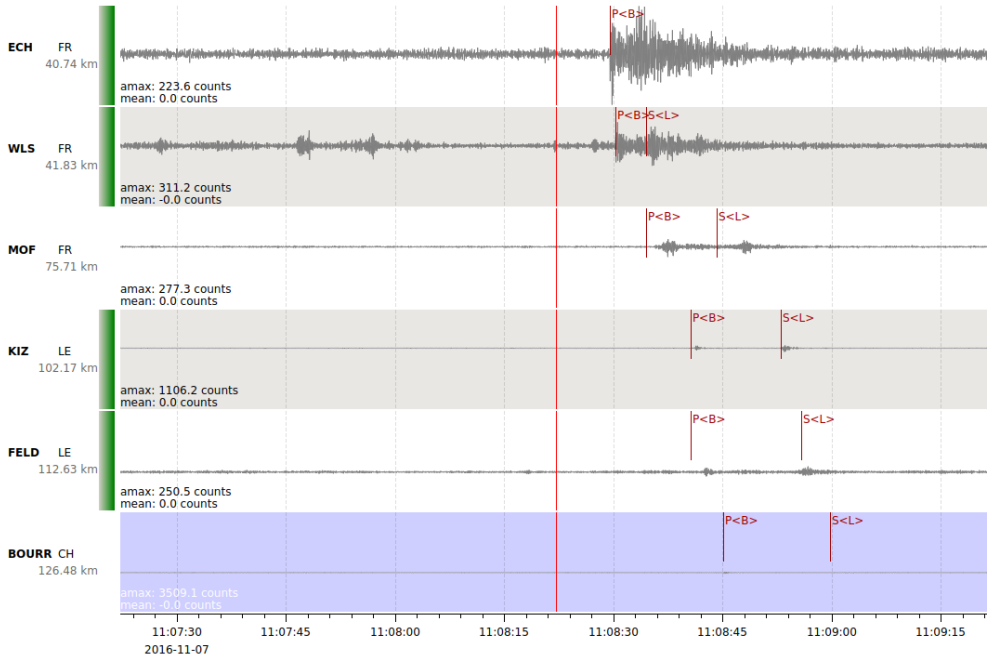


FIGURE 2.13: Exemple de tir de carrière enregistré au niveau de la carrière de Raon-l-Etape dans les Vosges (MLv 1.7).

Par exemple, le réseau national de surveillance sismique français (BCSF-RéNaSS) détecte actuellement majoritairement près de 50% d'événements autres que les séismes, dont 43% de tirs de carrières (Figure 2.9).

Si les seuils de détection étaient abaissés et, en tenant compte du nombre de stations utilisées par le BCSF-RéNaSS, la proportion de tirs de carrière détectés pourraient s'élever à près de 60%, avec en moyenne 400 faux événements supplémentaires détectés par jour, qui viennent encombrer le système d'alerte des événements.

2.2.4 Troisième limitation : un nettoyage des catalogues sismiques chronophage

Avec la diminution du seuil de détection et/ou la densification des réseaux de stations, la tâche quotidienne de nettoyage du catalogue de séismes apparaît donc difficile à réaliser entièrement manuellement. Sous la contrainte du temps, les événements vont donc être identifiés avec un inévitable compromis entre la vitesse de réalisation de la tâche à accomplir et la précision nécessaire à atteindre pour réussir cette dernière.

Il est par ailleurs impossible en temps réel de revoir manuellement des centaines de faux événements par jour. La fatigue naturelle physiologique propre à l'humain, liée ici à un afflux d'événements à examiner en un temps court, engendre une désensibilisation telle que des vrais événements (séismes et tirs de carrière principalement) peuvent être facilement traités par le cerveau humain comme des faux. L'effet produit est équivalent à celui de "crier au loup" : à force de fausses alertes, les vraies alertes finissent par passer plus facilement inaperçues (Figure 2.14). Sans oublier que les analystes doivent, en plus de la discrimination, nettoyer tous les vrais événements contaminés par du bruit.

Au final, par exemple, pour l'année 2014, seulement 8% des 6 000 000 de détections opérées par le système de surveillance international (IMS) ont été incluses dans le centre de données international (IDC). Le reste des détections, c'est-à-dire les 92% restants, sont en fait des faux événements. De plus, 39% des détections présentes dans le bulletin de l'IDC sont effectivement modifiées postérieurement par des analystes (Draelos et al., 2018).

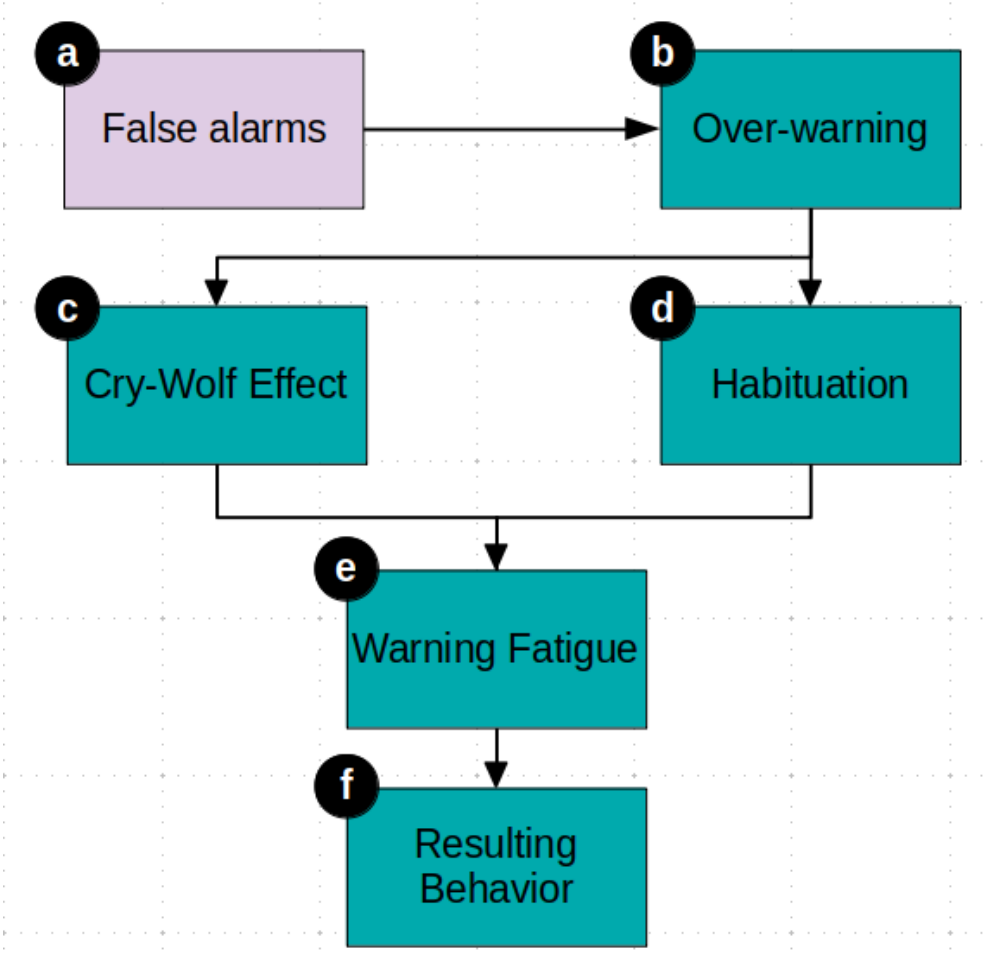


FIGURE 2.14: Fatigue physiologique liée aux fausses alertes. (a) Lorsque les seuils de détection sont abaissés, des centaines de fausses alarmes (engendrées par la détection de faux événements) sont émises par jour. (b) Les analystes (ou tout autre expert) sont submergés par une quantité trop importante d'événements à traiter (c) Une perte de vigilance physiologique est à craindre avec un effet équivalent à celui de "crier au loup" : à force de fausses alertes, les vraies alertes passent plus facilement inaperçues. (d) En conséquence, la diminution de la réponse physiologique des analystes aux stimulus provoqués par les alertes répétées d'événements engendre un phénomène d'habituation. (e) La fatigue s'installe alors. Les analystes exposés à des fausses alertes récurrentes ne répondent plus correctement à toutes les alertes d'événements. (f) De nombreuses vraies alertes ne sont plus traitées et seront supprimées. Près de 30% des vrais événements à identifier peuvent manquer dans les catalogues finaux.

Le temps de traitement manuel de grands volumes de données produites est non seulement conséquent, mais ces données massives saturent aussi les espaces de stockage qui archivent finalement une majorité de données superflues. Les efforts fournis pour obtenir un catalogue de séismes de plus faible magnitude de complétude apparaissent rapidement contre-productifs si l'on se réfère aux résultats finaux obtenus. En définitive, beaucoup de séismes vont inévitablement manquer dans des catalogues pollués par de faux événements.

Or, les systèmes de détection standards, basés sur le pointé des temps d'arrivées des ondes sismiques, ne reposent pas sur l'exhaustivité d'une base de données des événements, comme c'est le cas des techniques de détection basées sur les formes d'onde (E. J. Lee et al., **2020**). Par conséquent, ils offrent la possibilité de décoder de nouveaux signaux sismiques, enfouis dans la masse de données sismologiques désormais disponibles.

Seulement, trois grandes limitations, qui ont été décrites dans les paragraphes précédents, viennent fortement réduire la performance de ces systèmes de détection standards vis-à-vis des petits séismes.

La problématique de recherche qui rythmera ce travail de thèse s'intéressera alors à répondre à la question suivante :

Comment lever les limitations à la détection des petits séismes ?

2.3 Lever les limitations à la détection des petits séismes

2.3.1 Problématique de recherche

La performance des systèmes de détection standards est donc réduite par l'existence de trois grands limitations à lever. Ces trois limitations nécessitent de répondre aux trois questions de recherche suivantes :

(1) comment réduire la détection de ces très nombreux petits séismes contaminés par du bruit ?

(2) comment restreindre la détection de milliers de faux événements venant diluer l'information portée par les centaines de séismes détectés ?

(3) comment diminuer efficacement la charge conséquente de discrimination manuelle des séismes et des autres événements (principalement les tirs de carrière) ?

Les systèmes de détection standards ne sont actuellement ni complètement automatisés, ni complètement humanisés. Probablement parce que les deux acteurs de la détection des séismes, l'Homme et la machine (via les algorithmes), se complètent : leurs performances respectives compensent leurs propres limitations. Alors que les algorithmes de détection sont capables de traiter des données rapidement avec cohérence et objectivité, les humains présentent une expertise scientifique sur ces données qui est inégalable. Cette interaction Homme-machine est donc centrale, et mérite à être approfondie.

En effet, comment à partir de l'expertise humaine pourrait-on affiner le fonctionnement des algorithmes de détection pour (1) réduire la quantité de séismes contaminés par le bruit et (2) limiter les détections parasitaires ? Et comment à partir de la machine pourrait-on diminuer la charge de discrimination manuelle de l'ensemble des événements du catalogue (3) ?

L'approche qui guidera mes réponses aux différentes questions de recherche se basera donc sur :

Une optimisation de l'interaction Homme-machine pour une détection plus performante des petits séismes.

2.3.2 Choix qui vont guider le développement de la procédure de détection des petits séismes

Les choix qui vont guider le développement de la procédure de détection des petits séismes sont à relier avec les propriétés indispensables que doit avoir un système de détection qui opère en temps réel ; à savoir, sa rapidité, son évolutivité, sa flexibilité et sa longévité.

La **rapidité** de la procédure est un premier critère indispensable pour maintenir les opérations de surveillance sismologique en temps réel et faciliter la tâche quotidienne des analystes.

L'**évolutivité** de la procédure est également essentielle afin que cette dernière puisse être opérable à toutes les échelles possibles de détection des événements (locaux, régionaux, téléséismes) sans ajout excessif de complexité, ni de coût de calcul (C. E. Johnson, 2020).

La **flexibilité** de la procédure est un atout majeur pour que celle-ci soit adaptable aux besoins particuliers des services de surveillance sismique. Cela implique que le système de détection produit puisse être amélioré facilement une fois mis en opération, et donc être un outil de développement disponible (Yeck et al., 2019).

Enfin, le critère de **longévité** de la procédure est aussi important, pour garantir l'homogénéité d'un catalogue sur une longue période de temps. Cette procédure doit en effet susciter l'adhésion et l'implication de la communauté. Elle doit donc pouvoir être adoptée facilement : mise à jour facile, faible apprentissage, fiabilité, réponse aux besoins évolutifs.

Afin d'atteindre ces quatre objectifs de performance, j'ai donc utilisé les outils standards de détection déjà disponibles. En effet, même s'ils présentent des lacunes, ces outils sont déjà fortement implantés dans les observatoires sismologiques, travaillent déjà en temps réel et sont dès lors capables de traiter de grands volumes de données.

Je me suis donc intéressée à la mise en place d'une procédure plus performante en corrigeant les lacunes existantes des systèmes de détection utilisés. De par sa facilité d'implémentation, cette procédure pourrait plus facilement susciter l'adhésion de la communauté (familiarité du protocole, peu d'apprentissage nécessaire, facilité d'utilisation, langage algorithmique équivalent).

Ce travail de recherche privilégie les pistes de développement qui améliorent efficacement la détection des petits séismes avec les outils standards de détection actuels, tout en approfondissant l'interaction Homme-machine.

2.3.3 Comment réduire la détection des petits séismes contaminés par le bruit à partir de l'interaction Homme-machine ?

Le fait que de vrais événements détectés (principalement des séismes et des tirs de carrière) soient contaminés par du bruit est à relier au processus d'association lui-même. En effet, comme cela a été évoqué précédemment, le principe d'association se base sur les temps d'arrivées des pointés, et non sur le signal lui-même. Par conséquent, pour une fenêtre temporelle donnée, des temps d'arrivée correspondant aux arrivées des différentes phases sismiques (P et S) peuvent être groupés avec des temps d'arrivée qui ne sont reliés qu'à du bruit.

Trois pistes sont envisageables pour limiter le groupement de vrais pointés avec des faux pointés. Il est possible d'agir directement au niveau des pointés et/ou au niveau du processus même d'association et/ou au niveau de l'origine créée de chaque événement.

- **Agir au niveau des pointés ?**

Réduire la quantité de faux pointés produite offre l'avantage de désengorger le processus d'association et de limiter les fausses associations. En effet, l'utilisation des caractéristiques du signal (via les spectrogrammes ou les formes d'onde) sur une fenêtre temporelle centrée sur les temps d'arrivée de tous les pointés effectués permettrait de distinguer plus spécifiquement un temps d'arrivée, qui renseigne une phase sismique, d'un temps d'arrivée, qui signale du bruit. De cette manière, une labélisation des phases sismiques reconnues pourrait faciliter la suppression de tous les pointés générés par le bruit. Des études de reconnaissance des différentes phases sismiques ont été effectivement réalisées en utilisant par exemple l'intelligence artificielle, plus particulièrement les méthodes basées sur les réseaux neuronaux (Mousset et al., 1996 ; Gentili et al., 2006 ; Ross, M.-A. Meier et al., 2018 ; Y. Zhou et al., 2019).

Alors que l'implémentation de ces méthodes est une perspective intéressante pour la surveillance sismique globale comme c'est le cas à l'Institut d'Etudes Géologiques des Etats-Unis (USGS, Yeck et al., 2019), son implémentation reste plus délicate dans le cadre de la détection régionale de la sismicité. En effet, l'abaissement des seuils de détection amène à décoder des sismogrammes jusqu'à des rapports signal/bruit très faibles. Par conséquent, il y a un haut risque que des phases sismiques qui se détachent à peine du niveau de bruit moyen soient identifiées comme étant du bruit (McBrearty et al., 2019 ; Fu et al., 2019).

De plus, si les niveaux de bruit enregistré aux différentes stations sont élevés, comme c'est le cas dans les environnements urbains, des pointés correspondant à des signaux impulsifs transitoires de bruit, des pulses d'étalonnage ou des pointes de bruit peuvent être facilement identifiés comme étant des vrais pointés à conserver (Ross, M.-A. Meier et al., 2018).

L'objectif étant de développer une procédure qui puisse détecter les petits séismes avec de faibles rapport signal/bruit, cette dernière piste n'est pas privilégiée dans ce travail.

En revanche, la piste de travail qui est plutôt envisagée est celle d'agir directement sur les vrais pointés qui vont gouverner la création des vraies associations. En effet, une paramétrisation plus affinée du processus de pointé automatique améliorerait à la fois la qualité de l'estimation des temps d'arrivée des phases sismiques et la reconnaissance de l'ensemble des phases qui interviennent dans la création des événements.

Seulement, cela nécessite de comprendre comment ces pointés automatiques sont générés et quels sont les facteurs critiques qui déclenchent (ou ne déclenchent pas) un pointé à une station donnée. L'expertise humaine est donc ici indispensable pour augmenter la performance des algorithmes de détection. La réponse à ce premier point sera développée dans le chapitre 4.1.

→**Réponse : chapitre 4.1**

- **Agir au niveau du processus d'association ?**

En définitive, suivant les procédures de détection standards actuelles, les faux événements ne sont pas générés directement à partir des pures propriétés du bruit détecté, mais à partir d'une combinaison de temps d'arrivée qui sont à relier avec des sources complètement indépendantes (Tyler et al., 2018). Agir au niveau du processus même d'association est donc une piste indispensable à considérer.

Les faux pointés étant toujours générés, améliorer le processus d'association limiterait la création systématique de groupements de vrais pointés avec de faux pointés. Néanmoins, cela demande de comprendre comment fonctionnent les algorithmes d'association implémentés dans les procédures de détection standard et quels sont les paramètres décisifs qui contrôlent la qualité du processus d'association. L'expertise humaine est là-encore essentielle pour accroître l'efficacité des algorithmes d'association. La réponse à ce deuxième point sera développée dans le chapitre 4.2.

→ **Réponse : chapitre 4.2**

•Agir au niveau de l'origine créée de chaque événement ?

Agir au niveau des événements détectés est une opération délicate car cela implique que la fausse association ait déjà été créée et que l'événement ait été localisé avec cette présence de bruit à 1 ou 2 stations, voire plus dans les cas les plus difficiles. Une première piste qui peut être évoquée est d'utiliser les caractéristiques du signal pour distinguer de manière automatique ce qui révèle du bruit transitoire impulsif à une station donnée ou de phases sismiques à une autre station.

Seulement, les signaux sismiques, provenant d'une seule source décorrèlent, même avec de légères déviations du chemin parcouru par les ondes émises à partir de cette source (Harris, 2006 ; Dickey et al., 2019). Par conséquent, pour un même événement donné, la variabilité des signaux sismiques enregistrés à plusieurs stations peut significativement dégrader la possibilité de distinguer clairement tous les signaux associés à cet événement, de ceux associés à uniquement du bruit transitoire impulsif.

Cette tâche de discrimination peut être aussi d'autant plus difficile que le bruit d'origine anthropique présente des amplitudes et un contenu fréquentiel similaires à ceux des signaux sismiques régionaux (Hutton et al., 2010 ; Inbal et al., 2018 ; Perol et al., 2018). Pour ce travail de thèse, cette piste de recherche n'est pas sélectionnée.

Elle n'est pas privilégiée également car la procédure de détection standard qui est utilisée produit un catalogue d'événements à multiples origines. Chaque événement dans le catalogue contient plusieurs origines comme décrit dans la Figure 2.6. Même si une origine préférentielle est fixée automatiquement par défaut, il est alors probable d'agir sur la sélection préférentielle de cette origine de façon à éviter toute origine contaminée par du bruit.

C'est donc sur l'optimisation de cette sélection que ce travail de thèse se penche. Seulement, cela implique d'identifier des critères de sélection qui soient différents de ceux déjà disponibles par défaut, et qui soient décisifs pour le choix optimal de cette origine préférentielle. L'expertise humaine est une nouvelle fois indispensable à l'amélioration des algorithmes qui interviennent dans la sélection des origines préférentielles d'un catalogue multi-origine. La réponse à ce troisième point sera développée dans le chapitre 4.3.

→ **Réponse : chapitre 4.3**

2.3.4 Comment réduire la détection des faux événements à partir de l'interaction Homme-machine ?

Le contexte de la zone d'étude (une zone urbaine par exemple), le niveau de densification du réseau sismique et/ou la valeur du seuil de détection sont autant de facteurs qui vont contribuer à générer une importante quantité de pointés automatiques qui ne vont pas uniquement correspondre à des arrivées d'ondes de volume (P et S). Un faux événement est le produit d'une association de faux pointés qui ne sont pas reliés à un temps d'arrivée des ondes sismiques.

Deux options sont possibles pour diminuer le taux de faux événements : ou bien agir avant le processus d'association, en se focalisant sur les faux pointés, ou bien agir après le processus d'association, en cherchant à éliminer les faux événements détectés.

- **Agir au niveau des faux pointés ?**

Comme décrit précédemment, l'implémentation d'un processus de reconnaissance et d'élimination des faux pointés à partir des caractéristiques du signal est une opération délicate. Les faibles rapports signal/bruit qui sont utilisés pour détecter des petits séismes augmentent fortement les risques d'erreur d'identification des phases sismiques de très faible amplitude notamment. Cette piste de travail n'est donc toujours pas considérée.

- **Eliminer les faux événements ?**

L'Homme a l'expertise physique d'éliminer les faux événements détectés, en inspectant essentiellement l'aspect du signal enregistré aux stations qui sont intervenues dans le processus de fausse association. Il repère assez aisément un ensemble non cohérent et aléatoire de signaux. Seulement, face à des centaines de faux événements détectés par jour, suite à un abaissement des seuils de détection et une augmentation des volumes de sismogrammes à traiter, l'expertise humaine seule ne suffit plus.

Une automatisation du processus de reconnaissance des faux événements détectés allégerait donc l'opération de revue et d'élimination de ces événements qui parasitent les catalogues de séismes produits.

En sismologie, les outils de l'apprentissage machine ont été largement utilisés pour classer une diversité d'événements depuis les années 1990 (Dowla et al., 1990 ; J. Wang et Teng, 1995 ; Tiira, 1999 ; Maggi et al., 2017 ; Perol et al., 2018 ; Linville et al., 2019 ; Rouet-Leduc et al., 2019 ; Zhu et al., 2019). Ces outils ont également un faible coût opérationnel de calcul, peuvent analyser d'importants volumes de données en temps réel (M. Meier et al., 2019) et ont déjà fait leur preuve dans la détection routinière des signaux sismo-volcaniques (par exemple Malfante et al., 2018).

C'est donc naturellement vers ces outils que je me dirige. Cependant, construire des classifieurs automatiques d'événements performants demande de comprendre avant tout comment définir de façon robuste un faux événement relativement à un vrai événement.

Pour une optimisation du processus de discrimination automatisée des faux événements, l'expertise humaine sera donc une ressource précieuse et un guide nécessaire pour construire un classifieur fiable basé sur l'apprentissage machine. De plus, il faudra ajouter à cette classification automatique des faux événements, un processus d'élimination de ces derniers, de façon à désengorger la base de données d'information parasite. La réponse à ces deux derniers points sera développée dans le chapitre 5.

→ **Réponse : chapitre 5**

2.3.5 comment diminuer efficacement la charge de discrimination manuelle des événements du catalogue ?

Si la question des faux événements détectés est résolue dans la sous-section 2.3.4, il n'en demeure pas moins que le catalogue généré contient aussi de très nombreux vrais événements à identifier. En effet, dans le cas de la détection des petits séismes, la détection des faux événements représente 96% du total des détections. La charge de revue manuelle des événements est donc conséquemment allégée avec l'introduction potentiel d'un classifieur automatique de faux et de vrais événements, supprimant l'effet du "cri du loup".

En revanche, il reste encore ces milliers de vrais événements émis qui ne sont pas encore identifiés. Parmi ces vrais événements, on compte principalement des séismes et des tirs de carrière. La tâche de discrimination de ces deux derniers types d'événements n'est en fait pas toujours aisée et peut donc être très coûteuse en temps.

Les principaux critères usuellement utilisés pour discriminer les séismes des tirs de carrière sont la proximité de l'événement localisé à un site de carrière (Figure 2.15), le jour et l'heure de l'événement (Figures 2.16 et 2.17) ainsi que la similarité des formes d'onde (Figure 2.18) (Voyles et al., 2019).

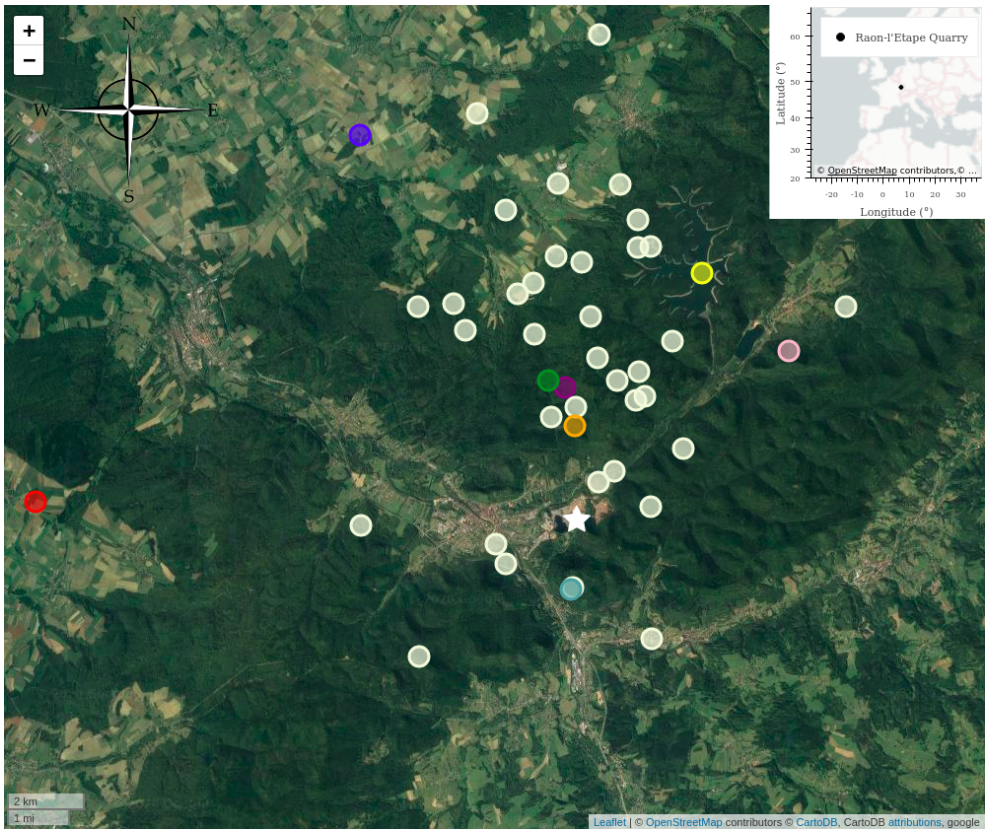


FIGURE 2.15: Localisations des tirs de la carrière de Raon-l'Etape dans les Vosges de juillet à décembre 2016. L'emplacement de la carrière est figuré par une étoile blanche et la localisation des tirs est représentée par des cercles. Les cercles de couleur (rouge, bleu, turquoise, violet, orange, vert, jaune et rose) correspondent aux tirs qui sont utilisés pour montrer les formes d'onde associées, enregistrées à la station ECH (voir Figure 2.18).

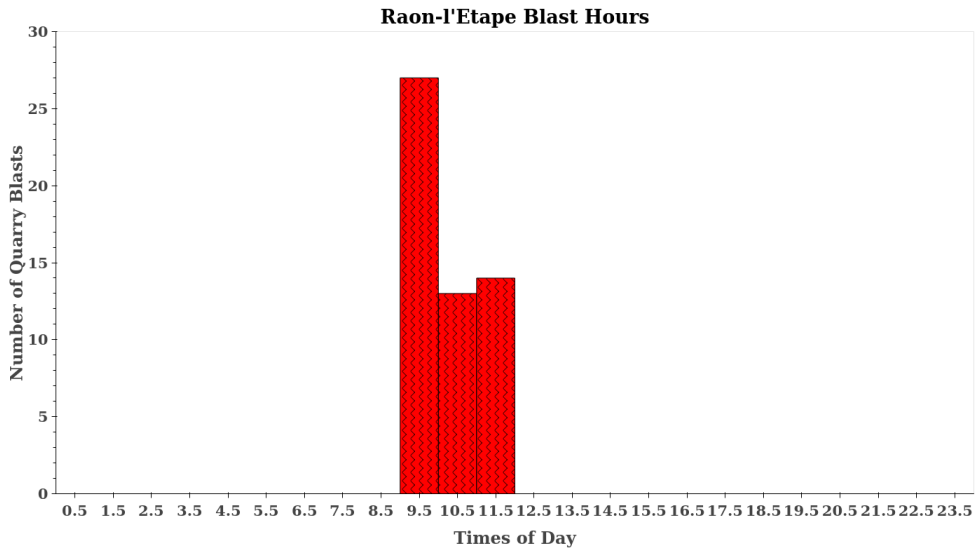


FIGURE 2.16: Exemple de distribution du nombre de tirs de carrière en fonction des heures de la journée pour la carrière de Raon-l'Etape dans les Vosges de juillet à décembre 2016.

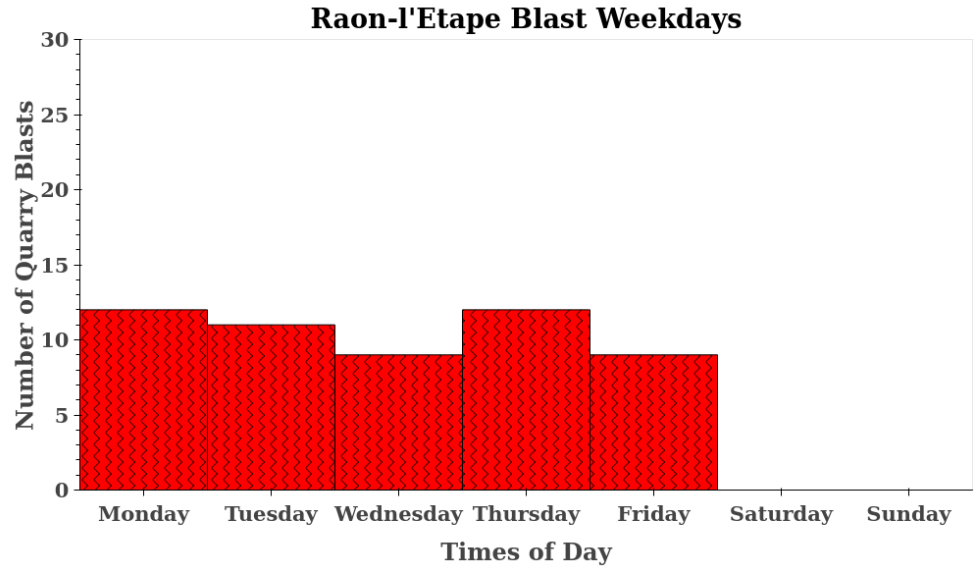


FIGURE 2.17: Exemple de distribution du nombre de tirs de carrière effectués en fonction du jour de la semaine pour la carrière de Raon-l'Etape dans les Vosges de juillet à décembre 2016.

Si les séismes ont lieu et se localisent plutôt aléatoirement, les tirs de carrière sont normalement reliés à leur lieu de production. Ils dépendent donc très fortement des heures et jours d'ouverture de la carrière ainsi que du calendrier des tirs. De plus, en se basant sur les formes d'onde, il est possible d'identifier chaque tir à une carrière donnée. En effet, des sismogrammes enregistrés à la même station, et correspondant à des tirs provenant d'une même carrière, sont visuellement similaires (Figure 2.18) (Israelsson, 1990). Ainsi, l'ensemble de tous ces arguments permettent de donner un diagnostic assez sûr pour repérer un tir.

Cependant, si cela est vrai pour les carrières très actives et qui génèrent des signaux qui se distinguent aisément du bruit ambiant enregistré, le diagnostic peut en fait s'avérer très complexe. Plusieurs raisons à cela peuvent être évoquées :

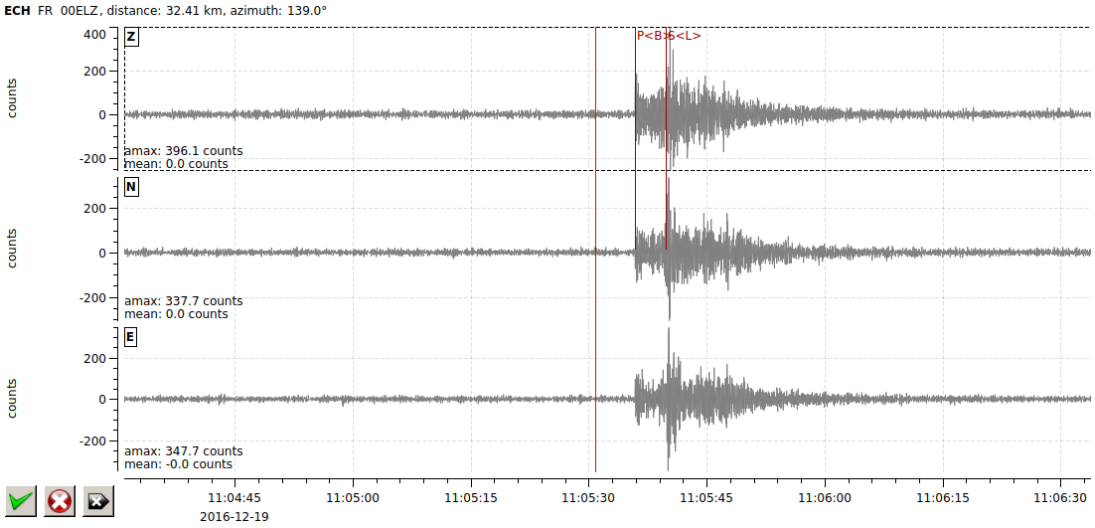
- La mémorisation difficile de tous les signaux associés à chaque carrière sans base de données conséquente à laquelle se référer (Figures 2.19, 2.20, and 2.21) ;
- La revue manuelle des événements chronologique, avec une impossibilité d'effectuer des aller-retour dans la base de données sans y passer beaucoup de temps ;
- La probabilité plus élevée de détecter des signaux de faible amplitude associés à des carrières très peu actives et/ou très peu connues, voire inconnues, lorsque les seuils de détection sont abaissés ;
- La localisation variable des tirs au sein même d'une seule carrière engendrant des dissimilarités dans les formes d'ondes, qui sont fortement influencées par les effets du milieu de propagation (Figure 2.22) ;
- La dissimilarité des signaux enregistrés à différentes stations pour un même tir de carrière, fortement influencés là encore par les effets du milieu de propagation (Figure 2.23) ;
- La probabilité non négligeable de séismes enregistrés proches des sites de carrières et/ou pendant les heures ouvrées et/ou possédant des caractéristiques du signal similaires aux tirs de carrière, du fait par exemple de la faible profondeur de leur source (Figure 2.24).

Face à la grande diversité des formes d'ondes, la distinction entre séismes et tirs de carrière peut devenir donc très subjective et demande une très grande expertise. Par conséquent, discriminer des milliers de vrais événements en temps limité devient difficile. La piste envisagée pour alléger cette charge laborieuse de discrimination des événements est donc l'automatisation du processus de classification des séismes et des tirs de carrières en utilisant l'apprentissage machine (rapidité, objectivité, évolutivité).

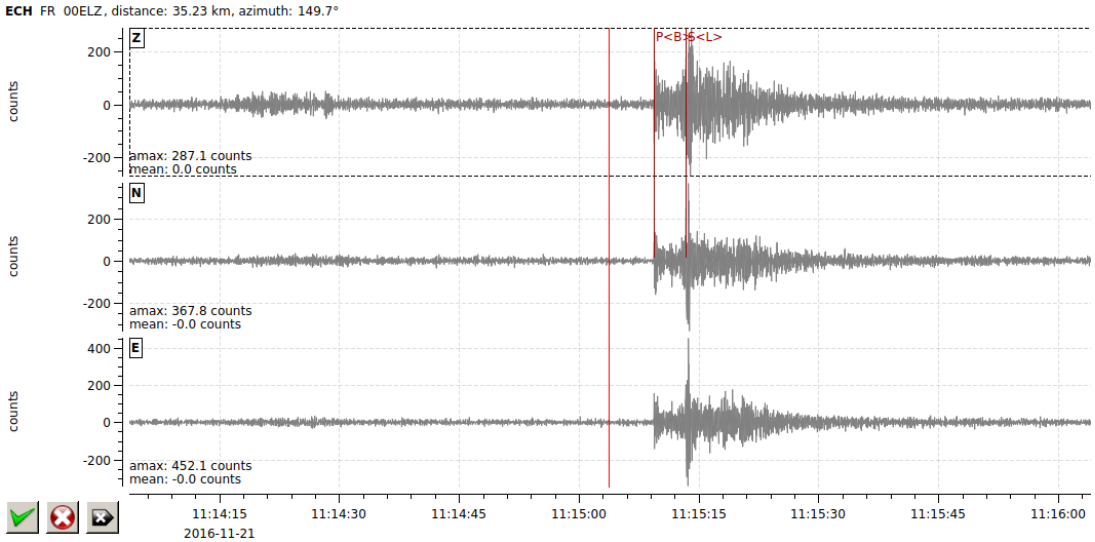
Seulement, pour obtenir une classification fiable et robuste, cela demande de déterminer précisément quels sont les critères qui vont permettre de distinguer de façon univoque un séisme d'un tir de carrière. L'expertise humaine sur les propriété des signaux enregistrés apparaît alors indispensable pour améliorer la performance des algorithmes de classification basés sur l'apprentissage machine. La réponse à ce dernier point sera développée dans le chapitre 5.

→**Réponse : chapitre 5**

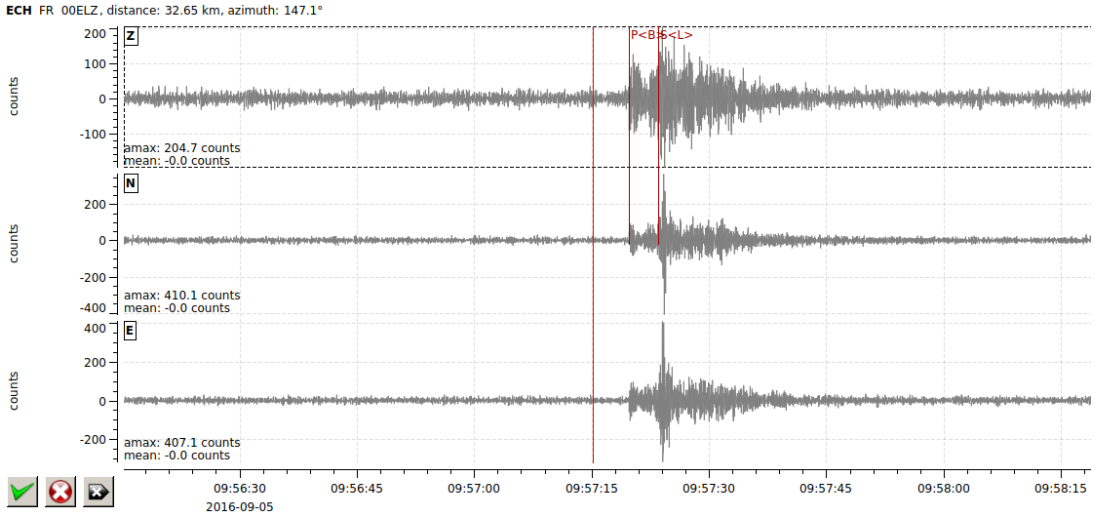
Le chapitre suivant (chapitre 3) est réservé à la présentation de l'objet d'étude. Chaque sous-chapitre présentera en quoi cet objet est le terrain idéal pour le développement d'une méthode qui puisse lever les plus grandes limitations à la détection des petits séismes.



a) 19 décembre 2016 à 11h05 (MLv 2.0). Cercle violet dans la Figure 2.15



b) 21 novembre 2016 à 11h15 (MLv 1.9). Cercle bleu dans la Figure 2.15



c) 05 septembre 2016 à 09h57 (MLv 1.5). Cercle jaune dans la Figure 2.15

FIGURE 2.18: Exemples de similarité de formes d'onde enregistrées à la station ECH pour différents tirs de la carrière de Raon-l'Etape dans les Vosges.

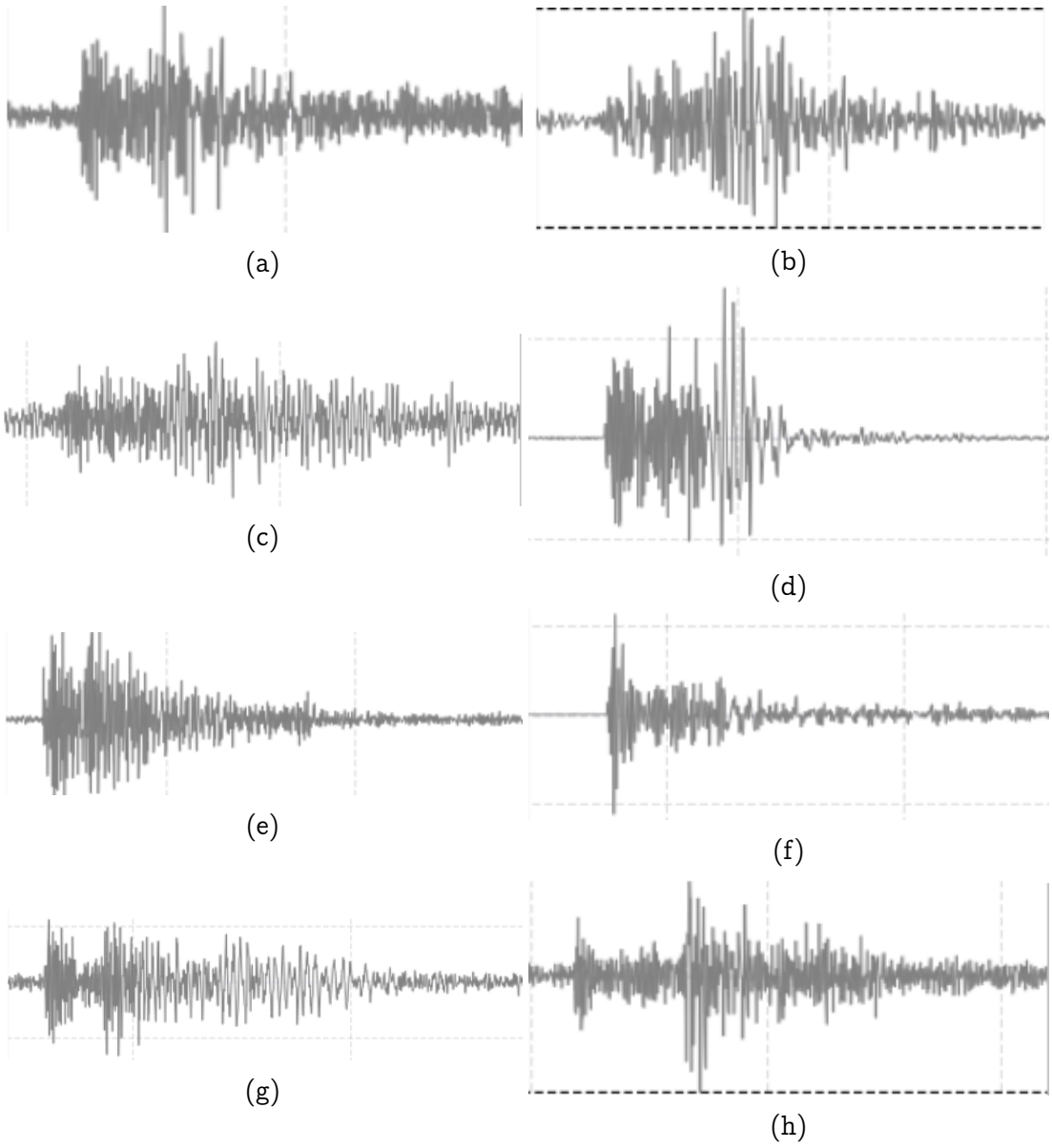


FIGURE 2.19: Exemples de formes d'ondes enregistrées à la première station sur la composante verticale pour différentes carrières de juillet à décembre 2016 : (a), (b), (c) carrières d'Arcey, de Chaffois et de Berche dans le Doubs en France, (d) Carrière de Bernécourt dans le département de la Meurthe-et-Moselle en France (e) carrière d'Attiswil dans le canton de Bern en Suisse, (f) carrière de Groß-Bieberau dans la région de Hesse en Allemagne, (g), (h) carrières de Dotternhausen et de Dunningen dans la région de Baden-Württemberg en Allemagne.

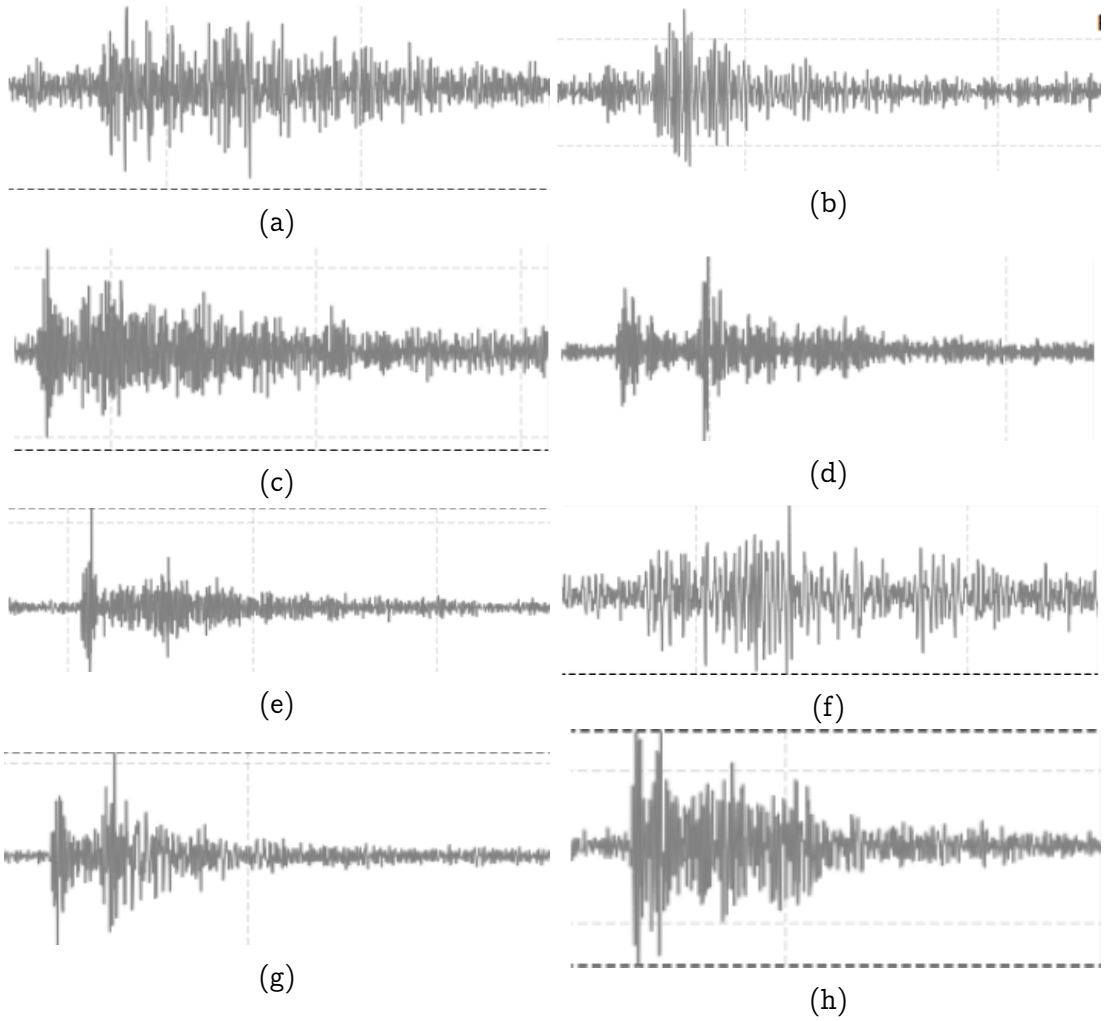


FIGURE 2.20: Exemples de formes d'ondes enregistrées à la première station sur la composante verticale pour différentes carrières de juillet à décembre 2016 : (a), (b), (c), (d) carrières d'Efringen-Kirchen, d'Ehingen, de Hausach-Dorf et de Mauer dans la région de Baden-Württemberg en Allemagne, (e) carrière de Gerbamont dans le département des Vosges en France, (f) carrière de La Heutte dans le Jura bernois en Suisse, (g) carrière de Lepuix-Gy dans le Territoire de Belfort en France, (h) carrière de Marchaux dans le département du Doubs en France.

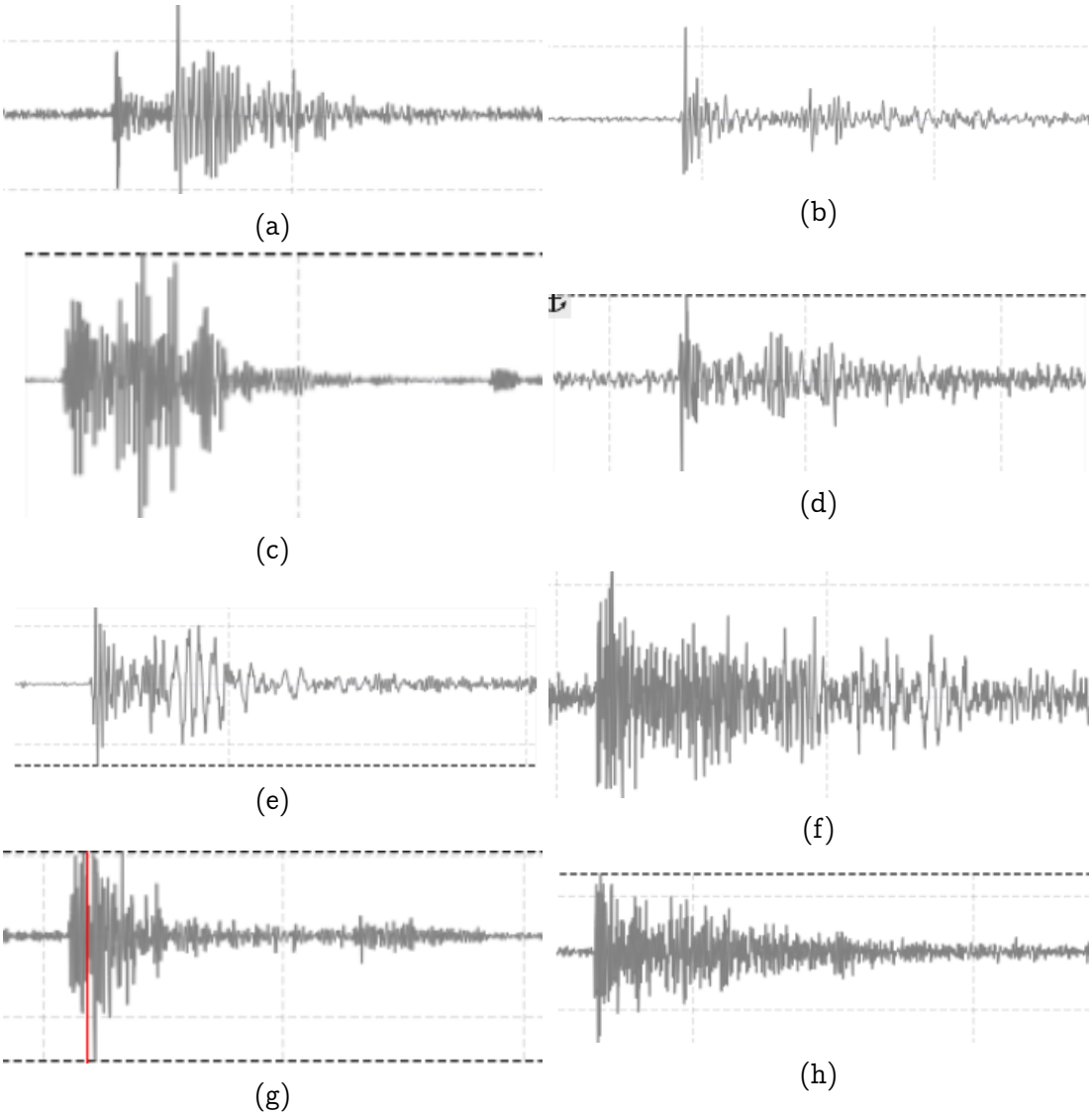


FIGURE 2.21: Exemples de formes d'ondes enregistrées à la première station sur la composante verticale pour différentes carrières de juillet à décembre 2016 : (a), (b), (c), (d), (e) carrières de Trochtelfingen, de Rems-Murr, de Schelklingen-Vohenbronnen, de Schuttertal et de Seebach dans la région de Baden-Württemberg en Allemagne, (f) carrière de Pagny-sur-Meuse dans le département de la Meuse en France, (g) carrière de Saint-Amé dans le département des Vosges en France, (h) carrière de Villigen dans le canton d'Aargau en Suisse.

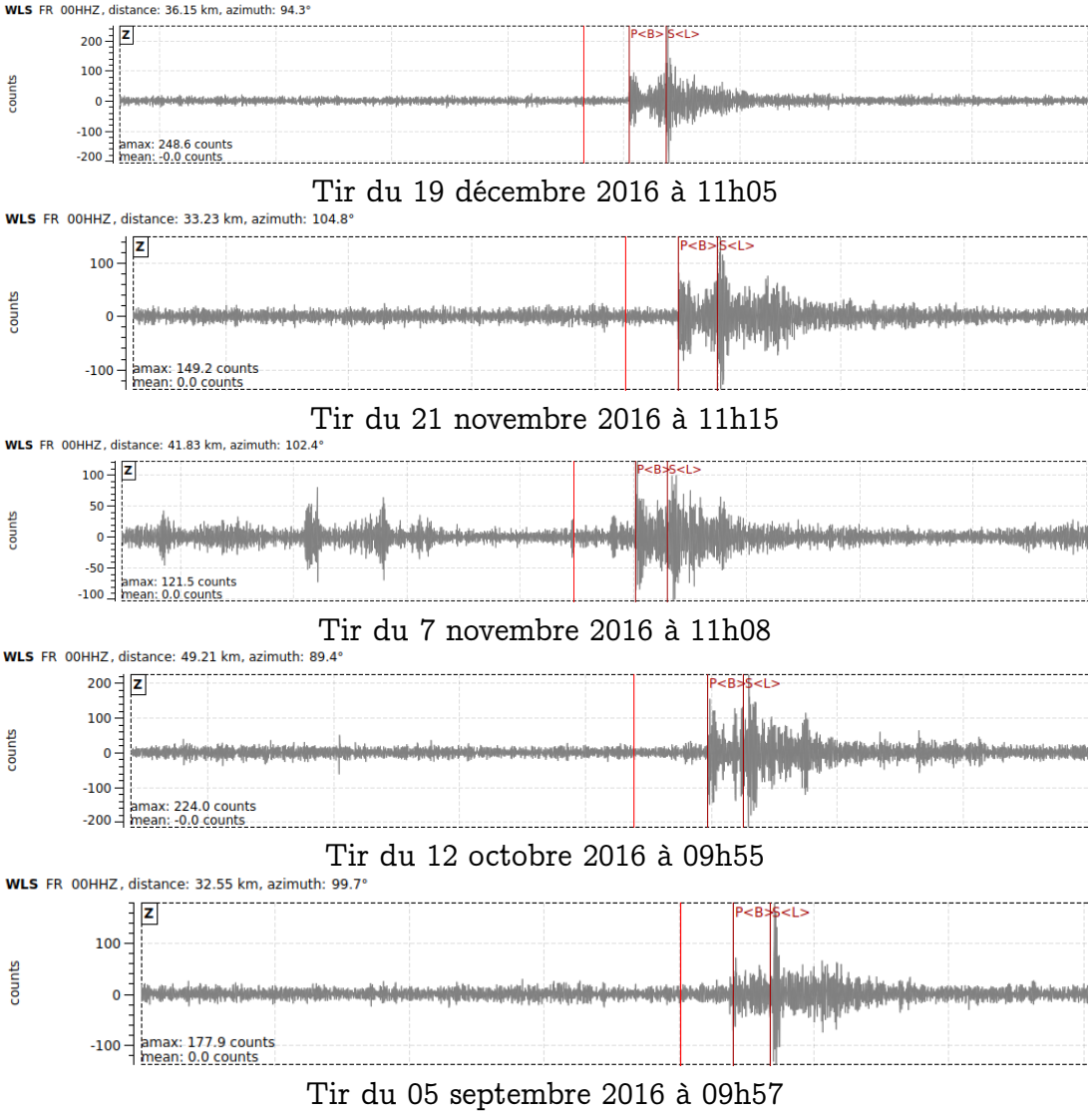
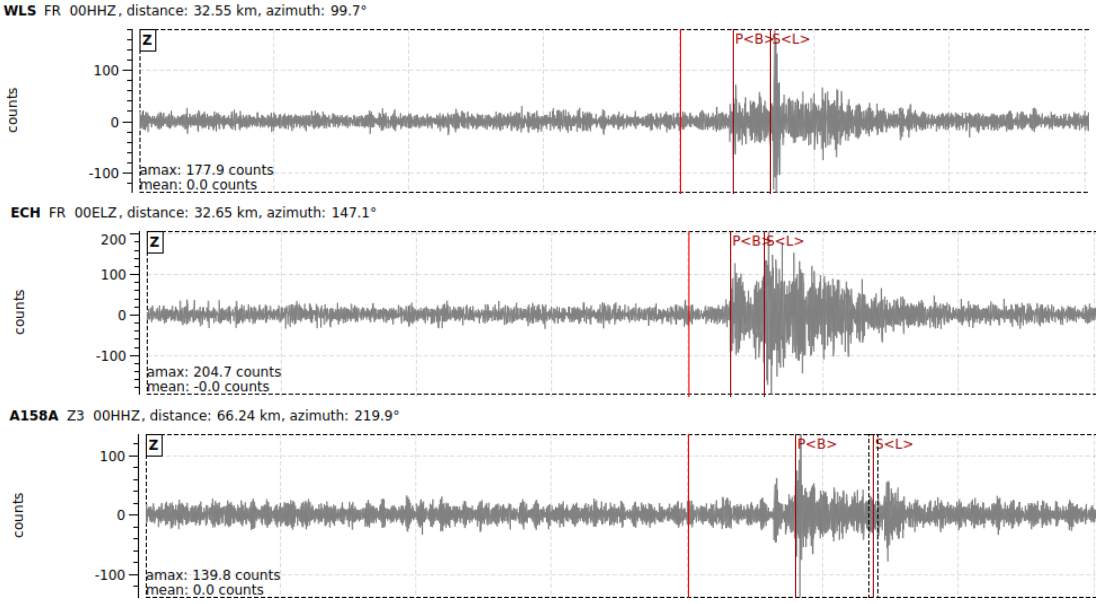
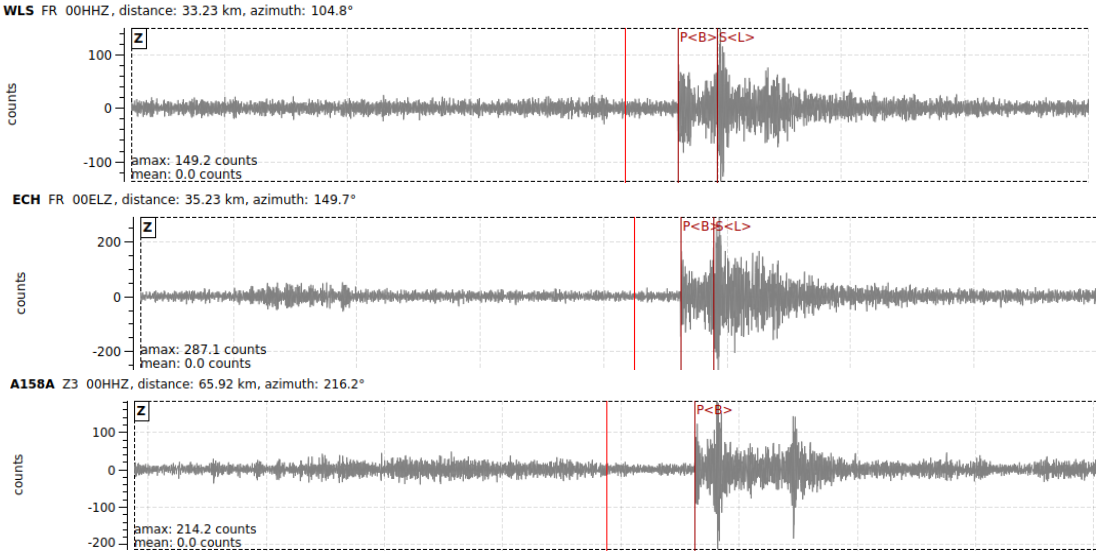


FIGURE 2.22: Exemples de variations dans les formes d'onde enregistrées à la station FR.WLS sur la composante verticale pour des tirs ayant eu lieu à la carrière de Raon-l'Etape dans les Vosges.

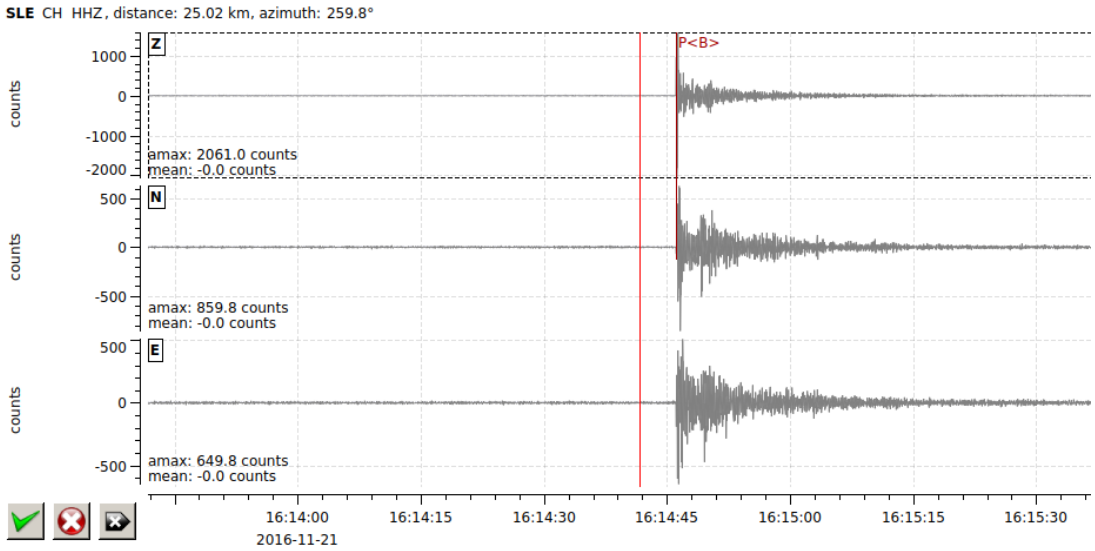


a) Tir du 5 septembre 2016 à 09h57

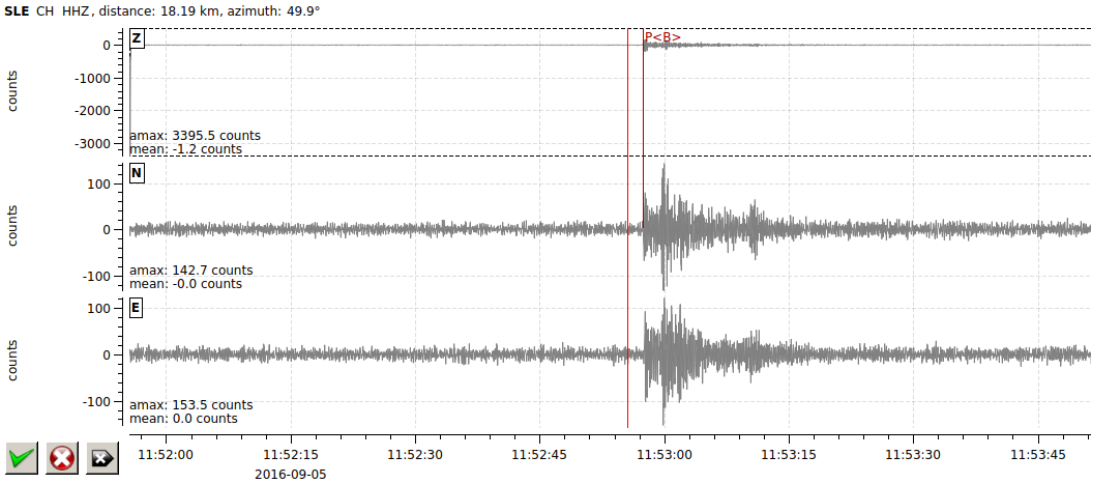


b) Tir du 21 novembre 2016 à 11h15

FIGURE 2.23: Exemples de formes d'ondes enregistrées à différentes stations pour deux tirs ayant eu lieu à la carrière de Raon-l'Etape dans les Vosges : à gauche, tir du 05 septembre 2016 à 09h57 et à droite, tir du 21 novembre à 11h15.



(a) Signal correspondant à un séisme ayant eu lieu le 21 novembre 2016 à 16h14 dans le région de Singen dans le Sud de l'Allemagne.



(b) Signal correspondant à un tir ayant eu lieu le 05 septembre 2016 à 11h52, localisé à 40 km au sud-ouest du séisme précédent près de la carrière de Waldshut-Tiengen-Detzel.

FIGURE 2.24: Exemples de signaux difficilement discriminables enregistrés à la station SLE pour un séisme (a) et un tir de carrière (b).

Chapitre 3

Un objet d'étude idéal pour développer la détection des séismes de faible magnitude

Sommaire

3.1 Une zone d'étude située au coeur d'un domaine intraplaque continental

3.1.1 Une zone géologiquement complexe

La région située au Nord-Est de la France et au delà des frontières est d'abord une zone géologiquement complexe (Figure 3.1). Elle souligne une tectonique complexe marquée par une histoire géologique ancienne et variée. Elle regroupe d'importants massifs paléozoïques appartenant à la chaîne varisque d'Europe de l'ouest : le Massif des Vosges, le Massif de la Forêt Noire, une partie du Massif Central (plus particulièrement le Massif du Morvan), le Massif de l'Ardenne et du Brabant, ainsi que le Massif de Rhenish.

Elle contient également des grands bassins sédimentaires épicontinentaux d'âge Méso-Cénozoique : la partie Est du Bassin Parisien et une partie du bassin sédimentaire au Sud de la Bavière allemande.

Cette région est traversée par deux segments majeurs du système de rifts Cénozoïques Ouest-Européen qui sont disposés concentriquement autour du front alpin : le graben de la Hesse et du Rhin Supérieur orientés NNE-SSO puis les fossés d'effondrement du Massif Central (les Limagnes) et de la Bresse orientés N-S. Cette zone inclut d'ailleurs une partie de la chaîne alpine. Associée à cette chaîne alpine, la zone renferme aussi un bassin flexural synorogénique, le Bassin Molassique Suisse.

Enfin, cette zone comporte le Massif du Jura d'âge Miocène (Jura Français, Jura Suisse, Jura Souabe).

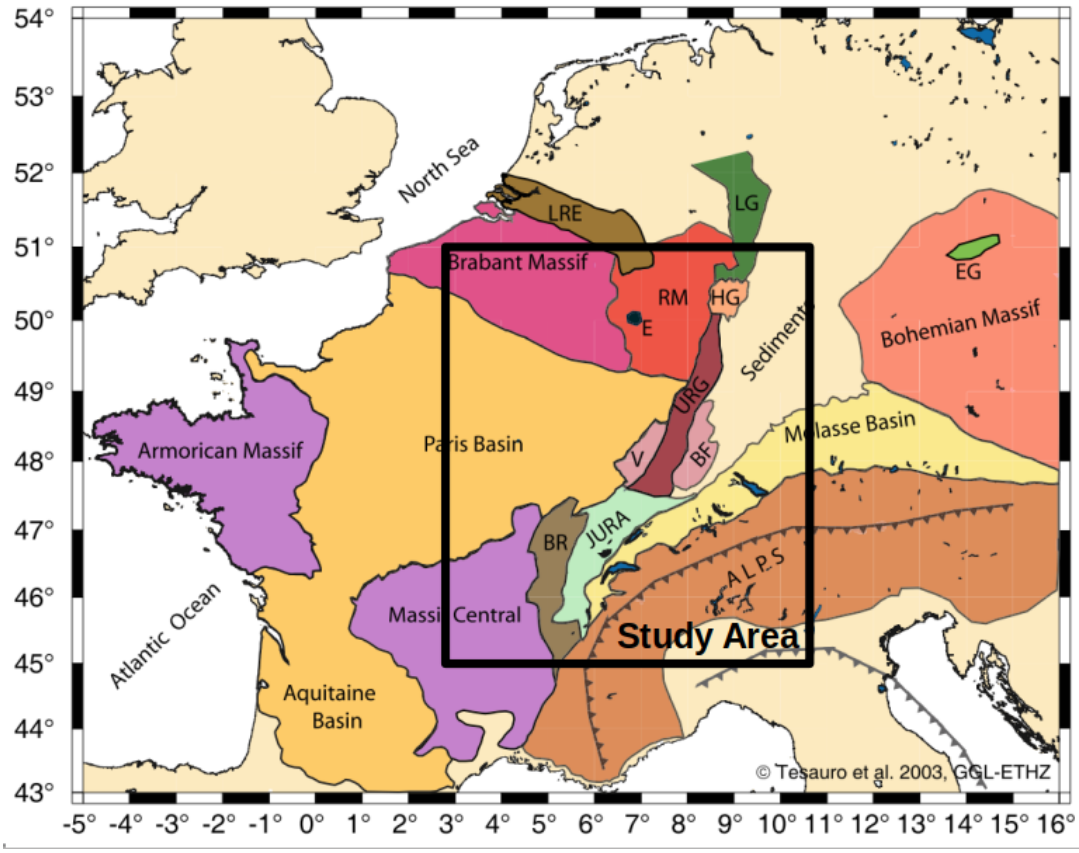


FIGURE 3.1: Principales unités géologiques du centre Ouest de l'Europe. La zone d'étude est marquée par un cadre noir. BR : Fossé de la Bresse, V : Vosges, BF : Forêt Noire, URG : Graben du Rhin Supérieur, HG : Graben de la Hesse, LG : Graben du Leine, RM : Massif de Rhenish, E : Eifel, LRE : Système d'effondrement du Rhin Inférieur, EG : Graben de l'Eger. Modifié d'après Tesauro et al. (2005).

3.1.2 Une zone continentale stable

La zone d'étude est une zone continentale intraplaque relativement stable. Les vitesses de surface horizontales et verticales, déterminées grâce aux données GNSS montrent des valeurs très faibles, de l'ordre de l'incertitude des mesures : une valeur moyenne de 0.37 $\pm$ 0.30 mm/yr pour les vitesses horizontales dans un référentiel Eurasie et une valeur moyenne absolue de l'ordre de 0.4 $\pm$ 0.52 mm/yr pour les vitesses verticales (Henrion et al., 2020).

L'étude du champ de vitesses horizontales montre des directions de ces vitesses hétérogènes avec de faibles amplitudes de vitesse sur toute la zone d'étude, à l'exception de la région comprise entre le front du Jura et les Alpes. En effet, un mouvement léger est observé en direction du Nord (vitesses horizontales de l'ordre de $0.49mm \pm 0.33mm/an$, Henrion et al., 2020).

L'analyse supplémentaire du tenseur des taux de déformation, établi à partir du tenseur des gradients de vitesse, met en évidence un raccourcissement NW-SE à NNW-SSE entre le front Alpin et le Jura, de l'ordre de 2.86 $\pm$ 0.2e-09 par an (Rabin et al., 2018) à 7e-09 par an (Henrion et al., 2020). En revanche, l'analyse de ce tenseur ne montre pas de déformation géodésique apparente claire au Nord du front jurassien, c'est-à-dire sur tout le reste de la zone d'étude. Cependant, la quasi-absence de mouvements tectoniques mesurables actuellement par la géodésie ne signifie pas une absence de déformation de cette région.

Cette zone continentale intraplaque stable enregistre effectivement quotidiennement une sismicité de faible magnitude qui reste encore difficile à expliquer sous des conditions actuelles de déformation très faible, voire négligeable, comme c'est le cas pour nombreuses autres zones intracontinentales à l'échelle du globe (Gallen et al., 2018 ; Bezada et al., 2019 ; Leclere et al., 2019).

3.1.3 Une zone sismique de faible magnitude

La zone d'étude est donc principalement caractérisée par une sismicité de faible magnitude. Comme de nombreuses autres zones intraplaques continentales, cette sismicité semble diffuse au premier ordre à l'échelle du réseau complexe de failles, qui est encore peu connu (Bowman et al., 1990 ; Gagnepain-Beyneix et al., 1982 ; Tuttle et al., 2002 ; Camelbeeck et al., 2007 ; Terrinha et al., 2009 ; Martinez-Garzon et al., 2019).

De plus, cette zone d'étude a épisodiquement hébergé une sismicité de plus forte magnitude, comme l'atteste l'ensemble de la sismicité instrumentale ainsi que la sismicité historique de la zone (voir Annexe A pour la zone du Graben du Rhin Supérieur). En effet, plusieurs séismes de magnitude modérée ont d'abord été sporadiquement enregistrés. Par exemple, le séisme d'Albstadt de 1978, ayant eu lieu dans le Jura Souabe, affiche une magnitude locale de 5.7 (Haessler, P. Hoang-Trong et al., 1980), les séismes de Remiremont de 1984 et de Rambervillers de 2003, ayant eu lieu dans les Vosges, présentent respectivement une magnitude locale de 4.8 (Haessler et H. Hoang-Trong, 1985) et 5.4 (Audin et al., 2002), et le séisme de Corrençon, situé dans les Alpes de l'Ouest près de Grenoble, a une magnitude locale estimée à 5.3 (Thouvenot et al., 2003).

De plus larges séismes ont également été répertoriés historiquement comme le séisme de Bâle de 1356 dont la magnitude de moment (Mw) a été estimée entre 6 et 7.1 (Meghraoui et al., 2001 ; Fah, Gisler et al., 2009 ; Shipton et al., 2017) ou le séisme de Visp plus au Sud datant de 1855 (Mw 6.2 ; Fah, Moore et al., 2012).

Si l'on s'intéresse à l'activité sismique régulière de la zone d'étude depuis 2012, date à partir de laquelle les événements autres que les séismes ont été intégrés au catalogue de sismicité, la majeure partie des séismes qui sont détectés par le système de détection du Réseau National de Surveillance Sismique (BCSF-RéNaSS) ont des magnitudes locales (calculées sur la composante verticale) comprises entre 1 et 3 (Figure 3.2).

La distribution des magnitudes locales MLv des séismes pour la période 2012-2020 montre que 70% des événements ont une magnitude inférieure à 1.5 et la quasi-totalité des séismes enregistrés ont une magnitude inférieure à 2.0 (Figure 3.3).

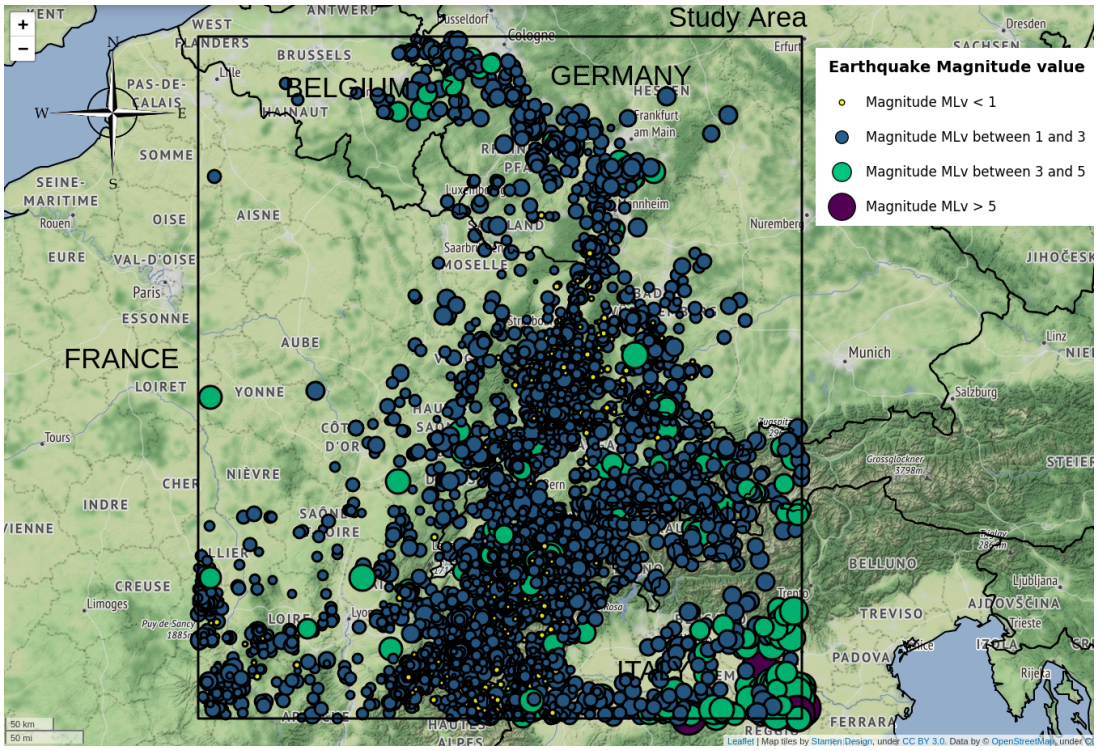


FIGURE 3.2: Distribution des séismes détectés par le Réseau National de Surveillance Sismique (RéNaSS) français pour la période janvier 2012-juillet 2020. Localisations extraites de la base de données RéNaSS selon un protocole FDSN à l'adresse http://renass-sc1.u-strasbg.fr:8080.

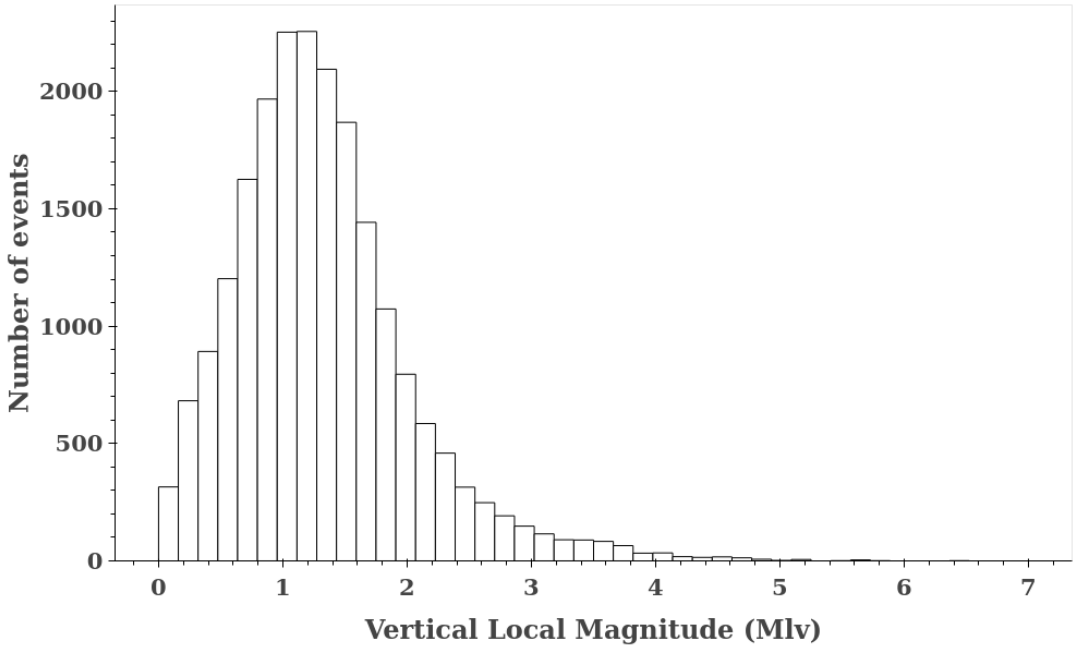


FIGURE 3.3: Distribution des magnitudes des séismes détectées par le Réseau National de Surveillance Sismique (RéNaSS) français pour la période janvier 2012-juillet 2020. La magnitude estimée est une magnitude locale calculée sur la composante verticale (MLv).

A partir de la représentation graphique de la distribution cumulative fréquence-magnitude des séismes détectés pour cette même période 2012-2020, il est possible d'estimer une valeur de magnitude de complétude environ égale à 1.2. (Figure 3.4).

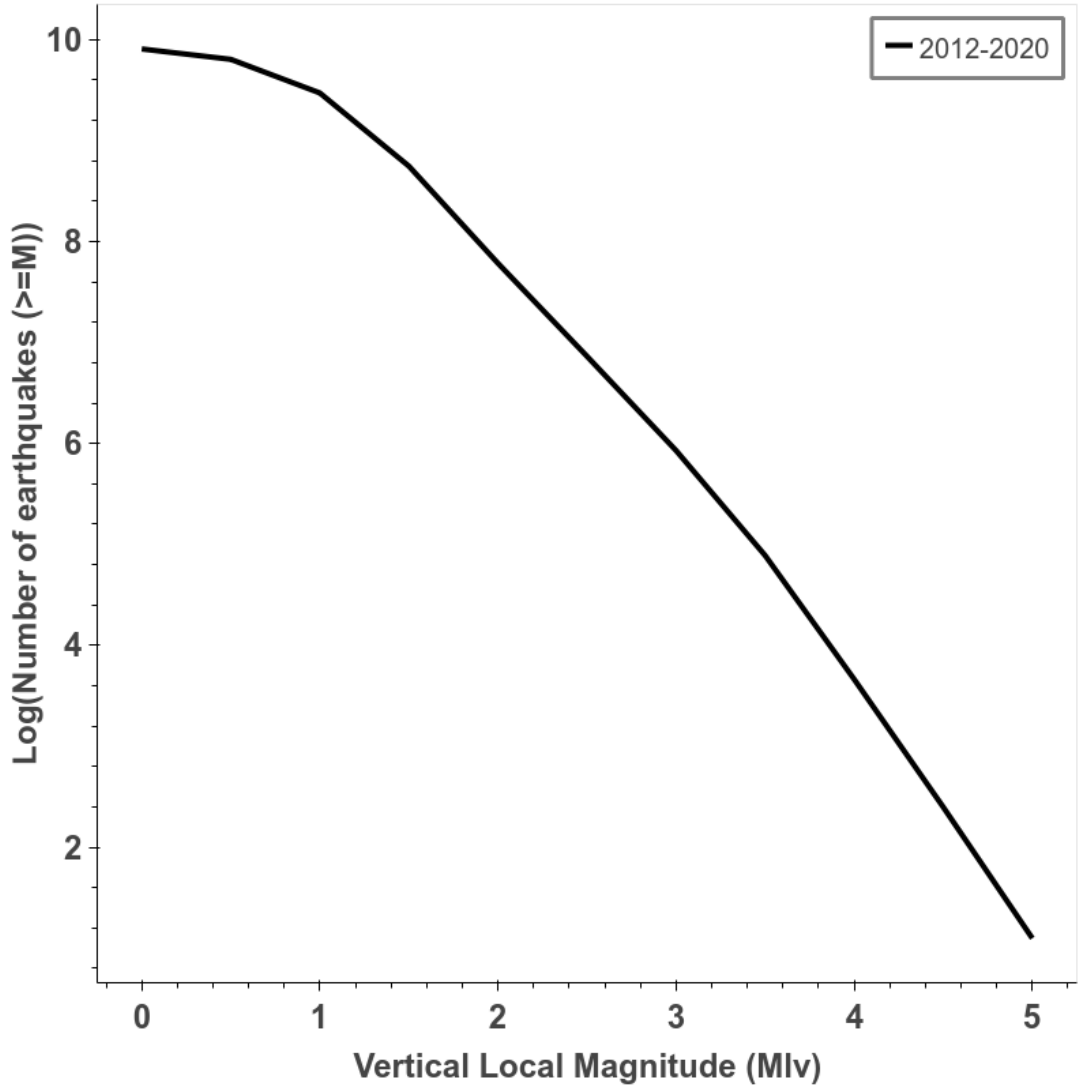


FIGURE 3.4: Distribution cumulative fréquence-magnitude des séismes détectés par le Réseau National de Surveillance Sismique (BCSF-RéNaSS) français pour la période janvier 2012-juillet 2020.

3.1.4 Une zone à activité anthropique régulière

Très urbanisée et économiquement active, cette zone enregistre quotidiennement des signaux qui sont reliés à une activité d'origine anthropique, principalement des tirs de carrière, mais aussi une sismicité induite par la géothermie profonde ainsi qu'une très faible activité sismique reliée à l'exploitation minière. (Figure 3.5).

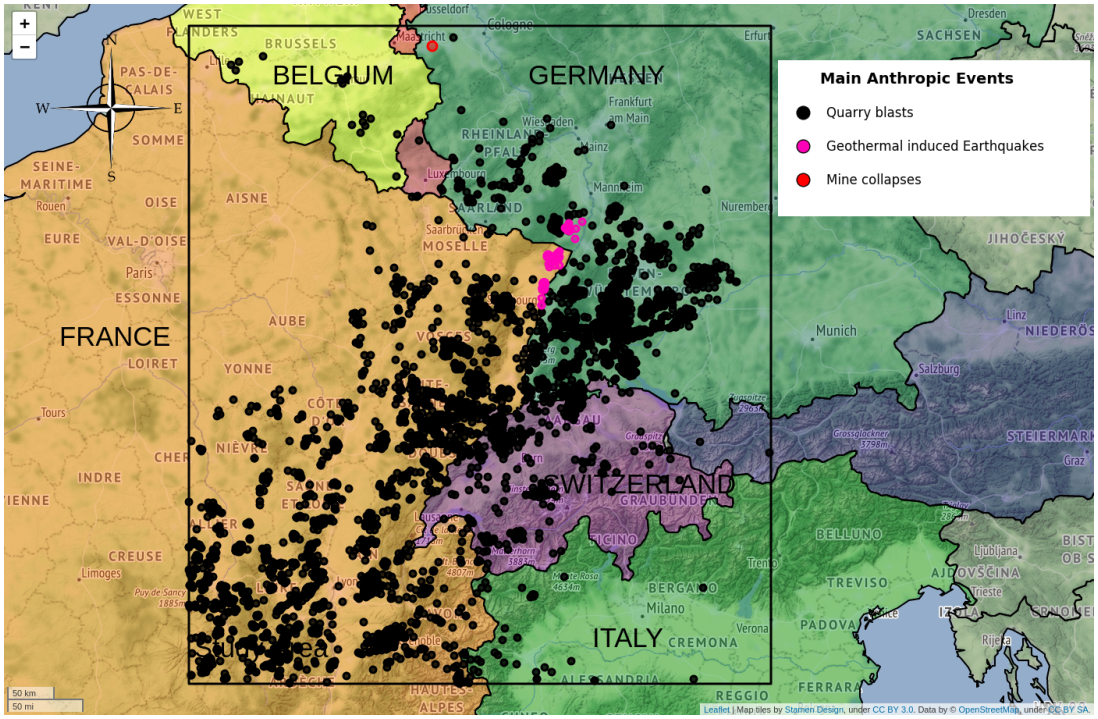


FIGURE 3.5: Distribution et répartition des événements d'origine anthropique, majoritairement des tirs de carrière, détectés par le Réseau National de Surveillance Sismique (BCSF-RéNaSS) français pour la période janvier 2012-juillet 2020.

Une carte de la distribution des sites de carrières, des sites géothermiques ainsi que quelques sites miniers, corrélant avec l'activité d'origine anthropique détectée précédemment, montre une large prépondérance de l'activité de carrière à travers tout le site d'étude (Figure 3.6). Un peu plus de 96% des événements d'origine anthropique détectés par le BCSF-RéNaSS correspondent à des tirs de carrières.

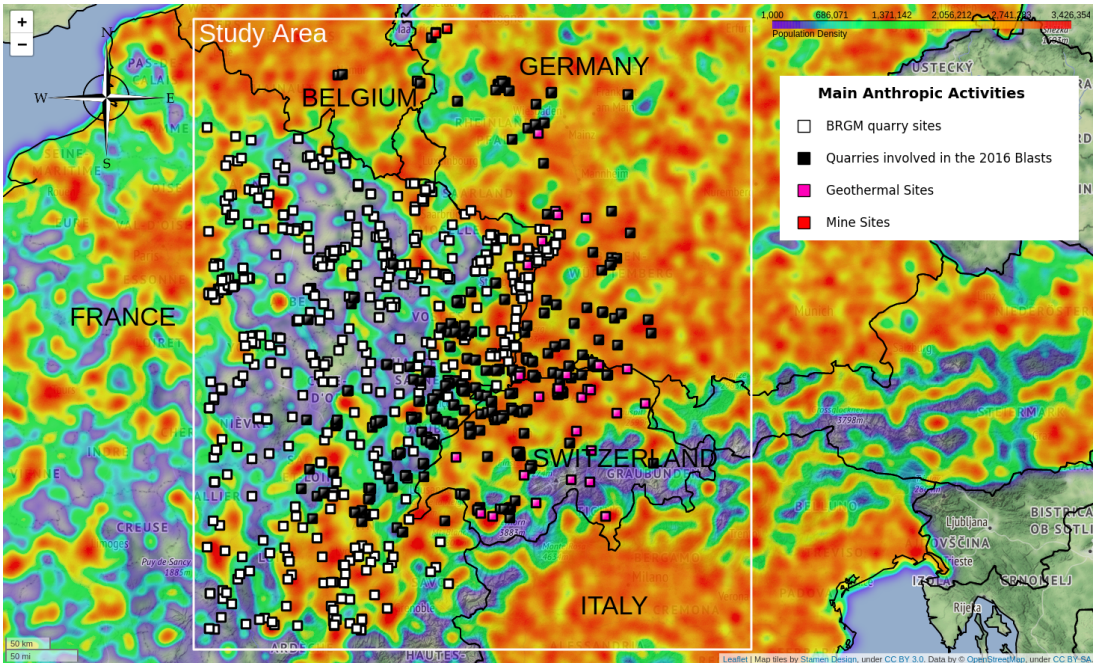


FIGURE 3.6: Densité de population et distribution des sites de carrière, de géothermie profonde et de quelques mines dans la zone d'étude. D'après https://public.opendatasoft.com/explore/dataset/geonames-all-cities-with-a-population-1000/table/?disjunctive.country pour la base de données sur la densité de population, d'après http://geoservices.brgm.fr/odmgm pour la base de données des carrières du Bureau de Recherches Géologiques et Minières (BRGM), d'après http://www.seismo.ethz.ch/en/knowledge/things-to-know/geothermal-energy-earthquakes/geothermal-energy-in-switzerland/ pour les sites de géothermie en Suisse, d'après www.geotis.de pour les sites de géothermie en Allemagne et d'après http://www.energies-renouvelables.org pour les sites de géothermie en France. Les carrières représentées en noir correspondent aux carrières dont les tirs ont été identifiés et détectés au cours de l'année 2016.

De plus, les magnitudes locales (MLv) de l'ensemble des événements étiquetés comme tirs de carrière présentées sur la Figure 3.5 pour la période 2012-2020 sont de même ordre de grandeur que celles des séismes naturels enregistrés (Figure 3.7). La totalité des tirs enregistrés ont une magnitude locale inférieure à 2.8 et environ 48% d'entre eux ont des magnitudes locales comprises entre 1.5 et 1.6.

La probabilité d'enregistrer plus de tirs de carrière, lorsque la détection des petits séismes est accentuée, est donc grande.

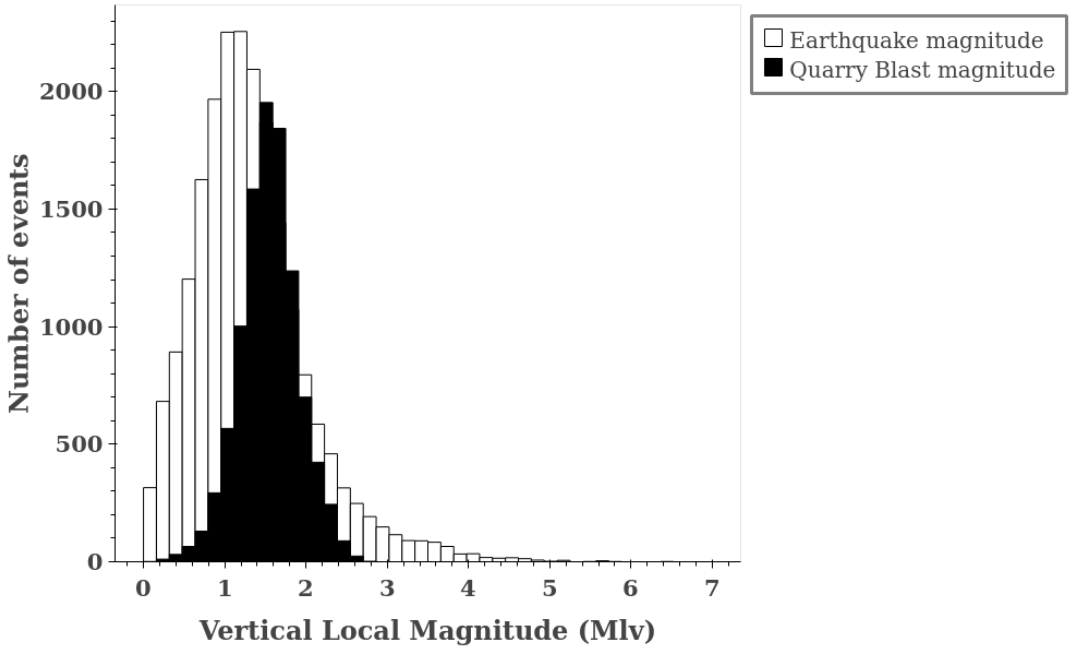


FIGURE 3.7: Distribution des magnitudes locales calculées sur la composante verticale (Mlv) pour l'ensemble des tirs de carrière détectés pour la période janvier 2012-juillet 2020.

Par ailleurs, le Réseau National de Surveillance Sismique (BCSF-RéNaSS) français met à disposition une base de données de séismes et de tirs de carrière qui sont robustement discriminés depuis 2016 par des analystes. Cette zone est donc particulièrement intéressante pour monter un protocole de discrimination automatique des tirs de carrière et des séismes.

Elle donne également la possibilité de comprendre finement les caractéristiques dominantes qui vont solidement différencier les tirs de carrière des séismes. En effet, cette zone d'étude est une zone géologique et structurale complexe comme peut le témoigner la variabilité pétrographique de ses roches. Par conséquent, à l'échelle de la zone d'étude, l'ensemble des carrières vont exploiter une diversité de matériaux produits à partir de roches sédimentaires, de roches volcaniques, de roches plutoniques et de roches métamorphiques (Figure 3.8).

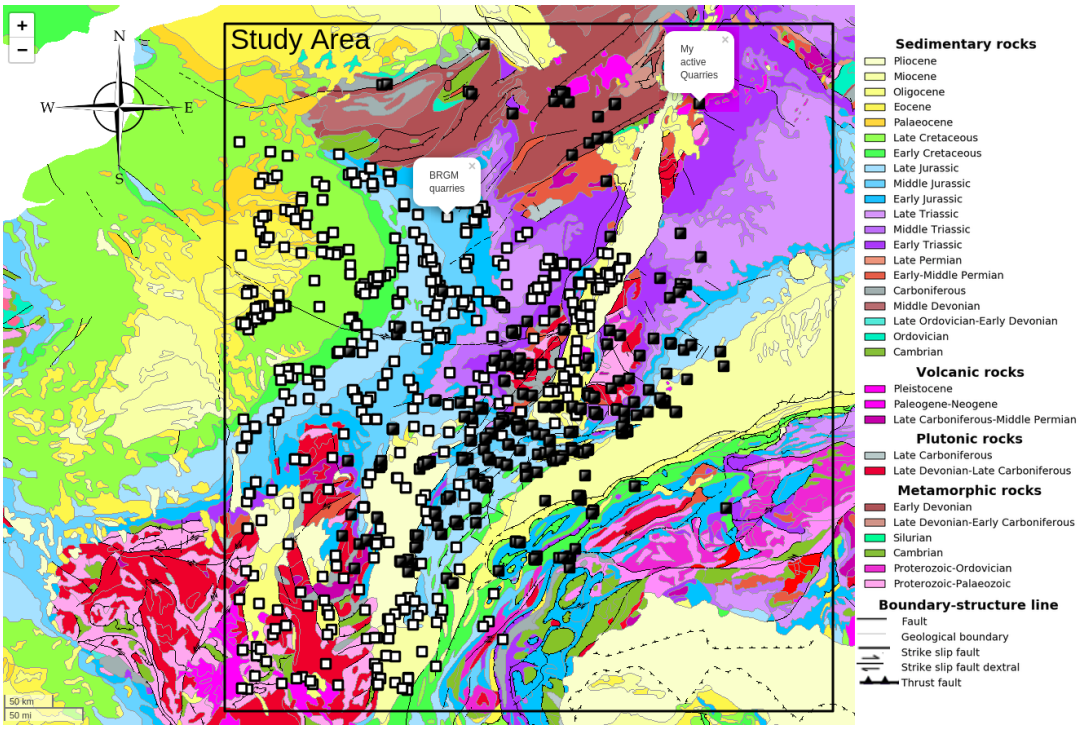


FIGURE 3.8: Distribution des carrières dans la zone d'étude en fonction de la nature géologique des terrains d'extraction. Les carrières représentées en noir correspondent aux carrières dont les tirs ont été identifiés et détectés au cours de l'année 2016. D'après https://services.bgr.de/wms/geologie/igme5000/? pour la carte géologique et la représentation des structures et d'après http://geoservices.brgm.fr/odmgm pour la base de données des carrières du Bureau de Recherches Géologiques et Minières (BRGM).

Roches sédimentaires. Par exemple, la carrière de Hauteville-Lompnes située au Sud de Bourg-en-Bresse dans le Massif du Jura français exploite un calcaire apparenté au marbre du Jurassique Supérieur. La carrière de Pagny-sur-Meuse située à proximité de la ville de Nancy dans le Bassin Parisien exploite un calcaire gélif corallien ou oolithique du Jurassique Supérieur. Les carrières d'Heidenheim et de Schelklingen-Vohenbronnen situées dans le Jura Souabe en Allemagne, à proximité d'Ulm, exploitent un calcaire pur du Jurassique Supérieur. La carrière de la Heutte située dans le Massif du Jura Bernois en Suisse exploite un calcaire marneux du Jurassique Supérieur (Figure 3.9 : 1, 2, 3, 4 et 5).

La carrière de Chaffois située dans le Massif du Jura, près de la ville de Pontarlier, la carrière d'Epagny située près de la ville de Dijon à l'extrême bord Sud-Est du Bassin Parisien ainsi que la carrière de Bainville-sur-Madon située à proximité de Nancy dans le Bassin Parisien exploitent le calcaire à polypiers du Jurassique Moyen (Figure 3.9 : 6, 7 et 8).

La carrière de Cielle située près de la ville de Rendeux en Belgique exploite le grès du Permien (Figure 3.9 9).

Roches volcaniques. La carrière de Trapp de Raon-l'Etape située dans les Vosges exploite du basalte porphyrique calco-alcalin d'âge Dévonien à Carbonifère Inférieur. La carrière de Lepuix-Gy située au sud des Vosges dans le Territoire de Belfort exploite la rhyodacite du Carbonifère Figure 3.9 : 10 et 11).

Roches plutoniques. La carrière de Saint-Amé située à proximité de la ville de Remiremont dans le Massif des Vosges exploite du granite de la fin du Carbonifère. La carrière de Waldhambach située près de la ville de Landau en Allemagne et la carrière de Seebach située dans le Massif de la Forêt Noire en Allemagne exploitent également du granite d'âge Carbonifère. La carrière de Jettenbach située en Bavière allemande extrait quant à elle de la microdiorite. La carrière de Groß-Bieberau située dans le canton de Hesse en Allemagne exploite du gabbro Figure 3.9 : 12, 13, 14, 15 et 16).

Roches métamorphiques. La carrière d'Heppenheim située au sud de Frankfurt dans le canton de Hesse en Allemagne exploite de la granodiorite avec des inclusions d'amphibolite de la fin du Carbonifère. La carrière de Dotternhausen située au bord du Massif du Jura Souabe en Allemagne exploite du calcaire mais également des schistes bitumineux du Jurassique Moyen. Enfin, la carrière de Hausach située dans le Massif de la Forêt Noire en Allemagne exploite du gneiss du Carbonifère (Figure 3.9 : 17, 18 et 19).

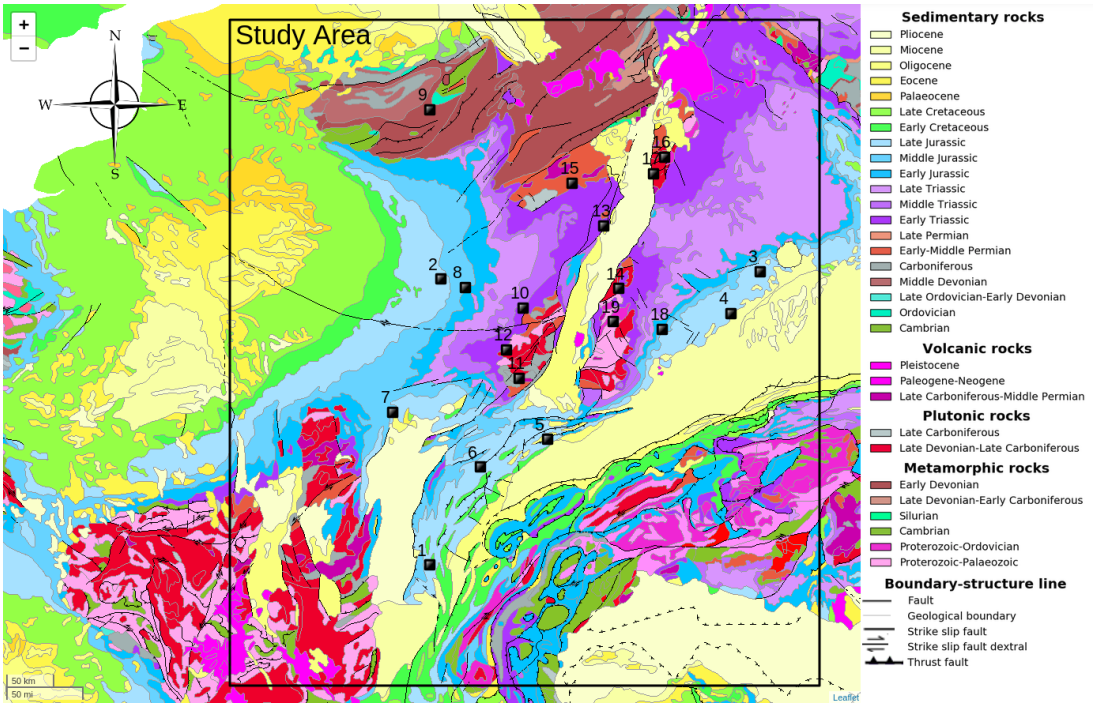
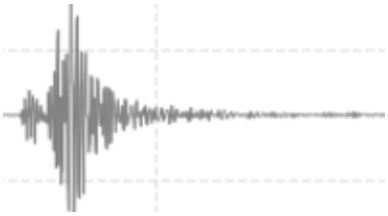


FIGURE 3.9: Distribution des exemples de carrière évoquées dans le paragraphe précédent pour la diversité globale du matériel qui est extrait dans la zone d'étude : roches sédimentaires calcaires de (1) Hauteville-Lompnes, (2) Pagny-sur-Meuse, (3) Heidenheim, (4) Schelklingen-Vohenbronnen, (5) La-Heutte, (6) Chaffois, (7) Epagny, et (8) Bainville-sur-Madon ; roches sédimentaires gréseuses de (9) Rendeux ; Roches volcaniques basaltiques de (10) Raon-l'Etape et (11) rhyodacitiques de Lepuix-Gy ; roches plutoniques granitiques de (12) Saint-Amé, (13) Waldhambach, (14) Seebach ; roches plutoniques microdioritiques de (15) Jettenbach et gabbroïques de (16) Groß-Bieberau ; roches métamorphiques à passées amphibolitiques de (17) Heppenheim, rpches métamorphiques schisteuses de (18) Dotternhausen et roches métamorphiques gneissiques de (19) Hausach.

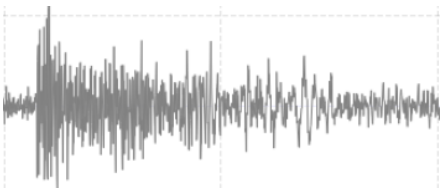
Produits dans des milieux d'une extrême diversité pétrographique et lithologique, ces tirs engendrent une variabilité de signaux qui sont détectables par les réseaux de stations. Cette variabilité observée dépend à la fois de la nature pétrographique du matériel extrait (Figures 3.10, 3.11, 3.12 et 3.13), des pratiques de dynamitage mais également de l'orientation du front de taille (Stump et al., 2001).

De plus, même si les formes d'ondes enregistrées aux mêmes stations se ressemblent très fortement pour des tirs ayant eu lieu dans une même carrière, les différents emplacements possibles pour ces tirs peuvent entraîner des variations sensibles dans les formes d'onde (Bonner et al., 2003). En effet, en plus d'une orientation variable du front de taille, des variations dans la nature du gisement pour un même site (présence d'un filon, dureté ou porosité différentielle, etc.) peuvent également modifier substantiellement les propriétés des roches formant ce front de taille. Enfin, la couverture azimutale des stations ainsi que les effets du milieu de propagation vont aussi fortement influencer la variabilité observée de ces formes d'onde enregistrées.

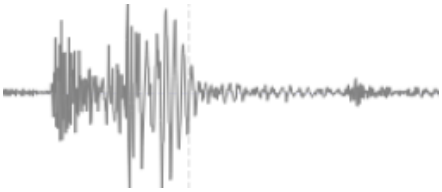
De part la variabilité intrinsèque observée des signaux associés aux tirs de carrière dans la zone d'étude, cette zone est donc idéale pour comprendre comment efficacement discriminer les tirs de carrière des séismes. C'est justement cette variabilité de formes d'onde observées dans cette zone d'étude que je souhaite exploiter pour solidement définir les caractéristiques fortes qui vont aider à distinguer avec une grande précision et exactitude les tirs de carrière des séismes.



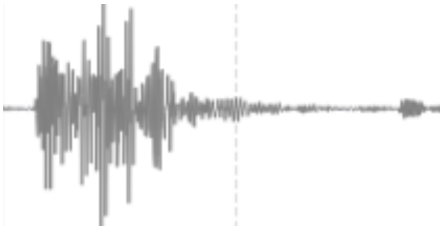
(a) Calcaire marbrier de Hauteville-Lompnes (station OG35 : 7.24 km)



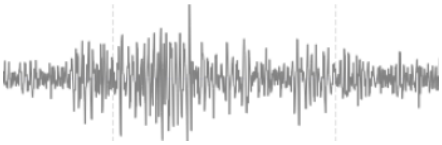
(b) Calcaire gélif et friable de Pagny-sur-Meuse (station A210A : 18.84 km)



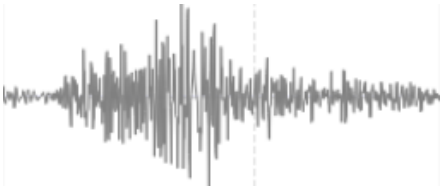
(c) Calcaire pur de Heidenheim (station A360A : 6.22 km)



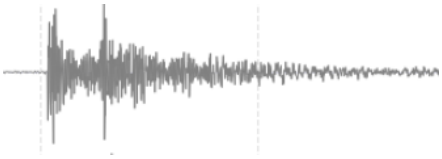
(d) Calcaire pur de Schelklingen-Vohenbronnen (station A108A : 3.67 km)



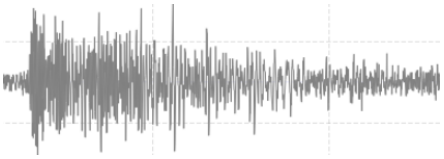
(e) Calcaire marneux de La-Heutte (station BOURR : 24.86 km



(f) Calcaire pur sublithographique de Chaffois (station BRANT : 14.23 km)



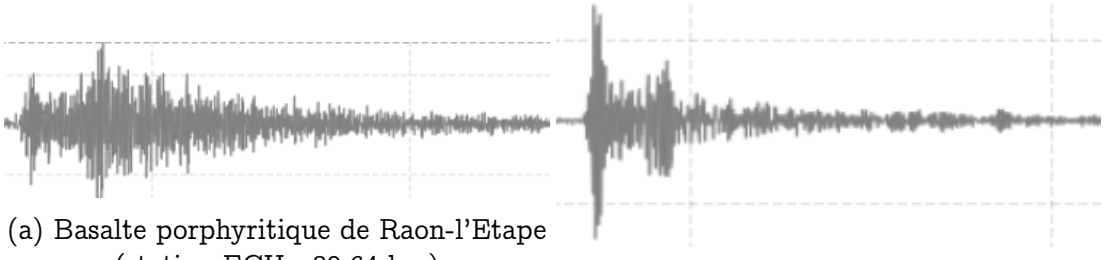
(g) Calcaire dalleux d'Epagny (station A160A : 28.29 km)



(h) Calcaire stratifié et diaclasé de Bainville-sur-Madon (station A210A : 27.34 km)

(i) Grès quartzitique de Rendeux (station RCHB : 25.03 km)

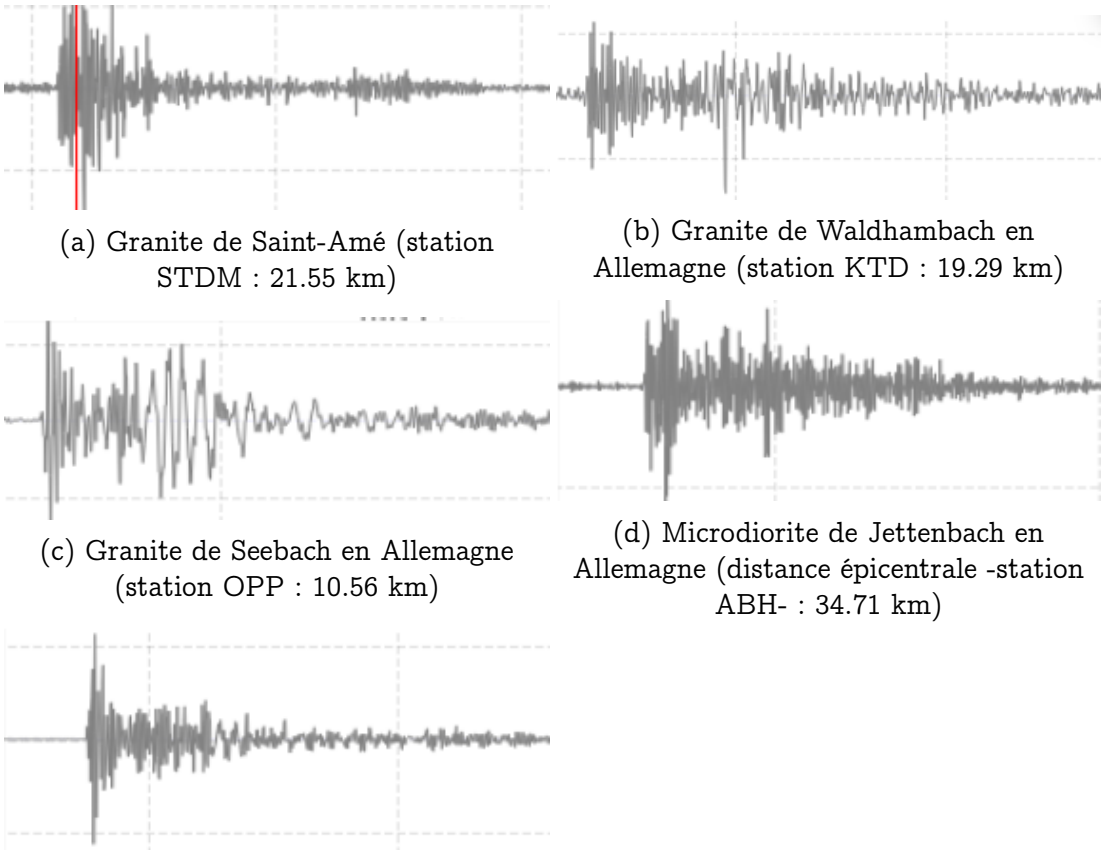
FIGURE 3.10: Exemples de variabilité des formes d'ondes enregistrées sur la composante verticale à la première station pour des tirs de carrière ayant eu lieu dans des roches sédimentaires.



(a) Basalte porphyritique de Raon-l'Etape (station ECH : 30.64 km)

(b) Rhyodacite de Lepuix-Gy (station RONF : 11.44 km)

FIGURE 3.11: Exemples de formes d'ondes enregistrées sur la composante verticale à la première station pour des tirs de carrière ayant eu lieu dans des roches volcaniques.



(a) Granite de Saint-Amé (station STDM : 21.55 km)

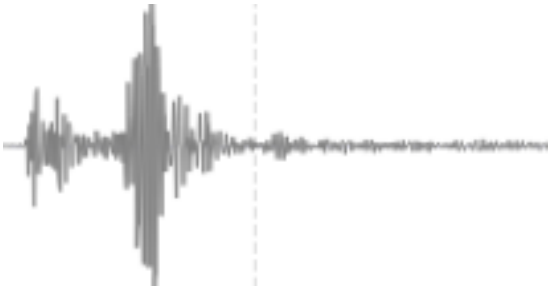
(b) Granite de Waldhambach en Allemagne (station KTD : 19.29 km)

(c) Granite de Seebach en Allemagne (station OPP : 10.56 km)

(d) Microdiorite de Jettenbach en Allemagne (distance épicentrale -station ABH- : 34.71 km)

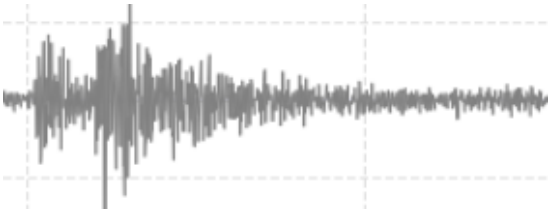
(e) Gabbro de Groß-Bieberau en Allemagne (station A110A : 8.74 km)

FIGURE 3.12: Exemples de formes d'ondes enregistrées sur la composante verticale à la première station pour des tirs de carrière ayant eu lieu dans des roches plutoniques.



(a) Granodiorite avec passées amphibolitiques de Heppenheim en Allemagne (station A113A : 11.04 km)

(b) Calcaires et schistes bitumineux de Dotternhausen en Allemagne (station GUT : 18.55 km)



(c) Gneiss de Hausach en Allemagne (station OPP : 23.18 km)

FIGURE 3.13: Exemples de formes d'ondes enregistrées sur la composante verticale à la première station pour des tirs de carrière ayant eu lieu dans des roches métamorphiques.

3.2 Des données volumineuses et de qualité

3.2.1 Un réseau sismologique récemment densifié

•Un apport supplémentaire de stations permanentes de qualité.

Le réseau sismologique de la zone d'étude a grandement été densifié depuis ces dernières années grâce à différents projets. On notera tout d'abord le construction du Réseau Large Bande Permanent dans le cadre dans le cadre de l'infrastructure de recherche RESIF (https://www.allenvi.fr/groupes-transversaux/infrastructures-de-recherche/resif). En particulier le volet Large Bande (RESIF-CLB) a permis la construction d'un réseau de 160 stations large bande sur l'ensemble du territoire français métropolitain.

Plus particulièrement dans notre zone d'étude, ce projet a permis la modernisation du réseau courte période existant depuis les années 1980 en réseau large bande (3-composantes, large bande, transmission temps réel 3G/4G-ADSL, etc.). Parmi l'ensemble de ces stations, on compte 6 stations construites avec des capteurs installés en fond de puits (5m) en 2016, évoluant vers 11 en 2020.

En complément, un projet d'instrumentation sismologique a été mené par l'École et Observatoire des Sciences de la Terre (EOST) et Électricité de Strasbourg (projet EGS, ADEME, 2016-2020) pour densifier le réseau en Alsace, et

améliorer la détection de séismes de faible magnitude. Ce réseau permet notamment d'être en mesure de mieux surveiller les éventuelles séquences sismiques induites par l'activité industrielle, l'activité géothermique notamment.

Les stations construites sont instrumentées à la fois par des capteurs accélérométriques et des capteurs vélocimétriques moyenne et large bande. La difficulté d'instrumenter cette région avec des stations de qualité réside dans le fait que celle-ci est très urbanisée comme précisé ci-dessus, mais aussi dans la nature même du sol en plaine d'Alsace, où la couverture sédimentaire peu consolidée perturbe les signaux sismiques. Ainsi, certains capteurs ont été installés jusqu'à 45 m pour limiter les effets de site.

En parallèle, les instituts de surveillance sismologiques en Allemagne ont également densifié et modernisé leur réseau de l'autre côté de la frontière. L'ensemble des signaux sont quotidiennement partagés grâce aux différents centres d'archivage et de distribution des données, dont l'initiative EIDA (European Integrated Data Archive), par chacun des organismes en charge de la surveillance sismologique de chacun des pays ou des landers.

- **Une forte densification du réseau par des stations temporaires.**

En plus de la densification des stations permanentes depuis 2015, un réseau temporaire AlpArray-Fr a pu compléter le réseau permanent jusqu'en 2020 (Figure 3.14). L'édification de ce réseau temporaire a été inscrit dans le cadre du projet européen AlpArray qui a permis d'impliquer plusieurs pays européens, dont la France, pour densifier le réseau permanent autour de la chaîne alpine (Hetényi et al., 2018).

De ce fait, la période 2016-2019 correspond à un maximum de couverture des stations à l'échelle de la zone d'étude (apport des stations permanentes françaises, allemandes, suisses, belges puis des stations temporaires AlpArray). Elle constitue donc une période propice pour développer cette méthodologie de détection des petits séismes dans la zone.

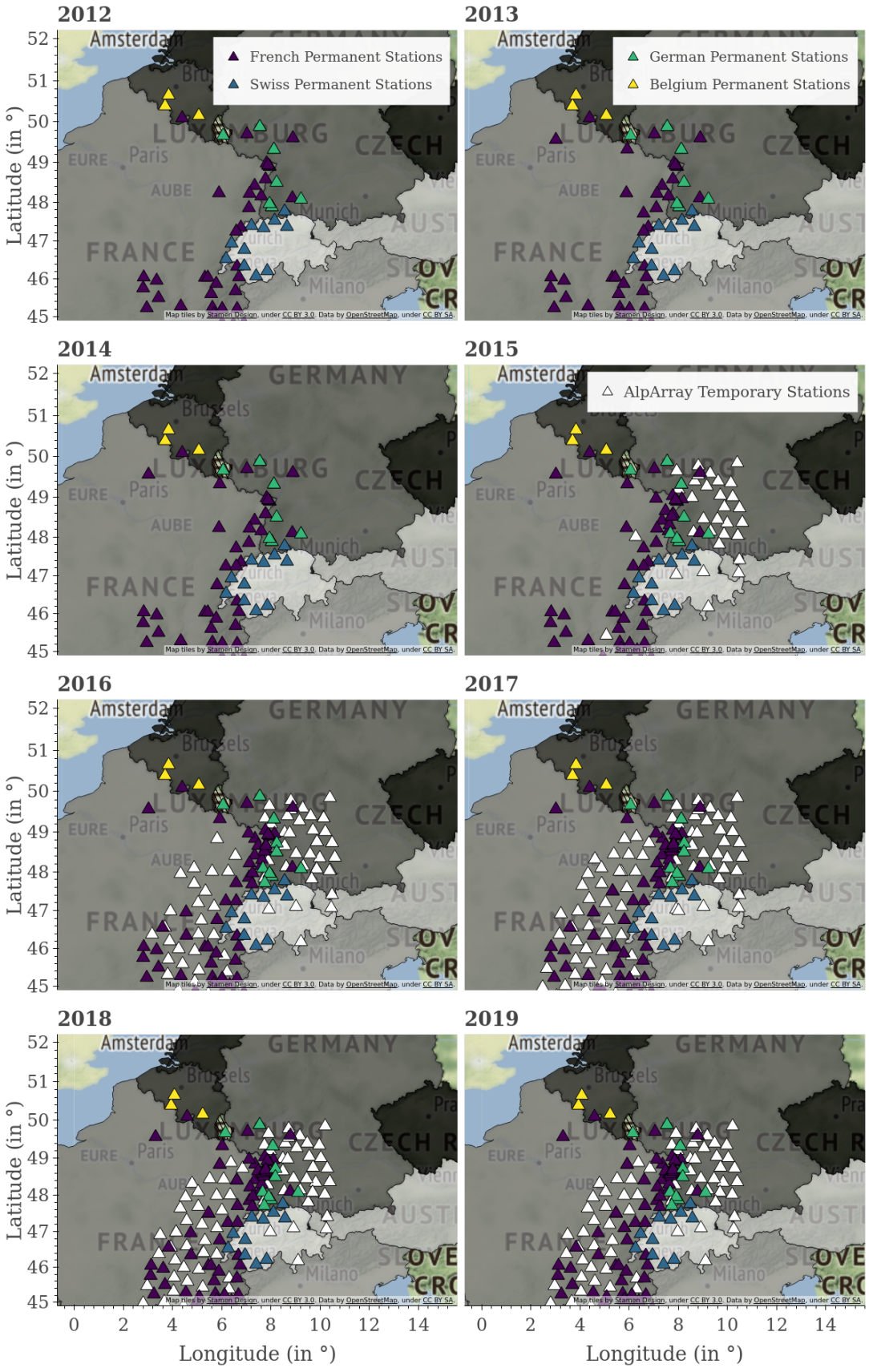


FIGURE 3.14: Evolution de la couverture de stations dans la zone d'étude depuis 2012.

•Potentiel de détection du réseau densifié

Le suivi de l'évolution de la détection de la sismicité naturelle par le BCSF-RéNaSS pour la période 2012-2020 montre un nombre de petits séismes qui augmentent en fonction de la densification progressive du réseau de stations (Annexe A). La distribution des magnitudes des séismes autour de 1 a été effectivement multipliée par 8 depuis 2012. De plus, un nombre de 2.5 fois plus de séismes de magnitude inférieur à 1 ont été détectés depuis 2017 ; ce qui correspond au maximum de couverture de stations de la zone d'étude (Figure 3.15).

Si l'on estime approximativement la magnitude de complétude à partir de la distribution cumulative fréquence-magnitude des séismes pour chaque année depuis 2013, celle-ci varie très peu et reste autour de 1.2. Par conséquent, même si le nombre de séismes de magnitude locale MLv inférieure à 1.0 a augmenté très fortement ces 3 dernières années, ce nombre supplémentaire de séismes détectés n'a pas d'incidence majeure sur la valeur de la magnitude de complétude globale de la zone étudiée, malgré les incertitudes estimées sur le calcul des petites magnitudes (Figures 3.16 et 3.17).

La détection des événements de magnitude inférieure à 1.2 reste donc encore sous-exploitée par le système de détection actuel du BCSF-RéNaSS. Il est important de noter que les stations AlpArray n'ont pas été intégrées au système de détection automatiques pendant toute leur période d'activité puisque les signaux n'étaient pas transmis en temps réel. Ainsi, elles n'ont été utilisées que pour localiser les événements. Inclure ce réseau dans une procédure de détection des séismes de faible magnitude offre une opportunité majeure pour augmenter la détectabilité des petits événements.

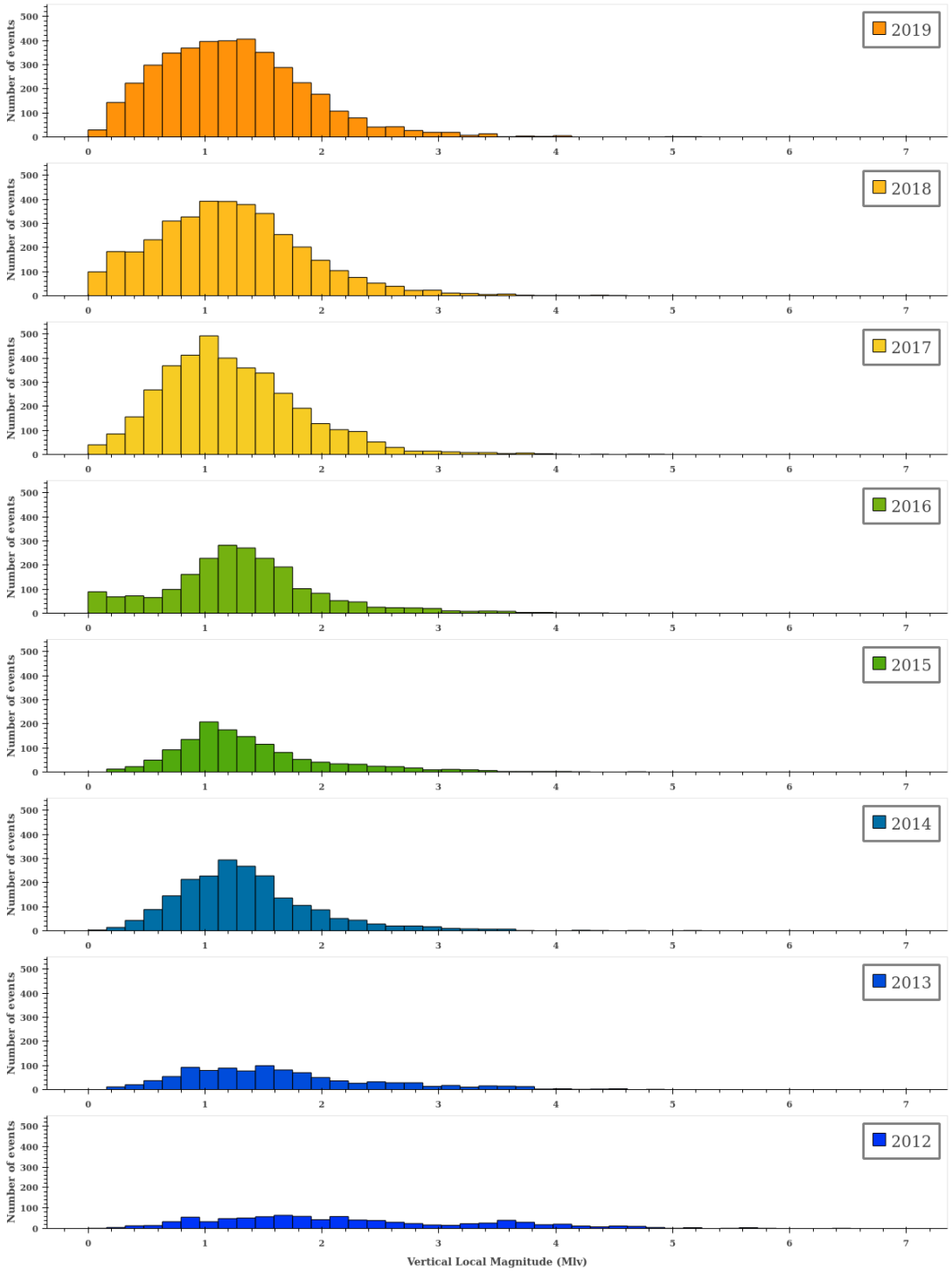


FIGURE 3.15: Évolution de distribution des magnitudes pour la période 2012-2019.

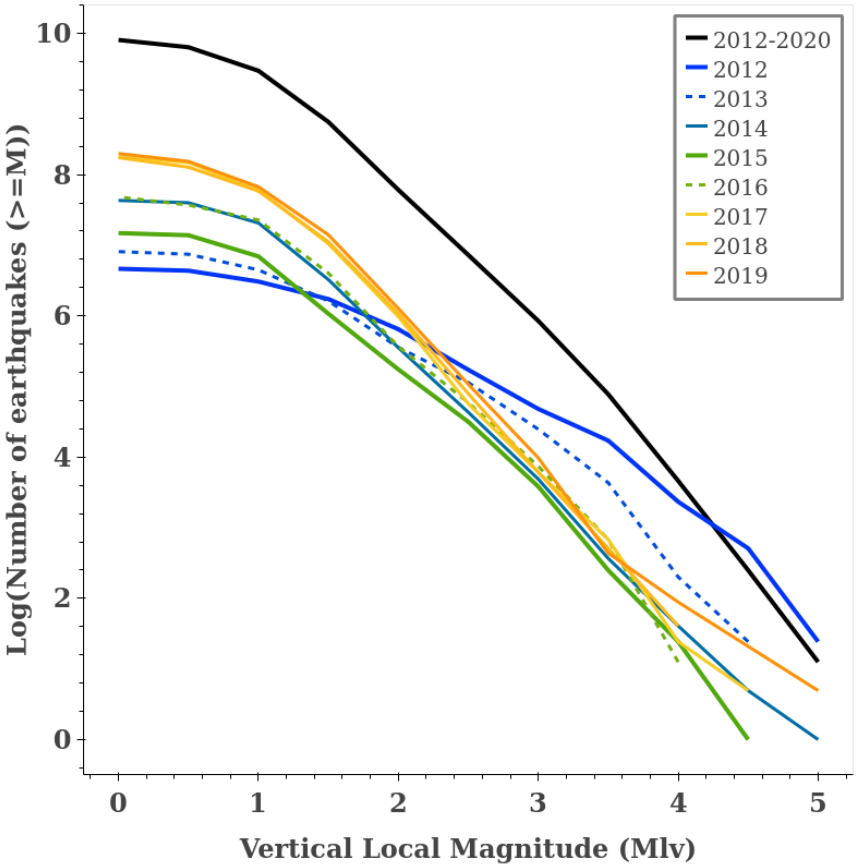


FIGURE 3.16: Distributions cumulatives fréquence-magnitude annuelles des séismes détectés par le réseau de stations utilisé par le BCSF-RéNaSS pour la période 2012-2019.

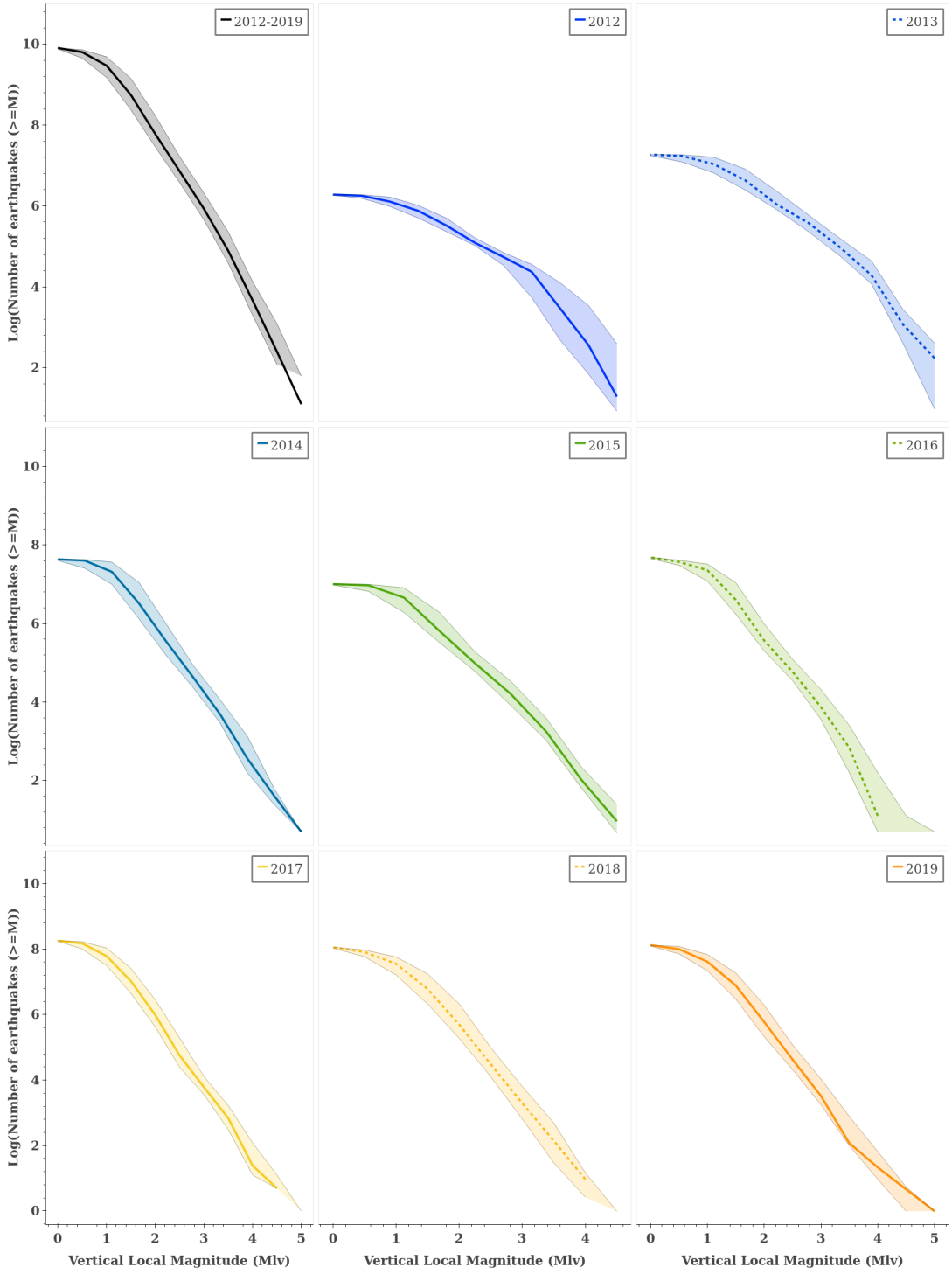


FIGURE 3.17: Distributions cumulatives des séismes détectés par le réseau de stations utilisé par le BCSF-RéNaSS pour la période 2012-2019 et incertitudes associées.

La période 2016-2019 est donc encore une fois la période idéale pour développer une procédure de détection qui puisse exploiter au maximum les capacités de détection du réseau (réseau à maillage plus fin, seuils de détection plus bas). De cette manière, un volume de 4 TéraBytes de sismogrammes échantillonnés à 100 Hz sur 3 canaux (2 composantes horizontales et une composante verticale) est donc disponible pour cette période d'étude.

3.2.2 Un réseau plus sensible au bruit d'origine anthropique

- **Caractérisation des stations impliquées dans la détection des faux événements.**

Si la période comprise entre 2016 et 2019 est la meilleure période pour détecter les petits séismes dans la zone d'étude, il s'avère que l'ajout de stations supplémentaires, combiné à un seuil de détection plus bas, augmente fortement le taux de fausses détections, comme expliqué précédemment (Chapitre 2). En effet, le résultat d'un test de détection automatique sur 4 mois (septembre 2016-décembre 2016) engendre environ 48 000 faux événements.

Une proportion de 26 % des stations temporaires AlpArray utilisées (soit 18 stations) sont effectivement impliquées dans la création d'au moins 10% des faux événements. Par exemple, des stations telles que A117A et A102A interviennent dans la détection de plus de 20 % des faux événements (Figure 3.18).

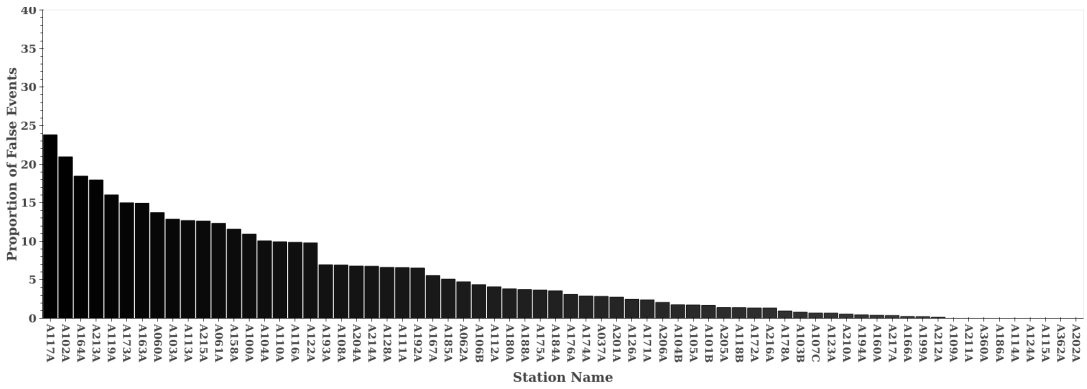


FIGURE 3.18: Taux d'implication des stations temporaires AlpArray (en %) dans la création de faux événements.

Si l'on compare le taux d'implication des stations AlpArray dans la création de faux événements à celui des stations permanentes, le taux d'implication diminue à 13% pour les stations permanentes. Parmi celles-ci, 2 stations (GIMEL et OGSI) interviennent dans la création de plus de 25 % des faux événements (Figure 3.19).

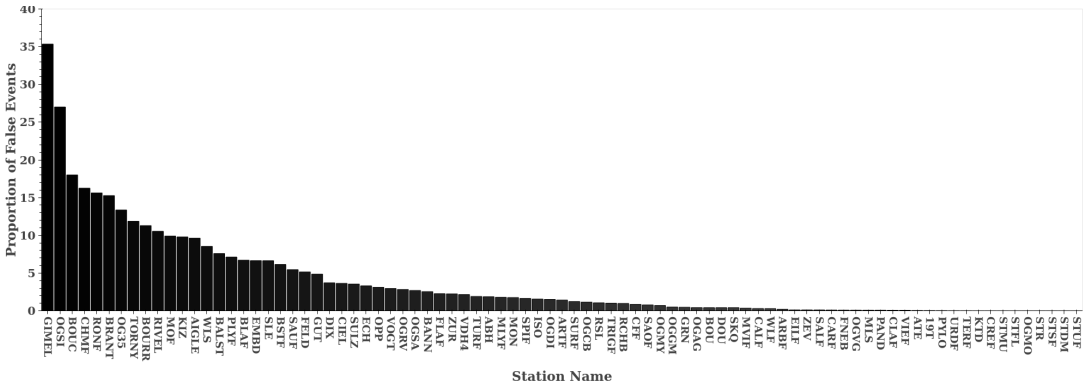


FIGURE 3.19: Taux d'implication des stations permanentes (en %) dans la création des faux événements pour la période septembre 2016-décembre 2016.

Cependant, les stations qui sont en fait le plus impliquées dans la génération des faux événements sont également celles qui sont largement utilisées pour la détection automatique des vrais événements.

Par exemple, la station A102A est une des stations AlpArray les plus engagées dans l'élaboration des faux événements (de l'ordre de 21%) pour la période test septembre 2016-décembre 2016, mais elle est également la station AlpArray la plus utilisée dans la création des vrais événements pour cette même période (de l'ordre de 40%). En définitive, l'ensemble des stations AlpArray qui sont impliquées dans la création de faux événements interviennent dans 4 à 40% des vrais détections (Figure 3.20).

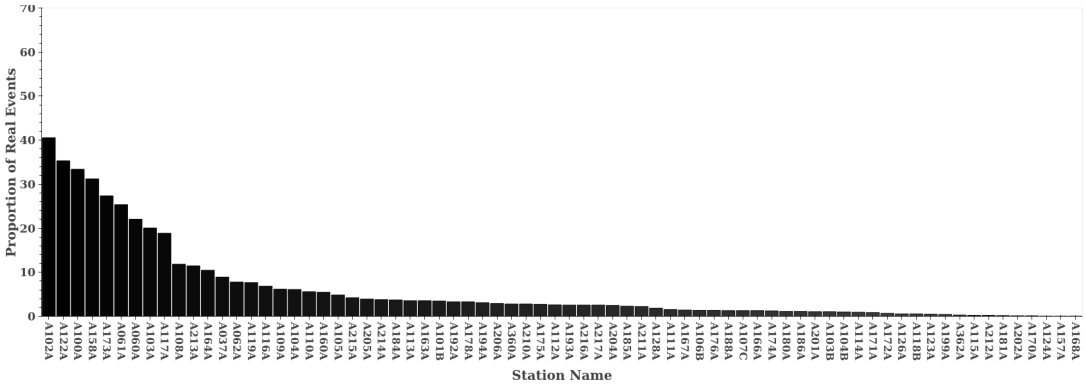


FIGURE 3.20: Taux d'implication des stations temporaires AlpArray (en %) dans la création des vrais événements pour la période septembre-décembre 2016.

Quant aux stations permanentes utilisées pour cette même période test, si le même constat peut également être fait, la part dominante d'implication de ces stations est plutôt réservée à la création de vrais événements. Des stations comme KIZ, SLE, GUT, CHMF, FELD ou ECH sont impliquées dans la création de plus de 50% des vrais événements. En revanche, elles sont utilisées pour la détection des faux événements à la hauteur de "seulement" 6 à 16% d'entre eux (Figure 3.21).

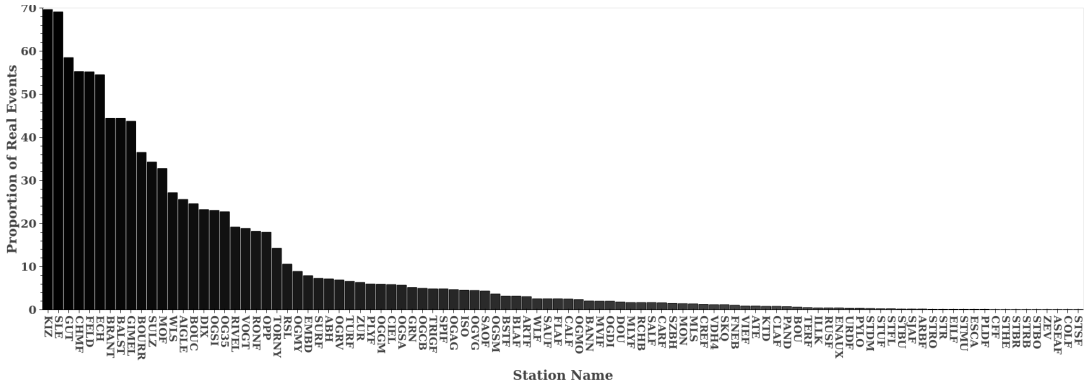


FIGURE 3.21: Taux d'implication des stations permanentes (en %) dans la création des vrais événements pour la période septembre-décembre 2016.

De ce fait, la détection des petits séismes s'accompagne inexorablement d'un fort taux de fausses détections car ce sont approximativement les mêmes stations qui interviennent dans la génération des vrais et fausses détections.

•**Caractérisation du bruit détecté aux stations**

Un niveau de bruit élevé aux hautes fréquences ($>$ 1 Hz). La diminution du seuil de détection, combinée à un taux élevé d'enregistrement de signaux impulsifs variés à des stations localisées proches de centres d'activité anthropique, augmente les probabilités de détection d'autres signaux que ceux associés aux séismes.

Par exemple, la station A102A est localisée près d'une route à proximité d'un centre équestre et à environ 2 km de la carrière de Sigmaringen dans le Sud de l'Allemagne.

L'analyse de la fonction de densité spectrale de puissance pour évaluer le niveau de bruit de fond de la station A102A montre que la puissance du bruit aux gammes de fréquences typiques du bruit d'origine anthropique, c'est-à-dire comprises entre 1 et 10 Hz, est variable et peut augmenter d'environ 20 décibels par rapport à la puissance minimale. Cette puissance de bruit atteint alors des probabilités plus fortes d'occurrence (de l'ordre de 20 %) par rapport au modèle de bruit bas (NLNM). Cette station est donc très sensible au bruit impulsif transitoire d'origine anthropique.

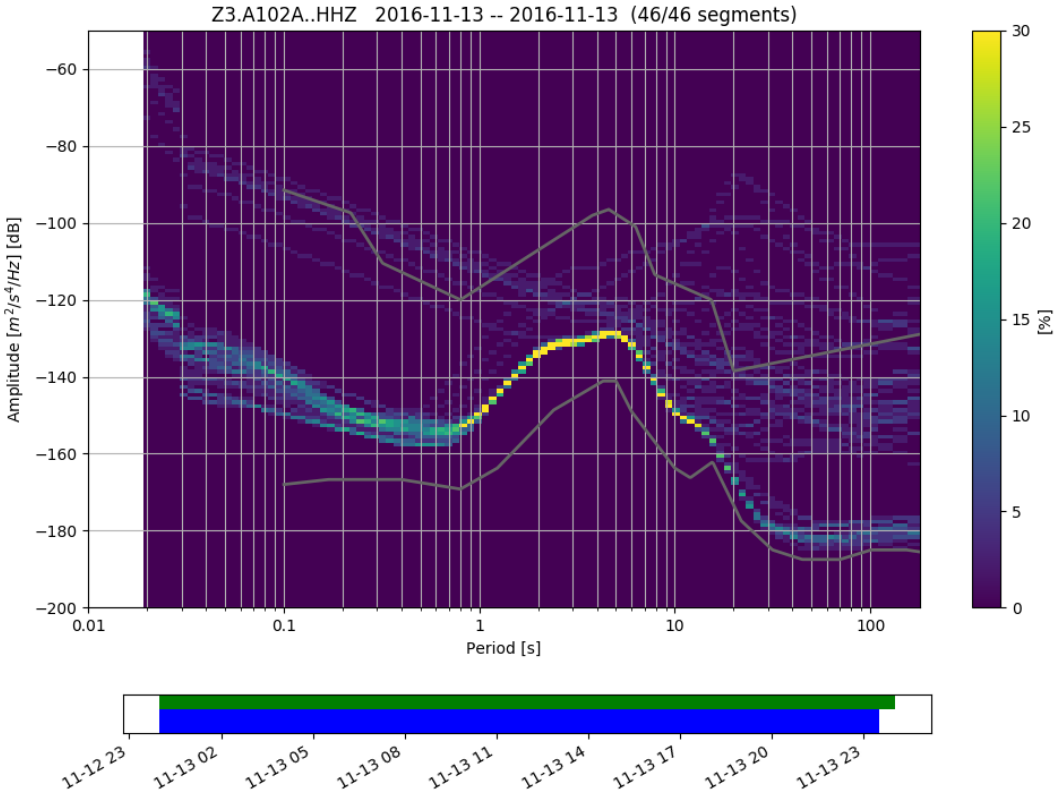


FIGURE 3.22: Densité spectrale de puissance probabiliste calculée pour la station A102A. Les courbes grises correspondent aux modèles de bruit standard (courbe supérieure = modèle de bruit élevé [NHNM] et courbe inférieure = modèle de bruit bas [NLNM] (PETERSON, 1993). Les niveaux de bruit de la station sont estimés sur une large gamme de fréquences de 0.01 Hz à 16 Hz (soit une période de 100 secondes à 0.0625 secondes). En bas du graphique sont affichés les données qui ont servi au calcul de cette fonction. Le rectangle vert représente les données disponibles et le rectangle bleu montre l'étendue des données qui ont servi au calcul. Ces spectres ont été obtenus via le package ObsPy de Python suivant la méthode de MCNAMARA et al., 2004.

Il en est de même pour les stations A213A ou A117A par exemple. La station A213A est située au Nord-Est de la région de Dijon en France à 200 m d'un moto club, et A117A est localisée dans une exploitation agricole, à proximité d'une petite industrie textile, au Nord-Est de la région de Stuttgart en Allemagne. La puissance du bruit aux gammes de fréquence caractéristiques du bruit d'origine anthropique atteint des probabilités d'occurrence très élevées jusqu'à près de 20% pour les 2 stations, et s'éloigne fortement des valeurs estimées pour le modèle de bruit bas (Figures 3.24 et 3.23). Un pic autour de la gamme de fréquence 10-20 Hz est observé sur la station A213A, correspondant à la puissance du bruit d'un trafic autoroutier (Riahi et al., 2015 ; Díaz et al., 2017 ; Xiao et al., 2020). Si l'autoroute est située à 25 km de cette station, il est plus probable que ce pic corresponde à l'activité du moto club situé à proximité.

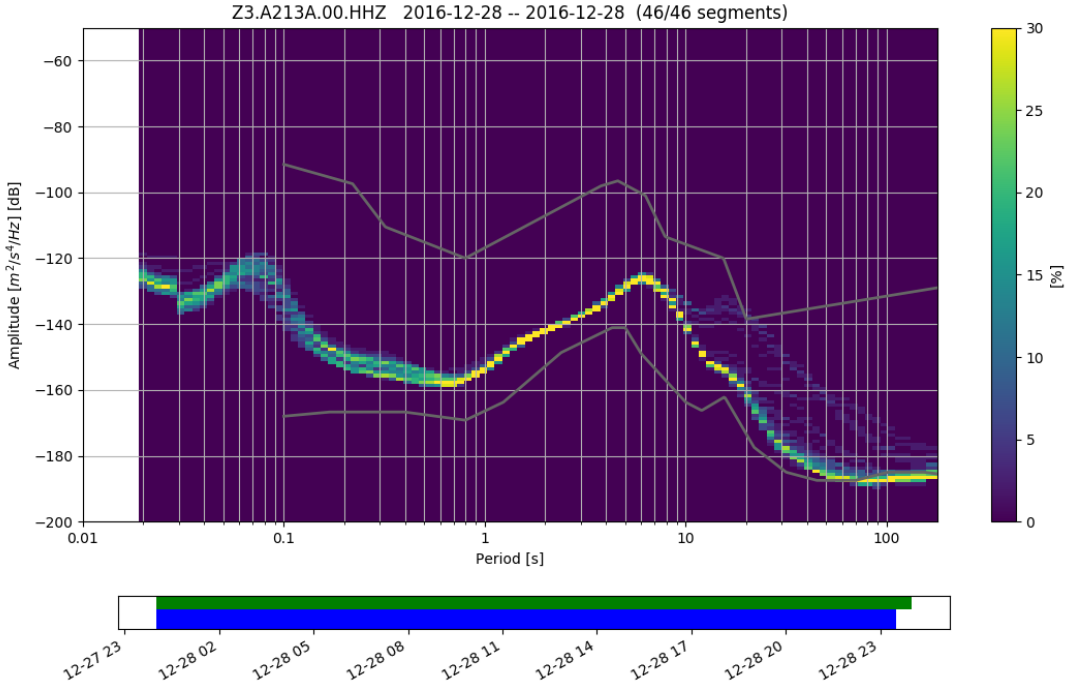


Figure 3.23: Densité spectrale de puissance probabiliste calculée pour la station A213A. Les courbes grises correspondent aux modèles de bruit standard (courbe supérieure = modèle de bruit élevé [NHNM] et courbe inférieure = modèle de bruit bas [NLNM] (Peterson, 1993). Les niveaux de bruit de la station sont estimés sur une large gamme de fréquences de 0.01 Hz à 16 Hz (soit une période de 100 secondes à 0.0625 secondes). En bas du graphique sont affichées les données qui ont servi au calcul de cette fonction. Le rectangle vert représente les données disponibles et le rectangle bleu montre l'étendue des données qui ont servi au calcul.Ces spectres ont été obtenus via le package ObsPy de Python suivant la méthode de McNamara et al., 2004.

La station A117A affiche un niveau de bruit globalement plus élevé que celui de la station A102A pour l'ensemble des gammes fréquentielles présentées. La puissance du bruit longue période tend notamment à se rapprocher des valeurs du modèle de bruit élevé avec une probabilité plutôt forte (de 20 à 30%). De cette façon, si cette station est la station AlpArray la plus impliquée dans la création des faux événements pour la période considérée, elle est proportionnellement beaucoup moins impliquée dans la création de vrais événements que la station A102A.

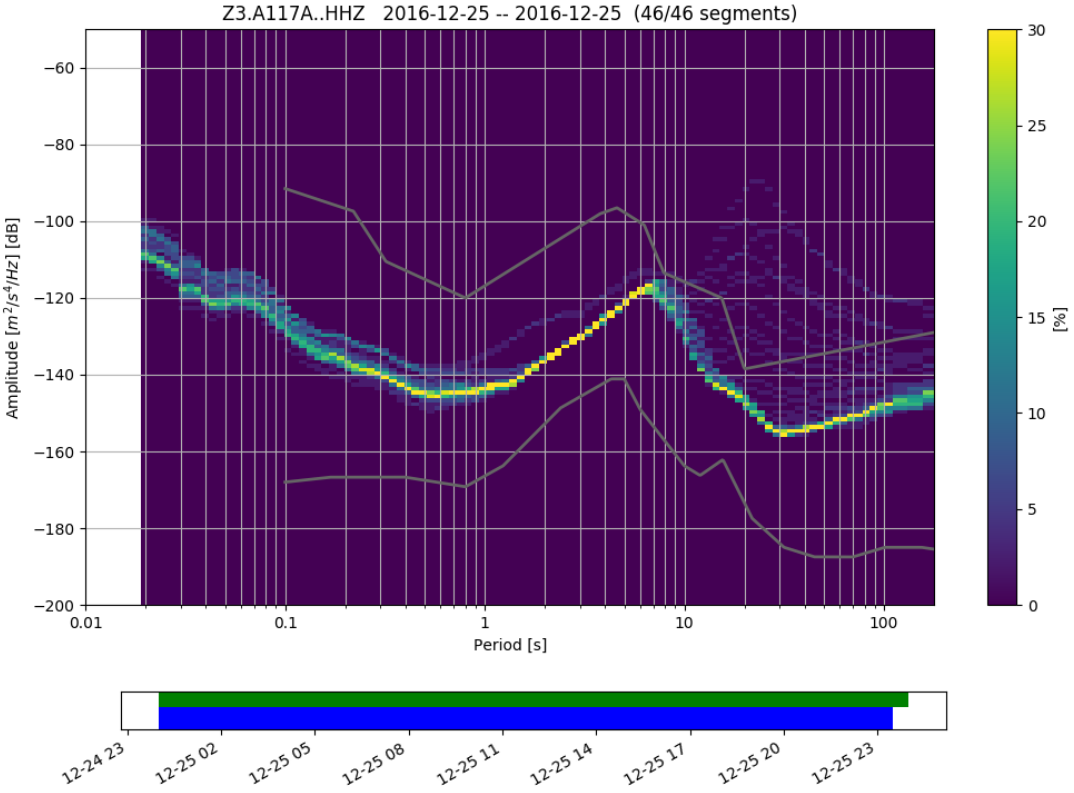


FIGURE 3.24: Densité spectrale de puissance probabiliste calculée pour la station A117A. Les courbes grises correspondent aux modèles de bruit standard (courbe supérieure = modèle de bruit élevé [NHNM] et courbe inférieure = modèle de bruit bas [NLNM] ; PETERSON, 1993. Les niveaux de bruit de la station sont estimés sur une large gamme de fréquences de 0.01 Hz à 16 Hz (soit une période de 100 secondes à 0.0625 secondes). En bas du graphique sont affichées les données qui ont servi au calcul de cette fonction. Le rectangle vert représente les données disponibles et le rectangle bleu montre l'étendue des données qui ont servi au calcul.Ces spectres ont été obtenus via le package ObsPy de Python suivant la méthode de MCNAMARA et al., 2004.

Le même constat peut être effectué pour les stations permanentes. C'est la cas par exemple de la station GIMEL qui est la station permanente intervenant le plus dans la création de faux événements, et qui fait partie des stations permanentes les plus utilisées pour la création des vrais événements. Cette station est située dans le Jura Vaudois en Suisse à proximité d'une route et

d'une activité de Gravière (Figure 3.25).

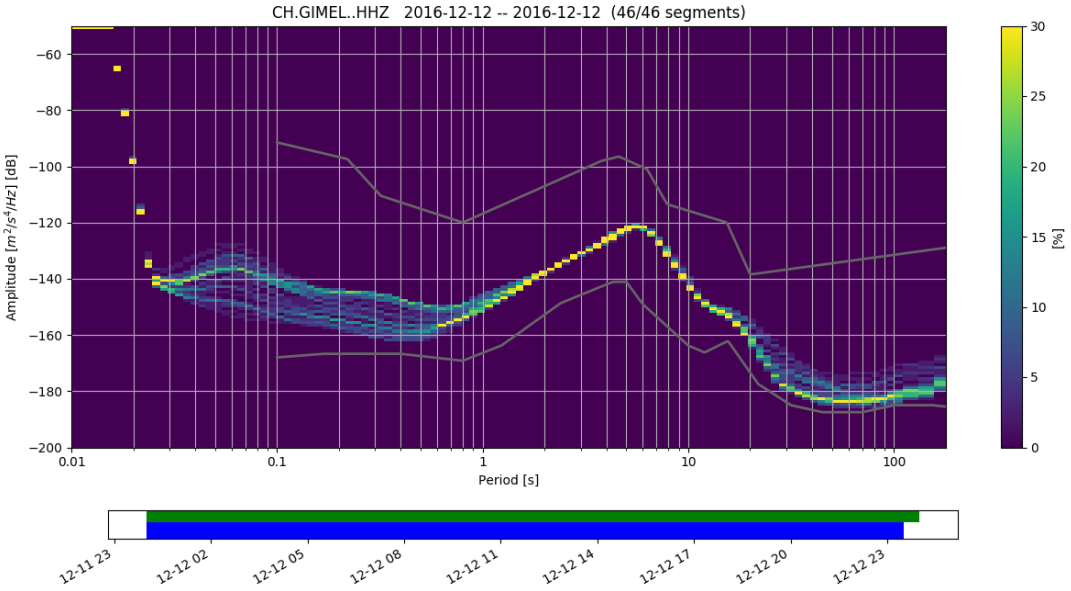


FIGURE 3.25: Densité spectrale de puissance probabiliste calculée pour la station GIMEL. Les courbes grises correspondent aux modèles de bruit standard (courbe supérieure = modèle de bruit élevé [NHNM] et courbe inférieure = modèle de bruit bas [NLNM] ; PETERSON, 1993). Les niveaux de bruit de la station sont estimés sur une large gamme de fréquences de 0.01 Hz à 16 Hz (soit une période de 100 secondes à 0.0625 secondes). En bas du graphique sont affichées les données qui ont servi au calcul de cette fonction. Le rectangle vert représente les données disponibles et le rectangle bleu montre l'étendue des données qui ont servi au calcul. Ces spectres ont été obtenus via le package ObsPy de Python suivant la méthode de MCNAMARA et al., 2004.

Ainsi, les stations les plus impliquées dans la création de vrais événements sont donc également celles qui sont sensibles au bruit impulsif d'origine anthropique. Les faux événements qui sont générés sont donc majoritairement reliés à ce type de bruit.

Une détection maximale aux heures d'activité humaine intense

L'étude de la répartition de la totalité des faux événements détectés en fonction des heures de la journée pour la période septembre-décembre 2016 montre que celle-ci se concentre effectivement autour des heures qui correspondent aux pics d'activité humaine (Figure 3.26), ce qui est une des caractéristiques du bruit d'origine anthropique (SHEEN et al., 2009).

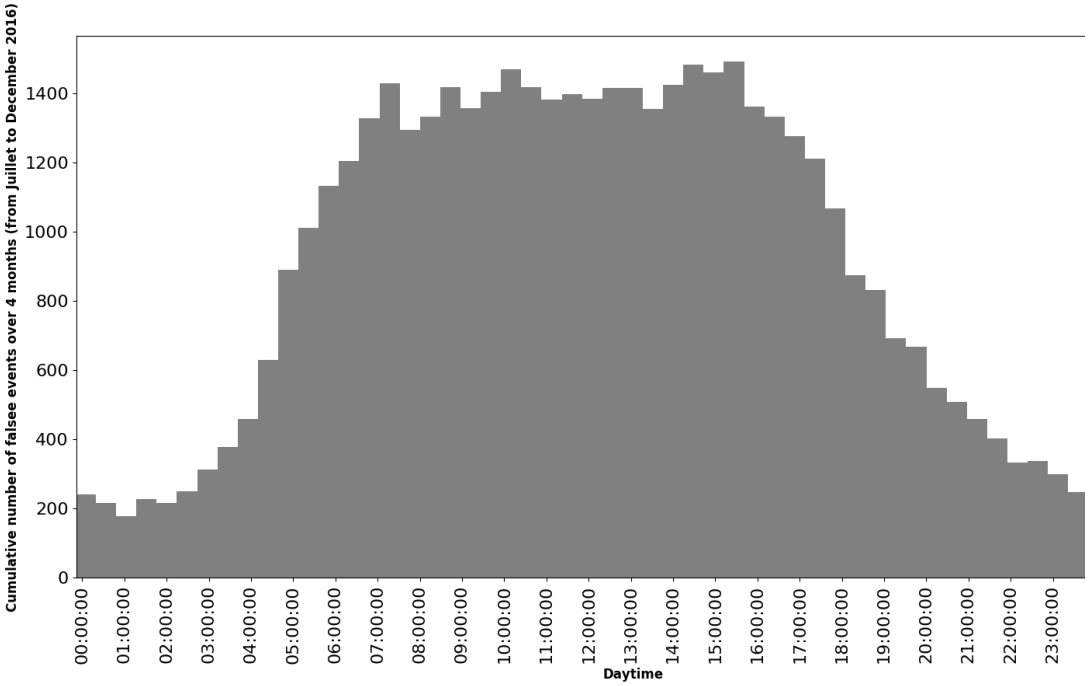


FIGURE 3.26: Distributions des faux événements détectés par une procédure automatique de détection incluant l'ensemble du réseau de stations disponible pour la période juillet 2016-décembre 2016 en fonction des heures de la journée.

Des signaux associés au bruit anthropique détectés avec les mêmes amplitudes, durées et contenus fréquentiels que les signaux sismiques associés aux séismes et aux tirs de carrière. L'analyse des spectrogrammes des signaux ayant engendré les faux événements montre une intensité maximale du signal dans les bandes de fréquence typique du bruit d'origine anthropique, c'est-à-dire principalement concentrée entre 1 et 10 Hz, mais pouvant s'étendre jusqu'à 20 Hz (Figures 3.27 à 3.29). Les variations diurnes d'amplitude du signal entre 1 et 20 Hz sont effectivement reconnues comme étant associées au bruit d'origine anthropique (BUNGUM et al., 1971 ; GURROLA et al., 1990 ; YOUNG et al., 1996 ; ATEF et al., 2009 ; LEWIS et al., 2012 ; LOER et al., 2018).

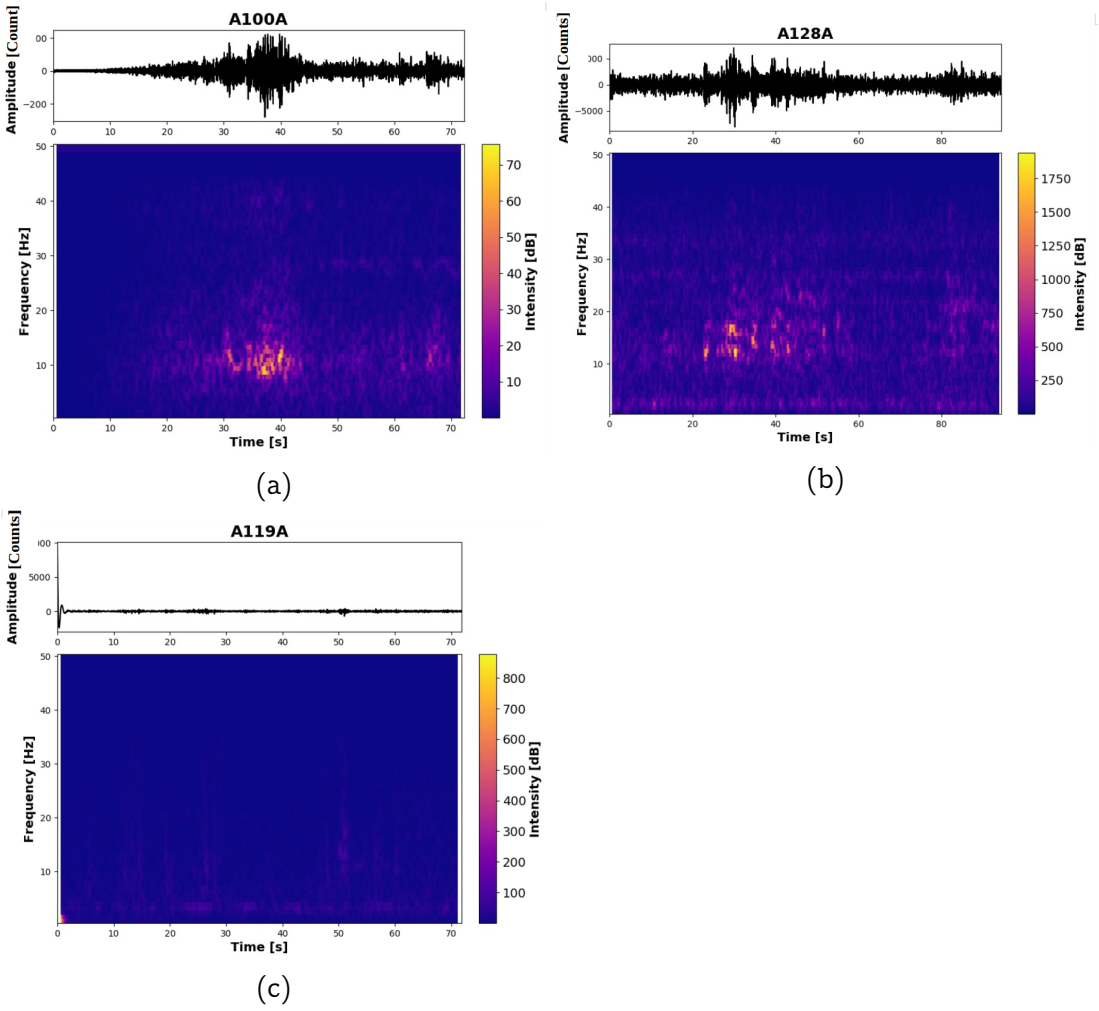


FIGURE 3.27: Formes d'ondes et spectrogrammes correspondant aux signaux enregistrés sur la composante verticale des stations (a) A100A, (b) A128A et (c) A119A. Ces signaux sont reliés à un faux événement détecté le 24 décembre 2016 à 11h14.

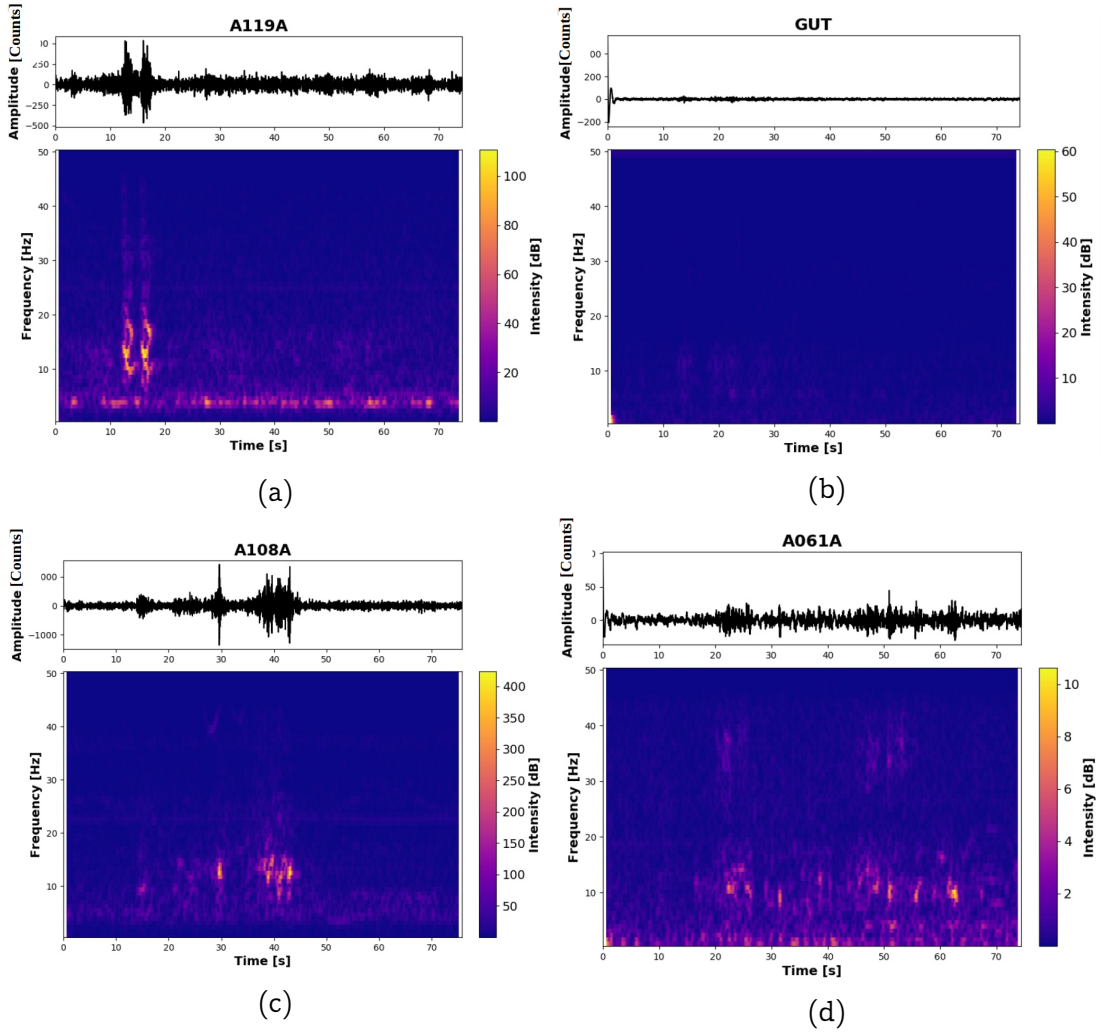


FIGURE 3.28: Formes d'onde et spectrogrammes correspondant aux signaux enregistrés sur la composante verticale des stations (a) A119A, (b) GUT, (c) A108A et (d) A061A. Ces signaux sont reliés à un faux événement détecté le 12 novembre 2016 à 09h29.

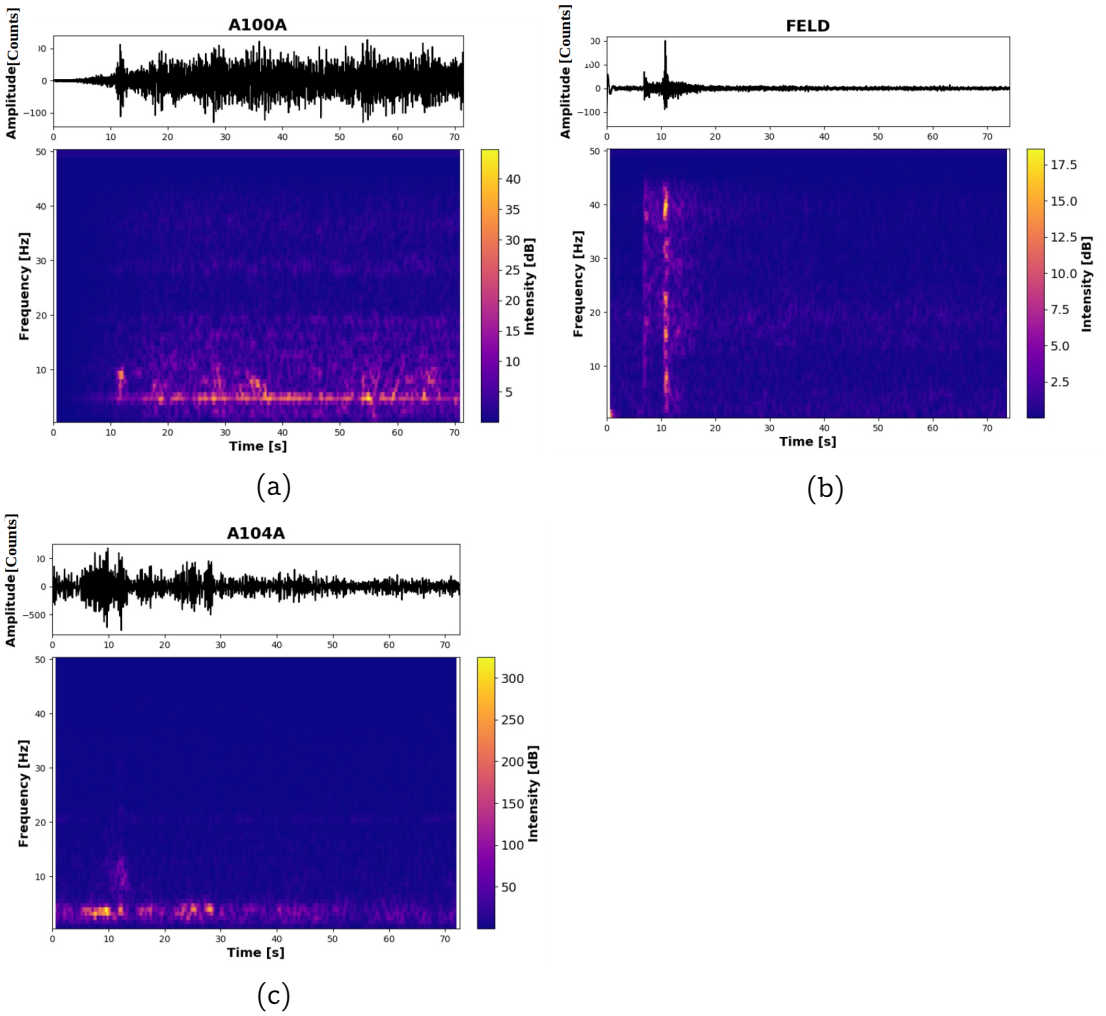


FIGURE 3.29: Formes d'onde et spectrogrammes correspondant aux signaux enregistrés sur la composante verticale des stations (a) A100A, (b) FELD et (c) A104A. Ces signaux sont reliés à un faux événement détecté le 27 octobre 2016 à 11h19. Ce faux événement contient dans son association un signal sismique isolé enregistré à la station FELD.

Or, ces signaux d'origine anthropique véhiculent un maximum d'énergie dans la bande de fréquences de 1 à 10 Hz qui est souvent utilisée pour observer l'activité microsismique (Hutton et al., 2010 ; Riahi et al., 2015 ; Inbal et al., 2018). En effet, l'analyse de quelques spectrogrammes de signaux correspondant à des séismes de très faible magnitude (ici MLv de 0.3 et 1.4, Figures 3.30 et 3.31) montre une concentration de l'intensité du signal dans cette gamme fréquentielle, avec un maximum autour de 5 à 10 Hz. Quelques pics d'intensité apparaissent jusqu'à 20 Hz et semblent correspondre à l'arrivée des ondes P.

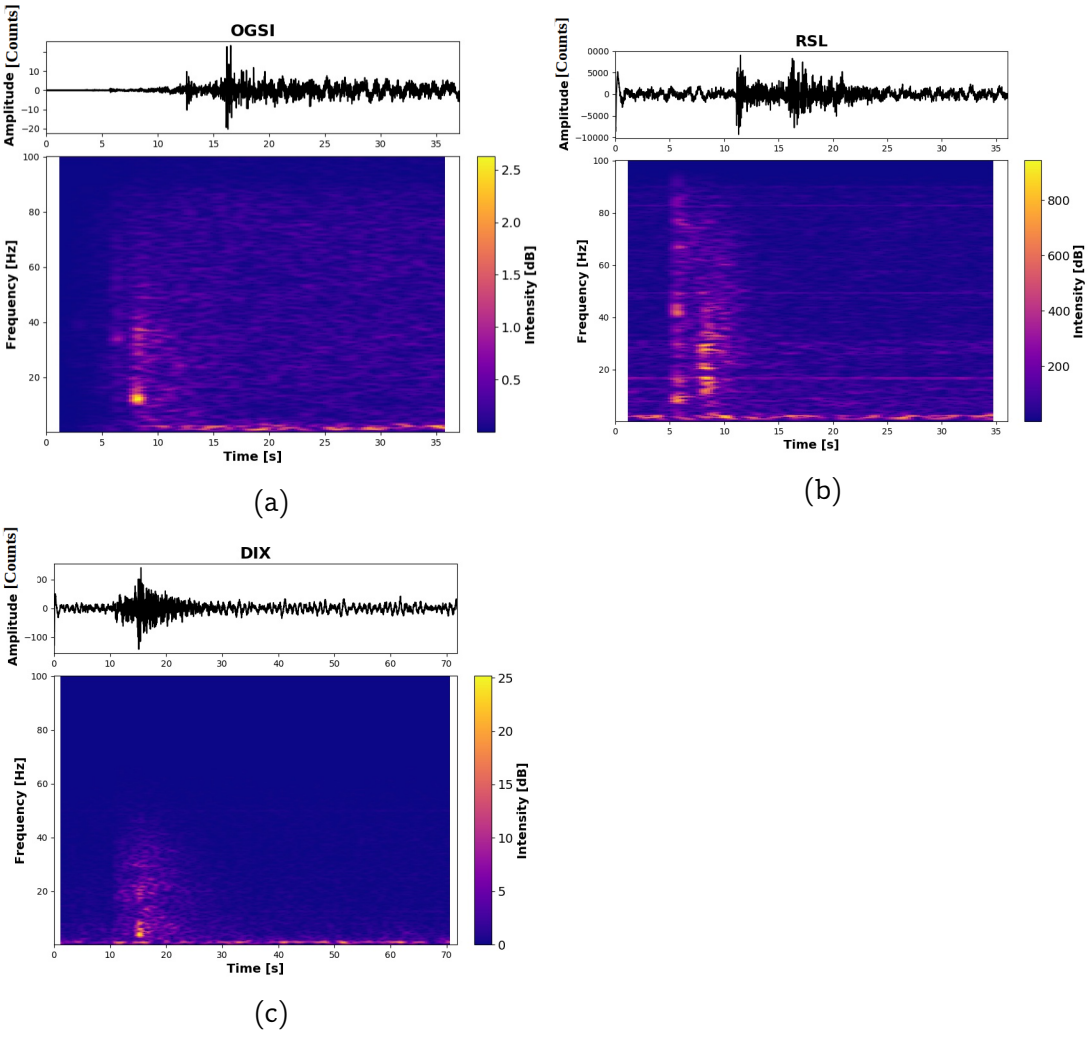


Figure 3.30: Formes d'onde et spectrogrammes correspondant aux signaux enregistrés sur la composante verticale des stations (a) OGSI, (b) DIX, et (c) RSL. Ces signaux sont reliés à un séisme ayant eu lieu le 03 décembre 2016 à 20h00 dans la région de Chamonix dans les Alpes françaises (MLv=0.3).

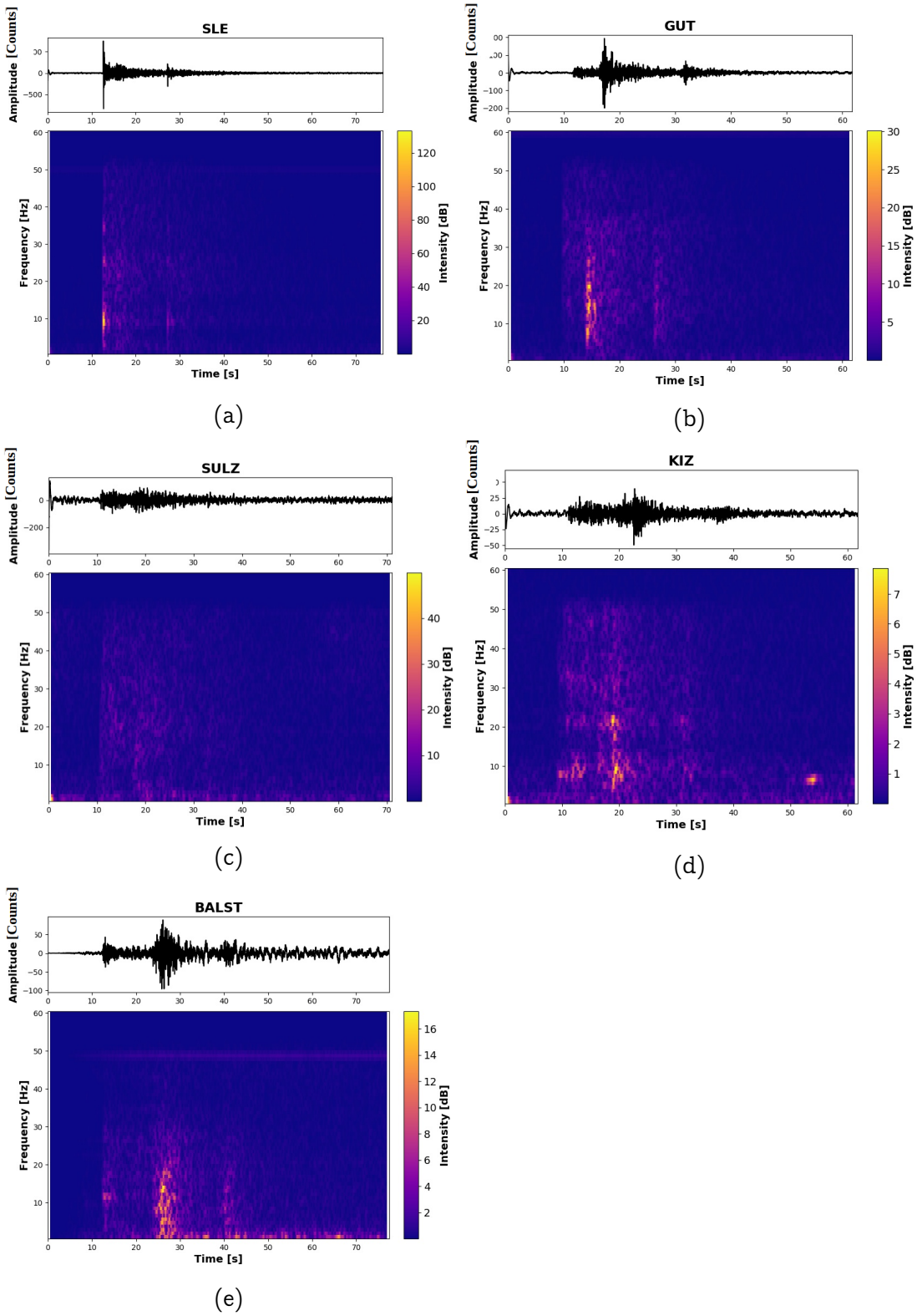


FIGURE 3.31: Formes d'onde et spectrogrammes correspondant aux signaux enregistrés sur la composante verticale des stations (a) SLE, (b) GUT, (c) SULZ, (d) KIZ et (e) BALST. Ces signaux sont reliés à séisme ayant eu lieu le 20 novembre 2016 à 20h08 dans le Sud de l'Allemagne, près de la frontière Suisse (MLv=1.4).

Pour les séismes de plus forte magnitude (c'est-à-dire MLv $>$ 1.5 dans les exemples proposés), même si l'intensité du signal se concentre également aux gammes fréquentielles caractéristiques du bruit d'origine anthropique, cette forte intensité s'étale à plus forte fréquence c'est-à-dire un peu au-delà de 20Hz, en particulier pour la partie du signal qui semble correspondre à l'arrivée des ondes S (Figures 3.32 et 3.33).

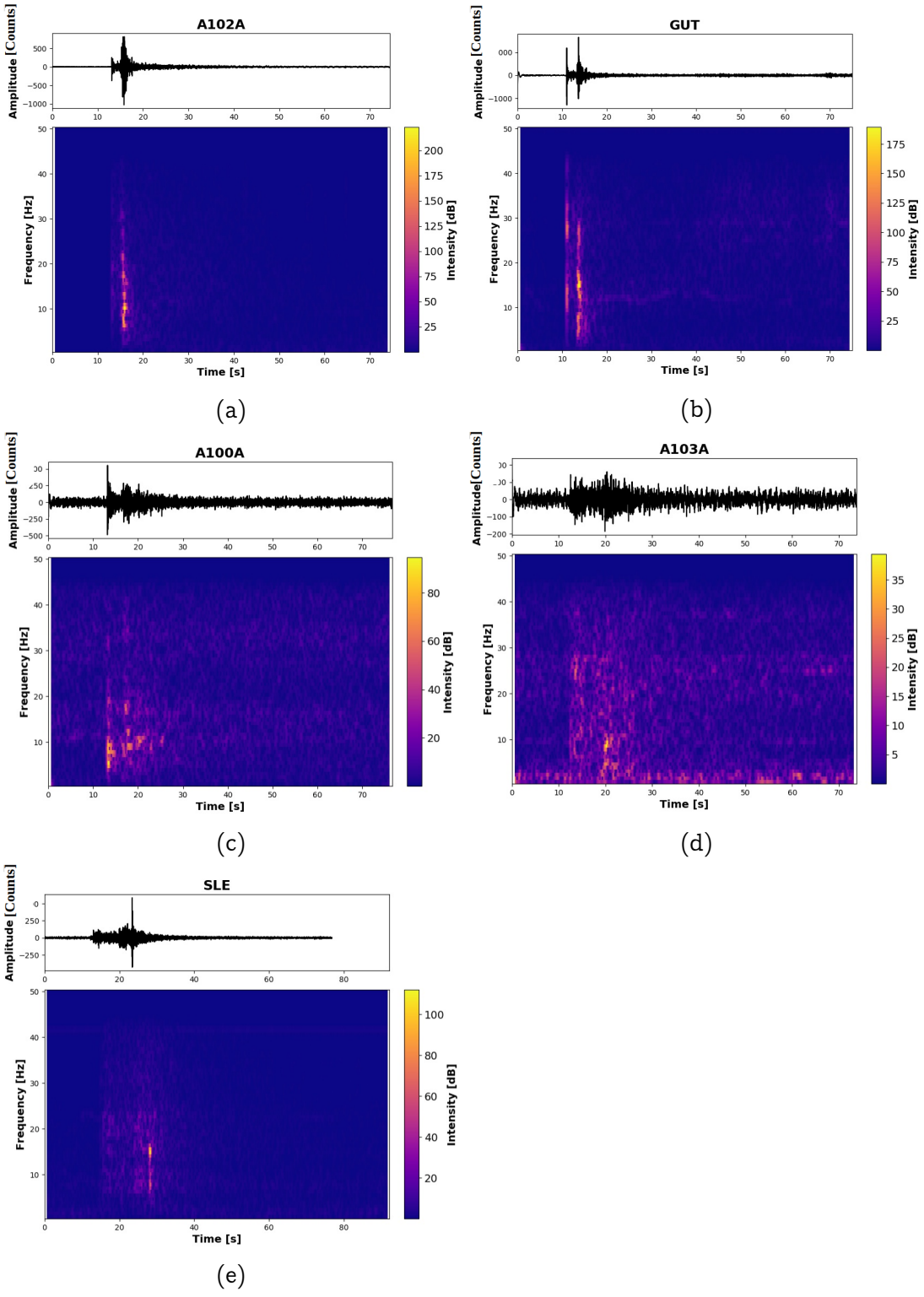


FIGURE 3.32: Formes d'onde et spectrogrammes correspondant aux signaux enregistrés sur la composante verticale des stations (a) A102A, (b) GUT, (c) A100A, (d) A103A et (e) SLE. Ces signaux sont reliés à un séisme ayant eu lieu le 12 octobre 2016 à 18h33 dans le Jura Souabe (MLv=1.5).

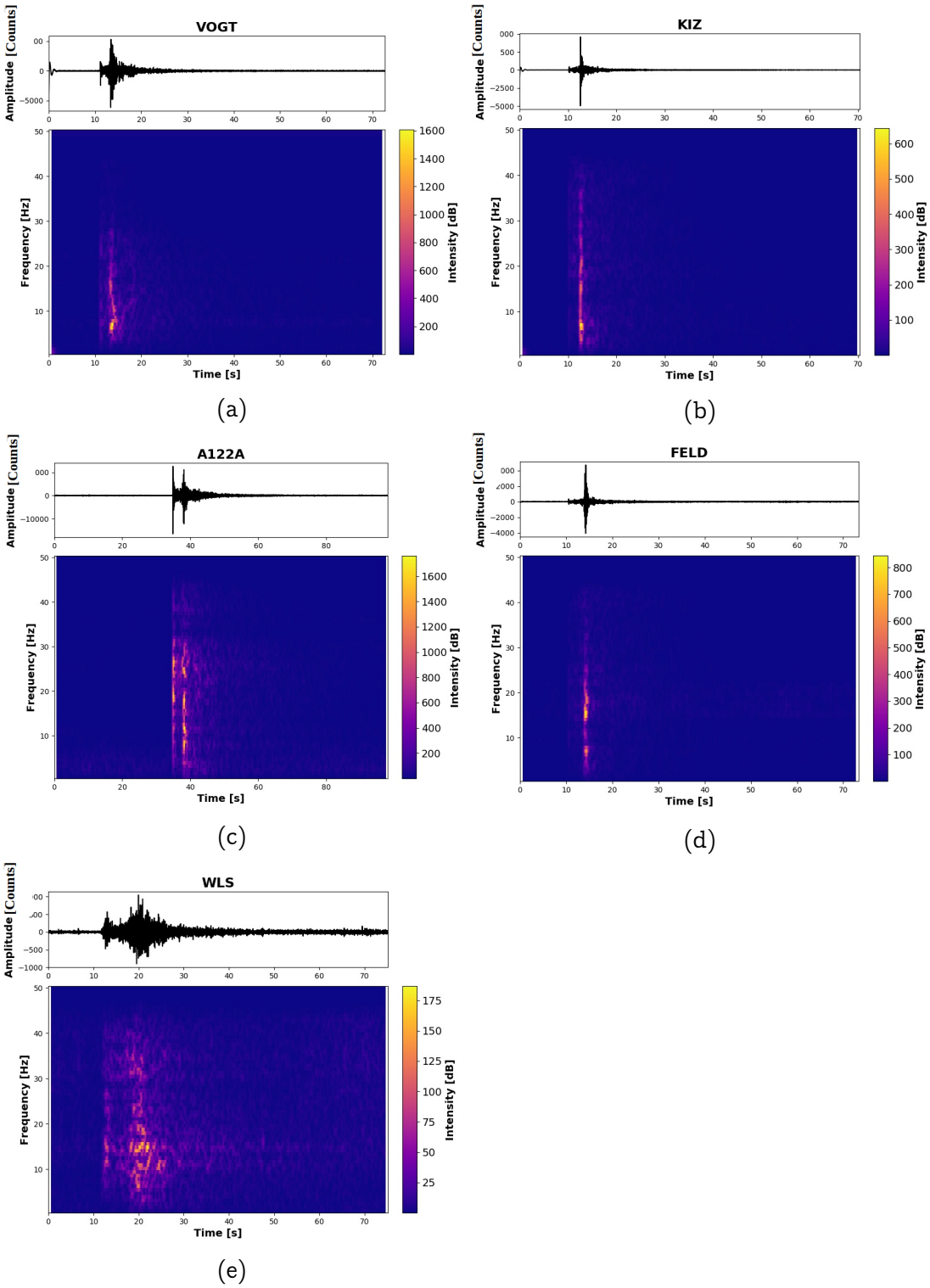


FIGURE 3.33: Formes d'onde et spectrogrammes correspondant aux signaux enregistrés sur la composante verticale des stations (a) VOGT, (b) KIZ, (c) A122A, (d) FELD et (e) WLS. Ces signaux sont reliés à un séisme ayant eu lieu le 07 septembre 2016 à 06h58 au Nord de Freiburg en Allemagne (MLv=2.1).

En ce qui concerne les signaux qui sont reliés aux tirs de carrière, l'analyse de leurs spectrogrammes montre aussi une intensité plus forte du signal dans la bande fréquentielle typique des signaux d'origine anthropique. Seulement, celle-ci est plutôt concentrée vers les plus basses fréquences, particulièrement entre 1 et 5 Hz. Quelques pics d'intensité peuvent être notés à plus haute fréquence, entre 10 à 15 Hz, et semblent être corrélés avec l'arrivée des ondes P (Figure 3.34 et figure 3.35).

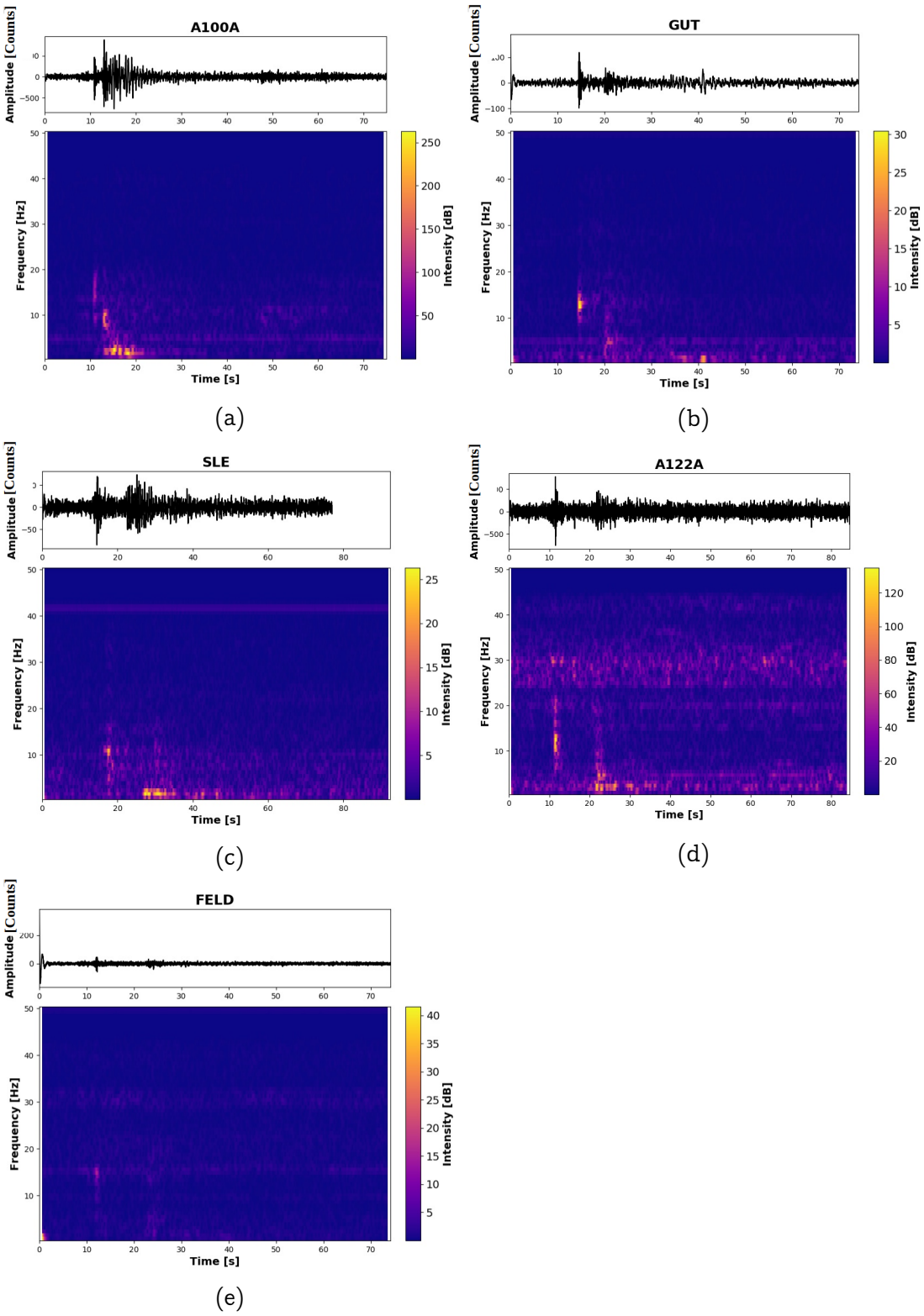


FIGURE 3.34: Formes d'onde et spectrogrammes correspondant aux signaux enregistrés sur la composante verticale des stations (a) A100A, (b) GUT, (c) SLE, (d) A122A et (e) FELD. Ces signaux sont reliés à un tir de la carrière de Haigerloch-Weildorf située à 60 km au Sud de Stuttgart en Allemagne et ayant eu lieu le 02 décembre 2016 à 08h37 (MLv=1.3).

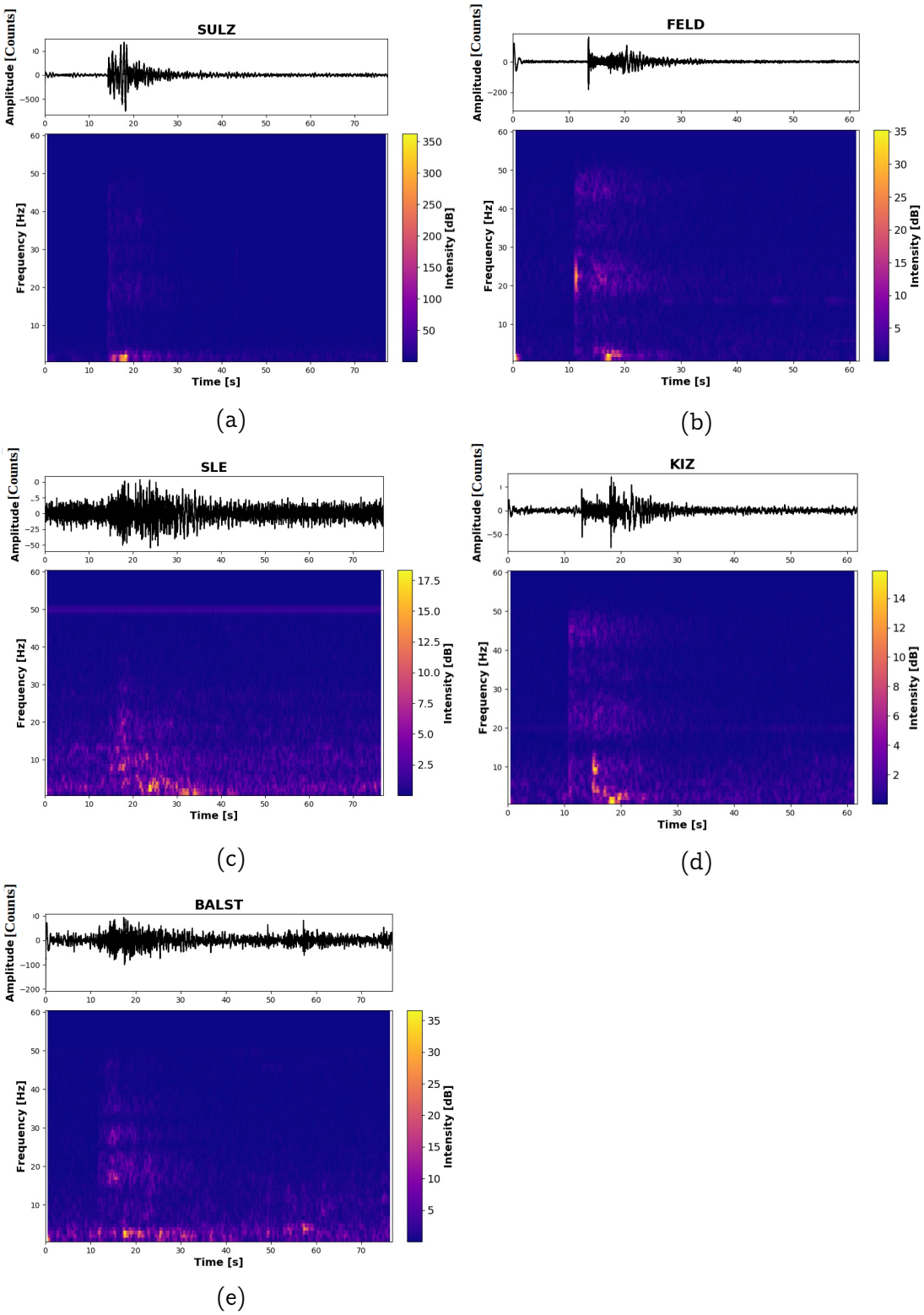


FIGURE 3.35: Formes d'onde et spectrogrammes correspondant aux signaux enregistrés sur la composante verticale des stations (a) SULZ, (b) BALST, (c) FELD, (d) KIZ et (e) SLE. Ces signaux sont reliés à un tir de la carrière de Rheinfelden située à l'Est de Bâle et ayant eu lieu le 26 octobre 2016 à 13h53 (MLv=1.2).

La zone d'étude est donc une zone caractérisée par un fort déploiement de stations entre 2016 et 2019, notamment du fait de l'installation des stations temporaires AlpArray. Cette période est la période qui est sélectionnée pour détecter les séismes de faible magnitude.

Seulement, l'inclusion de la totalité des stations dans le protocole de détection automatique, combinée à une diminution du seuil de détection, engendre des milliers de faux événements. Or, les stations (permanentes ou temporaires) qui interviennent le plus dans la création des vrais événements sont en fait aussi celles qui sont le plus impliquées dans la génération des faux événements, car très sensibles au bruit transitoire impulsif d'origine anthropique.

L'information véhiculée par les séismes est donc aisément diluée dans un flot d'information d'origine anthropique, impossible à décoder manuellement. Que les faux événements constituent un facteur limitant majeur des capacités réelles des systèmes de détection est confirmé. Cette zone d'étude représente donc un terrain idéal pour comprendre comment les dépasser efficacement.

En définitive, il s'agit de comprendre comment distinguer de manière solide un faux événement d'un vrai événement, en s'interrogeant sur les paramètres univoques qui vont automatiquement permettre de supprimer ce flux constant, mais inexorable, de faux événements.

3.2.3 Une base de données bien discriminée

Depuis 2016, une attention particulière est portée à la discrimination manuelle des événements, nous l'avons vu plus haut. Le catalogue rendu disponible par le BCSF-RéNaSS est donc soigneusement labélisé depuis cette date.

Au début de ce travail de thèse, j'ai revu manuellement l'ensemble des événements détectés pour l'année 2016 en introduisant les stations temporaires AlpArray pour la localisation. J'ai alors pointé 28079 phases (17707 phases P -Pg et Pn- et 10372 phases S -Sg et Sn) réparties sur 1134 événements (351 tirs de carrière, 774 séismes et 9 séismes induits par l'activité géothermique profonde). Les stations AlpArray les plus pointées correspondent aux mêmes stations qui sont impliquées dans la détection automatique des vrais et des faux événements. La performance de stations AlpArray telles que A060A, A061A, A100A, A102A, A103A, A158A ou bien A160A, peut donc être confirmée pour la période septembre 2016-décembre 2016 (Annexe B).

Ce travail de pointés, de localisation et de discrimination manuelle m'a permis de construire une base de données de carrières actives, en plus de celle fournie par le Bureau de Recherches Géologiques et Minières (BRGM). Au total 438 carrières ont été répertoriées en France, en Allemagne et en Suisse. Cette base de données contient le nom de la carrière et ses coordonnées géographiques. J'y ai ajouté les formes d'onde associées aux premières stations pour environ 180 carrières (celles actives pour la période septembre 2016-décembre 2016).

- **Des épicentres localisés plus précisément**

Cette base de données a pu être solidement construite car l'ajout des stations AlpArray dans la localisation a mieux contraint les épicentres des événements, en diminuant les incertitudes latitudinales et longitudinales de l'ordre de 1.5 km en moyenne, pour environ 50 % des événements (Figure 3.36). Sur l'ensemble des événements détectés, environ 15 % ont été localisés avec un nombre de pointés (P et S) supérieurs à 35 en présence des stations AlpArray contre 8 % sans inclusion de ces stations temporaires.

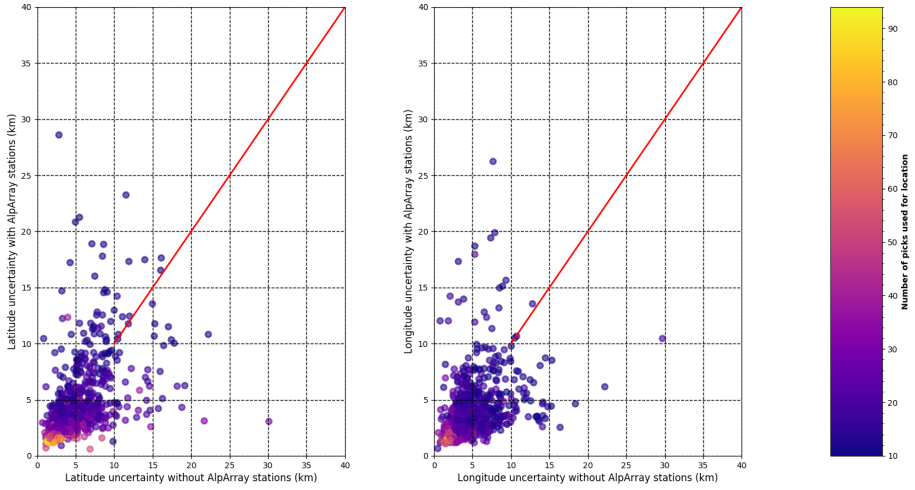


FIGURE 3.36: Comparaison des incertitudes latitudinales (à gauche) et longitudinales (à droite) obtenues des épicentres des événements détectés au cours de l'année 2016, avec et sans inclusion des stations AlpArray.

De plus, en guise d'exemple, la comparaison des épicentres des tirs de la carrière de Raon-l-Etape pour l'année 2016 montre un déplacement des épicentres vers le centre de la carrière lorsque les stations AlpArray sont incluses dans la localisation, facilitant un peu mieux le diagnostic de discrimination pour certains tirs (Figure 3.37).

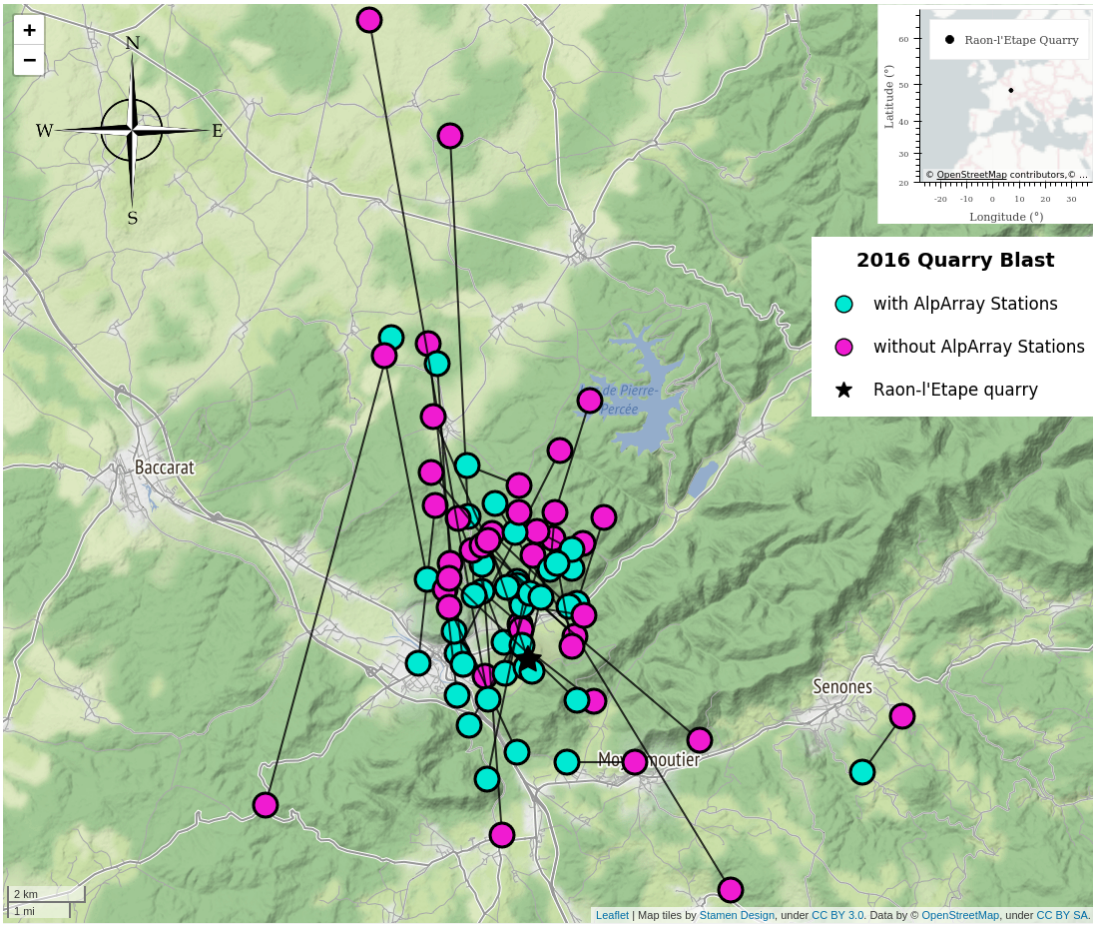


FIGURE 3.37: Comparaison des localisations épicentrales des tirs de la carrière de Raon-l'Etape détectés au cours de l'année 2016, avant et après inclusion des stations temporaires AlpArray.

En revanche, les profondeurs des événements n'ont pas été un critère retenu pour constituer la base de données de carrières actives. En incorporant les stations AlpArray, les profondeurs n'ont pas été fixées après relocalisation des événements. Autrement dit, aucun événement n'a été fixé à une profondeur donnée (Figure 3.38).

De cette façon, si l'ensemble des événements semblent avoir des profondeurs plus faibles lorsque les stations AlpArray sont incluses, ces mêmes événements semblent plutôt concentrés autour de 5 et de 10 km (ici les séismes) quand les stations AlpArray ne sont pas incluses, soulignant le fait que les localisations déterminées par le BCSF-RéNaSS soient souvent fixées. Ce phénomène réduit alors probablement artificiellement les incertitudes de localisation hypocentrale calculées par l'algorithme de localisation LOCSAT.

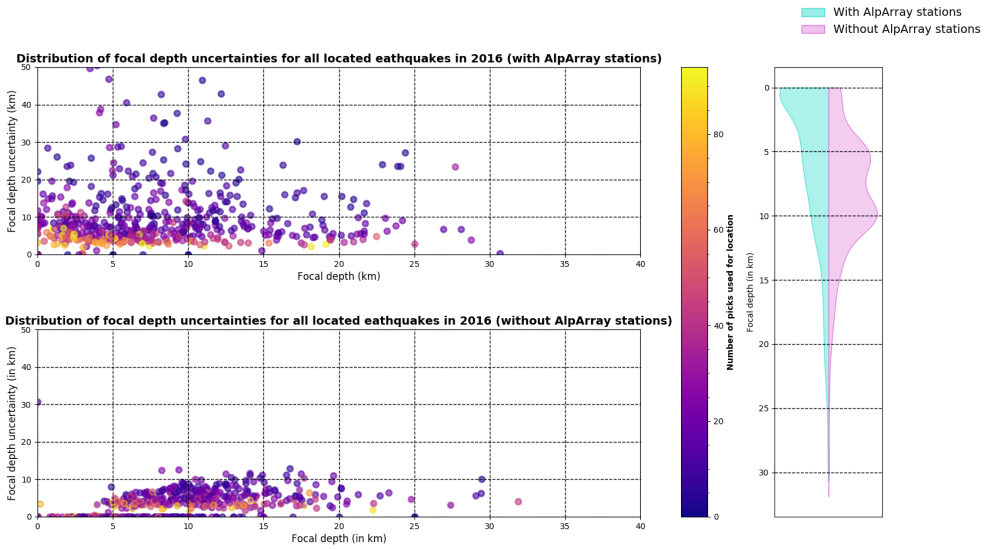


FIGURE 3.38: Comparaison des localisations hypocentrales ainsi que des incertitudes associées de l'ensemble des événements détectés au cours de l'année 2016, avant et après inclusion des stations temporaires AlpArray. Les incertitudes de localisation hypocentrales sont estimées par l'algorithme de localisation LOCSAT qui calcule une ellipsoïde de confiance pour chaque origine de chaque événement à partir de la diagonalisation d'une matrice de covariance 3D.

Par conséquent, les profondeurs laissées libres, les incertitudes hypocentrales augmentent alors, soulignant indirectement les incertitudes liées aux modèles de vitesse à 3 couches utilisés. Ces modèles 1D très simples, même si efficaces pour détecter, ne tiennent pas compte des variations d'épaisseur de la couche sédimentaire ou des discontinuités lithologiques latérales par exemple. Le modèle de vitesse régional le plus utilisé dans la zone d'étude est le modèle d'Haslach (Figure 3.39).

Enfin, la forte proportion de tirs de carrière positionnés librement autour de 1-2 km, ainsi que le fort pourcentage d'événements fixé artificiellement à 2 km par LOCSAT lorsque ce dernier ne converge pas vers une solution hypocentrale stable, ne sont pas des arguments solides pour considérer de façon fiable les profondeurs évaluées automatiquement par LOCSAT dans le diagnostic de discrimination.

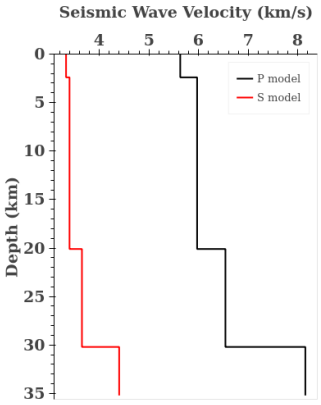


FIGURE 3.39: Modèle de vitesse, dit modèle de "Haslach", le plus utilisé pour la détection et les localisations des événements dans la zone d'étude. Les vitesses ont été déduites de l'étude des tirs de la carrière située près de la ville d'Haslach, au coeur de la Forêt Noire en Allemagne. Modifié d'après ROTHE et al., 1950.

•Des distances épicentrales minimales plus petites

L'étude des formes d'onde a été un autre critère fondamental pour consolider cette base de carrières, notamment pour 180 d'entre elles. Grâce à l'observation de ces formes d'onde, certains tirs de carrières ont pu être révélés bien distinctement grâce aux stations AlpArray.

En effet, l'ajout de ces stations temporaires a diminué la distance épicentrale minimale pour environ 60 % d'entre eux. La moyenne des distances épicentrales minimales est alors descendue à 25 km, au lieu de 33 km sans les stations AlpArray. De plus, 73% des événements détectés au cours de l'année 2016 ont désormais des distances minimales épicentrales de moins de 30 km, contre 62 % sans les stations AlpArray (Figure 3.40).

De ce fait, certains tirs ont pu être clairement identifiés à une carrière bien spécifique grâce à la première station AlpArray la plus proche. C'est le cas de carrières telles que la carrière de Rochefort-sur-Nenon dans le Jura français avec la station A213A, la carrière de Gerbamont dans les Vosges avec la station A158A, les carrières de Bernécourt, Bainville-sur-Madon, Pagny-sur-Meuse ou bien Barville, situées dans la région de Nancy au coeur du Bassin Parisien, avec la station A210A.

De même, en Allemagne, les stations AlpArray ont permis de révéler des formes d'onde particulières associées à plusieurs carrières situées dans le Jura Souabe (A100A, 102A, A108A, A109A, A360A), le Massif de Rhenish (A110A, A112), les terrains escarpés du Trias à l'Est du Massif de la Forêt Noire (A113A, A117A, A119) et dans le Massif de la Forêt Noire lui-même (A122A).

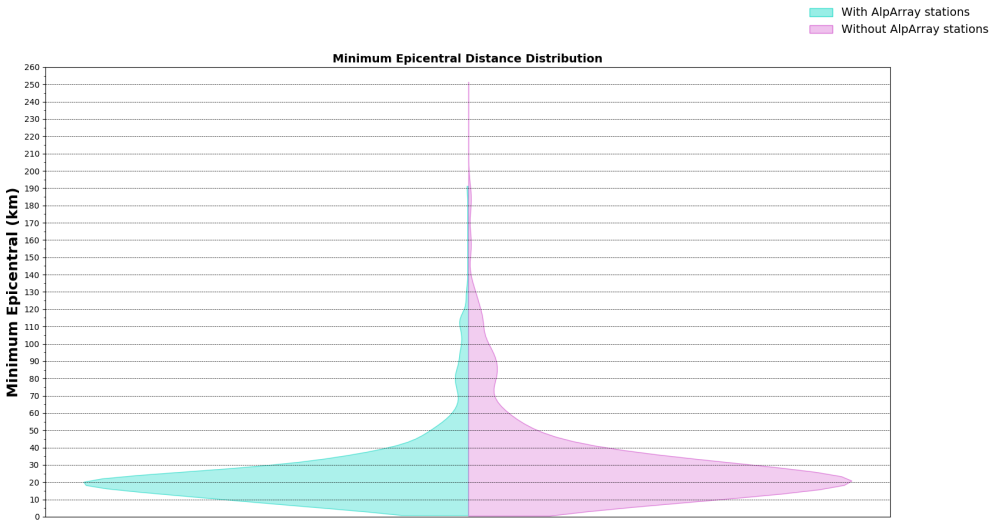


FIGURE 3.40: Comparaison des distances épicentrales minimales des événements détectés pour l'année 2016, avec et sans inclusion des stations AlpArray.

Ainsi, associé à une discrimination plus solide des événements détectés depuis 2016, ce recueil de carrières et de formes d'ondes apporte un ensemble de données disponibles de qualité, permettant d'évaluer la performance future de la classification automatique des événements, majoritairement séismes et tirs de carrière, avec plus de certitude.

3.3 Des outils disponibles de haute performance

3.3.1 Un système de détection mondialement utilisé avec un code source en libre accès

Les principaux modules formant le système de détection de SeisComP3 ont leur code source mis librement à disposition https://github.com/SeisComP3/seiscomp3. Ce code source offre une mine d'or possible de développement, notamment en offrant des bibliothèques exploitables pour le traitement du signal (filtrage, taperisation, déconvolution, etc.), des outils mathématiques (dérivation, intégration, transformée de Fourier, rotation, métriques statistiques, etc.) ainsi que des outils propres à la sismologie (calcul des magnitudes, calcul des temps de trajet, etc.). La procédure de détection peut donc être facilement mise en place en utilisant les fonctionnalités entières du logiciel SeisComP3.

Ensuite, développer la future procédure de détection au sein de SeisComP3 garantit un accès aux données directement en base (formes d'onde, métadonnées des stations, catalogue d'événements) sans nécessité de téléchargements préalables superflus, venant encombrer les espaces de stockage.

De plus, le code de SeisComP3 est écrit en langage C++ mais des modules peuvent être entièrement écrits en Python, tout en utilisant les fonctionnalités entières de SeisComP3. En effet, l'utilisation d'un compilateur (SWIG) aide à créer une interface d'accès aux déclarations C++ à partir du langage Python, via des bibliothèques compatibles avec ce dernier langage. L'intérêt de cette interface est donc de pouvoir combiner la performance et la rapidité du langage C++ avec la simplicité, la diversité et l'universalité du langage Python. C'est donc la voie que j'ai choisie.

Enfin, d'un point de vue opérationnel, intégrer la procédure de détection des petits séismes directement dans SeisComP3, peut assurer facilement son transfert automatique intégral en temps réel. Par conséquent, cela permet de traduire instantanément des résultats scientifiques (les paramètres univoques qui permettent de détecter proprement un séisme avec une haute probabilité) en une opération de surveillance sismique qui va produire efficacement des catalogues de séismes encore plus complets.

3.3.2 Des superordinateurs à haute performance de calcul

Un volume de 4 Térabytes de sismogrammes est disponible pour la période 2016-2019. Afin de traiter efficacement ces données volumineuses dans des temps raisonnables, c'est-à-dire plus rapides que le temps réel, l'emploi des superordinateurs, mis à disposition par le centre de Haute Performance de Calcul (HPC) de l'Université de Strasbourg, est particulièrement utile. Seulement, cela nécessite de comprendre comment transférer sur ces superordinateurs la procédure de détection développée sous SeisComP3 pour un fonctionnement optimal.

La création d'un conteneur SINGULARITY sera donc une étape importante pour assurer le fonctionnement autonome de la procédure de détection sur un cluster HPC. Cette encapsulation isolante permettra le déploiement de plusieurs instances SeisComP3 en parallèle, accélérant alors le processus de détection (Figure 3.41. Opération qui est pour l'instant impossible à réaliser avec le système de détection actuel de SeisComP3 sans développement méthodologique.

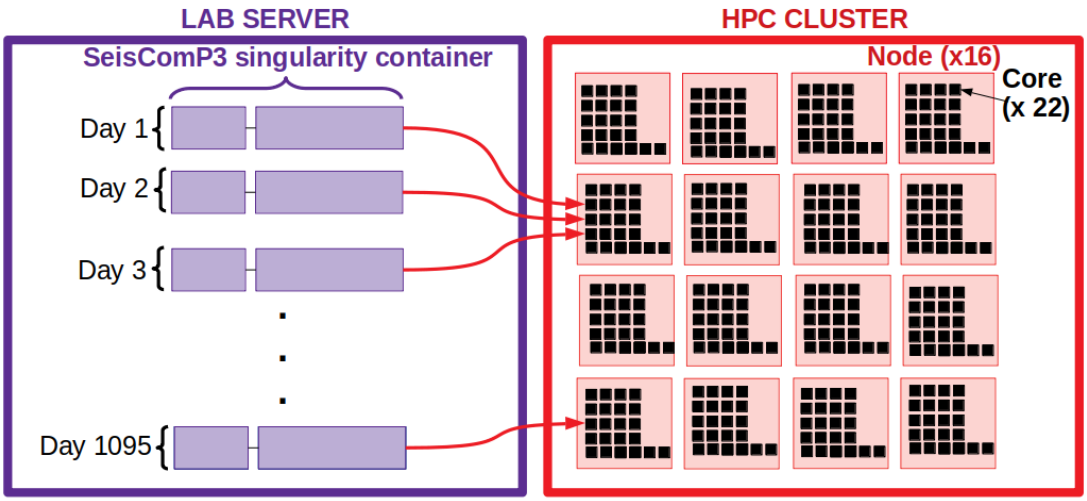


FIGURE 3.41: Schéma simplifié d'un déploiement de commandes multiples sur un cluster à Haute Performance de Calcul (HPC). Dans cet exemple, chaque commande activée exécute un conteneur Singularity sur un processeur (coeur) d'un ordinateur (noeud). Chaque conteneur encapsule une instance de détection SeisComP3 qui traite ici 1 jour de données. Si 1095 jours sont traités (c'est-à-dire l'équivalent de 3 ans de données), 1095 commandes seront exécutées sur 16 ordinateurs (noeuds), chacun contenant 24 coeurs.

J'ai alors tous les facteurs pour développer une procédure de détection optimale des petits séismes : une forte probabilité d'occurrence des petits séismes, une intense activité anthropique détectée régulièrement par les réseaux de stations, un fort taux de faux événements inexorablement détectés également.

J'ai aussi des données disponibles (un volume de 4 térabytes pour la période 2016-2019) grâce au récent apport de nouvelles stations permanentes et l'important déploiement des stations temporaires AlpArray.

J'ai enfin des outils disponibles performants : un système de détection qui peut être optimisé facilement grâce à un code source en libre accès, un cluster de calcul de haute performance intégré dans un des centres de calcul les plus puissants de France.

Par conséquent, je dispose d'un objet d'étude solide qui permettra de résoudre dans les chapitres suivants les importantes questions de recherche évoquées dans le chapitre précédent. A savoir, comment limiter la détection des très nombreux petits séismes contaminés par du bruit ? Comment réduire de façon conséquente la détection de milliers de faux événements ? Et comment efficacement discriminer les séismes des tirs de carrière ?

Chapitre 4

Comment limiter la détection des séismes contaminés par du bruit ?

Sommaire

4.1 Améliorer la qualité des pointés

4.1.1 Comment fonctionne le processus de pointés dans le système de détection ?

•Pointé automatique des ondes P

Une première estimation du temps d'arrivée des ondes P est établie grâce à un algorithme qui détecte les phases sismiques, en se basant sur la méthode STA/LTA. Comme évoqué dans le chapitre 1, cet algorithme recherche des anomalies dans le signal sous la forme de changements d'amplitude en calculant un rapport moyen STA/LTA (STA = fenêtre temporelle courte sensible aux événements sismiques, LTA = fenêtre temporelle longue fournissant des informations sur l'amplitude temporelle du bruit sismique à une station donnée). Un pointé est émis dès que la valeur du rapport STA/LTA dépasse une valeur seuil de référence préalablement définie.

Quatre paramètres principaux vont gouverner la fréquence d'occurrence de ces premiers pointés émis : les tailles des deux fenêtres temporelles STA et LTA, la valeur du seuil de déclenchement d'un pointé, ainsi que la valeur minimale de rapport STA/LTA à atteindre après qu'un pointé ait été émis, pour de nouveau activer une opération de pointé.

Un éventail de valeurs de ces paramètres ont été testées empiriquement sur la détection automatique des événements pour les mois de juillet et août 2016. Des valeurs comprises entre 0.1s et 2s ont été testées pour la fenêtre temporelle STA et entre 10s et 80s pour la fenêtre temporelle LTA. En ce qui concerne la valeur seuil de déclenchement d'un pointé et la valeur minimale de rapport STA/LTA à atteindre après qu'un pointé ait été émis, celles-ci ont été affinées à partir des valeurs de référence obtenues par Grunberg et al., 2018 sur la zone du Graben du Rhin Supérieur (qui étaient respectivement de 2.2 et de 2.7).

Ainsi, les paramètres finaux qui ont permis d'aboutir à un taux de détection optimal (meilleure qualité des pointés automatiques P et S, nombre d'événements détectés automatiquement comparativement à ceux détectés par le BCSF-RéNaSS pour la même période, nombre d'événements nouvellement détectés en plus), correspondent à :

- **une taille de fenêtre STA égale à 0.5 seconde**, ajustée de sorte à augmenter la sensibilité de l'algorithme aux événements locaux, mais ceci implique une augmentation du taux de faux pointés liés à du bruit transitoire impulsif d'origine anthropique (Trnkoczy, 1999) ;
- **une taille de fenêtre LTA égale à 40 secondes** définie par rapport aux fluctuations importantes et irrégulières du bruit d'origine anthropique enregistré aux stations, c'est-à-dire pas trop grande pour accommoder en continu les valeurs du rapport STA/LTA aux changements graduels de bruit enregistré ;

— **une valeur seuil de déclenchement plutôt basse, égale à 2.4** de sorte à élever la probabilité de détecter de plus faibles événements ;
— **une valeur référence de réactivation du pointé également basse, 2.0**, pour capturer entièrement la coda du signal pointé précédemment.

L'amplitude et le type de bruit sismique enregistrés aux stations influencent fortement le paramétrage de la valeur du seuil de déclenchement d'un pointé. En effet, un bruit sismique statistiquement stationnaire va permettre une valeur de seuil plus basse, alors qu'un comportement irrégulier de bruit sismique nécessite de choisir de plus hautes valeurs. La valeur du seuil de détection a été dans ce travail choisie particulièrement bas, induisant alors un nombre important de faux pointés.

Le pointé des temps d'arrivée des ondes P est ensuite affiné à partir d'une fenêtre temporelle autour de la détection émise par la méthode précédente. L'algorithme qui est choisi pour cette affinage se base sur le calcul d'une fonction caractéristique de l'enveloppe du signal qui utilise en plus une métrique statistique, à savoir la variance (Figure 4.1, pour plus de détails, voir KRADOLFER et al., 1987). Cette méthode est nommée méthode BK. En effet, (KRADOLFER et al., 1987) ont modifié la fonction enveloppe d'Allen (R. ALLEN, 1978) en l'élevant au carré et en implémentant la variance de cette enveloppe. Un pointé P est émis quand la valeur de le fonction caractéristique excède un certain seuil $\gamma = 10$ (KUPERKOCH et al., 2012). De plus, la variance est continuellement mise à jour afin d'accommoder le calcul de la fonction caractéristique aux variations temporelles du niveau de bruit enregistré, sauf lorsque les valeurs de cette fonction caractéristique excède un second seuil dynamique $\delta = 2 \times \gamma$.

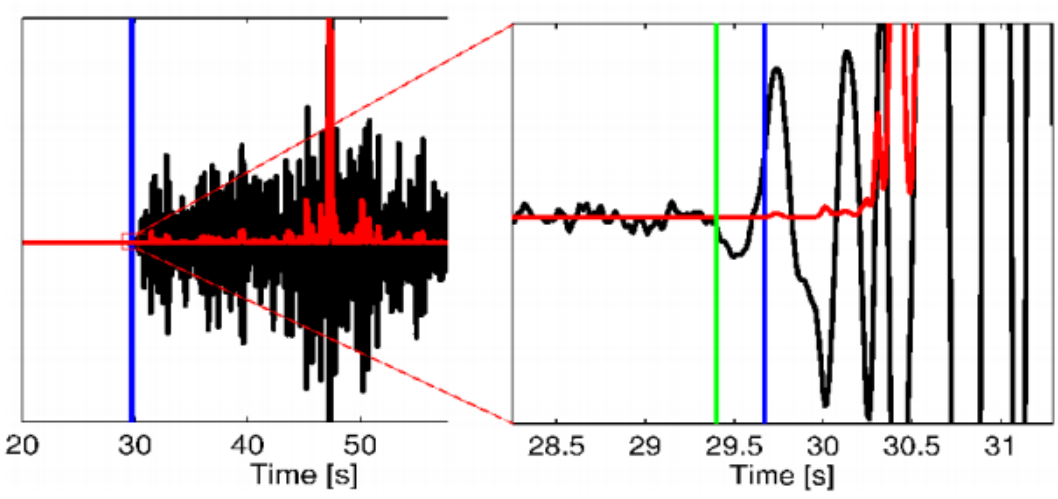


FIGURE 4.1: Exemple de fonction caractéristique (CF, représentée en rouge) de KRADOLFER et al., 1987 calculée pour une forme d'onde correspondant à un événement local (en noir). La ligne verticale bleue indique le pointé automatique de la phase P, la ligne verticale verte la lecture manuelle de la première arrivée des ondes P. D'après KUPERKOCH et al., 2012.

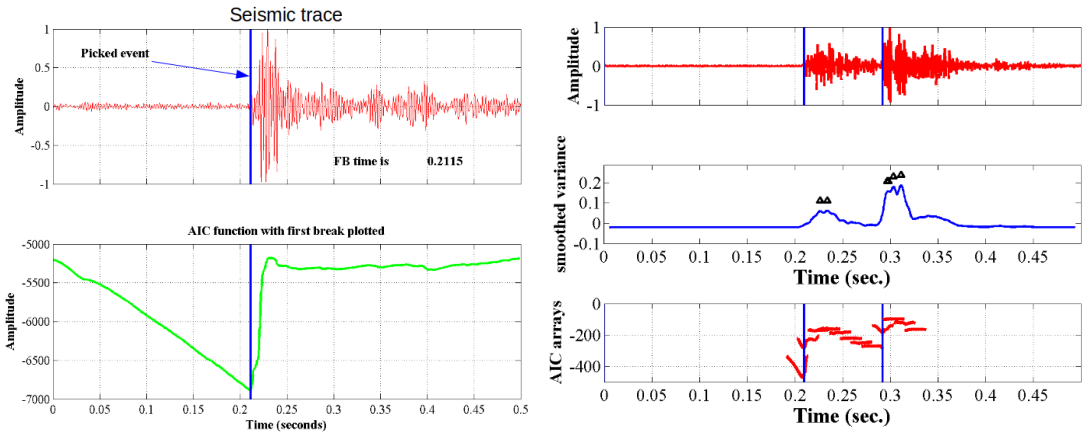
•Pointé automatique des ondes S

Un algorithme basé sur le critère d'information d'Akaike (AIC, Akaike, 1971) détecte les phases sismiques S, une fois que les pointés des phases sismiques P sont émis. Par conséquent, les pointés des temps d'arrivée des ondes S ne sont effectués que s'il y a eu au préalable une détection des temps d'arrivée des ondes P sur la composante verticale de la station. Les pointés des ondes S sont déterminés sur la somme vectorielle des composantes horizontales.

L'apparition d'une phase (ici S) sur une trace sismique peut être déterminée en modélisant le bruit et le signal sismique dans des fenêtres temporelles de taille pré-établie. L'algorithme de pointé automatique des ondes S se base effectivement sur l'hypothèse que la trace peut être divisée en segments temporels avec des caractéristiques de stationnarité spécifiques (Maeda, 1985). Si deux segments consécutifs ont des caractéristiques de stationnarité différentes (segment de signal correspondant uniquement à du bruit, suivi par un segment de signal transitoire impulsif par exemple), cela souligne alors l'émergence d'une phase sismique correspondant à la première arrivée des ondes S (Sleeman et al., 1999).

Le critère AIC, se basant sur le calcul continu d'une fonction caractéristique qui utilise la variance des amplitudes de chaque segment de signal, est donc utilisé pour marquer le point de deux fenêtres temporelles adjacentes qui ont des propriétés statistiques différentes (Figure 4.2). Un pointé S est donc émis lorsque la valeur du critère AIC a atteint sa valeur minimale, c'est-à-dire au moment où la variance du signal sismique non-stationnaire enregistré augmente soudainement, se détachant nettement du bruit de fond plutôt stationnaire.

Le développement qui va suivre met alors en évidence les critères principaux qui vont conditionner fortement la qualité des pointés P et S émis à chaque station. Ces critères principaux qui vont optimiser la qualité du pointé automatique des phases P et S sont ceux qui sont reliés aux caractéristiques de bruit communément enregistré à une station donnée ainsi qu'à la localisation de cette dernière. En effet, si le paramétrage des algorithmes de pointés des temps d'arrivée des ondes sismiques P et S n'est pas adapté au contenu fréquentiel des signaux enregistrés aux stations, il y a un risque accru de ne pas activer l'émission d'un pointé, de retarder fortement ou activer trop précocement cette émission.



(a) (b)

FIGURE 4.2: Principe de l'utilisation du critère AIC pour pointer les premières arrivées des phases sismiques. (a) Trace sismique et fonction AIC calculée. Un pointé est émis lorsque le critère AIC est minimisé (point de contact entre deux segments de trace consécutifs, caractérisé par un changement marqué de la variance du signal).(b) Trace sismique avec pointés des phases P et S (en haut), évolution de la variance du signal correspondant à cette trace sismique (au milieu) et calcul du critère AIC sur 60 fenêtres temporelles définies sur la trace sismique (en bas). Deux des fenêtres AIC ont détecté les premières arrivées des phases sismiques P et S. Les lignes bleues verticales correspondent aux premières arrivées des phases sismiques et les lignes rouges horizontales correspondent aux fenêtres pour lesquelles le critère AIC est évalué. Modifié d'après ST-ONGE, 2011.

4.1.2 S'adapter aux caractéristiques de bruit des stations et à leur localisation

•**Pour les pointés automatiques des ondes P**

La taille de la fenêtre STA qui a été choisie pour récupérer la valeur instantanée du signal sismique, est relativement courte (0.5 seconde). Par conséquent, celle-ci devient plus sensible aux pointes de bruit, en particulier pour les stations implantées dans des sites très pollués par du bruit transitoire impulsif. Dans cette configuration, les pointés des ondes P peuvent être anticipés du fait de bruit parasite précédent le signal sismique cible (Figure 4.3).

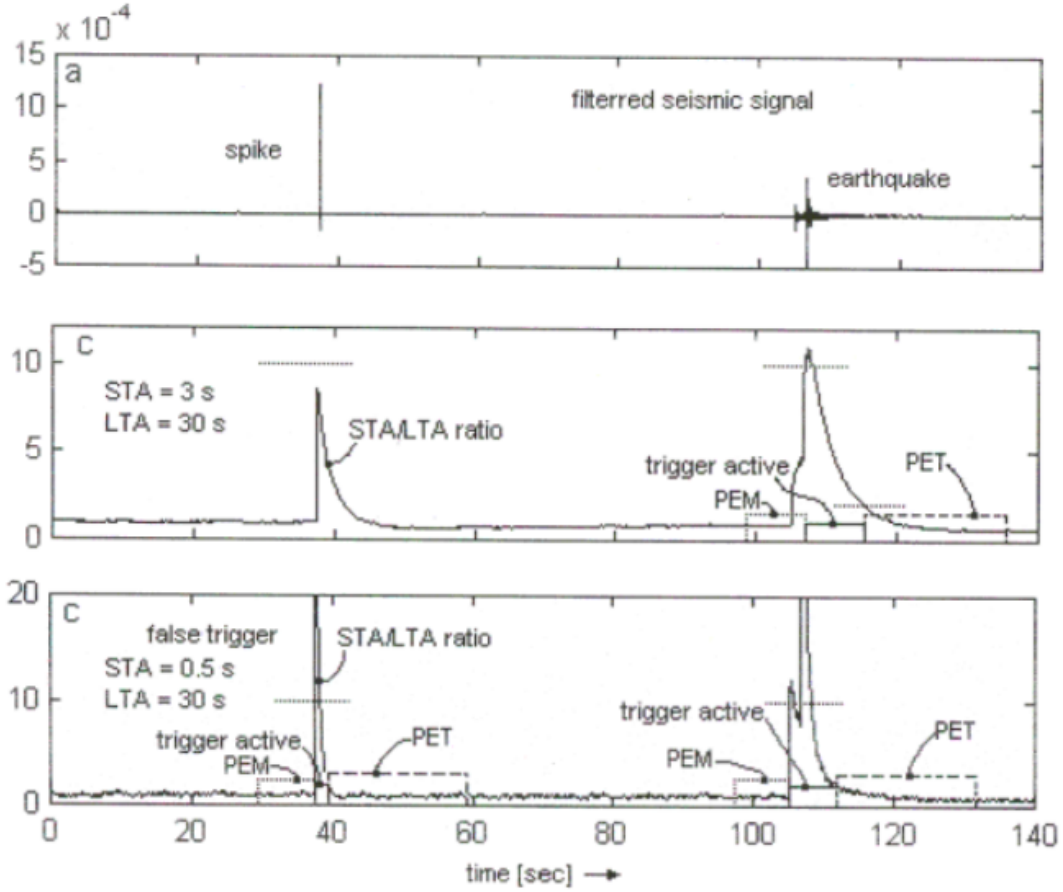


FIGURE 4.3: Influence de la durée de la fenêtre STA sur la sensibilité de l'algorithme de détection des ondes P. (a) Signal correspondant à un tir de la carrière précédé d'un artefact de bruit de courte durée. (b) Évolution du rapport STA/LTA associée à la trace sismique pour une fenêtre STA égale à 3 s et une fenêtre LTA égale à 30 s. Un pointé est déclenché au bon endroit.(c) Évolution du rapport STA/LTA associée à la trace sismique pour une fenêtre STA égale à 0.5 s et une fenêtre LTA égale à 30 s. Un pointé supplémentaire anticipé est émis au début du pic de bruit. Modifié d'après TRNKOCZY, 1999.

De même, la taille de la fenêtre LTA est relativement courte (40 secondes). Ainsi, dans cette configuration, face à des ondes P de très faible amplitude, il y a un risque accru que cette phase sismique passe inaperçue, d'autant plus si le niveau de bruit de fond est élevé. Si aucun pointé P n'est alors déclenché, l'arrivée des ondes P non détectées vient augmenter l'amplitude du bruit sismique enregistré, diminuant la sensibilité du déclenchement d'un futur pointé au moment où des ondes S plus énergétiques arrivent. De cette façon, ou bien un pointé P retardé est émis au moment où ce sont les ondes S qui arrivent, mais avec un faible rapport signal/bruit, ou bien aucun pointé n'est déclenché car les signaux sismiques sont de trop faible amplitude (Figure 4.4).

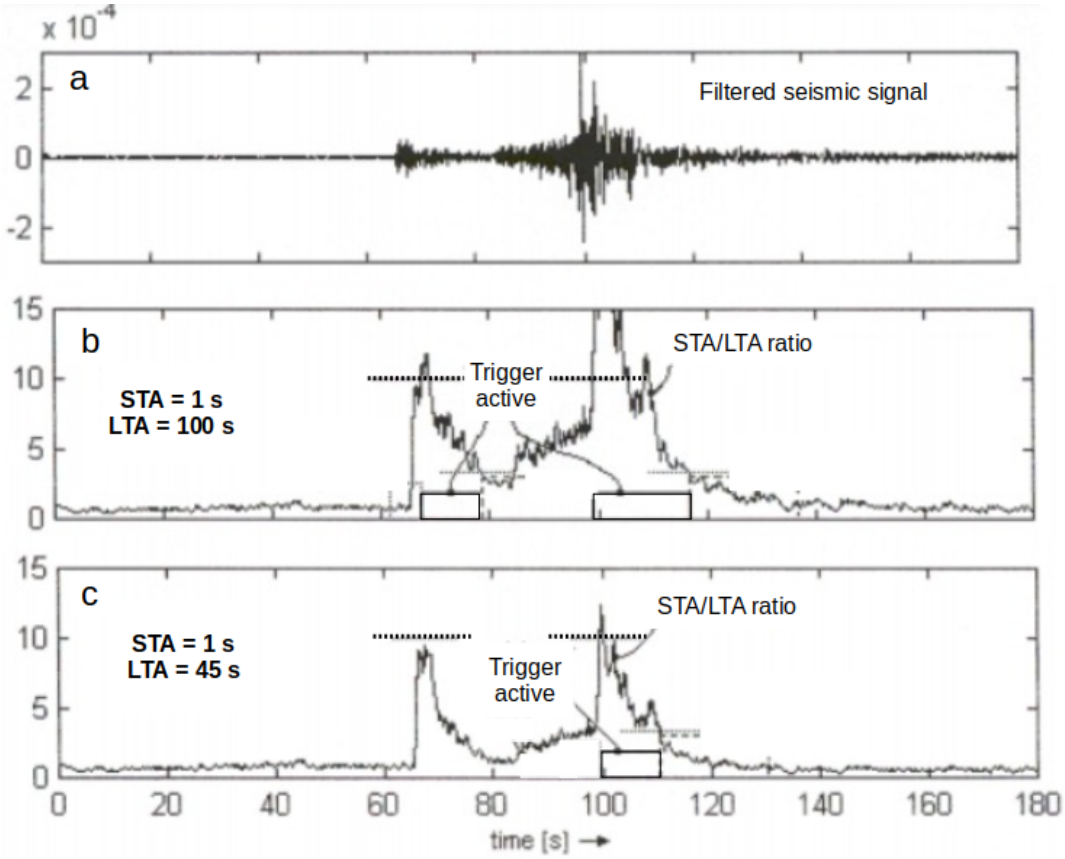


FIGURE 4.4: Influence de la durée de la fenêtre LTA sur la sensibilité de l'algorithme de détection des ondes P. (a) Signal avec des ondes P de faible amplitude, correspondant à un séisme local. (b) Évolution du rapport STA/LTA associée à la trace sismique pour une fenêtre STA égale à 1 s et une fenêtre LTA égale à 100 s. Un pointé est déclenché au bon endroit.(c) Evolution du rapport STA/LTA associée à la trace sismique pour une fenêtre STA égale à 1 s et une fenêtre LTA égale à 45 s. Un pointé P retardé est émis au moment où des ondes S de plus forte amplitude arrivent. Modifié d'après TRNKOCZY, 1999.

Le paramétrage de l'affinage du pointé P par la méthode BK est alors indispensable pour éviter la propagation de tels pointés P erronés.

Les paramètres qui vont améliorer la qualité de ces pointés sont en fait deux paramètres qui sont reliés indirectement aux caractéristiques du bruit enregistré à la station. Ces deux paramètres sont utilisés pour calculer la fonction caractéristique nécessaire à l'émission d'un pointé selon la méthode BK, à savoir : la fenêtre temporelle définie autour du premier pointé P déterminé par la méthode STA/LTA ainsi que le filtrage du signal utilisé.

- Adapter la fenêtre temporelle pour calculer la fonction caractéristique

Le début de la fenêtre temporelle choisie pour calculer la fonction caractéristique de l'enveloppe du signal selon la méthode BK est définie à partir du déclenchement du pointé P émis par la méthode STA/LTA. Par défaut elle est de -20 s à partir de cette détection initiale.

Seulement, les niveaux de bruit enregistrés varient temporellement pour une station donnée et spatialement en fonction de la localisation de cette station. Par conséquent, un paramétrage unique de la fenêtre temporelle utilisée pour pointer le temps d'arrivée des ondes P ne tient pas compte des variations spatio-temporelles des niveaux de bruit enregistrés aux stations.

De cette façon, afin de comprendre l'impact de la valeur du début de cette fenêtre temporelle sur le pointé des temps d'arrivée des ondes P à chaque station, différentes valeurs ont été testées empiriquement sur l'ensemble des stations impliquées dans la détection des événements pour les mois de juillet-août 2016 et janvier 2017.

Il a été alors constaté que, pour les stations qui enregistrent des niveaux de bruit assez élevés, avec des soubresauts répétés de bruit non-stationnaire de courte durée qui se détachent du niveau de fond, cette fenêtre s'initiera plus tardivement. Ceci évite effectivement une pollution du calcul de la fonction caractéristique par du signal parasite, enregistré avant l'arrivée des ondes P, comme c'est le cas dans la Figure 4.5.

En effet, pour illustrer ce propos, si je prends l'exemple d'une station particulièrement sensible au bruit comme la station FELD, située sur le sommet le plus élevé du Massif de la Forêt Noire en Allemagne, près de 4 tours de communication et non loin d'une station de ski, le début de la fenêtre temporelle a été placée quelques secondes avant le pointé P initié, c'est-à-dire à -2 s. Ceci limite alors la probabilité de passer sous silence l'arrivée des ondes P qui serait dans le sillage de la fenêtre de traitement d'un précédent faux pointé P (Figure 4.6).

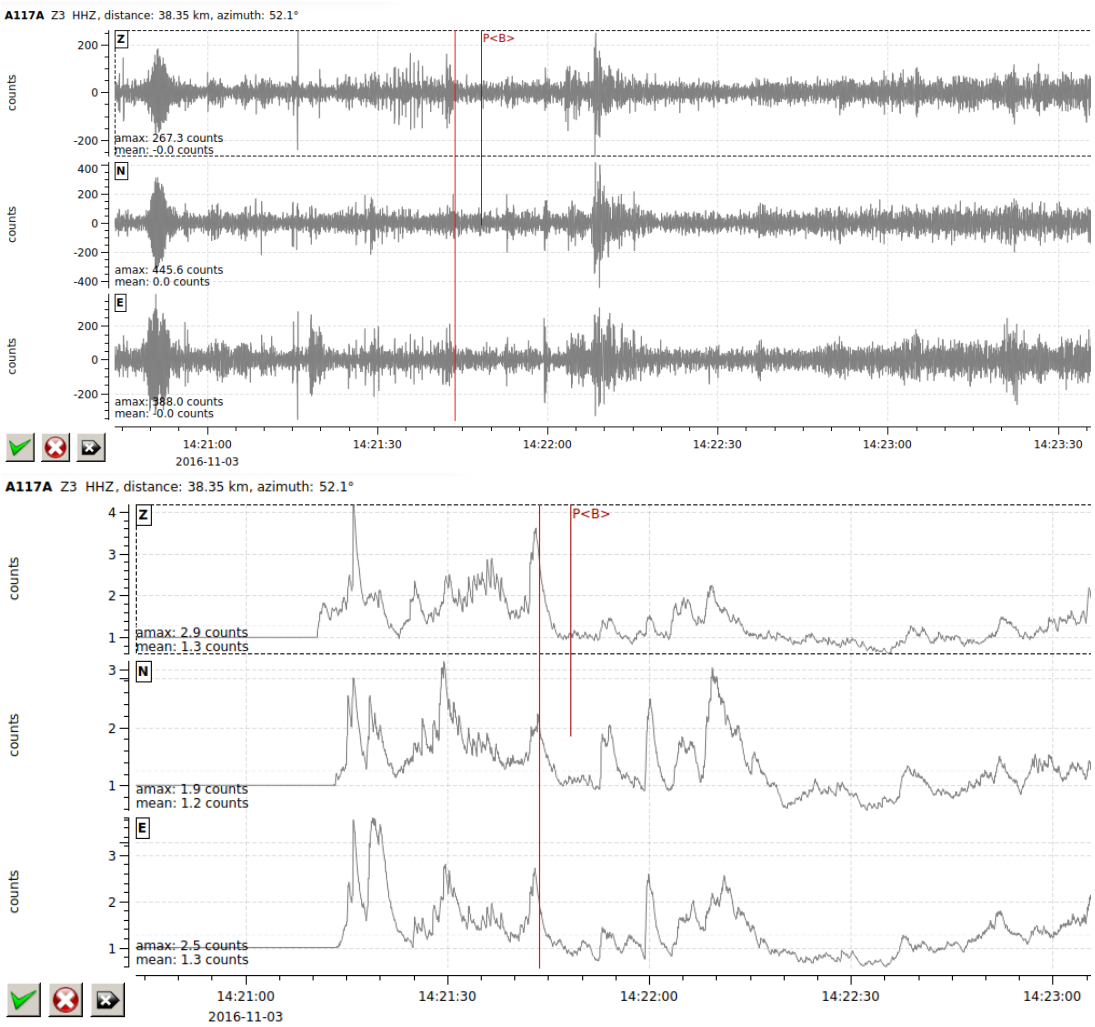
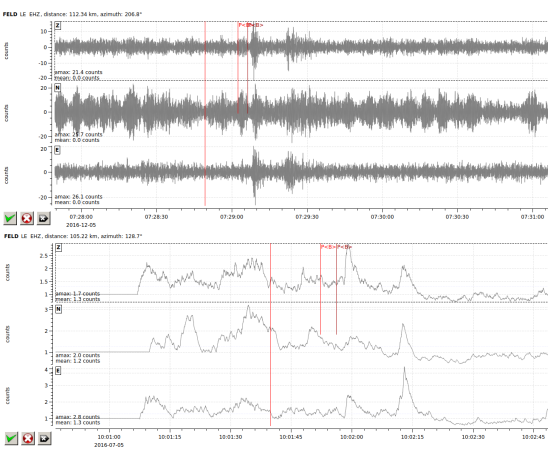
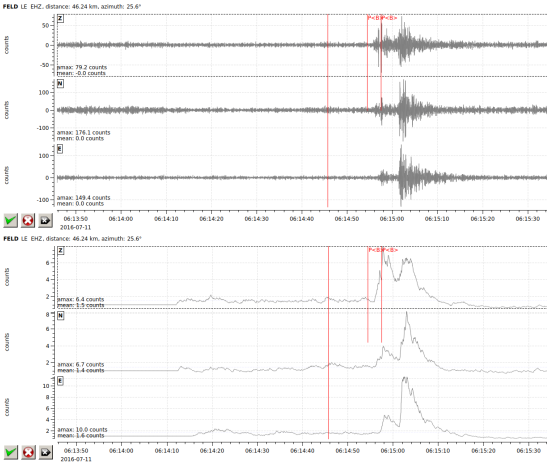


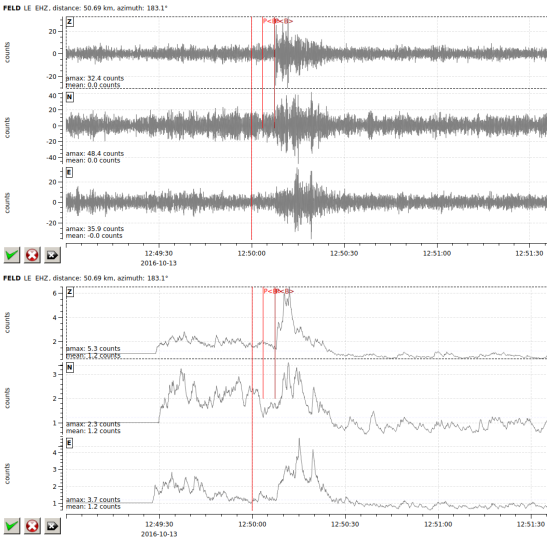
FIGURE 4.5: Sismogramme enregistré sur la composante verticale d'une station bruitée (A117A) et fonction STA/LTA correspondante. Un pointé P a été créé de façon anticipée quelques secondes avant les premières arrivées des ondes P émises par un tir de la carrière de Satteldorf-Crailsheim ayant eu lieu le 03 novembre 2016 à 14h21 en Allemagne (MLv 1.6).



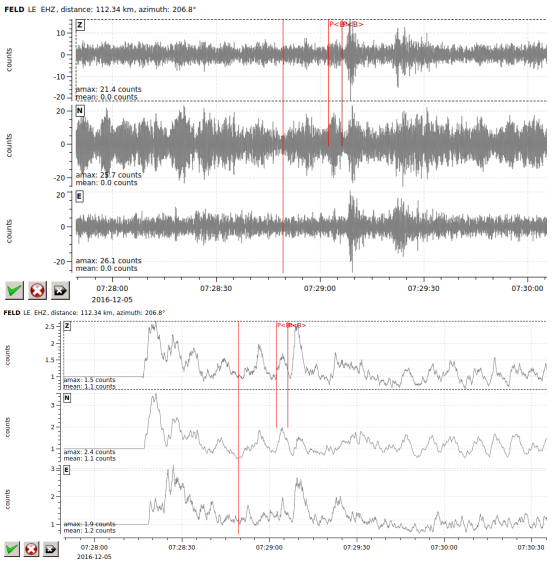
(a) Sismogramme et fonction STA/LTA correspondante (tir de la carrière de Raon-l'Etape, identifié le 05 juillet 2016 à 10h01 dans les Vosges, MLv 1.6). Le trait rouge foncé équivaut au pointé P défini à partir d'une fenêtre temporelle débutant à -2 s et le trait rouge clair à une fenêtre temporelle débutant à -6 s.



(b) Sismogramme et fonction STA/LTA correspondante (séisme identifié dans la région de Bâle en Suisse, le 11 juillet 2016 à 06h14, MLv 1.2). Le premier trait vertical rouge équivaut au pointé P défini à partir d'une fenêtre temporelle débutant à -6 s et le deuxième à une fenêtre temporelle débutant à -2 s.



(c) Sismogramme et fonction STA/LTA correspondante (tir de la carrière de Schuttertal dans le Massif de la Forêt Noire en Allemagne, identifié le 13 octobre 2016 à 12h49, MLv 1.5). Le trait vertical rouge clair équivaut au pointé P défini à partir d'une fenêtre temporelle débutant à -6 s et le trait vertical rouge foncé à une fenêtre temporelle débutant à -2 s.



(d) Sismogramme et fonction STA/LTA correspondante (tir de la carrière de Magstadt à L'Ouest de Stuttgart en Allemagne, identifié le 05 décembre 2016 à 07h28, MLv 1.8). Le trait vertical rouge clair équivaut au pointé P défini à partir d'une fenêtre temporelle débutant à -6 s et le trait vertical rouge foncé à une fenêtre temporelle débutant à -2 s.

FIGURE 4.6: Exemple de signaux enregistrés à la station FELD et pointés P automatiquement émis pour deux fenêtres temporelles différentes : une débutant à -6 s et une autre à -2 s.

En revanche, pour les stations qui enregistrent un bruit de fond continu plus ou moins élevé, avec peu de signaux de bruit transitoire impulsif, une fenêtre temporelle débutant plus précocement est privilégiée. Ceci évite ainsi les pointés P retardés, du fait de premières arrivées d'ondes P émergentes, se détachant très peu du niveau de bruit de fond, comme c'est le cas dans l'exemple de la figure 4.7. Le calcul de la fonction caractéristique sur une fenêtre temporelle plus précoce peut capter effectivement plus facilement les changements subtils de phase et/ou d'amplitude et/ou de contenu fréquentiel associés à l'arrivée de ces faibles ondes P, se détachant à peine du bruit.

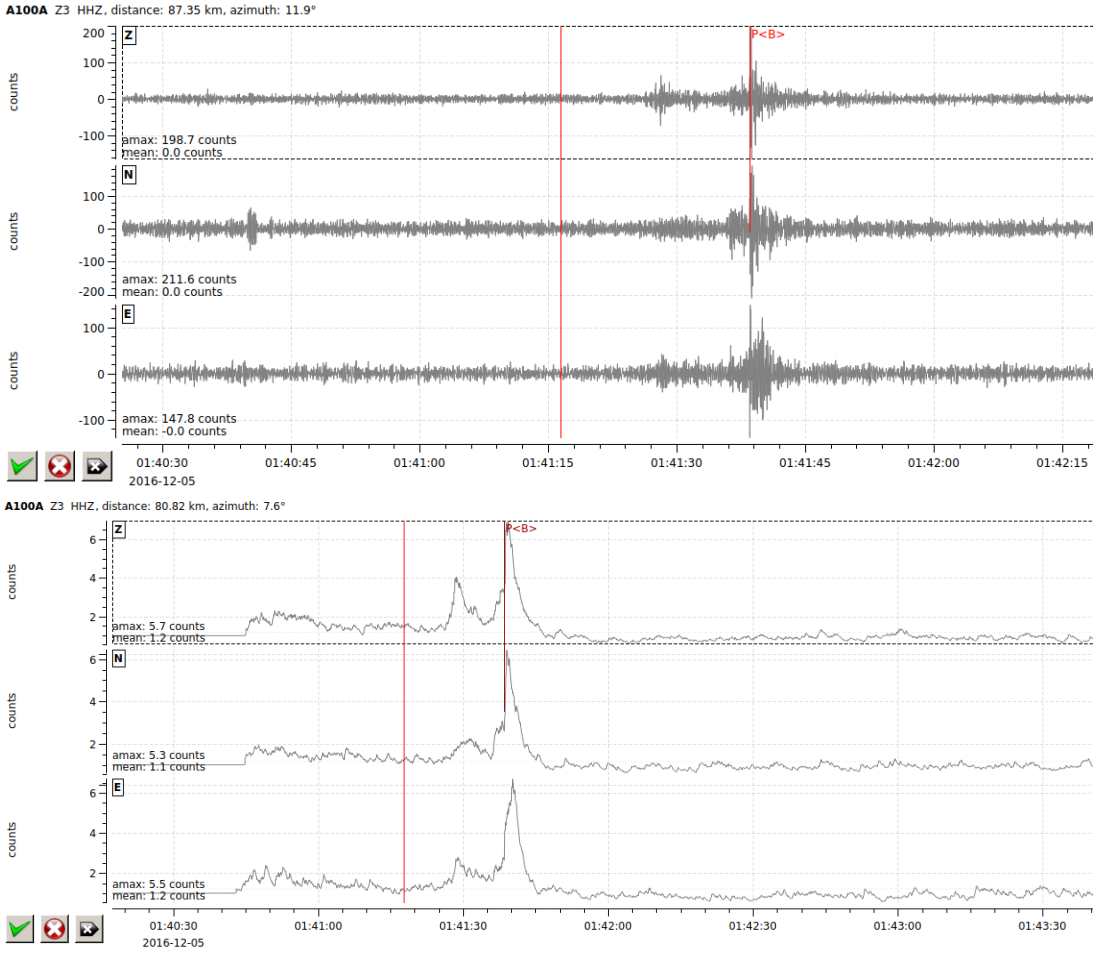


FIGURE 4.7: Sismogrammes enregistrés à la station A100A et fonctions STA/LTA associées. Un faux pointé P a été créé de façon retardée, c'est-à-dire une dizaine de secondes après les premières arrivées des ondes P émises par un séisme identifié au Sud de l'Allemagne, près du lac Konstanz, le 05 décembre 2016 à 01h41 (MLv 1.6).

De cette façon, si je prends l'exemple de la station EMBD, située dans la région du Valais Suisse, près d'une station de ski et non loin d'une voie ferrée, une fenêtre à -25 s captera la faible arrivée des ondes P d'un signal de faible amplitude alors qu'une fenêtre initiée à -6 s ne produira aucun pointé (Figure 4.8).

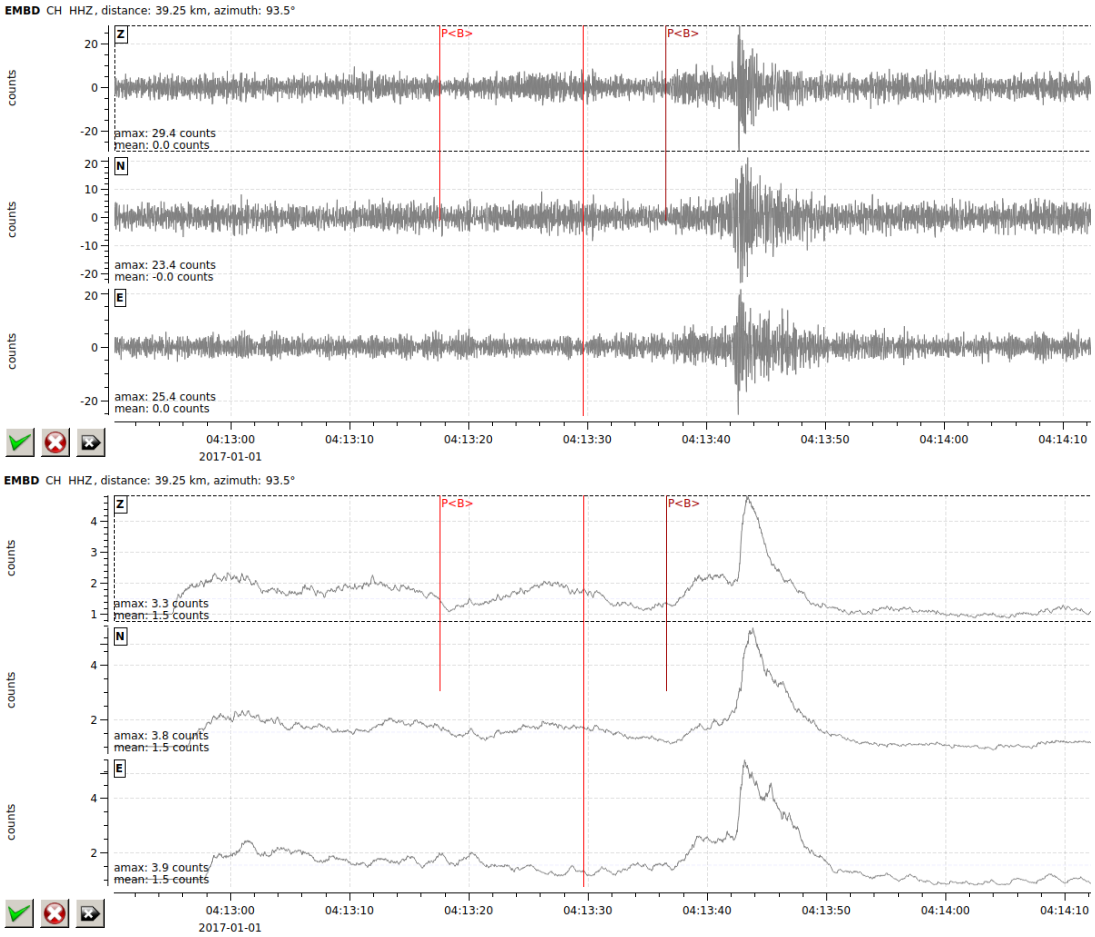


FIGURE 4.8: Sismogrammes enregistrés à la station EMBD et fonctions STA/LTA associées. Les signaux affichés sur ces sismogrammes correspondent à un séisme identifié dans la région de Sion en Suisse le 01 janvier 2017 à 04h13 (MLv 1.1). Deux pointés automatiques P ont été émis à partir d'une fenêtre temporelle débutant à -25s : un premier faux pointé P anticipé (trait vertical rouge clair) et un deuxième pointé P captant les subtils changements de phase et d'amplitude liés à l'arrivée des ondes P (trait vertical rouge foncé). Une deuxième fenêtre temporelle initiée à -6 s n'a produit aucun pointé.

De même, pour la station BRANT, située au coeur du Massif du Jura Suisse, une fenêtre temporelle initiée à -11 s captera plus facilement les variations d'amplitude associées à l'arrivée des ondes P qu'une fenêtre débutant à -3 s (Figure 4.9).

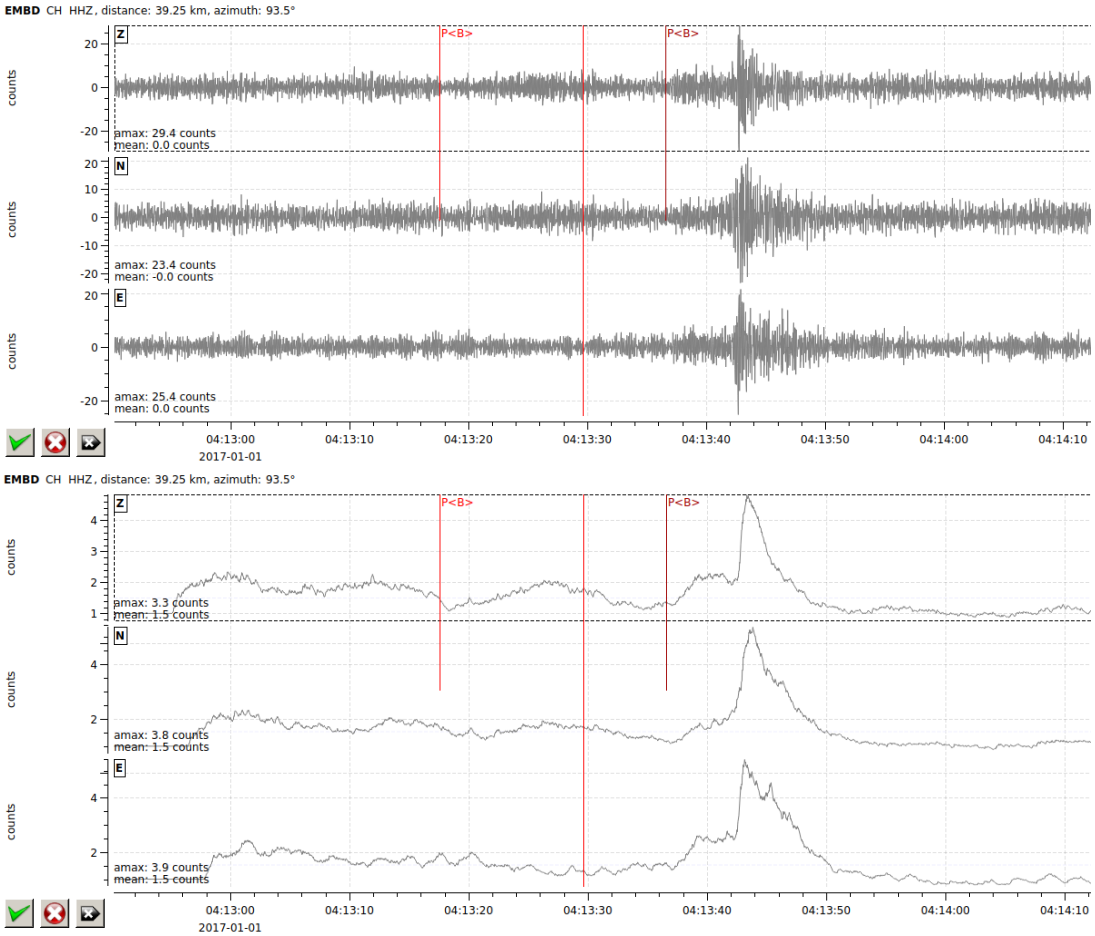
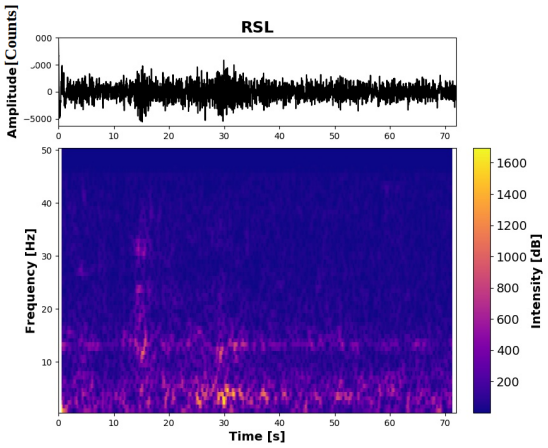
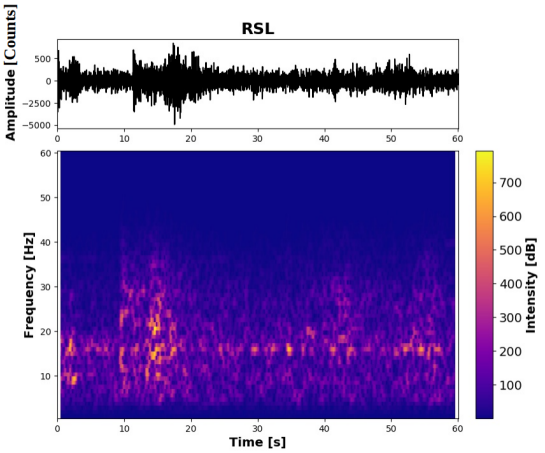


FIGURE 4.9: Sismogrammes enregistrés à la station BRANT et fonctions STA/LTA associées. Les signaux affichés sur ces sismogrammes correspondent à un tir de la carrière de Chéniaz, située en Suisse au Sud du Lac Leman, ayant eu lieu le 05 janvier 2017 à 04h25 (MLv 1.0). Un premier pointé automatique P a été émis à partir d'une fenêtre temporelle débutant à -11s (trait vertical rouge foncé) et un deuxième faux pointé P retardé a été produit à partir d'une fenêtre temporelle commençant à -3 s (trait vertical rouge clair). Une fenêtre temporelle plus longue permet le calcul d'une fonction caractéristique qui capte plus clairement les variations ténues de phase et/ou d'amplitude et/ou contenu fréquentiel associés à l'arrivée des ondes P.

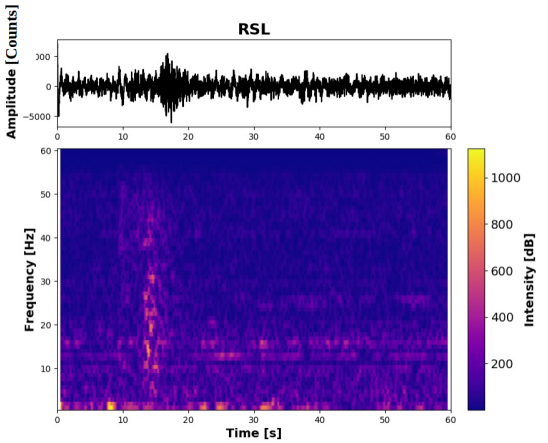
Seulement, face à la diversité des signaux émis et des conditions fluctuantes du niveau de bruit enregistré à une même station, il a été souvent plus judicieux d'établir plusieurs valeurs d'initiation de cette fenêtre temporelle. En effet, par exemple, la station RSL, située à quelques mètres du barrage de Roselend dans les Alpes françaises, enregistre quotidiennement un bruit de fond de forte intensité (500 à 700 décibels) autour de 12 Hz (Figure 4.10).



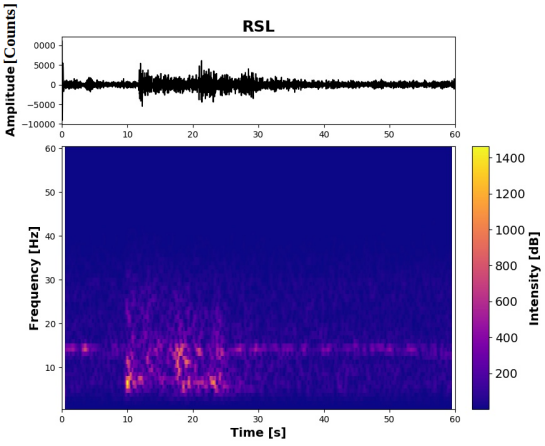
(a) Tir de la carrière de Montalieu-Vercieu à l'Est de Lyon, identifié le 04 août 2016 à 08h06 au Sud-Est de Chambéry dans les Alpes françaises (MLv 1.0).



(b) Séisme ayant eu lieu le 10 septembre 2016 à 07h28 près de Vallorcine dans les Alpes françaises (MLv 0.4).



(c) Séisme ayant eu lieu le 02 octobre 2016 à 06h28 à l'Ouest du Massif de l'Argentière dans les Alpes françaises (MLv 0.6).



(d) Séisme ayant eu lieu le 29 décembre 2016 à 03h48 au Sud du Lac Léman dans les Alpes françaises (MLv 1.1).

FIGURE 4.10: Exemples de spectrogrammes pour quelques signaux enregistrés sur la composante verticale de la station RSL. Cette station enregistre en continu un bruit autour de 12 Hz.

Ce niveau de bruit presque continu se trouve dans les mêmes gammes de fréquences que les ondes P. Si des signaux sismiques de faible amplitude par rapport au bruit de fond sont enregistrés, les rapport signal/bruit évalués vont alors être très faibles. De plus, si les ondes P sont tout juste émergentes, une fenêtre temporelle qui débute à -20 s permettra de capter plus facilement ces faibles arrivées d'ondes P à la station RSL. La fonction caractéristique sera effectivement calculée sur une fenêtre temporelle de bruit plus longue, et évaluera donc mieux les changements subtils de phase et/ou d'amplitude liés à l'arrivée des ondes P, malgré des changements fréquentiels peu perceptibles (Figure 4.11).

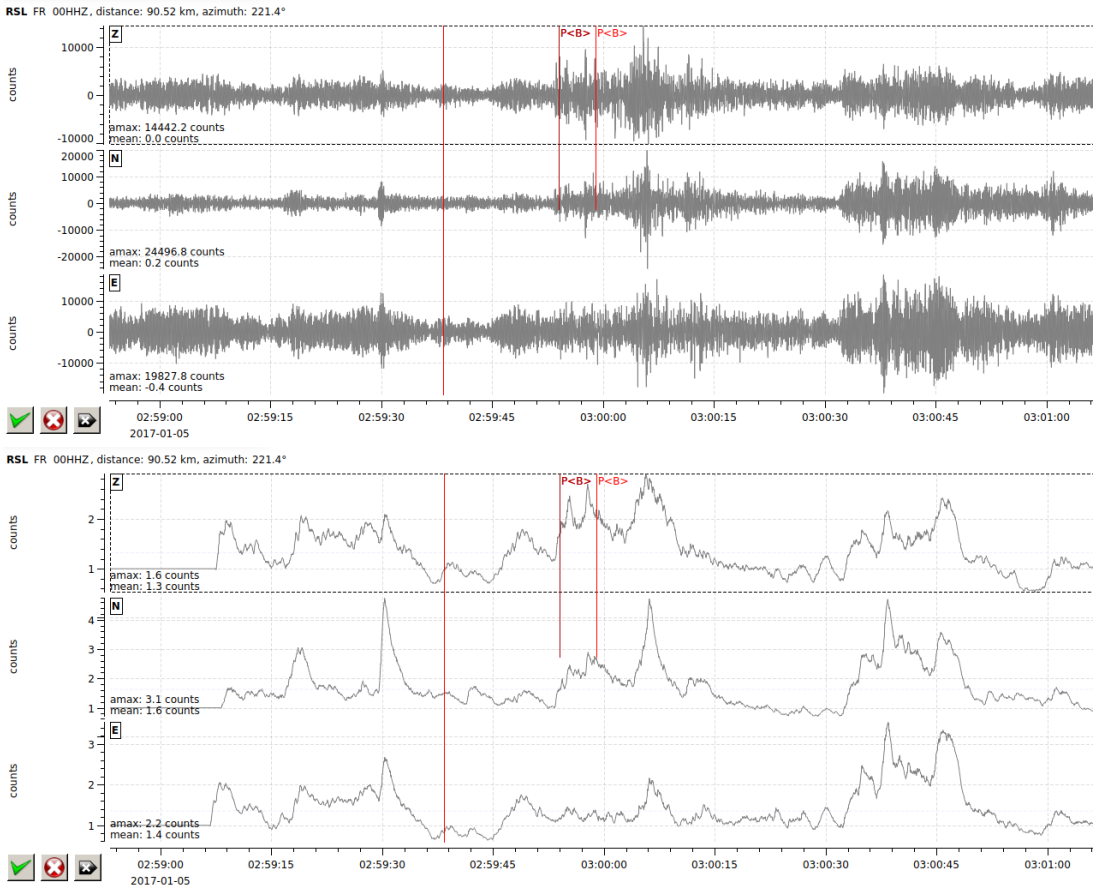


FIGURE 4.11: Sismogrammes enregistrés par la station RSL et fonctions STA/LTA correspondantes. Les sismogrammes affichent un signal qui correspond à un séisme ayant eu lieu le 05 janvier 2017 à 02h59 dans la région de Sion en Suisse (MLv 2.2). Le premier trait vertical rouge correspond à un premier pointé P qui a été redéfini à partir d'une fenêtre temporelle débutant à -20 s. Le deuxième trait vertical rouge correspond à un deuxième pointé P qui a été affiné à partir d'une fenêtre temporelle initiée à -5 s. Ce dernier pointé est un faux pointé retardé.

En revanche, lorsque cette même station enregistre plus périodiquement du bruit impulsif transitoire, de même ordre d'amplitude et contenu fréquentiel que les ondes P, une fenêtre temporelle initiée plus tardivement (ici -5 s) sera préférable. En effet, le calcul de la fonction caractéristique raccourcira l'enregistrement de ces fluctuations importantes de bruit transitoire qui précèdent l'arrivée des ondes P, rendant alors plus visibles les changements de phase associés à l'arrivée de ces ondes (Figure 4.12).

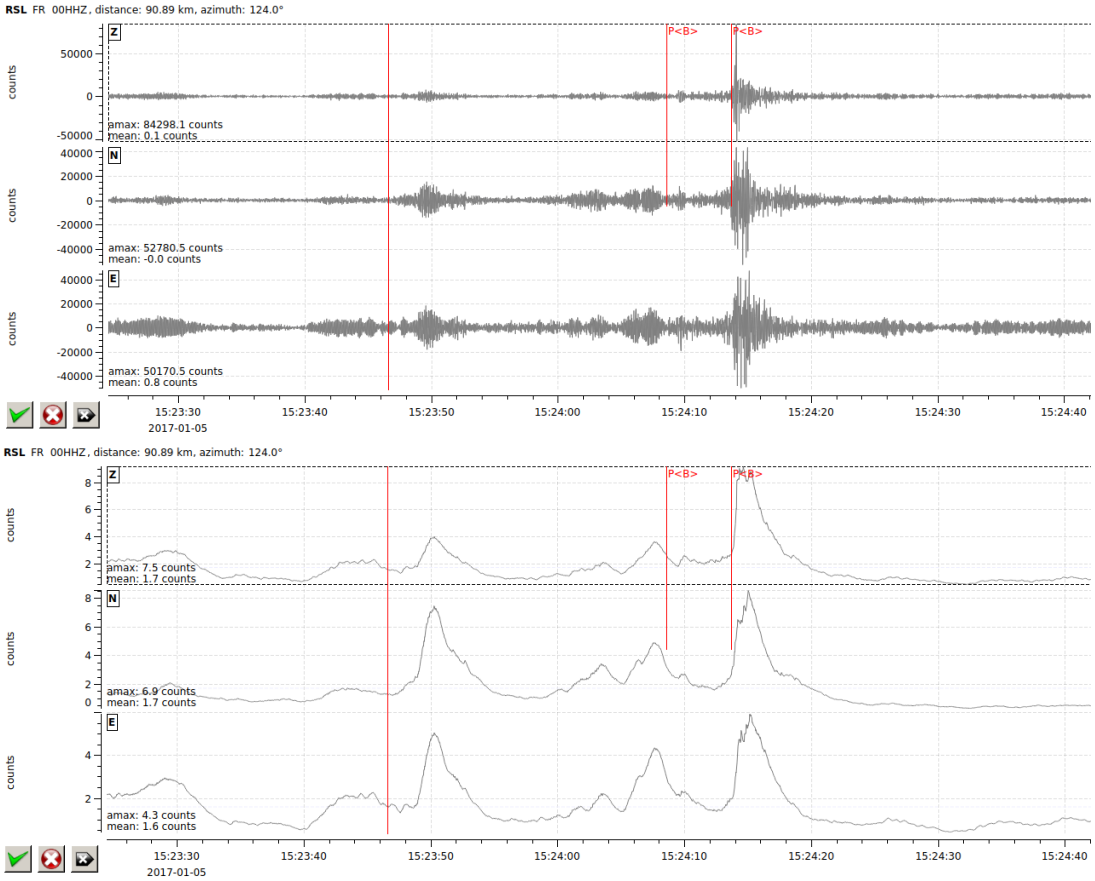
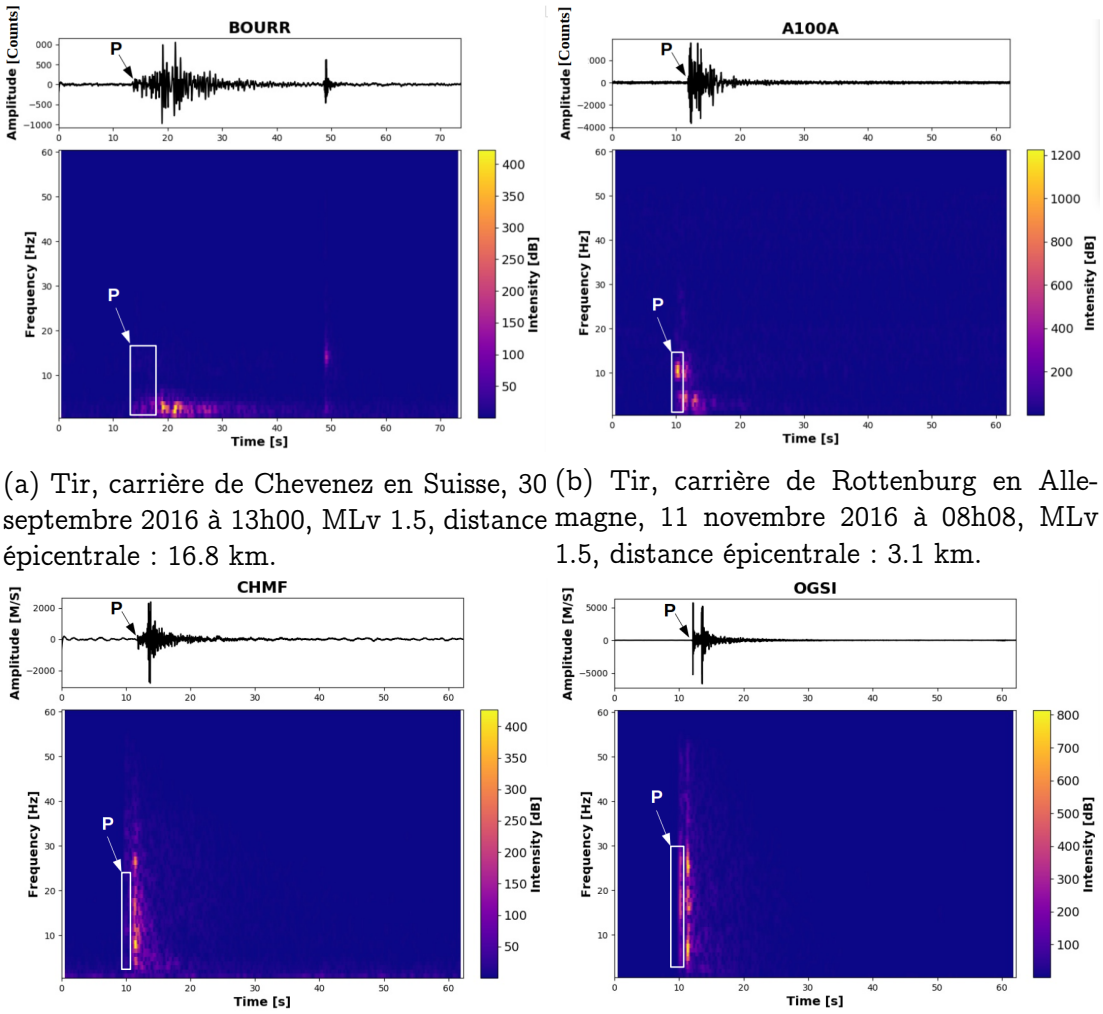


FIGURE 4.12: Sismogrammes enregistrés par la station RSL et fonctions STA/LTA correspondantes. Les sismogrammes affichent un signal qui correspond à un séisme ayant eu lieu le 05 janvier 2017 à 15h24 dans les Alpes françaises (MLv 0.64). Le premier trait vertical rouge correspond à un premier pointé P qui a été redéfini à partir d'une fenêtre temporelle débutant à -5 s. Le deuxième trait vertical rouge correspond à un deuxième pointé P qui a été affiné à partir d'une fenêtre temporelle initiée à -20 s. Ce dernier pointé est un faux pointé retardé.

- Optimiser le filtrage du signal

Le filtre qui a été utilisé pour calculer la fonction caractéristique sur la fenêtre temporelle préalablement définie autour de la détection de l'arrivée des ondes P par la méthode STA/LTA est un filtre passe-bande de Butterworth d'ordre 2 avec fréquences de coupures comprises entre 4 et 20 Hz. En effet, l'intensité du signal associée à l'arrivée des ondes P est concentrée en moyenne dans cette gamme de fréquences (Figure 4.13).



(a) Tir, carrière de Chevenez en Suisse, 30 septembre 2016 à 13h00, MLv 1.5, distance épicentrale : 16.8 km.

(b) Tir, carrière de Rottenburg en Allemagne, 11 novembre 2016 à 08h08, MLv 1.5, distance épicentrale : 3.1 km.

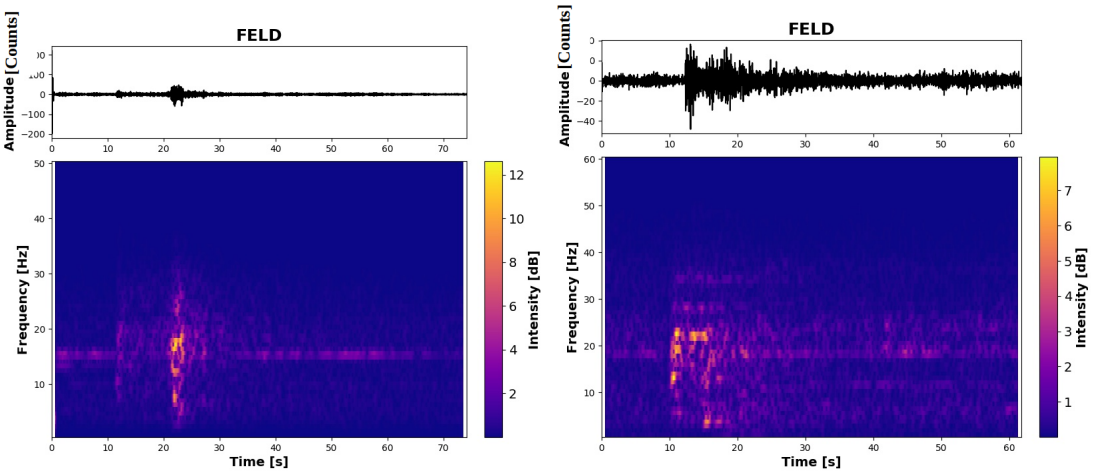
(c) Séisme dans le Jura français, 09 septembre 2016 à 22h39, MLv 1.5, distance épicentrale : 11.1 km.

(d) Séisme dans les Alpes françaises, 02 octobre 2016 à 05h46, MLv 2.2, distance épicentrale : 7.5 km.

FIGURE 4.13: Exemples de spectrogrammes pour quelques signaux enregistrés sur la composante verticale des stations. L'intensité du signal équivalent aux ondes P est plus élevée entre 4 et 20 Hz en moyenne.

Seulement, cette intervalle de fréquences n'est pas efficace pour toutes les stations, et ceci en raison du contenu fréquentiel des signaux enregistrés qui diffèrent en fonction de la localisation des stations. Une étude plus précise du contenu fréquentiel des signaux enregistrés temporellement aux différentes stations utilisées dans cette étude est donc nécessaire pour spécifiquement adapter le filtrage nécessaire à un pointé des ondes P de qualité.

En effet, par exemple, la station FELD est installée au coeur du Massif de la Forêt Noire sur un socle métamorphique, majoritairement des gneiss. Celle-ci enregistre quasi-systématiquement des ondes P qui arrivent avec des fréquences plus élevées, comprises entre 6 et 25 Hz (Figure 4.14). Ce phénomène est probablement dû à l'effet de la propagation des ondes P dans un milieu qui atténue moins rapidement les hautes fréquences. Par conséquent, un filtre avec des fréquences de coupures comprises entre 6 et 25 Hz a été choisi.



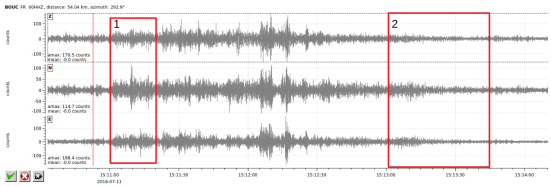
(a) Séisme ayant eu lieu le 18 décembre 2016 à 10h08 dans le Massif des Vosges (MLv 1.8, distance épicentrale : 80.9 km).

(b) Tir de la carrière de Villigen en Suisse ayant eu lieu le 06 octobre 2016 à 09h41 (MLv 1.3, distance épicentrale : 40.5 km.

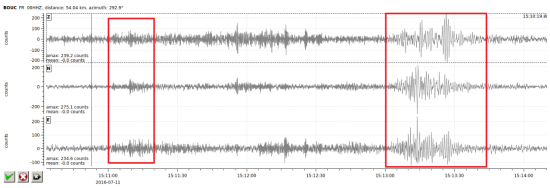
FIGURE 4.14: Exemples de spectrogrammes pour quelques signaux enregistrés sur la composante verticale de la station FELD. L'intensité du signal équivalent aux ondes P est concentrée à des fréquences plus élevées que la moyenne, à savoir comprises entre 6 et 25 Hz.

De même, le filtrage utilisé dépend également du contenu fréquentiel des différents types de bruit enregistrés aux différentes stations. Par exemple, la station BOUC, située en France, dans la périphérie de la ville de Besançon, au bord d'une route départementale, à 6 km d'une voie ferrée, et à 3 km de la carrière de Gonsans, est soumise régulièrement à du bruit haute fréquence ($>$ 10 Hz). L'utilisation du filtre passe-bande de 4 à 20 Hz empêche alors la capture des ondes P, plus particulièrement celles de même ordre d'amplitude que le bruit enregistré, car noyées dans le bruit haute fréquence. Par conséquent, un filtre passe-bande avec des fréquences de coupure plus basses, à savoir entre 3 et 12 Hz, a réduit fortement l'amplitude du bruit haute fréquence, mettant

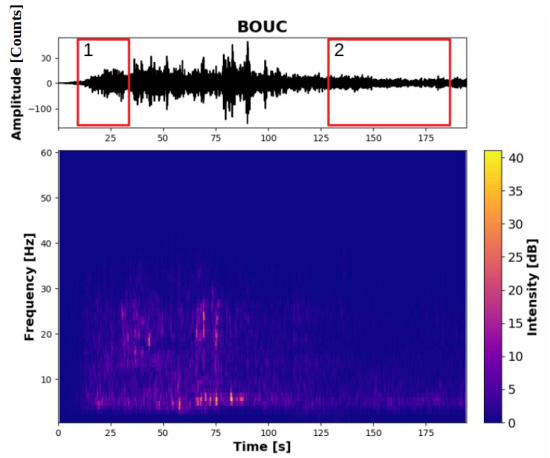
en évidence plus aisément les changements d'amplitude et de fréquence liés à l'arrivée des ondes P (Figure 4.15).



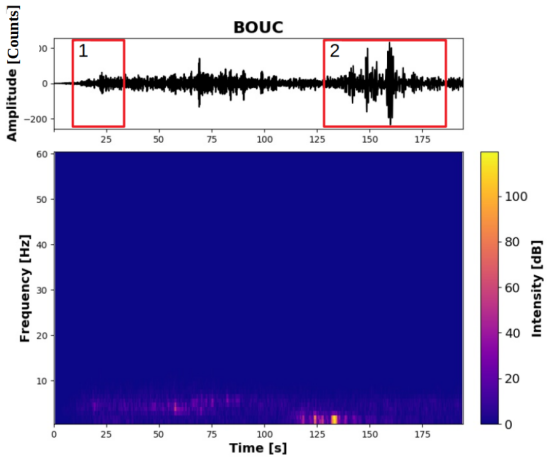
(a) Signaux filtrés (1 et 2) avec une bande passante des fréquences comprises entre 4 et 20 Hz.



(b) Signaux filtrés (1 et 2) avec une bande passante des fréquences comprises entre 3 et 12 Hz.



(c) Spectrogramme des signaux filtrés avec une bande passante des fréquences comprises entre 4 et 20 Hz.



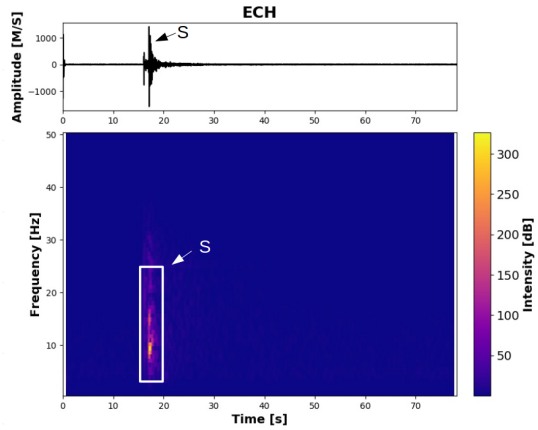
(d) Spectrogramme des signaux filtrés avec une bande passante des fréquences comprises entre 3 et 12 Hz.

FIGURE 4.15: Impact du filtrage sur la détection des signaux à la station BOUC. Le premier signal (1) correspond à un tir de la carrière de Fontaines identifié le 11 juillet 2016 à 15h10 (MLv 1.3, distance épicentrale : 54.0 km). Les spectrogrammes ont été définis à partir des signaux enregistrés sur la composante verticale de la station. Le filtrage bande-passante 3-12 Hz est un filtrage plus adapté pour affiner le pointé automatique P à cette station BOUC.

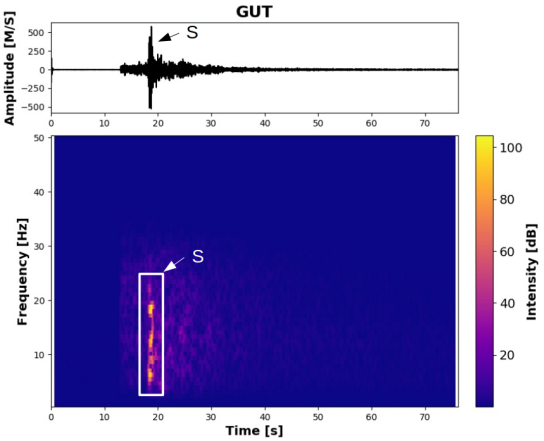
•Pour les pointés automatiques des ondes S

- Déterminer le filtrage du signal le plus adapté

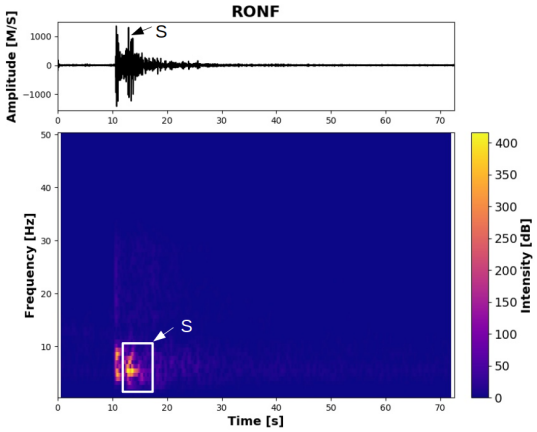
De la même façon, le filtrage du signal est un critère essentiel pour obtenir un pointé du temps d'arrivée des ondes S de qualité. Le filtrage qui a été principalement utilisé pour réaliser l'opération de pointé automatique des phases S est un filtre de Butterworth passe-bande d'ordre 4 avec des fréquences de coupure comprises entre 4 Hz et 25 Hz. En effet, cette gamme fréquentielle correspond à la gamme qui va le mieux capturer les ondes de volume S (Figure 4.16).



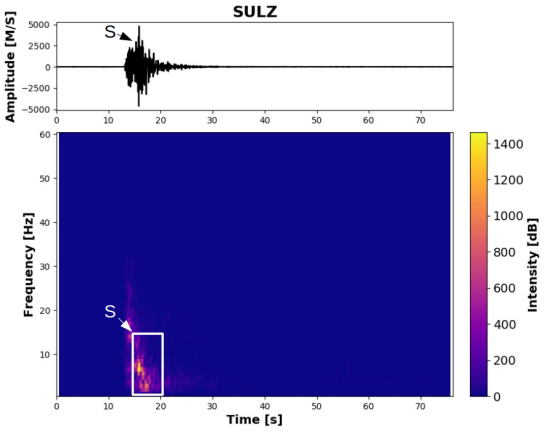
(a) Signal enregistré à la station ECH (distance épicentrale : 7.5 km), correspondant à un séisme identifié dans les Vosges, le 18 décembre 2016 à 10h08 (MLv 1.8).



(b) Signal enregistré à la station GUT (distance épicentrale : 40.0 km), correspondant à un séisme identifié en Allemagne près du lac Konstanz, le 05 décembre 2016 à 02h44 (MLv 2.0).



(c) Signal enregistré à la station RONF (distance épicentrale : 7.0 km), correspondant à un tir de la carrière Lepuix-Gy identifié au Nord de Belfort, le 25 novembre 2016 à 09h45 (MLv 2.0).



(d) Signal enregistré à la station SULZ (distance épicentrale : 17.5 km), correspondant à un tir de la carrière de Villigen identifié dans le Jura Suisse, le 15 septembre 2016 à 09h41 (MLv 1.9).

FIGURE 4.16: Exemples de spectrogrammes de quelques signaux détectés automatiquement. L'intensité du signal correspondant à la phase sismique S se concentre en moyenne entre 4 et 25 Hz.

Si l'intensité du signal équivalent à l'arrivée des ondes S reste forte sur des gammes fréquentielles plus larges (jusqu'à 25 Hz) pour les séismes, cette intensité reste concentrée autour de 4 et 10 Hz pour les tirs de carrière. De cette façon, utiliser une filtre passe-bande moins restrictif (bande 4-25 Hz) permet de capter plus facilement l'arrivée des ondes S qui sont beaucoup plus marquées pour les séismes.

Seulement, ce filtre de Butterworth n'est pas efficace pour toutes les stations. Pour optimiser le pointé automatique des ondes S, il a été nécessaire d'adapter le filtrage aux caractéristiques systématiques du bruit enregistré à ces stations. Une analyse des signaux enregistrés temporellement aux différentes stations a donc là aussi été nécessaire pour mettre en évidence l'impact des fréquences des différents bruits enregistrés sur la qualité des pointés S. En effet, si l'on prend l'exemple de la station AIGLE, celle-ci se situe en Suisse au Sud du Lac Léman à 600 m de la voie ferrée et à 1 km d'un réseau autoroutier.

L'analyse de la fonction de densité spectrale de puissance pour la station AIGLE montre que la puissance du bruit est élevée aux gammes de fréquence typiques du bruit d'origine anthropique, c'est-à-dire comprises entre 1 et 10 Hz. Pour ces gammes fréquentielles, cette puissance est variable et peut augmenter d'environ 20 décibels par rapport à la puissance minimale. De plus, elle atteint des probabilités fortes d'occurrence (de l'ordre de 15 à 20 %) par rapport au modèle de bruit bas (NLNM). A partir de 20 Hz, la puissance de bruit atteint des probabilités d'occurrence maximales élevées (30 %), s'éloignant radicalement du NLNM. Cette station, particulièrement bruitée, est donc très sensible au bruit haute fréquence. (Figure 4.17).

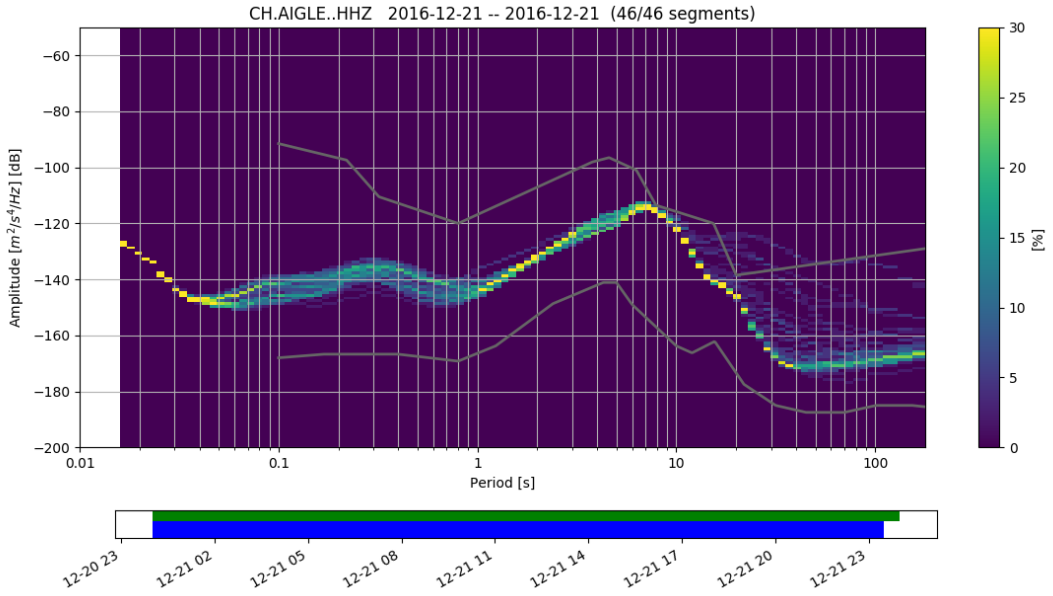
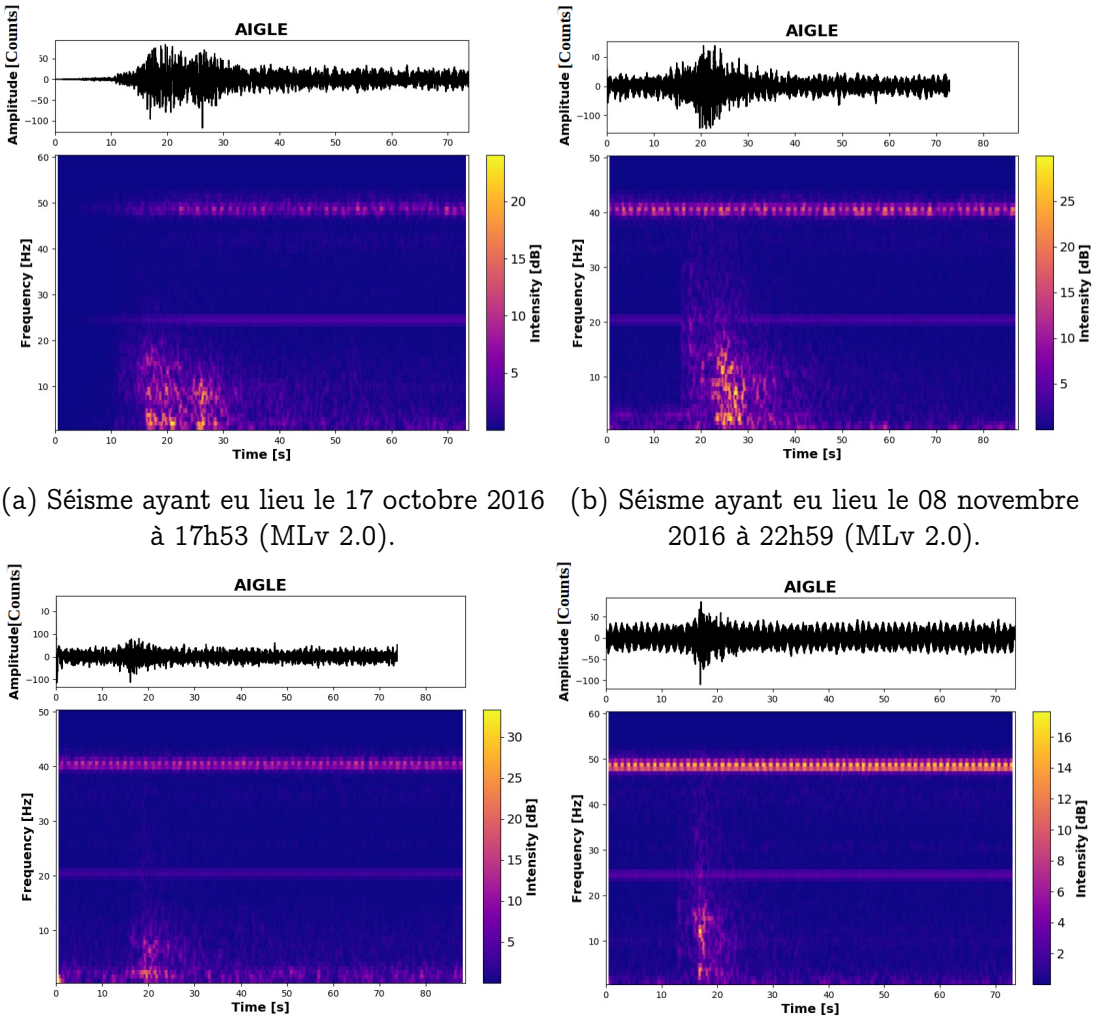


FIGURE 4.17: Densité spectrale de puissance probabiliste calculée pour la station AIGLE. Les courbes grises correspondent aux modèles de bruit standard (courbe supérieure = modèle de bruit élevé [NHNM] et courbe inférieure = modèle de bruit bas [NLNM] ; PETERSON, 1993). Les niveaux de bruit de la station sont estimés sur une large gamme de fréquences de 0.01 Hz à 16 Hz (soit une période de 100 s à 0.0625 s). En bas du graphique sont affichées les données qui ont servi au calcul de cette fonction. Le rectangle vert représente les données disponibles et le rectangle bleu montre l'étendue des données qui ont servi au calcul. Ces spectres ont été obtenus via le package ObsPy de Python suivant la méthode de MCNAMARA et al., 2004

L'observation de quelques spectrogrammes de signaux enregistrés à la station AIGLE montre effectivement deux bandes continues de haute fréquence quasi-systématiques : une bande à 25 Hz montrant une intensité du signal de l'ordre de 5 à 12 décibels et une autre à 50 Hz, affichant une intensité du signal de l'ordre de 7 à 20 décibels. La bande à 50 Hz équivaut à la fréquence fondamentale de l'alimentation électrique et la bande à 25 Hz probablement son harmonique inférieure, soulignant alors un artefact d'origine électrique (Figure 4.18).



(a) Séisme ayant eu lieu le 17 octobre 2016 à 17h53 (MLv 2.0).

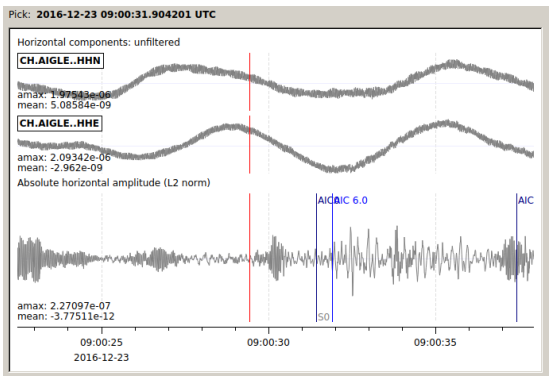
(b) Séisme ayant eu lieu le 08 novembre 2016 à 22h59 (MLv 2.0).

(c) Séisme ayant eu lieu le 23 décembre 2016 à 09h00 (MLv 0.7).

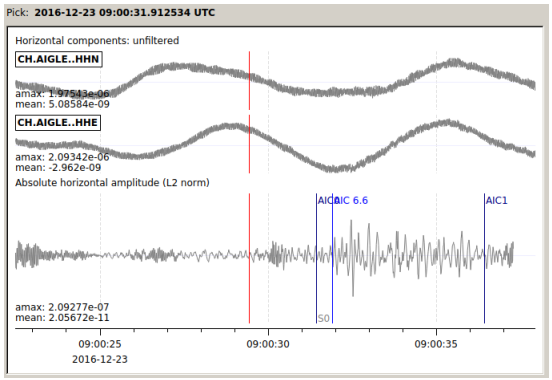
(d) Séisme ayant eu lieu le 01 janvier 2017 à 04h13 (MLv 2.0).

FIGURE 4.18: Spectrogrammes de quelques signaux enregistrés sur la composante verticale de la station AIGLE.

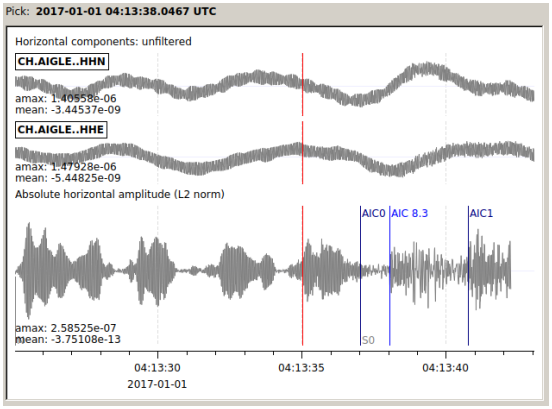
De cette façon, afin de capturer le maximum d'intensité du signal correspondant à l'arrivée des ondes S, un filtrage passe-bande de Butterworth avec des fréquences de coupure comprises entre 4 et 21 Hz a été sélectionné pour cette station AIGLE, au lieu de 4 et 25 Hz. Avec ce filtre spécifique, le bruit de fond haute-fréquence enregistré à cette station est fortement réduit. Ceci a pour effet d'augmenter corrélativement le rapport signal/bruit associé au pointé de la phase S (Figure 4.19a et b), et de diminuer la probabilité de pointer du bruit haute fréquence, qui se chevauche avec l'arrivée des ondes S, du fait d'une amplitude et d'un contenu fréquentiel équivalents (Figure 4.19b et c). Ces deux effets favorisent alors l'émission de pointés automatiques S de meilleure qualité.



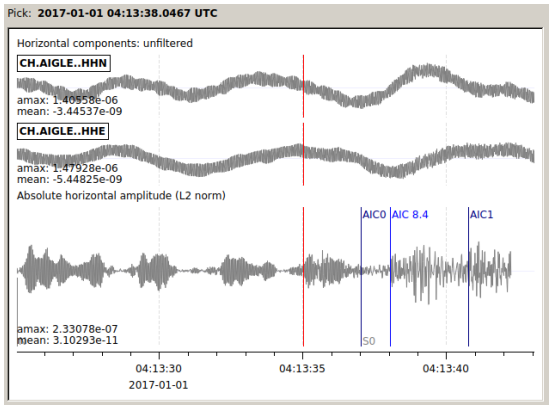
(a) Filtre passe-bande de Butterworth d'ordre 4 avec fréquences de coupure 4-25 Hz (séisme du 23 décembre 2016 à 09h00).



(b) Filtre passe-bande de Butterworth d'ordre 4 avec fréquences de coupures 4-21 Hz (séisme du 23 décembre 2016 à 09h00).



(c) Filtre passe-bande de Butterworth d'ordre 4 avec fréquences de coupures 4-25 Hz (séisme du 01 janvier 2017 à 04h13).



(d) Filtre passe-bande de Butterworth d'ordre 4 avec fréquences de coupures 4-21 Hz (séisme du 01 janvier 2017 à 04h13).

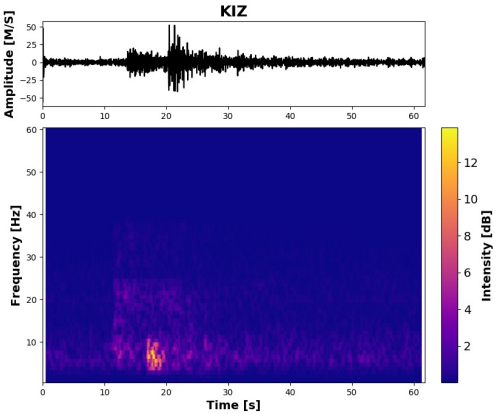
FIGURE 4.19: Impact du filtrage sur la qualité des pointés automatiques des phases sismiques S pour la station AIGLE. (a) + (b) Augmentation du rapport signal/bruit. (b) + (c) Réduction de l'effet parasite du bruit haute-fréquence. Le trait vertical rouge correspond au pointé P de référence ; les traits verticaux bleus surmontés de "AIC0" et "AIC1" indiquent respectivement le début et la fin du traitement du sismogramme pour réaliser le pointé S. Le traitement s'arrête lorsque le rapport signal/bruit minimum (SNR) et le nombre minimum à partir duquel le critère AIC minimal doit être rencontré sur des fenêtres temporelles adjacentes sont atteints pour pointer une phase S (ici SNR= 3.5 et nombre minimum= 2). Le ligne verticale bleue surmontée de AIC correspond au pointé S effectué et le nombre relate le rapport signal/bruit avec lequel il a été émis. Les deux composantes horizontales de la station sont utilisées pour pointer les phases S. La partie supérieure de chaque encadré montre les traces sismiques non filtrées des deux composantes horizontales. La partie inférieure correspond à la somme vectorielle des composantes horizontales (trace L2).

De même, un filtre plus restrictif avec fréquences de coupure comprises entre 3 et 15 Hz a permis de mieux détecter les arrivées des ondes S à la station KIZ, qui est située au Sud-Est de Freiburg au sein du Massif de la Forêt Noire, mais pour d'autres raisons. En effet, cette station est impliquée dans la détection de près de 3.5 fois plus de carrières et les distances épicentrales évaluées sont en moyenne de 103.22 km (médiane = 65.05 km). De plus, pour seulement 8 % des événements détectés, les distances épicentrales sont estimées à moins de 30 km.

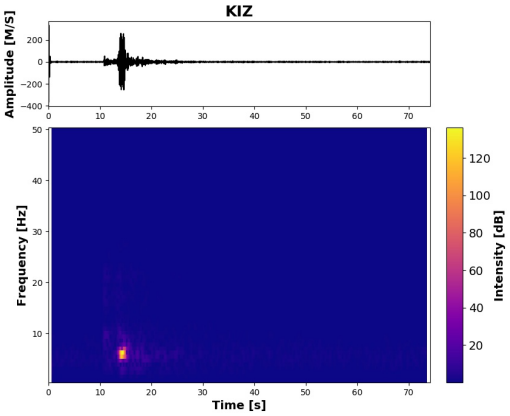
Par conséquent, au-delà des caractéristiques de bruit inhérentes à chaque station, le type de filtrage utilisé dépend également de la localisation de la station au regard de la probabilité d'occurrence des événements enregistrés (type d'événement et localisation de la source). En effet, ce filtrage reflète à la fois la probabilité plus élevée que des signaux enregistrés à la station KIZ soient reliés à des tirs de carrière émettant des ondes S dans des gammes de fréquence globalement plus faibles (< 15 Hz), et la probabilité plus grande que ces stations soient situées à des distances épicentrales plus grandes, enregistrant donc des ondes S plus atténuées en haute fréquence (Figure 4.20).

Seulement, si ce filtrage peut paraître au premier abord biaisé, il n'empêche pas le pointé correct des ondes S pour des signaux dont l'intensité du signal reste élevée jusqu'à 25 Hz (Figure 4.20e). En effet, si un peu de signal risque d'être perdu, la station KIZ affichant globalement un niveau de bruit de fond constant et minimal à des fréquences supérieures à 7 Hz (avec peu de bruit impulsif), il est possible de capter plus facilement le signal autour de 10 Hz avec moins d'interférences (Figure 4.21).

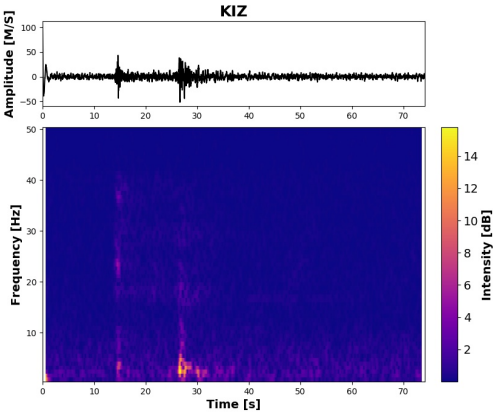
La configuration optimale du filtrage résulte donc à la fois du milieu de propagation des ondes, du site d'implantation de la station, ainsi que de l'orientation et de la distance de cette station à la source, soulignant là encore la diversité et la complexité de l'ensemble des signaux enregistrés.



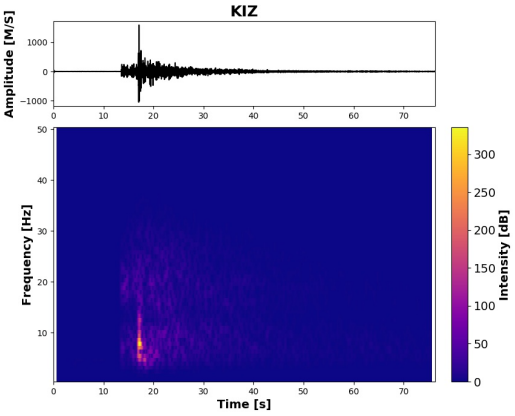
(a) Tir de la carrière de Villigen en Suisse (15 septembre 2016 à 09h41, MLv 1.9 , distance épicentrale : 57.3 km).



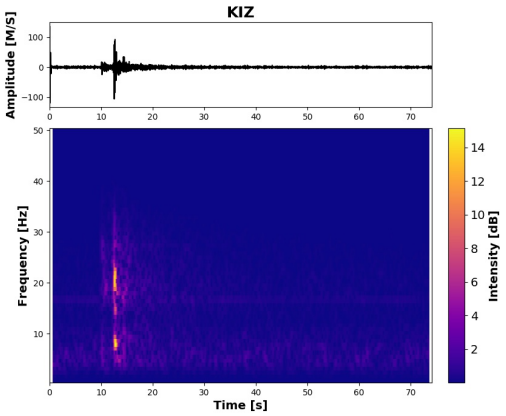
(b) Tir de la carrière de Bötzingen en Allemagne (21 septembre 2016 à 12h00, MLv 1.4 , distance épicentrale : 20.4 km.



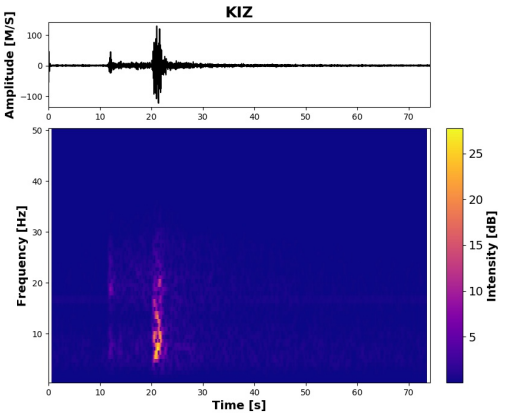
(c) Tir de la carrière de Raon-L'Etape dans les Vosges (10 octobre 2016 à 10h01, MLv 1.6 , distance épicentrale : 98.3 km.



(d) Séisme identifié au Nord-Est de Mulhouse dans le Fossé Rhénan (18 août 2016 à 00h51, MLv 2.1, distance épicentrale : 26.9 km).

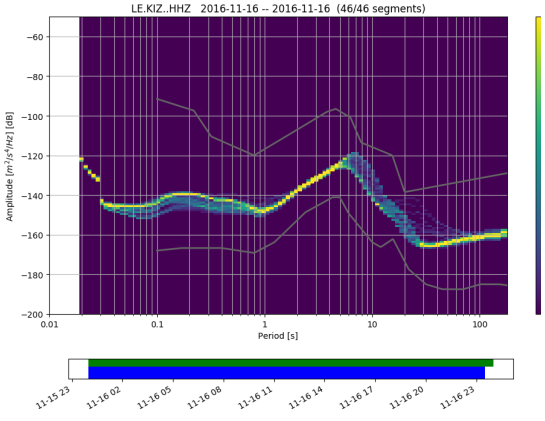


(e) Séisme identifié au Sud-Est de Freiburg en Allemagne (25 novembre 2016 à 14h41, MLv 0.8, distance épicentrale : 0.5 km).

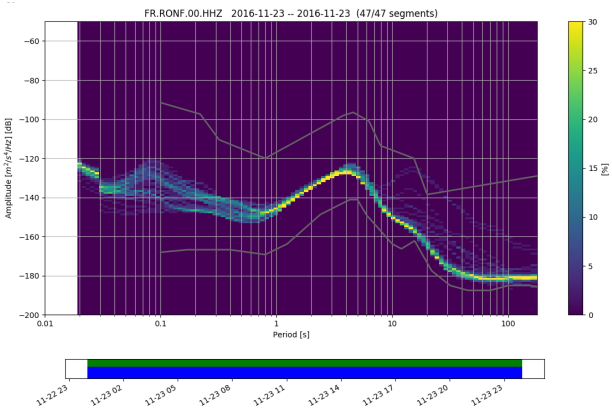


(f) Séisme identifié dans le Massif des Vosges (18 décembre 2016 à 10h08, MLv 1.8, distance épicentrale : 71.1 km).

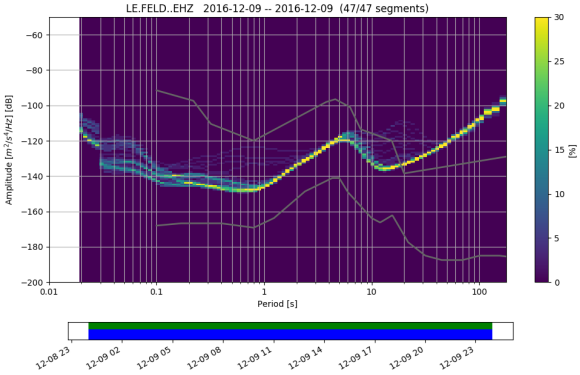
FIGURE 4.20: Exemples de spectrogrammes de signaux enregistrés sur la composante verticale de la station KIZ. Le filtrage utilisé pour le pointé automatique (3-15 Hz) est fonction de la distance épicentrale et du type d'événement.



(a) Station KIZ située au Sud-Est de Freiburg, à 2.5 km d'une route nationale et 3 km d'un aérodrome.



(b) Station RONF située dans les Vosges du Sud à 500 mètres d'une voie ferrée



(c) Station FELD située sur le plus haut sommet de la Forêt Noire, à quelques dizaines mètres de 3 tours de communication et à 2 km d'une station de ski très touristique.

FIGURE 4.21: Densité spectrale de puissance probabiliste calculée pour la station KIZ (a), comparée à celles des stations RONF (b) et FELD (c), plus sensibles au bruit transitoire impulsif. Les 3 stations sont sensibles au bruit d'origine anthropique (fréquences > 1 Hz). Contrairement aux stations RONF et FELD, la puissance du bruit à la station KIZ atteint un niveau de l'ordre de -145 dB en moyenne, avec une probabilité d'occurrence de l'ordre de 30 %, et affiche un seuil minimal d'amplitude aux plus hautes fréquences (7 à 50 Hz). En revanche, les stations RONF et FELD affichent des niveaux de bruit beaucoup plus variables, avec des sauts d'amplitude plus forts autour de 11 Hz pour RONF et autour de 15 Hz pour FELD, ainsi que des probabilités d'occurrence plus faibles (de l'ordre de 15%). Ces spectres ont été obtenus via le package ObsPy de Python suivant la méthode de MCNAMARA et al., 2004.

•Déterminer un rapport signal/bruit optimal

Un autre paramètre qui gouverne la qualité des pointés S est l'estimation du rapport signal/bruit minimal nécessaire pour accepter un pointé. Ce rapport signal/bruit a été paramétré globalement à 3.5. Cette valeur a été définie empiriquement à partir d'un jeu test de différentes valeurs comprises entre 1.5 et la valeur par défaut définie dans SeisComP3 égale à 5. Ces différents paramétrages ont été testés sur le pointé automatique des ondes S à partir de sismogrammes enregistrés entre juillet et août 2016, puis janvier 2017. La valeur de 3.5 a été obtenue pour un grand nombre de stations. Cette valeur correspond à la valeur minimale du rapport signal/bruit nécessaire pour obtenir un nombre maximal de pointés S de qualité, malgré un niveau de bruit enregistré élevé.

Seulement, pour des stations enregistrant régulièrement des fluctuations importantes de bruit transitoire d'origine anthropique, un rapport signal/bruit minimal plus élevé a été obtenu (autour de 4). Ceci limite effectivement la possibilité de pointer du bruit au lieu de la phase S, en particulier pour des signaux sismiques dont l'amplitude et le contenu fréquentiel s'approchent de ceux du bruit.

En revanche, un rapport signal/bruit minimal supérieur à 4 a par exemple été obtenu pour pointer les arrivées des ondes S aux stations permanentes telles que GIMEL (à 50 m d'une route circulante), BRANT (à 3 km d'une voie ferrée et d'une autoroute), MOF (à 4 km d'une route nationale et 800 m d'une route départementale) ou EMBD (à 300 m d'une station de ski et à 400 m d'une voie ferrée). Par ailleurs, ces stations correspondent aux mêmes stations qui sont fortement impliquées dans la génération de 10 à 35% des faux événements détectés (Figure 3.19).

Il en est de même pour les stations temporaires AlpArray telles que A117A (au sein d'une exploitation agricole et à 1 km d'une zone urbaine), A164A (au bord d'une petite route, à 1500 m d'une autoroute et à 500 m d'une route nationale), A113A (au coeur d'un village, à 2 km d'une voie ferrée) ou A116A (à 900 m d'une autoroute). Ces stations sont également impliquées dans la génération de 10 à 25 % des faux événements détectés (cf Figure 3.18).

En revanche, pour d'autres stations, la valeur du rapport signal/bruit minimal obtenue a été plus faible, c'est-à-dire autour de 3. En effet, ces stations, un peu plus éloignées des axes routiers et des centres d'activité urbaine, ont tendance à enregistrer des niveaux de bruit haute-fréquence moins élevés que les stations précédentes. Les ondes S ayant des gammes de fréquence et des amplitudes équivalentes à ceux du bruit enregistré (en particulier pour les événements de plus faible magnitude), les rapports signal/bruit sont alors ici plus élevés, permettant la diminution du rapport signal/bruit minimal à atteindre pour pointer les ondes S.

Parmi ces stations, on retrouve des stations permanentes comme GUT (située dans le Jura Souabe), BALST (située dans le Jura Suisse), ECH (située

dans les Vosges du Nord) ou RIVEL (située dans le Jura français) et des stations temporaires telles que A112A (située dans le Massif de Rhenish au Nord-Ouest de l'Allemagne), A158A (située dans les Vosges) ou A173A (située dans les Alpes du Nord françaises).

Ces stations affichent effectivement des puissances de bruit aux hautes fréquences ($>$ 10 Hz) qui sont inférieures de 20 à 40 dB aux puissances de bruit estimées pour les stations évoquées précédemment, pour des probabilités d'occurrence équivalentes (Figure 4.22).

Par ailleurs, ces stations sont également moins impliquées dans la création de faux événements, de l'ordre de 5 à 10% d'entre eux (cf Figures 3.19 et 3.18), mais elles font partie de celles qui sont le plus impliquées dans la création des vrais événements (cf Figures 3.21 et 3.20).

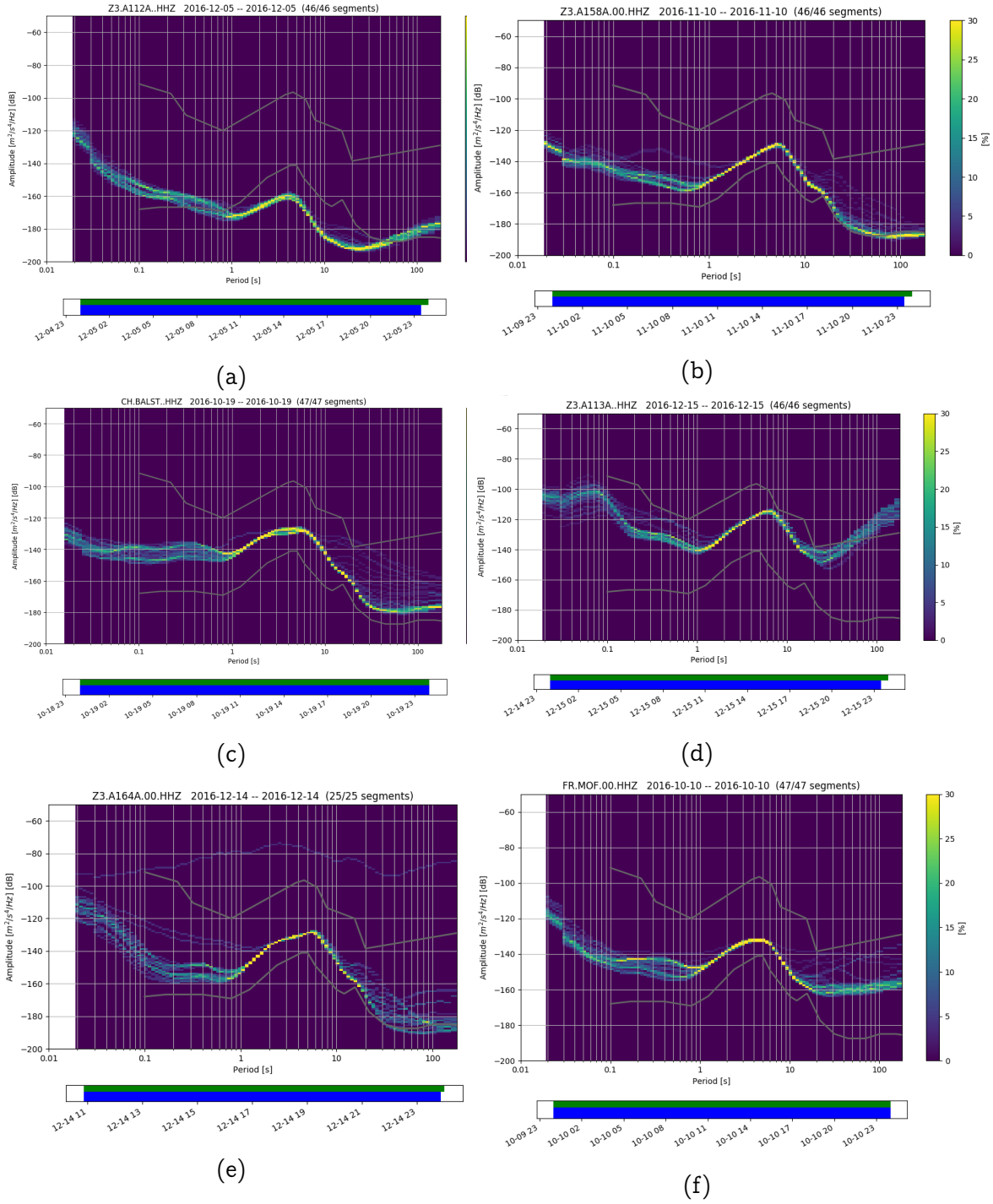


FIGURE 4.22: Comparaison des densités spectrales de puissance probabiliste des stations dont le rapport signal/bruit minimum pour activer un pointé S est inférieur ou égal à 3 (A112A (a), A158A (b) et BALST(c)) avec celles des stations dont le rapport signal/bruit minimum pour activer un pointé S est supérieur ou égal à 4 (A113A (d), A164A (e) et MOF (f)). Avec une probabilité d'occurrence équivalente, les stations A112A, A158A et BALST affichent une puissance du bruit de l'ordre de 20 à 40 dB inférieure aux puissances estimées aux stations A113A, A164A et MOF, pour les gammes de haute fréquence comprises entre 10 et 35 Hz. Ces spectres ont été obtenus via le package ObsPy de Python suivant la méthode de MCNAMARA et al., 2004.

•Optimiser le calcul du critère AIC

Trois paramètres sont utilisés pour définir le critère AIC minimal : la durée du signal choisie pour rechercher la valeur minimale du critère AIC, la taille des fenêtres temporelles utilisées sur ce signal pour calculer le critère AIC sur différents segments du signal, et le nombre minimum de fois que le critère AIC minimal doit être trouvé consécutivement.

Début du signal. Le début du signal qui est choisi pour entamer le calcul du critère AIC est défini à partir du pointé P qui sert de référence. Pour beaucoup de stations, la valeur qui est sélectionnée est la valeur qui correspond au temps minimal qui sépare l'arrivée des ondes P de l'arrivée des S dans la zone d'étude, c'est-à-dire 1.68 s. Cette valeur a été obtenue à partir de l'analyse statistique des différences de temps séparant les ondes P et S calculées pour l'ensemble des événements détectés par le BCSF-RéNaSS au cours de l'année 2016.

En revanche, pour certaines stations, il a fallu augmenter cette valeur, qui est en fait fonction des distances épicentrales moyennes estimées pour chaque station. Pour obtenir la valeur optimale, un ensemble de valeurs a été testé empiriquement sur l'ensemble des stations en évaluant leur impact sur la qualité du pointé automatique des ondes S à partir des sismogrammes enregistrés entre juillet et août 2016 ainsi que janvier 2017.

De ce fait, le début du signal a été placé à des valeurs comprises entre 2.5 et 3 s pour des stations comme SLE (située au Nord du Lac de Konstanz en Allemagne), ECH (située au Nord des Vosges) ou bien RIVEL (située dans le Jura français). Or, 75% des événements détectés par ces stations sont situés à des distances épicentrales supérieures à 50 Km. Ces stations sont alors plus impliquées dans la détection des événements à l'échelle régionale qu'à l'échelle locale. La probabilité d'enregistrer des arrivées d'ondes S plus retardées relativement aux ondes P de référence, est donc nécessairement plus élevée.

Taille des fenêtres temporelles De plus, la taille des fenêtres temporelles utilisées pour calculer le critère AIC le long du signal extrait est en moyenne de 0.96 s pour l'ensemble des stations. De la même façon, celle-ci a été définie empiriquement à partir d'un ensemble de valeurs possibles testées, dont l'impact sur la qualité des pointés automatiques des ondes S émis aux différentes stations a été évalué sur les périodes juillet-août 2016 et janvier 2017.

Cependant, cette valeur de 0.96s va être modifiée également en fonction de la localisation des stations par rapport aux événements détectés. En effet, par exemple 35% des signaux enregistrés à la station AIGLE ou à la station DIX correspondent à des événements (séismes ou tirs de carrière) qui sont situés à des distances épicentrales de moins de 30 km pour AIGLE (située au Sud du Lac Léman) et 40 km pour DIX (située au coeur du Valais Suisse). Dans ces cas-ci, des fenêtres temporelles plus courtes comprises entre 0.5 et 0.6 s a permis de mieux capter les arrivées précoces des ondes S relativement aux ondes P, du fait de la faible distance épicentrale.

Et ceci a d'ailleurs permis d'améliorer davantage la qualité des pointés des ondes S plutôt que de placer le début du signal à des valeurs inférieures à 1.68 s. L'avancée du début du signal pour calculer le critère AIC augmente effectivement la probabilité de pointer du signal haute fréquence dans la coda des ondes P, plutôt que les premières arrivées des ondes S.

Ainsi, rechercher des valeurs optimales pour les deux paramètres nécessaires au calcul optimal du critère AIC (début du signal et taille des fenêtres temporelles) permet d'adapter ce calcul aux distances épicentrales, et donc à l'arrivée différentielle des ondes S.

Deux tendances se dégagent en conséquence. La première tendance observée est que plus la distance épicentrale augmente, plus le début du signal sélectionné sera retardé par rapport à la valeur de référence (ici 1.68 s) et plus la fenêtre temporelle utilisée pour calculer le critère AIC sera longue. La deuxième tendance qui est constatée est que, pour un même début de signal (ici entre 1.68 s et 2.2 s), la fenêtre temporelle utilisée pour calculer la critère AIC augmente avec la distance épicentrale (Figure 4.23).

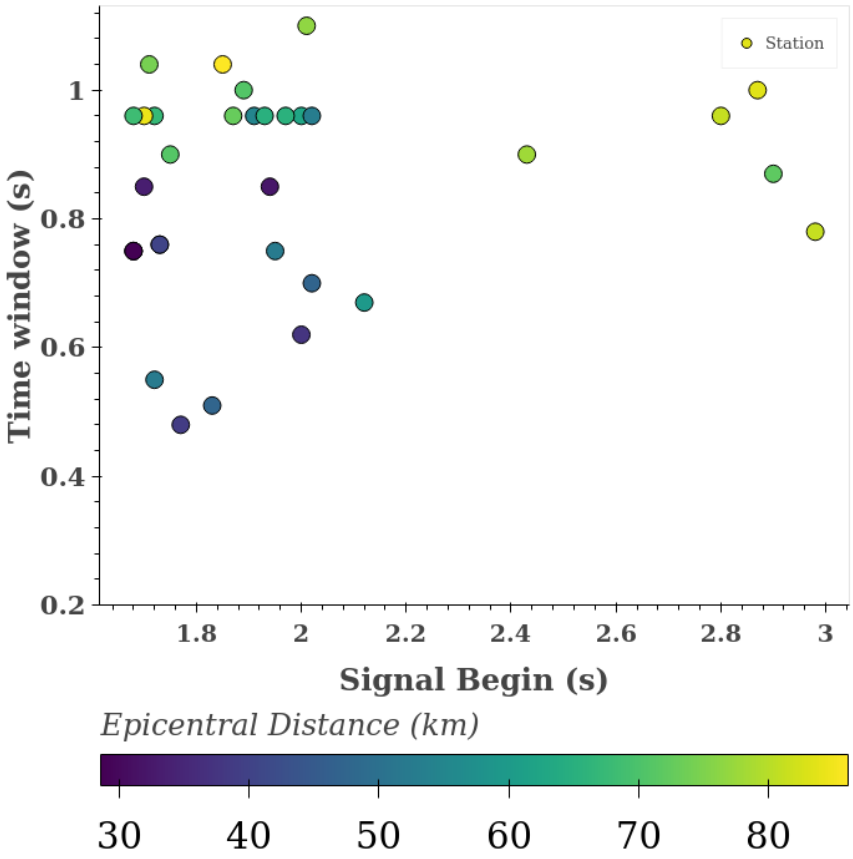


FIGURE 4.23: Évolution de la taille de la fenêtre temporelle utilisée pour calculer le critère AIC en fonction du début du signal sélectionné pour initier le calcul et de la distance épicentrale. Chaque point correspond à une station.

Enfin, le nombre de fois que le critère AIC minimum est trouvé sur deux fenêtres temporelles adjacentes pour activer un pointé automatique S, a été placé à 3, qui est la valeur définie par défaut dans SeisComP3.

Néanmoins, une valeur de 2 a parfois été paramétrée dans les cas où des stations performantes, qui ont tendance à détecter beaucoup d'événements, ont un début de signal pour calculer le critère AIC tardif par rapport à la valeur moyenne de référence (c'est-à-dire $>$ 1.68 s). Ainsi, si ces stations en question ont tendance à détecter plus fréquemment des événements locaux, comme par exemple AIGLE ou DIX, cette valeur de 2 offre la possibilité de détecter l'arrivée des ondes S pour des rapports signal/bruit plus élevés, c'est-à-dire pour des signaux moins pollués par le bruit haute fréquence que ces stations ont tendance à enregistrer. Si au contraire ces stations détectent plus fréquemment des séismes distants, comme c'est le cas des stations KIZ, ECH ou GIMEL, une valeur de 2 augmente les chances de pointer l'arrivée des ondes S pour des détections à des distances épicentrales plus faibles.

La qualité des pointés P et S dépend alors de deux principaux facteurs :

- **les caractéristiques du bruit enregistré aux stations, indirectement définies par l'amplitude et le contenu fréquentiel du signal enregistré ainsi que le rapport signal/bruit ;**
- **la localisation de ces dernières relativement à la probabilité d'occurrence spatiale des séismes, autrement dit la distance épicentrale.**

Par conséquent, plusieurs configurations des paramètres critiques à l'amélioration des pointés des ondes P et S apparaissent souvent nécessaires pour une même station.

De cette façon, j'ai implémenté 2 instances de pointés automatiques, fonctionnant simultanément. Chaque instance s'adapte spécifiquement aux conditions de bruit et à la localisation particulières des stations.

Seulement, pour une même station, les signaux enregistrés sont d'une grande diversité et correspondent à des distances épicentrales variables d'un événement à l'autre. De plus, les niveaux de bruit enregistrés ne sont jamais constants et sont fortement dépendants de l'environnement de la station ainsi que sa qualité instrumentale intrinsèque.

Le nombre d'instances à définir a été déterminé à partir de l'évaluation de la performance du pointé automatique des premières arrivées des ondes P et S. Cette performance a été estimée en comparant les pointés automatiques des ondes P et S aux pointés manuels effectués pour les mêmes événements détectés.

4.1.3 Quelle performance pour ces pointés automatiques des ondes P et S ?

•**Comparaison aux pointés manuels**

La comparaison des temps d'arrivée des ondes P estimés par les pointés manuels et par les pointés automatiques pour la période juillet-octobre 2016 montre que 70 % des pointés automatiques P diffèrent des pointés manuels de seulement $\pm 0.5s$, dont 19% sont identiques (Figure 4.24).

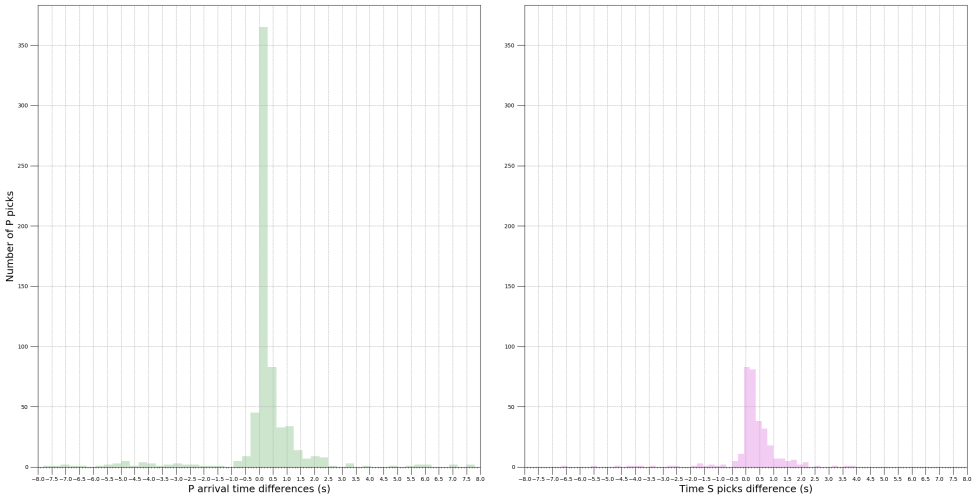


FIGURE 4.24: Distribution des temps d'arrivée différentiels entre les pointés manuels et automatiques pour des mêmes événements ayant été détectés pendant la période juillet-octobre 2016. Pointés automatiques P (à gauche) et pointés automatiques S (à droite).

De plus, 11% des pointés automatiques P ont des temps d'arrivée qui diffèrent de plus de $\pm 1.5s$ des pointés manuels, et 94 % d'entre eux diffèrent de moins de $\pm 3s$. Pour autant, on notera que ces pourcentages n'engagent pas de façon absolue la performance réelle du pointé automatique P. En effet, cette comparaison tient uniquement compte des pointés automatiques qui ont été sélectionnés pour créer les événements détectés.

Or, plusieurs pointés identifiés comme P peuvent être émis consécutivement à une même station, en fonction des conditions de bruit enregistré, mais un seul correspond à l'arrivée réelle des ondes P. Par conséquent, si, parmi le choix des pointés P émis à cette station, c'est finalement un pointé erroné qui est sélectionné dans le processus d'association, cela signifie que lorsque l'on compare ce pointé au pointé manuel équivalent, ce n'est pas la performance réelle de l'opération de pointé qui est évalué mais celle du processus d'association (Figure 4.25).

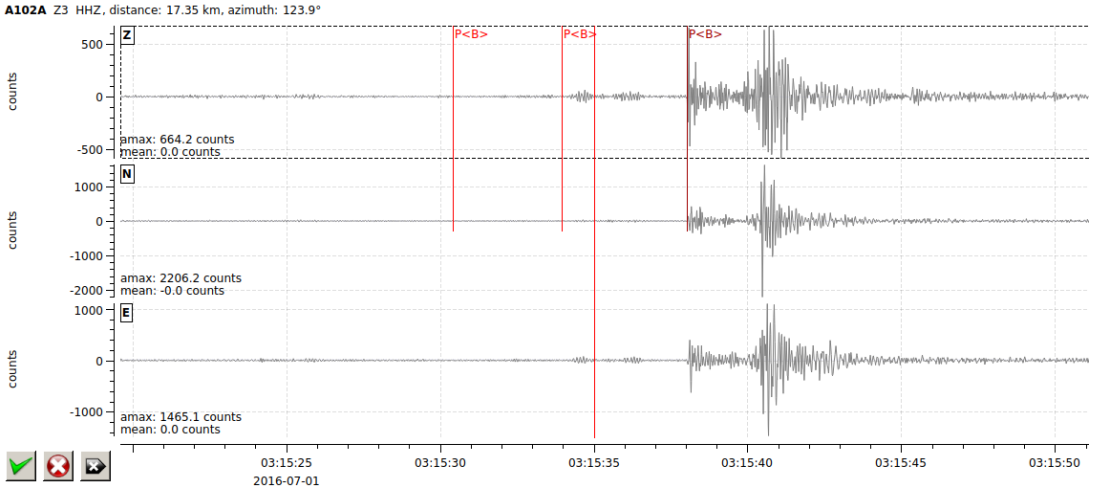


FIGURE 4.25: Exemple d'émission de plusieurs pointés automatiques P consécutifs à la station A102A. Signal correspondant à un séisme ayant eu lieu le 01 juillet 2016 à 03h15 dans la région d'Albstadt en Allemagne (MLv 1.4). Si le premier ou le deuxième pointé P (trait vertical rouge clair) était sélectionné dans le processus d'association, chacun aurait respectivement une différence de temps d'arrivée estimée des ondes P de -7.5 s et -4 s, alors que pourtant un vrai pointé P a été émis (troisième trait vertical rouge foncé).

De même, 65% des pointés automatiques S diffèrent des pointés manuels équivalents de seulement $\pm 0.5s$ et 97% diffèrent de moins de $\pm 3s$ (Figure 4.24). La même remarque peut être également établie concernant l'estimation de la performance réelle de l'opération de pointé des ondes S, au regard des multiples pointés S qui peuvent être aussi émis consécutivement à une même station. Cependant, étant donné que les pointés S sont émis une fois que les pointés P sont produits, la probabilité qu'un pointé S erroné soit sélectionné dans le processus d'association est supérieure si un pointé P sélectionné est lui-même erroné (Figure 4.26).

L'opération de pointé des ondes S est un processus plus délicat. En effet, la phase S arrive souvent dans la coda des ondes P et est parfois précédée par des phases sismiques converties (LOMAX, 2008). Une proportion moins élevée de pointés automatiques S, relativement aux pointés P, présente conséquemment des différentiels de temps d'arrivée plus petit que $\pm 0.5s$ par rapport aux pointés manuels. Ces pointés plus inexacts peuvent être le vecteur d'une plus grande incertitude dans la localisation future des événements détectés.

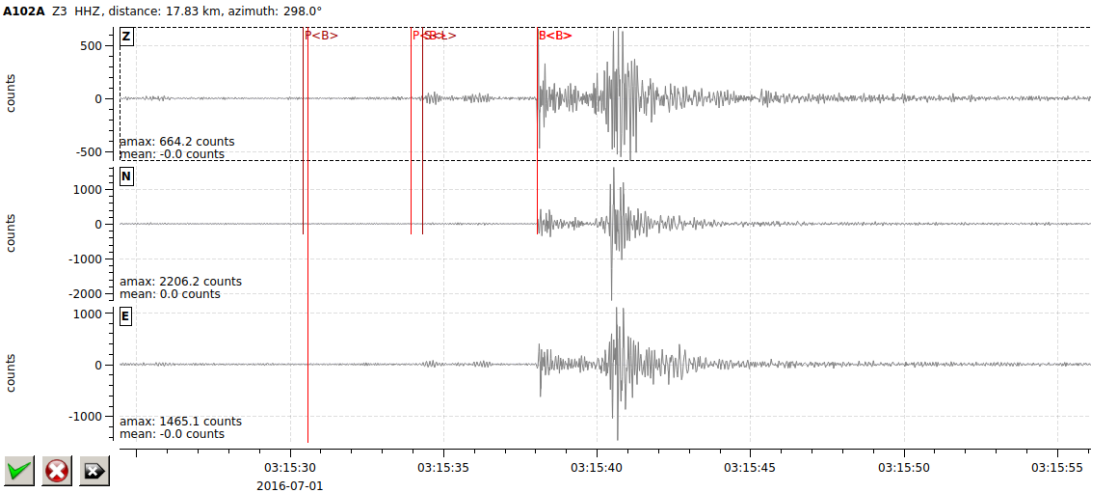


FIGURE 4.26: Exemple d'émission de plusieurs pointés automatiques P et S consécutifs à la station A102. Signal correspondant à un séisme ayant eu lieu le 01 juillet 2016 à 03h15 dans la région d'Albstadt en Allemagne (MLv 1.4). Les traits verticaux rouge foncé représentent les pointés automatiques erronés qui ont été sélectionnés pour cet exemple. Le pointé P est anticipé de 7 s et le pointé S résultant est anticipé de 5.5 s.

Ainsi, l'estimation de l'impact de l'incertitude des pointés des ondes S sur les localisations épi- et hypocentrale, confrontée aux incertitudes liées au modèle de vitesse, est un facteur important à considérer pour définir un niveau satisfaisant d'amélioration du pointé automatique.

•Impact sur les localisations épi- et hypocentrales

Le logiciel NonLinLoc (Lomax, Virieux et al., 2000) a été utilisé pour simuler l'impact des variations des temps d'arrivée des ondes S par rapport au temps d'arrivée référence, évalués manuellement, sur l'incertitude des localisations épicentrale et hypocentrale des événements au regard de 100 modèles de vitesses (50 modèles à 3 couches et 50 modèles à 12 couches). Ces modèles de vitesse ont été établis aléatoirement à partir d'une gamme de vitesses des ondes P comprises entre 3.5 km/s et 8.2 km/s et des rapports Vp/Vs compris entre 1.65 et 2.00.

La gamme de valeurs choisie pour les vitesses des ondes P correspondent aux gammes de vitesse qui peuvent être possiblement rencontrées dans les roches qui composent le milieu de propagation. L'intervalle de valeurs pour les rapports Vp/Vs a été défini à partir de la construction du diagramme de Wadati pour les événements détectés en 2016 par le BCSF-RéNaSS, et des résultats de l'étude réalisée par Rothe et al., 1950 sur les carrières souterraines d'Haslach en Allemagne pour élaborer le modèle de vitesse régional d'Haslach (Figure 4.27).

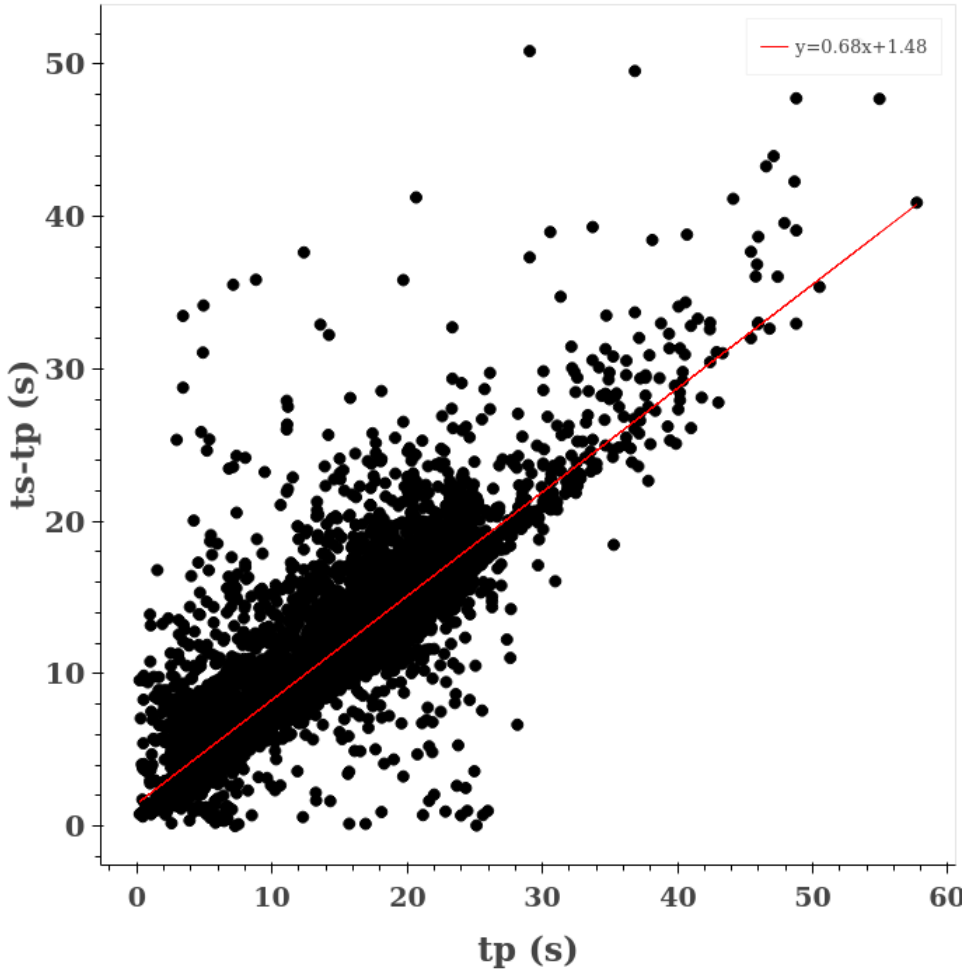


FIGURE 4.27: Diagramme de wadati réalisé à partir des temps d'arrivée des ondes S (ts) et des ondes P (tp) définis par les pointés manuels de l'ensemble des événements détectés en 2016 par le BCSF-RéNaSS. La valeur du rapport Vp/Vs est égale à 1.68

La procédure de localisation utilisée par le programme NonLinLoc détermine une fonction de densité de probabilités *a posteriori* sur toutes les solutions épi- et hypocentrales. En effet, cette localisation quantifie l'accord entre les temps d'arrivée observés et prédits en relation à toutes les incertitudes considérées (pointés, calcul des temps de trajet, géométrie du réseau) et forme une solution complète probabiliste qui représente la distribution de toutes les localisations possibles.

Cette fonction de densité de probabilité est calculée ici à partir de l'algorithme Oct-Tree (Lomax et Curtis, 2001). Cet algorithme utilise une subdivision récursive et un échantillonnage de cellules dans un espace 3-D pour générer une cascade de cellules échantillonnées, où la densité des cellules échantillonnées suit les valeurs de la fonction de densité de probabilités du centre de la cellule (Husen, Kissling et al., 2003). La valeur maximale de cette fonction est prise comme hypocentre préférentiel avec le maximum de vraisemblance (Figure 4.28).

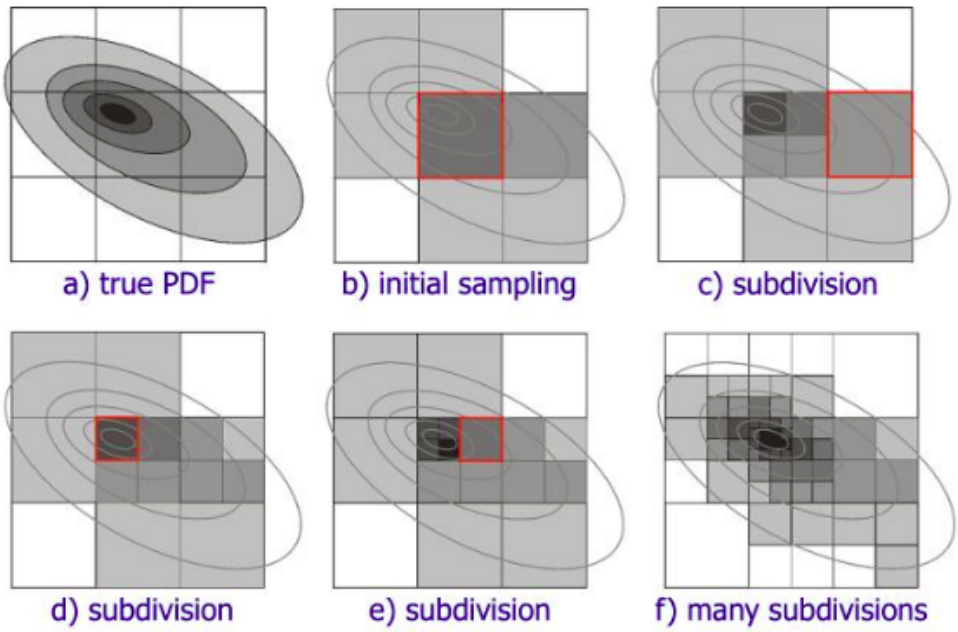


FIGURE 4.28: Procédure d'échantillonnage de l'algorithme Oct-Tree pour obtenir la fonction de densité de probabilités complète (a). Cette procédure est initialisée par un échantillonnage global de l'espace de recherche sur une grille grossière et régulière (b). La probabilité est calculée pour chaque cellule puis celle-ci est insérée dans la liste des probabilités à la position correspondant à la valeur de sa probabilité. La cellule avec la probabilité la plus grande (Pmax, carré rouge) est obtenue de la liste ordonnée des probabilités (b). Cette cellule est alors divisée en 8 cellules filles (c). La probabilité est calculée pour chacune des 8 cellules filles. Les 8 cellules filles sont insérées dans le liste ordonnée des probabilités selon la valeur de leur probabilité et ainsi de suite jusqu'à obtenir la fonction complète de densité de probabilités (d-f). L'ensemble de la figure représente les projections 2-D des échantillons 3-D. D'après LOMAX et CURTIS, 2001.

En plus des incertitudes de localisation incluses dans la solution probabiliste, le programme NonLinLoc produit des estimations traditionnelles gaussiennes telles que la localisation hypocentrale attendue et l'ellipsoïde de confiance à 68% (Lomax, Virieux et al., 2000). Cette ellipsoïde représente une approximation statistique gaussienne de la fonction de densité de probabilités, tronquée au niveau de confiance de 68%. Ceci signifie que si la fonction de densité de probabilités était parfaitement ellipsoïdale, alors il y aurait une probabilité de 68% que l'hypocentre soit à l'intérieur de cette ellipsoïde.

L'hypocentre attendu et l'ellipsoïde de confiance peuvent être interprétés comme des résultats obtenus par des algorithmes de localisation tels que HYPO-71 (W. H. K. Lee et al., 1972) ou HYPOELLIPSE (Lahr, 1989). Cependant, les incertitudes véhiculées par les estimations gaussiennes sont significatives uniquement lorsque la fonction de densité de probabilités exprime un minimum clair, unique et global (Husen, Kissling et al., 2003).

L'effet des variations des temps d'arrivée des ondes S sur les localisations épicentrales et hypocentrales dépend d'abord fortement du modèle de vitesse utilisé pour localiser (nombre de couches et vitesses de propagation des ondes). Seulement, quelques généralités transparaissent. De manière globale, en prenant comme référence les localisations émises à partir des temps d'arrivées des ondes S estimés manuellement, les solutions épicentrales et hypocentrales vont fortement se dégrader à partir de retards de temps d'arrivée des ondes S moyens, par rapport aux temps d'arrivée références, supérieurs à +1 s pour tous les modèles de vitesse, voire supérieurs à +2 s, pour 36% des modèles de vitesse multicouches (Figures 4.29 et 4.30). Cette dégradation des solutions épicentrales et hypocentrales se manifeste par des incertitudes plus grandes : un étalement spatial plus large des fonctions de densité de probabilités, un net allongement des ellipsoïdes de confiance et un plus fort éloignement des deux hypocentres (gaussien et maximum de vraisemblance).

Seulement, cette dégradation des solutions a tendance à s'atténuer avec l'augmentation du nombre de phases impliquées dans la localisation, pour la majorité des modèles de vitesse sélectionnés (Figures 4.31 et 4.32).

En ce qui concerne les variations négatives des temps d'arrivée des ondes S, c'est-à-dire des temps d'arrivée estimés en moyenne de façon anticipée par rapport aux temps d'arrivée définis manuellement, le même constat peut être fait. Si les temps d'arrivées des ondes S sont émis jusqu'à -1 s, voire jusqu'à -2 s pour 26% des modèles de vitesse multicouches, les solutions épicentrales et hypocentrales vont rester comparables aux solutions de référence, puis se dégrader au-delà de -1 et -2 s (Figures 4.33 et 4.34). De même, la dégradation des solutions épi- et hypocentrales a tendance à s'atténuer avec l'augmentation du nombre de phases (Figures 4.33 et 4.34).

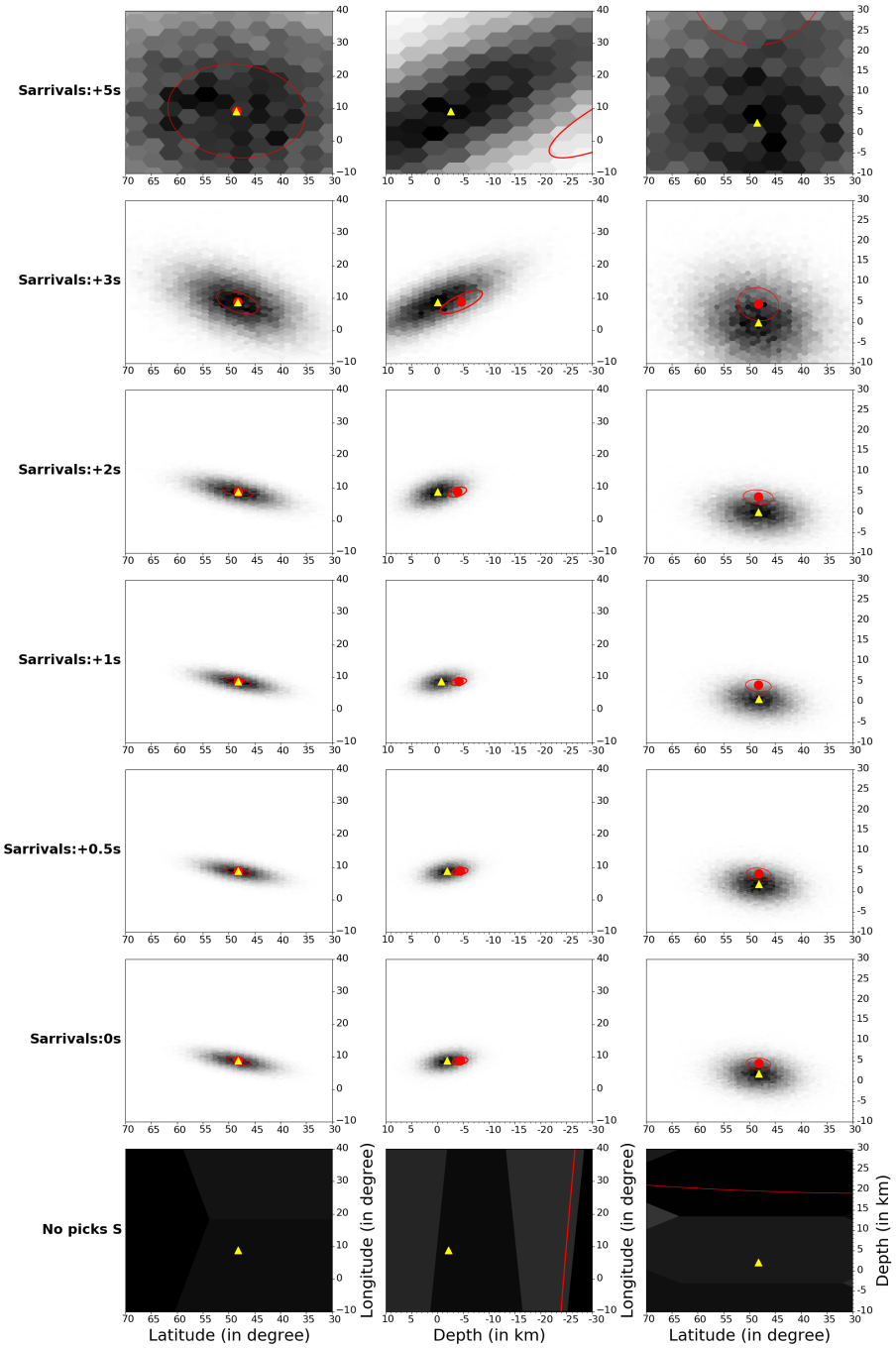


FIGURE 4.29: Solutions épi- et hypocentrales pour un tir de la carrière de Dotternhausen (MLv 1.7, 15/07/2016 10h25) en fonction des variations positives moyennes (de +0.5 s à +5s) des temps d'arrivée des ondes S relativement aux temps de référence estimés manuellement (= 0s). Solution épicentrale (à gauche) et solution hypocentrale en fonction de la longitude (au milieu) et la latitude (à droite). Le point rouge correspond à l'hypocentre gaussien et l'ellipsoïde rouge l'ellipsoïde de confiance à 68%. La fonction de densité de probabilités est représentée avec une palette de niveaux de gris et son hypocentre optimal de maximum de vraisemblance est défini par un triangle jaune. **Les localisations sont émises avec 23 phases (11 phases S) et un modèle de vitesse multicouche** (cf Annexe D.1).

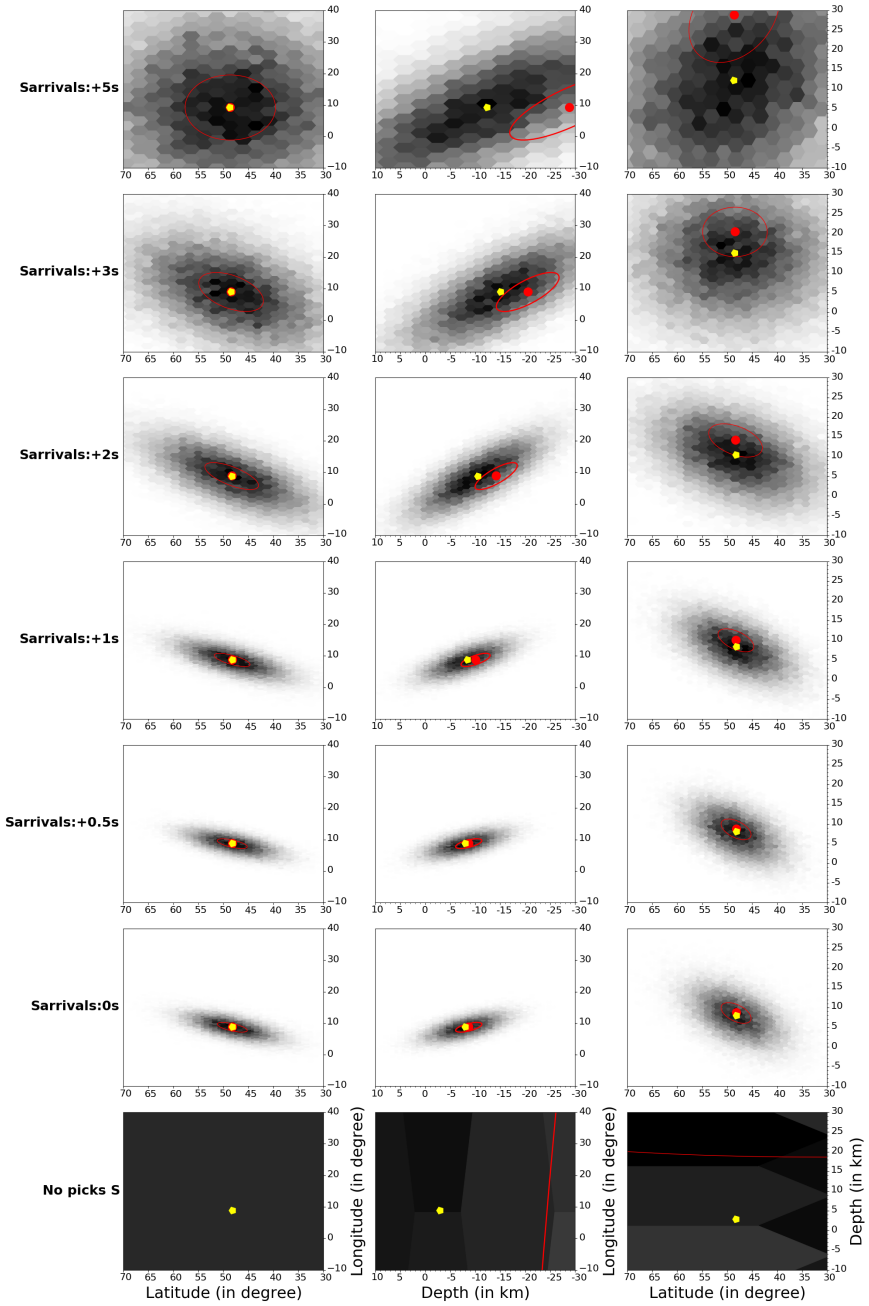


FIGURE 4.30: Solutions épi- et hypocentrales pour un tir de la carrière de Dotternhausen émis le 15 juillet 2016 à 10h25 (MLv 1.7) en fonction des variations positives moyennes (de +0.5 s à +5 s) des temps d'arrivée des ondes S relativement aux temps de référence estimés manuellement (= 0 s). Solution épicentrale (en haut) et solution hypocentrale en fonction de la longitude (au milieu) et la latitude (en bas). Le point rouge correspond à l'hypocentre gaussien et l'ellipsoïde rouge l'ellipsoïde de confiance à 68%. La fonction de densité de probabilités est représentée avec une palette de niveaux de gris et son hypocentre optimal de maximum de vraisemblance est définie par un hexagone jaune. **Les localisations sont émises avec 23 phases (11 phases S) et un modèle de vitesse à 3 couches** (cf Annexe D.2).

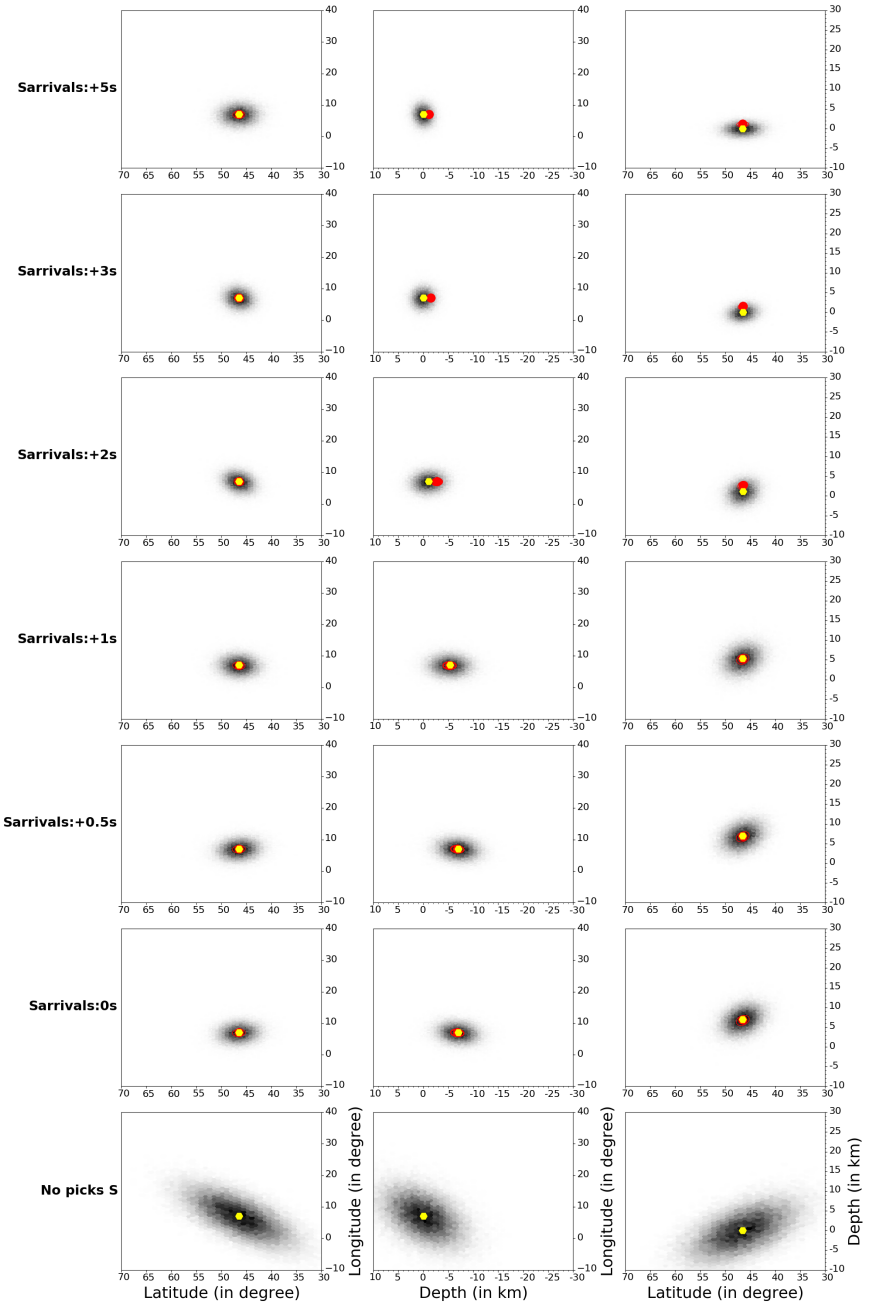


FIGURE 4.31: Solutions épi- et hypocentrales pour un séisme ayant eu lieu le 16 juillet 2016 à 02h36 dans les Pré-alpes Suisses (MLv 2.7) en fonction des variations positives moyennes (de +0.5 s à +5 s) des temps d'arrivée des ondes S relativement aux temps de référence estimés manuellement (= 0 s). Solution épicentrale (en haut) et solution hypocentrale en fonction de la longitude (au milieu) et la latitude (en bas). Le point rouge correspond à l'hypocentre gaussien et l'ellipsoïde rouge l'ellipsoïde de confiance à 68%. La fonction de densité de probabilités est représentée avec une palette de niveaux de gris et son hypocentre optimal de maximum de vraisemblance est définie par un cercle jaune. **Les localisations sont émises avec 52 phases (18 phases S) et un modèle de vitesse multicouche (cf Annexe D.3).**

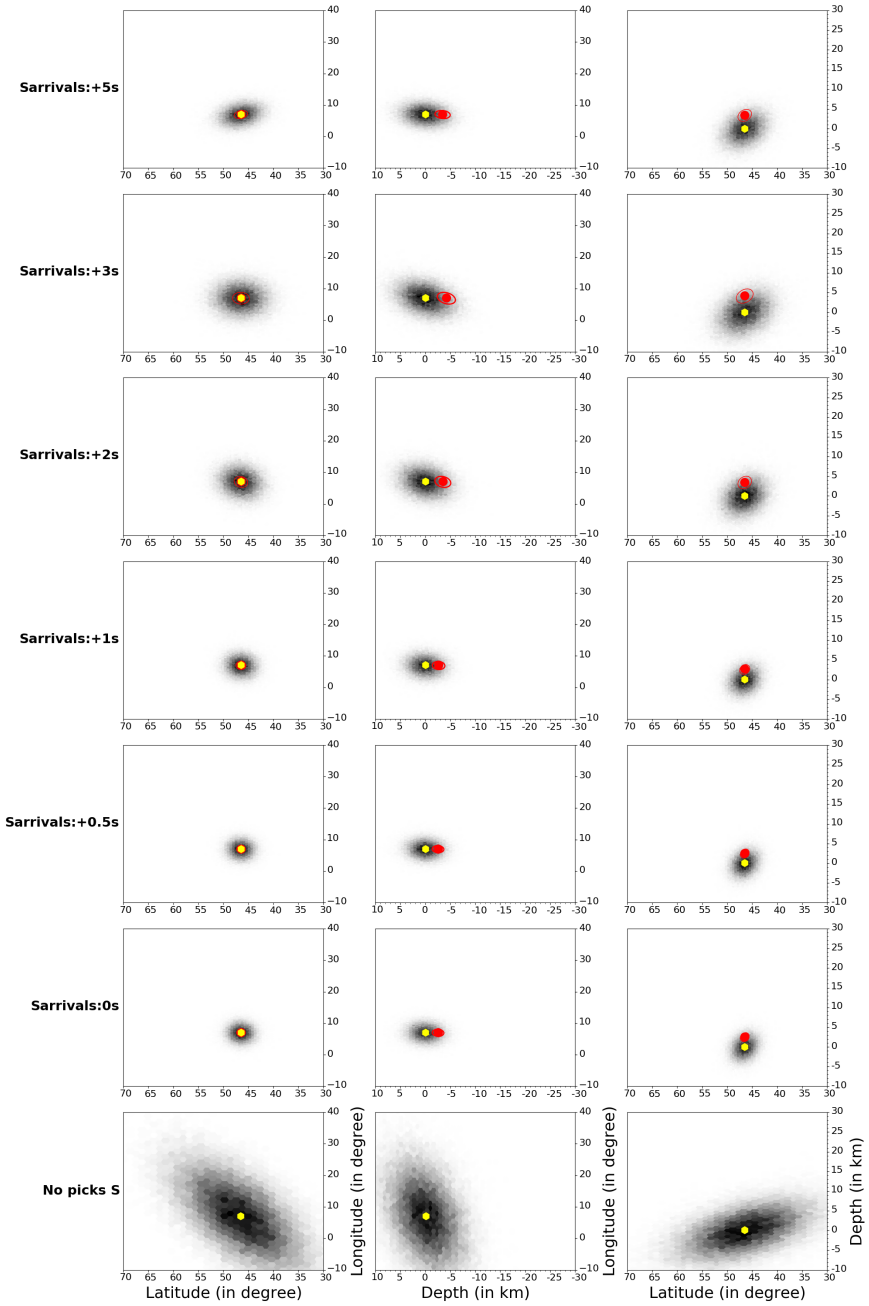


FIGURE 4.32: Solutions épi- et hypocentrales pour un séisme ayant eu lieu le 16 juillet 2016 à 02h36 dans les Pré-alpes Suisses (MLv 2.7) en fonction des variations positives moyennes (de +0.5 s à +5 s) des temps d'arrivée des ondes S relativement aux temps de référence estimés manuellement (= 0 s). Solution épicentrale (en haut) et solution hypocentrale en fonction de la longitude (au milieu) et la latitude (en bas). Le point rouge correspond à l'hypocentre gaussien et l'ellipsoïde rouge l'ellipsoïde de confiance à 68%. La fonction de densité de probabilités est représentée avec une palette de niveaux de gris et son hypocentre optimal de maximum de vraisemblance est définie par un cercle jaune. **Les localisations sont émises avec 52 phases (18 phases S) et un modèle de vitesse à 3 couches (cf Annexe D.5).**

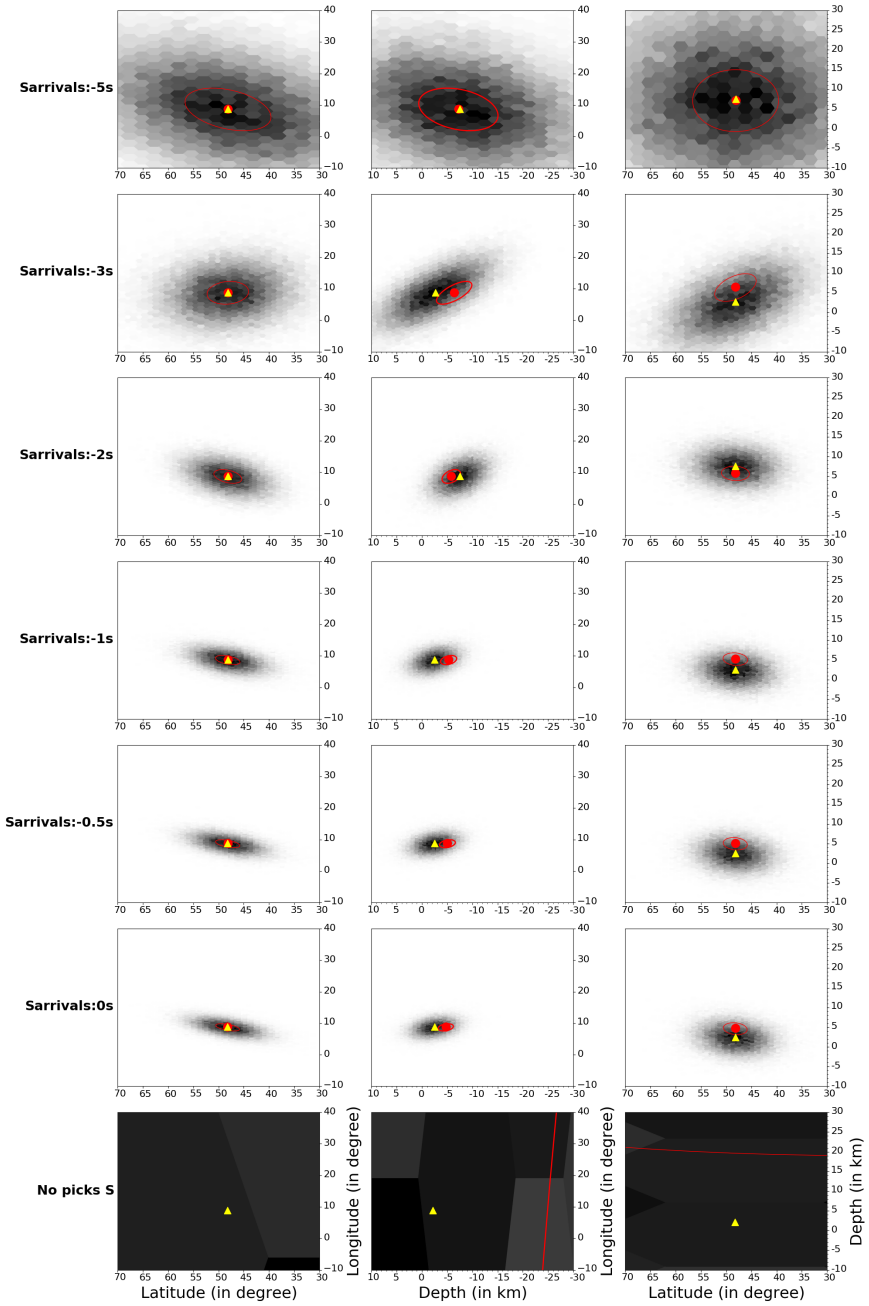


FIGURE 4.33: Solutions épi- et hypocentrales pour un tir de la carrière de Dotternhausen émis le 15 juillet 2016 à 10h25 (MLv 1.7) en fonction des variations négatives moyennes (de -0.5 s à -5 s) des temps d'arrivée des ondes S relativement aux temps de référence estimés manuellement (= 0 s). Solution épicentrale (en haut) et solution hypocentrale en fonction de la longitude (au milieu) et la latitude (en bas). Le point rouge correspond à l'hypocentre gaussien et l'ellipsoïde rouge l'ellipsoïde de confiance à 68%. La fonction de densité de probabilités est représentée avec une palette de niveaux de gris et son hypocentre optimal de maximum de vraisemblance est définie par un triangle jaune. **Les localisations sont émises avec 23 phases (11 phases S) et un modèle de vitesse multicouche (cf Annexe D.1).**

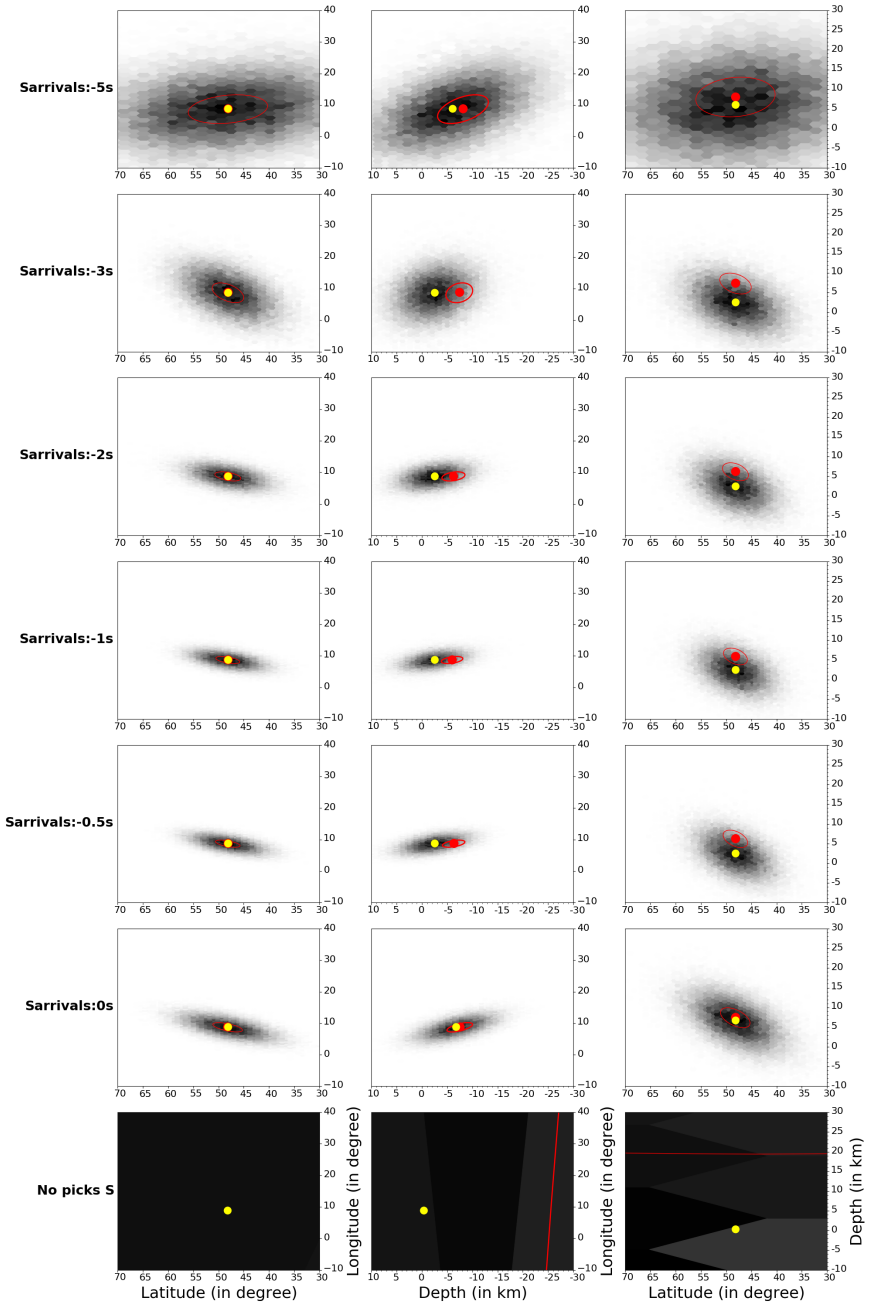


FIGURE 4.34: Solutions épi- et hypocentrales pour un tir de la carrière de Dotternhausen émis le 15 juillet 2016 à 10h25 (MLv 1.7) en fonction des variations négatives moyennes (de -0.5 s à -5 s) des temps d'arrivée des ondes S relativement aux temps de référence estimés manuellement (= 0 s). Solution épicentrale (en haut) et solution hypocentrale en fonction de la longitude (au milieu) et la latitude (en bas). Le point rouge correspond à l'hypocentre gaussien et l'ellipsoïde rouge l'ellipsoïde de confiance à 68%. La fonction de densité de probabilités est représentée avec une palette de niveaux de gris et son hypocentre optimal de maximum de vraisemblance est définie par un cercle jaune. **Les localisations sont émises avec 23 phases (11 phases S) et un modèle de vitesse à 3 couches (cf Annexe D.6).**

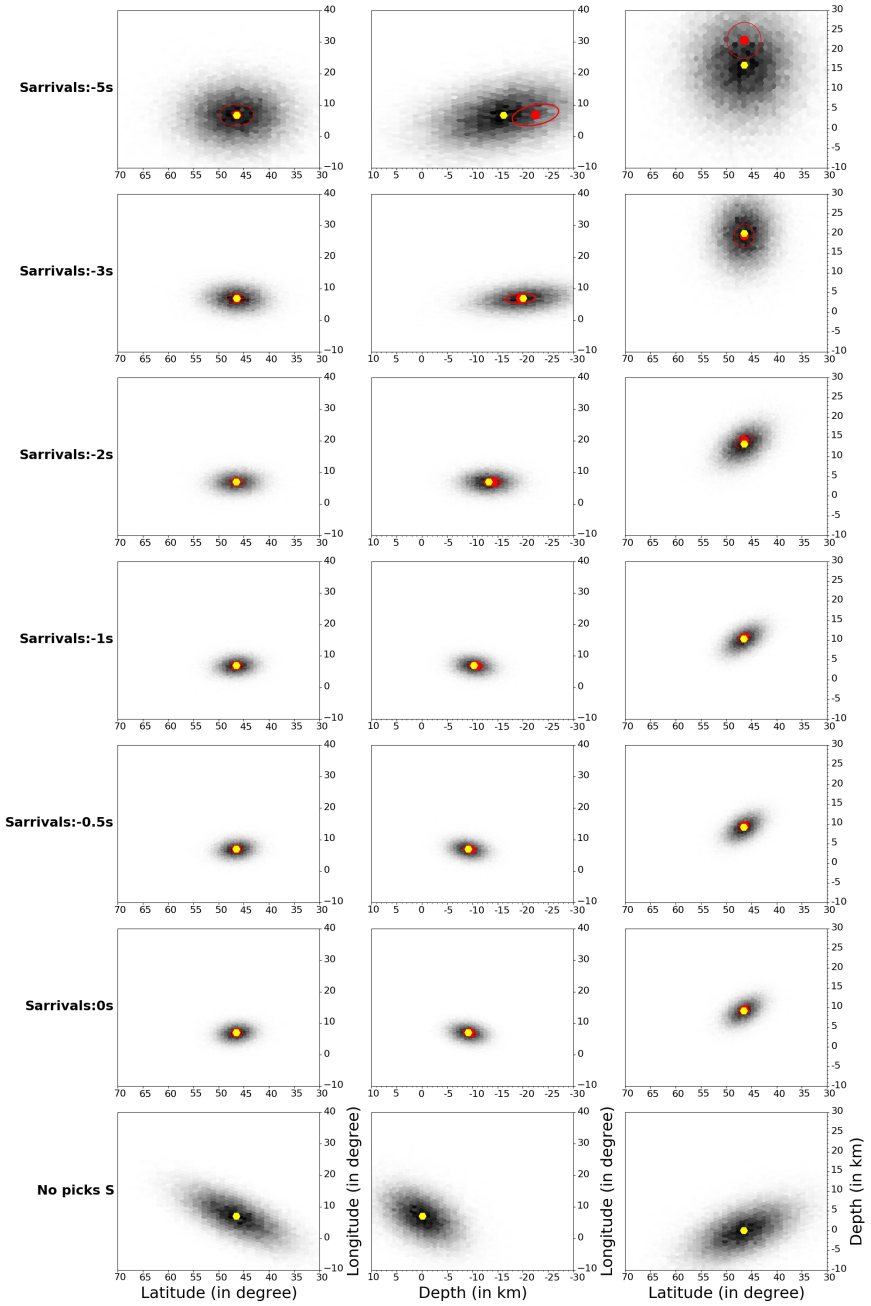


FIGURE 4.35: Solutions épi- et hypocentrales pour un séisme ayant eu lieu le 16 juillet 2016 à 02h36 dans les Pré-alpes Suisses (MLv 2.7) en fonction des variations négatives moyennes (de -0.5 s à -5 s) des temps d'arrivée des ondes S relativement aux temps de référence estimés manuellement (= 0 s). Solution épicentrale (en haut) et solution hypocentrale en fonction de la longitude (au milieu) et la latitude (en bas). Le point rouge correspond à l'hypocentre gaussien et l'ellipsoïde rouge l'ellipsoïde de confiance à 68%. La fonction de densité de probabilités est représentée avec une palette de niveaux de gris et son hypocentre optimal de maximum de vraisemblance est définie par un cercle jaune. **Les localisations sont émises avec 52 phases (18 phases S) et un modèle de vitesse multicouche (cf Annexe D.3).**

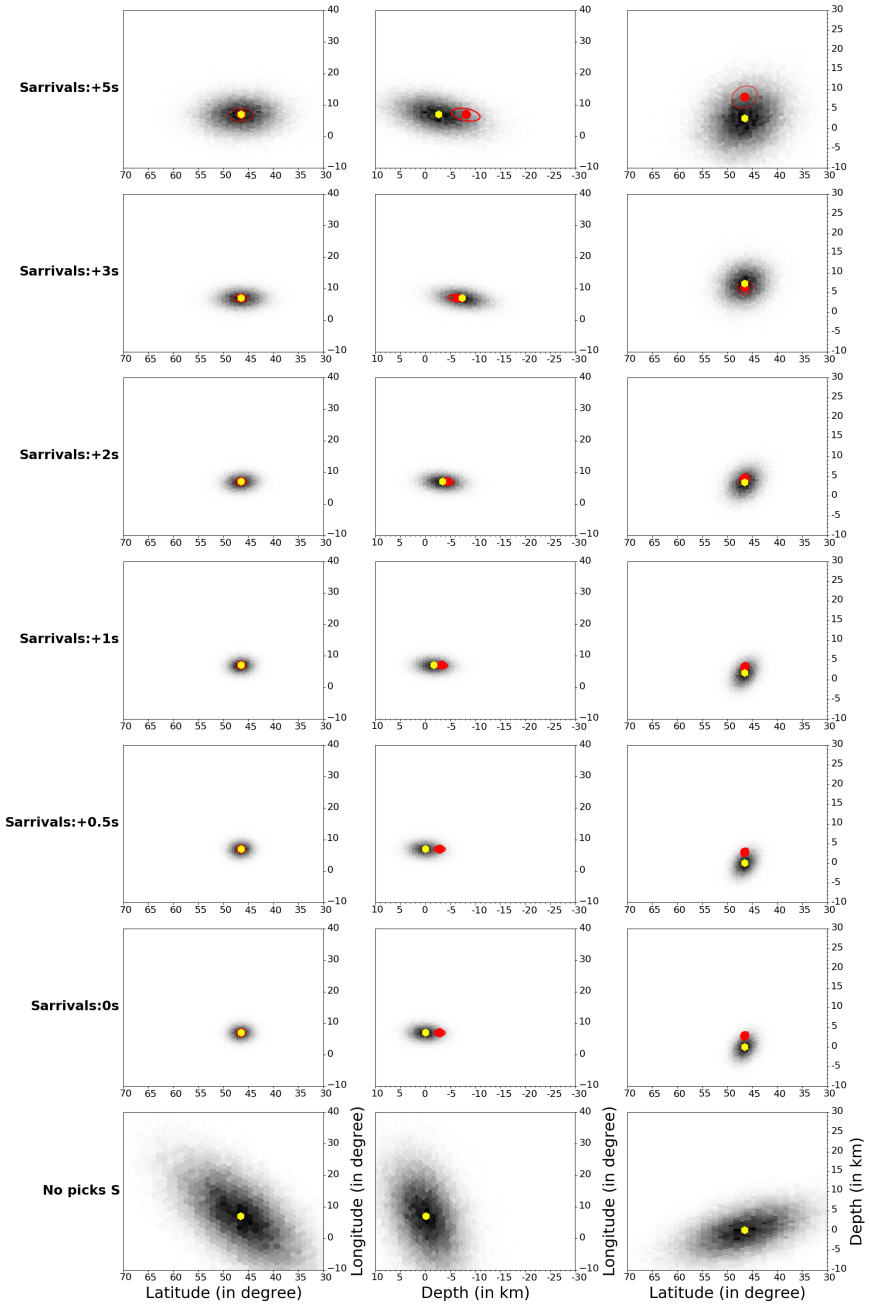


FIGURE 4.36: Solutions épi- et hypocentrales pour un séisme ayant eu lieu le 16 juillet 2016 à 02h36 dans les Pré-alpes Suisses (MLv 2.7) en fonction des variations négatives moyennes (de -0.5 s à -5 s) des temps d'arrivée des ondes S relativement aux temps de référence estimés manuellement (= 0 s). Solution épicentrale (en haut) et solution hypocentrale en fonction de la longitude (au milieu) et la latitude (en bas). Le point rouge correspond à l'hypocentre gaussien et l'ellipsoïde rouge l'ellipsoïde de confiance à 68%. La fonction de densité de probabilités est représentée avec une palette de niveaux de gris et son hypocentre optimal de maximum de vraisemblance est définie par un hexagone jaune. **Les localisations sont émises avec 52 phases (18 phases S) et un modèle de vitesse à 3 couches (cf Annexe D.5).**

Plus spécifiquement, si les localisations épicentrales et hypocentrales sont maintenant analysées par modèle de vitesse, celles-ci dépendent fortement des temps d'arrivée moyens des ondes S émis par rapport aux temps de référence. En effet, lorsque les modèles de vitesse multicouches présentent des vitesses assez élevées pour les ondes S dans les premières couches (environ 2.80 km/s en moyenne jusqu'à 5 km), les localisations hypocentrales seront plus approximatives si les pointés S émis sont anticipés. Si je prends l'exemple du tir de la carrière de Dotternhausen produit le 15 juillet 2016 à 10h25 (MLv 1.7) et localisé avec un de ces modèles de vitesse, il est possible de remarquer que l'hypocentre optimal de maximum de vraisemblance change soudainement de position et passe à une profondeur de l'ordre de 15 km lorsque les pointés des ondes S sont émis en avance (de -0.5 s à -5 s). L'élongation de la fonction de densité de probabilités ainsi que la séparation nette des deux hypocentres (gaussien et maximum de vraisemblance) de l'ordre de 5 km mettent en évidence une large incertitude hypocentrale, en partie biaisée par les temps d'arrivées des ondes S. (Figure 4.37).

En revanche, cette observation est moins marquée lorsque les variations des temps d'arrivée des ondes S sont positives, c'est à dire que les pointés S sont retardés. En effet, jusqu'à +2 s, la position de l'hypocentre optimal de maximum de vraisemblance est stable et située à environ 2 km. De plus, la fonction de densité de probabilités garde la même allure. En revanche, l'hypocentre gaussien de référence (c'est-à-dire celui émis à partir des temps d'arrivées des ondes S de référence) est placé initialement à 7 km et avec une ellipsoïde de confiance de l'ordre de 7 km de longueur. Au fur et à mesure que les temps d'arrivées estimées des ondes S sont retardés, cet hypocentre gaussien tend à se rapprocher de l'hypocentre optimal et l'aire de l'ellipsoïde de confiance diminue. L'hypocentre vrai étant situé à 0 km (tir de carrière), cet effet traduit alors nettement des incertitudes supplémentaires liées au modèle de vitesse, qui semble surestimer les vitesses des ondes S (Figure 4.37).

Avec un modèle de vitesse qui présente des vitesses des ondes S moins élevées (environ 2.72 km/s en moyenne pour les premières couches jusqu'à 5 km), les solutions hypocentrales restent beaucoup plus stables pour tous les cas de pointés S émis entre -1 et + 1 s autour du temps d'arrivée de référence (Figures 4.29 et 4.33). Les incertitudes de ces localisations hypocentrales sont d'ailleurs plus faibles comme en témoignent la contraction des fonctions de densité de probabilité, les petites surfaces des ellipsoïdes de confiance (longueur de 3 km) et le rapprochement des deux hypocentres calculés (gaussien et maximum de vraisemblance). De plus, l'hypocentre optimal de vraisemblance reste à une profondeur stable d'environ 1-2 km.

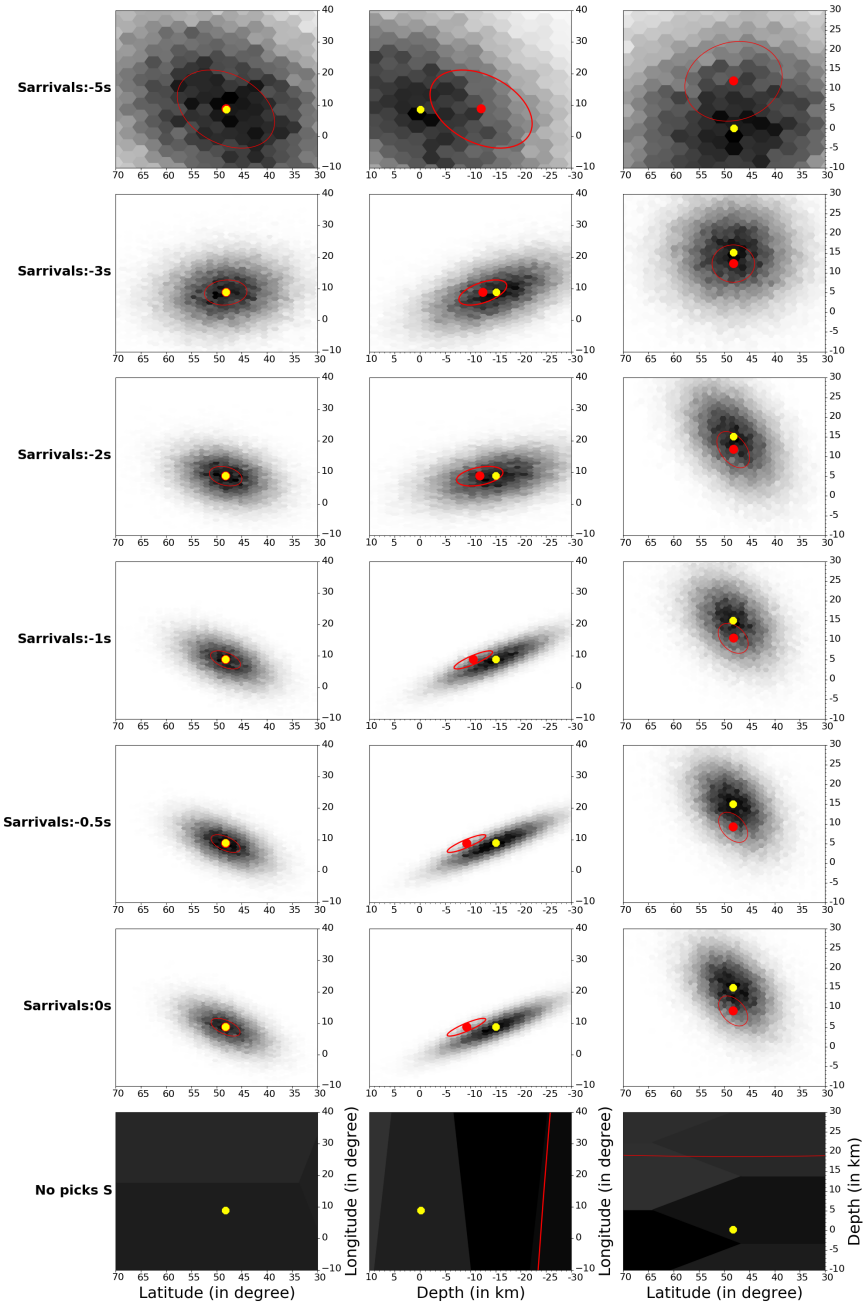


FIGURE 4.37: Solutions épi- et hypocentrales pour un tir de la carrière de Dotternhausen en Allemagne émis le 15 juillet 2016 à 10h25 (MLv 1.7) en fonction des variations négatives moyennes (de -0.5 s à -5 s) des temps d'arrivée des ondes S relativement aux temps de référence estimés manuellement (= 0 s). Solution épicentrale (en haut) et solution hypocentrale en fonction de la longitude (au milieu) et la latitude (en bas). Le point rouge correspond à l'hypocentre gaussien et l'ellipsoïde rouge l'ellipsoïde de confiance à 68%. La fonction de densité de probabilités est représentée avec une palette de niveaux de gris et son hypocentre optimal de maximum de vraisemblance est définie par un hexagone jaune. **Les localisations sont émises avec 23 phases (11 phases S) et un modèle de vitesse multicouche (cf Annexe D.7).**

Les mêmes observations peuvent être également effectuées pour les modèles de vitesse plus simples à 3 couches, à l'exception que l'incertitude liée aux temps d'arrivée des ondes S a plus d'implication dans l'estimation de l'incertitude des solutions épicentrales et hypocentrales. Par exemple, pour le même tir de la carrière de Dotternhausen du 15 juillet 2016 et pour des modèles de vitesse avec des vitesses moyennes des ondes S dans la première couche d'environ 2.60 km/s, l'étalement spatial des fonctions de densité de probabilités, l'allongement des ellipsoïdes de confiance et l'écartement entre les deux hypocentres estimés (gaussien et maximum de vraisemblance) augmentent avec le retard des temps d'arrivée des ondes S (par rapport aux temps de référence), même si l'hypocentre optimal de maximum de vraisemblance reste fixé à la même profondeur. Cette incertitude devient très grande dès +2 s de retard (Figure 4.38). Néanmoins, cet effet se manifeste beaucoup moins pour des temps d'arrivée des ondes S qui sont en moyenne pointés en avance par rapport au pointé manuel de référence, soulignant alors là encore l'impact fort du modèle de vitesse sur les incertitudes calculées (Figure 4.39).

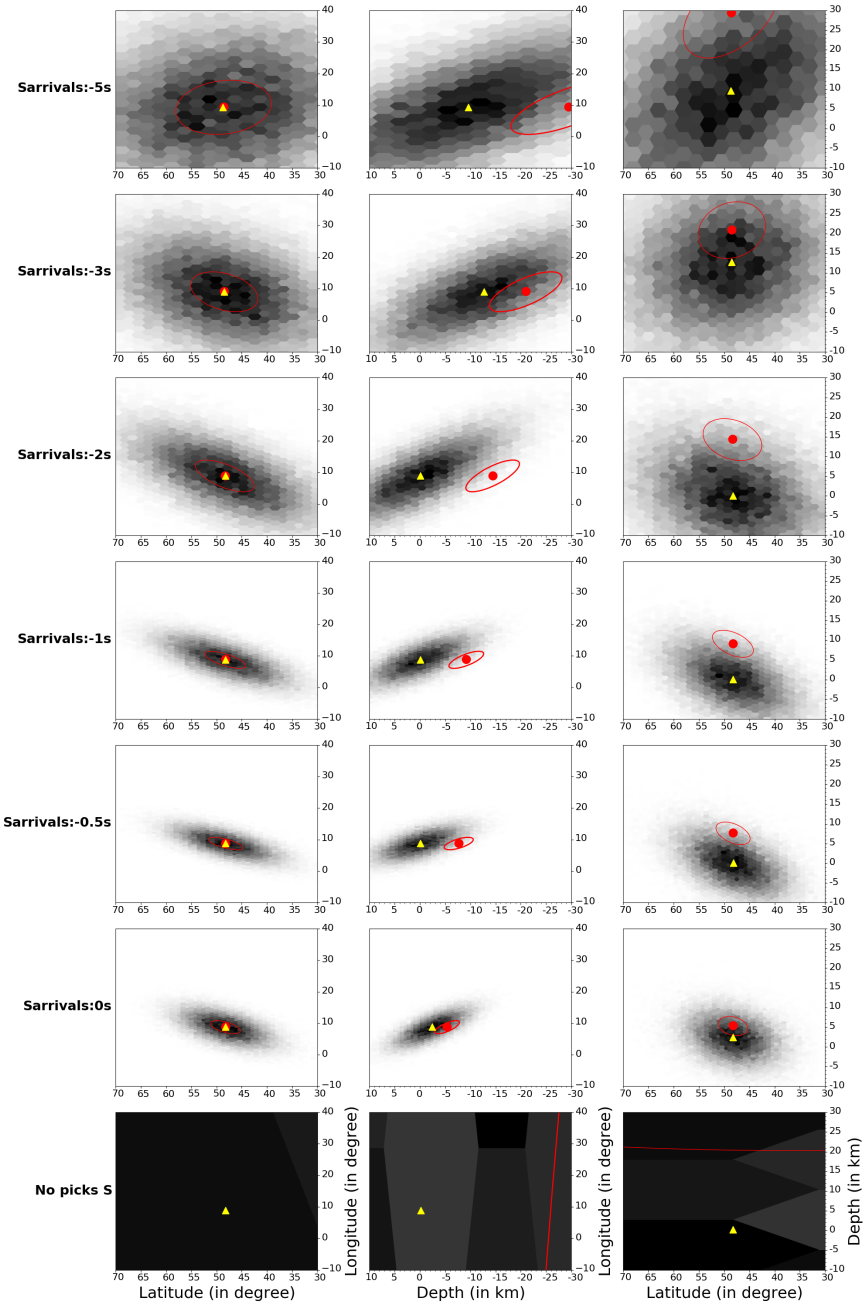


FIGURE 4.38: Solutions épi- et hypocentrales pour un tir de la carrière de Dotternhausen ayant eu lieu le 15 juillet 2016 à 10h25 en Allemagne (MLv 1.7) en fonction des variations positives moyennes (de +0.5 s à +5 s) des temps d'arrivée des ondes S relativement aux temps de référence estimés manuellement (= 0 s). Solution épicentrale (en haut) et solution hypocentrale en fonction de la longitude (au milieu) et la latitude (en bas). Le point rouge correspond à l'hypocentre gaussien et l'ellipsoïde rouge l'ellipsoïde de confiance à 68%. La fonction de densité de probabilités est représentée avec une palette de niveaux de gris et son hypocentre optimal de maximum de vraisemblance est définie par un triangle jaune. **Les localisations sont calculées à partir de 23 phases (11 phases S) et un modèle de vitesse à 3 couches (cf Annexe D.8).**

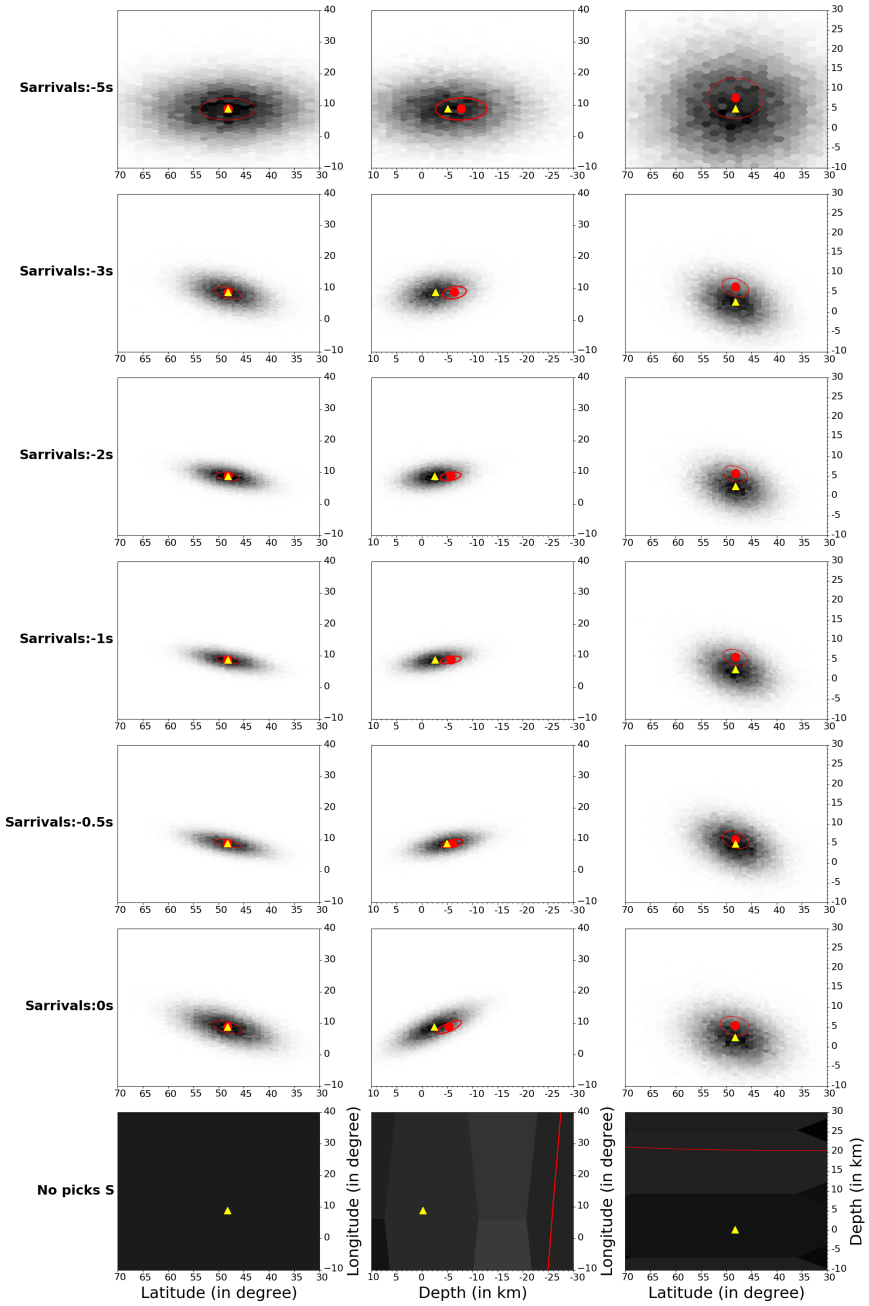


FIGURE 4.39: Solutions épi- et hypocentrales pour un tir de la carrière de Dotternhausen ayant eu lieu le 15 juillet 2016 à 10h25 en Allemagne (MLv 1.7) en fonction des variations négatives moyennes (de -0.5 s à -5 s) des temps d'arrivée des ondes S relativement aux temps de référence estimés manuellement (= 0 s). Solution épicentrale (en haut) et solution hypocentrale en fonction de la longitude (au milieu) et la latitude (en bas). Le point rouge correspond à l'hypocentre gaussien et l'ellipsoïde rouge l'ellipsoïde de confiance à 68%. La fonction de densité de probabilités est représentée avec une palette de niveaux de gris et son hypocentre optimal de maximum de vraisemblance est définie par un triangle jaune. Les localisations sont calculées à partir de 23 phases (11 phases S) et un modèle de vitesse à 3 couches (cf Annexe D.8).

De ce fait, lorsque les variations des temps d'arrivée des ondes S pour un même événement sont comprises entre -1 et 1 s par rapport au temps de référence, les incertitudes hypocentrales évaluées sont davantage dominées par les incertitudes liées à la structure du modèle de vitesse que par les variations des temps d'arrivée des ondes S. En revanche, si ces pointés S ont des marges d'erreur trop grandes, c'est-à-dire retardés de plus de 2 s et ou anticipés de plus de 2 s, les fonctions de densité de probabilités ont de plus larges distributions et les ellipsoïdes d'erreur ont des surfaces très grandes, soulignant une très grande incertitude des localisations hypocentrales et un plus grand impact des erreurs des pointés. Par ailleurs, l'utilisation de modèles de vitesse très simples (ici à 3 couches) met davantage en lumière les erreurs liées aux pointés des temps d'arrivée des ondes S, quelque soit leur ampleur.

Concernant les localisations épicentrales, celles-ci apparaissent beaucoup plus facilement contraintes que les localisations hypocentrales. En effet, quel que soient le modèle de vitesse utilisé et l'incertitude des pointés S, les deux épicentres (gaussien et maximum de vraisemblance) se chevauchent et ont des positions très stables. En revanche, cette position gagne rapidement en incertitude lorsque les pointés S sont émis avec une marge d'erreur très importante (inférieure à -3 s et supérieure à + 3 s), comme l'expriment l'élargissement de la fonction de densité de probabilités et de l'ellipsoïde de confiance.

Un total de 80% des événements détectés automatiquement pour la période juillet-octobre 2016 présente des temps d'arrivée moyens des ondes S compris entre -1.5 et + 1.5 s par rapport aux temps d'arrivée de référence estimés manuellement. Par exemple, pour le tir de la carrière de Dotternhausen effectué le 15 juillet 2016 à 10h25, la moyenne est évaluée à +1.15 s et pour le séisme des Pré-Alpes Suisses ayant du 16 juillet à 02h36, il est +0.77 s. Les incertitudes de localisation pour la majorité des événements vont donc être très sensibles aux incertitudes liées au modèle de vitesse, voire du nombre de phases et du type de phases disponibles (Tarantola et al., **1982**).

De plus, si des variations importantes des temps d'arrivée des ondes S entre le pointé automatique et le pointé manuel peuvent exister, celles-ci ne reflètent pas forcément la qualité du pointé automatique lui-même, mais peut mettre en évidence une défaillance du processus d'association. En effet, la structure du modèle de vitesse utilisé (par exemple des vitesses des ondes sismiques trop lentes) jouant sur la contrainte des solutions épicentrales et hypocentrales calculées, il est possible d'anticiper que, si plusieurs pointés ont été émis consécutivement, comme dans l'exemple de la Figure **4.25**, un faux pointé anticipé de quelques secondes pourrait être sélectionné au détriment du vrai pointé parce que son temps d'arrivée pourrait expliquer de façon plus cohérente une structure de vitesse avec des vitesse plus ralenties.

Enfin, l'ensemble de cette étude sur l'amélioration de la qualité des pointés automatiques P et S met en évidence clairement les critères principaux qui vont conditionner la qualité de ces pointés automatiques. L'analyse empirique des valeurs optimales obtenues pour chaque paramètre SeisComP3 configuré offre également une solide base d'étude pour le paramétrage futur des algorithmes de pointé automatique des phases sismiques P et S dans d'autres zones d'étude. De plus, le paramétrage manuel intense effectué dans ce travail de thèse a permis de mettre en lumière les deux facteurs fondamentaux qui contrôlent la valeur des différents paramètres SeisComP3 configurés, à savoir les distances épicentrales et les caractéristiques du bruit enregistré. De ce fait, ces deux facteurs fondamentaux vont être à considérer pour une mise en place future d'un paramétrage automatique dynamique des paramètres SeisComP3 exposés dans ce travail. Ce paramétrage dynamique aura la possibilité de s'ajuster automatiquement à la localisation statistique des stations par rapport aux événements détectés et aux caractéristiques du bruit enregistrés aux stations, en utilisant par exemple un apprentissage continu par renforcement comme dans l'étude établie par Draelos et al., **2018**.

4.2 Améliorer le processus d'association

4.2.1 Comment fonctionne le processus d'association dans le système de détection ?

•Un assemblage de pointés basé sur une recherche sur grille

Le premier algorithme d'association qui est utilisé se base sur la recherche de l'hypocentre optimal à partir d'une grille qui propose toutes les localisations et temps d'origine possibles. Cette grille constitue donc un jeu de points arbitraires qui échantillonne densément la zone d'intérêt (la zone d'étude). Chaque point de la grille correspond alors à un hypocentre hypothétique pour tous les pointés P à associer qui arrivent. Chaque pointé est rétro-projeté dans le temps pour chacun des points de la grille, à supposer que ce dernier corresponde à la première arrivée des ondes P.

Si le pointé équivaut bien à un temps d'arrivée des ondes P d'un événement sismique et si cet événement est enregistré à un nombre de stations suffisant, le nouveau pointé rétro-projeté est assemblé avec les pointés précédents compatibles qui proviennent du même événement. Le regroupement de ces pointés sera le plus dense autour du temps d'origine du point de la grille le plus proche de l'hypocentre optimal. Cependant, si un regroupement est identifié comme une potentielle origine, cela ne signifie pas nécessairement que tous les pointés qui y sont impliqués soient nécessairement des phases P. Ces pointés pourraient être aussi bien des faux pointés qui coïncident fortuitement, mais qui peuvent être regroupés du fait d'une maille grossière de la grille élaborée et/ou d'éventuelles contaminations liées au bruit enregistré.

Un programme de localisation (LocSAT, Bratt et al., 1988) est ensuite utilisé pour tenter une localisation et tester si le jeu de pointés regroupés correspond à un hypocentre cohérent. La qualité de l'hypocentre est évaluée à travers le meilleur accord entre les temps d'arrivée des ondes P calculés à chaque station et les temps observés pour la même station. Cet accord est estimé avec le calcul de la moyenne quadratique des résidus temporels (RMS des résidus des pointés). Si la valeur de cette RMS est trop grande, une amélioration est tentée en excluant chacun des pointés contributifs un à un pour vérifier s'il est possible de réduire la valeur de la RMS. Si la qualité de l'hypocentre estimée par le calcul de la RMS est validée, une origine est déclarée.

Seulement, l'origine déclarée (ou mise à jour) peut être encore contaminée par des phases faussement interprétées comme des phases P. Par conséquent, le rapport signal/bruit et les amplitudes pour chacun des pointés sont pris en compte pour affiner chaque origine. Un pic avec un rapport signal/bruit élevé est moins susceptible d'être associé à une salve de bruit transitoire qu'un pic dépassant simplement le seuil du rapport signal/bruit défini. De même, un pic

associé de manière absolue à une forte amplitude est plus susceptible de correspondre à un déclenchement sismique réel, notamment en cas d'observations simultanées de fortes amplitudes aux stations voisines.

Certains critères heuristiques sont en plus appliqués pour comparer les qualités des origines concordantes. Ces critères sont combinés en un score identifié pour chaque origine, qui est basé sur les propriétés des pointés eux-mêmes (valeurs des résidus, RMS, gap azimutal).

•Un clustering des pointés P et S

A chaque pointé qui arrive, le deuxième processus d'association vérifie si ce pointé peut être associé à une ou plusieurs origine(s) déjà identifiée(s) en calculant à chaque fois un score pour chaque origine. Ce score est une somme pondérée de 4 facteurs principaux : le nombre de pointés P et S associés ($pCount$ et $sCount$), le nombre de pointés P et S non associés ($p0Count$ et $s0Count$), la profondeur ($depthFactor$) et les résidus temporels ($residualFactor$). Chaque facteur est pondéré d'un poids ($score.weights.p$, $score.weights.p0$, $score.weights.s$, $score.weights.s0$, $score.weights.depth$ et $score.weights.residual$) qui peut être librement défini (équation 4.1).

Les valeurs définies pour les facteurs profondeur et résidus temporels (c'est-à-dire $depthFactor$ et $residualFactor$ dérivent des profondeurs et des résidus considérés, tout en tenant compte des valeurs maximales des profondeurs (> 50 km) et des résidus temporels (> 6s) à ignorer qui sont configurées au préalable.

$$\begin{aligned} score = {} & score.weights.p \times pCount + score.weights.p0 \times p0Count \\ & + score.weights.s \times sCount + score.weights.s0 \times s0Count \\ & + score.weights.depth \times depthFactor \\ & + score.weights.residual \times residualFactor \end{aligned} \quad (4.1)$$

C'est l'origine qui possède le score le plus élevé qui est sélectionnée. Cette origine est ensuite envoyée au module de gestion des événements uniquement si le score de la nouvelle origine excède le score de la dernière origine envoyée. Le score minimal à atteindre pour envoyer une origine est de 6.

Le poids affilié au nombre de pointés P (*score.weights.p*) et de pointés S (*score.weights.s*) associés est la valeur par défaut de 1. Le poids attribué au nombre de pointés P (*score.weights.p0*) et de pointés S (*score.weights.s0*) non associés est de 0. De même, le poids assigné à la profondeur est également de 0. Ces derniers critères ne sont donc pas considérés dans le calcul du score : le calcul de la profondeur étant très incertaine et le nombre de pointés P et S non associés n'étant pas un facteur critique pour la sélection d'une origine robuste. Au contraire, plus de poids a été alloué aux résidus temporels (poids de 5). Celui-ci a été défini empiriquement et correspond au poids qui a permis d'éliminer le maximum d'origines avec des pointés dont les résidus temporels étaient trop élevés. En effet, le score minimal à atteindre pour envoyer une origine étant de 6 et le nombre minimal de pointés à associer étant de 4, ceci favorise alors dans un premier temps la sélection d'origines définies à partir de résidus temporels plus faibles. Par conséquent, ceci évite au maximum les associations de pointés contenant de faux pointés, aboutissant à des origines correspondant à des événements pollués par du bruit d'origine anthropique pointé.

Dans le cas où le pointé ne peut pas être associé à une nouvelle origine, ce deuxième processus d'association va déterminer des nouvelles solutions hypocentrales en recherchant des clusters basés sur l'algorithme DBSCAN (Ester et al., 1996). Cet algorithme forme des clusters de pointés en cherchant si les stations voisines ont des pointés émis qui peuvent être regroupés dans le même cluster.

Plusieurs types de pointés sont identifiés dans le processus de clustering : les pointés centraux, les pointés atteignables et les pointés aberrants. Un pointé émis à une station P2 est considéré comme central si au moins un nombre minimal de pointés, incluant d'ailleurs le pointé émis à la station P2, se trouve à une distance temporelle inférieure ou égale à une distance de référence R autour du pointé émis à la station P2 (Figures 4.40a et 4.41 Step(1)).

Cette distance temporelle R est donc la distance de référence pour l'opération de clustering. Elle est équivalente à la somme vectorielle des différences de temps d'arrivée entre les pointés (Δt, en secondes) et des différences de temps de trajet entre les stations (en tt, secondes) : distance $= \sqrt{\Delta t^2 + tt^2}$. Les temps de trajet tt correspondent à $tt = \Delta x/v$ où x désigne la distance spatiale entre les stations (en kilomètres) et v la vitesse moyenne (en kilomètres/seconde) apparente horizontale des ondes P dans la croûte continentale.

Par conséquent, un pointé émis à la station P1 est directement atteignable depuis le pointé émis à la station P2 si ce pointé émis à la station P1 est à distance temporelle R du point central défini par le pointé émis à la station P2 (Figures 4.40b et 4.41 Step(2)). Tous les pointés qualifiés d'atteignables le sont à partir de pointés qualifiés de centraux. Par conséquent, un nouveau pointé émis à la station P5 est atteignable depuis le pointé émis à la station P2 s'il existe un chemin qui relie ces deux pointés entre eux au cours duquel chaque autre pointé émis par une station Pi+1 peut être directement atteignable depuis un autre pointé émis à une station Pi (Figures 4.40b-f et 4.41 Step(3)-Step(6)). Cela implique que chaque pointé initial et tous les pointés sur ce chemin doivent être des pointés centraux, avec l'exception possible du pointé émis à la station P5. Enfin tous les pointés qui ne sont atteignables depuis aucun autre pointé sont des pointés aberrants ou des pointés bruités (Figures 4.40g-h et 4.41 Step(7)-Step(8)).

Ainsi si le pointé émis à la station P2 est un pointé central, alors tous les pointés qui sont atteignables depuis ce pointé émis à la station P2 forment un cluster avec lui. Chaque cluster contient au moins un pointé central ; des pointés non centraux peuvent aussi faire partie du cluster mais ces derniers vont définir le contour du cluster, puisqu'ils ne peuvent pas être utilisés pour atteindre d'autres pointés (Figures 4.40 et 4.41).

Lorsque le cluster de pointés P contient le nombre minimal de phases qui est préalablement défini, l'algorithme d'association localise le cluster de pointés, crée une origine et associe des pointés P et S supplémentaires qui sont déjà présents, mais non encore associés. Si le pointé supplémentaire est un pointé P, l'algorithme tente directement d'associer le pointé basé sur la valeur de son résidu. Si cette association est un succès, la nouvelle solution est relocalisée. Dès que les pointés P sont associés, ce sont les pointés S qui vont ensuite être associés, ces derniers étant émis après les pointés P.

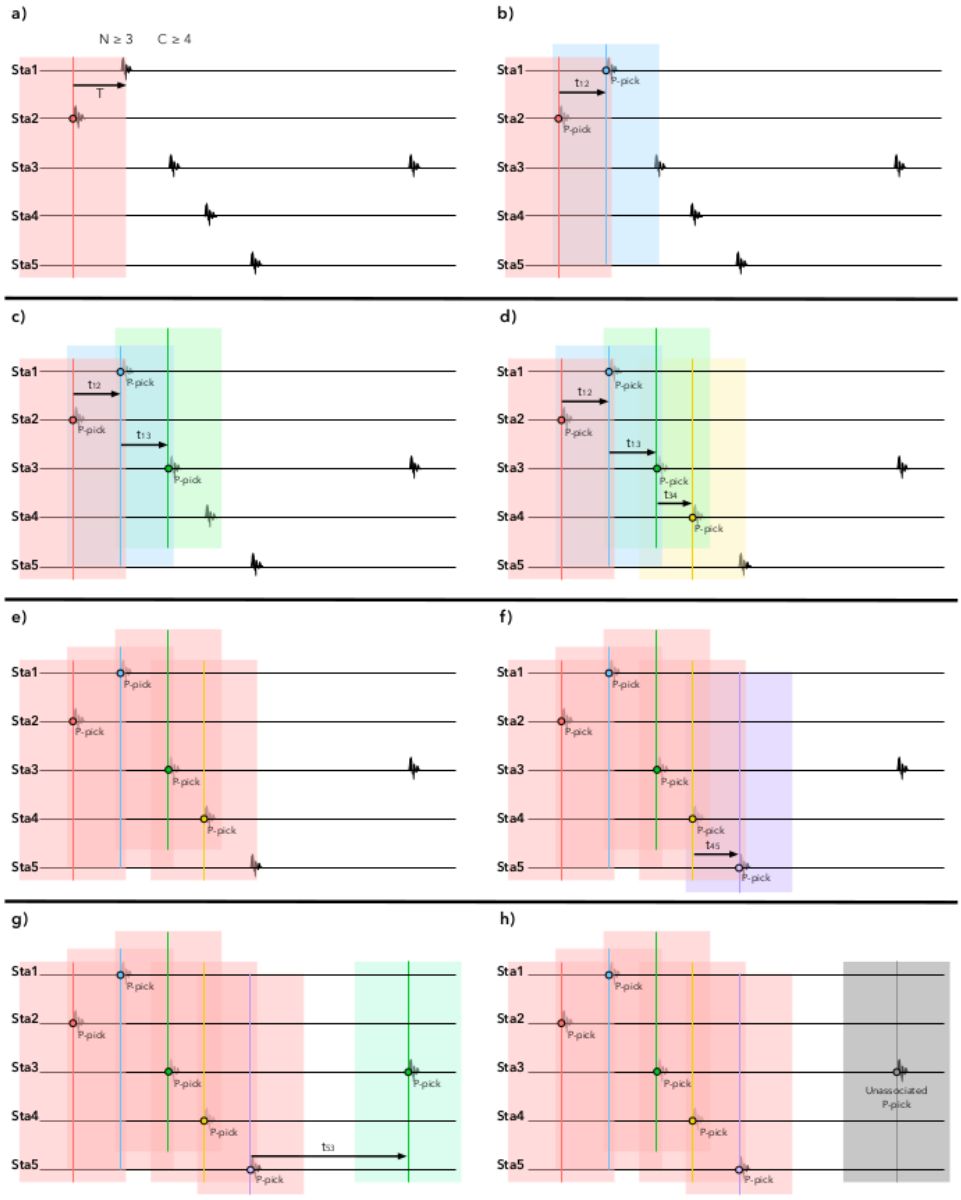


FIGURE 4.40: Principe de la méthode de clustering des pointés établi avec l'algorithme DBSCAN (voir Figure 4.41) pour plus de détails. Le nombre minimal de pointés défini pour classer un pointé comme pointé central est de 3 ($N \leq 3$) dans l'exemple de la figure. La lettre T représente la distance temporelle de référence et C désigne le nombre minimal de pointés nécessaires pour détecter un événement ($C \leq 4$). Les annotations t_ij constituent la différence de temps absolue ($t_ij = |t_i - t_j|$) entre les pointés émis à la station i et la station j. Chaque couleur différente représente à la fois un pointé (cercle de couleur) et une fenêtre temporelle (de distance temporelle de référence T) de recherche de pointés voisins atteignables (rectangle de couleur). Une fois qu'un cluster se forme, le pointé formant ce cluster prend la même couleur que le pointé central qui a été à l'origine de la formation de ce cluster (en l'occurrence rouge clair). D'après GRIGOLI, SCARABELLO et al., 2018.

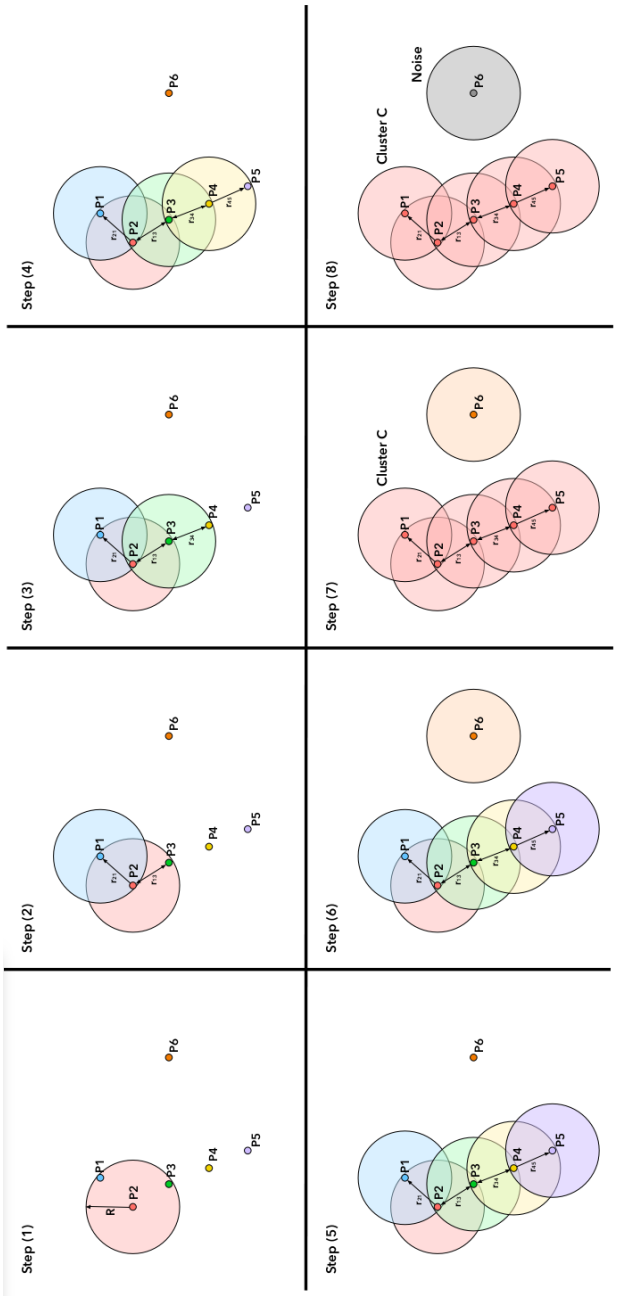


FIGURE 4.41: Principe de la méthode de clustering des pointés établi avec l'algorithme DBSCAN. P1, P2, P3, P4, P5, P6 désignent des pointés émis aux stations P1, P2, P3, P4, P5 et P6. R représente la distance temporelle de référence utilisée pour rechercher des pointés voisins. Les cercles de couleur représentent les surfaces de recherche de pointés voisins à partir d'un pointé central de référence. Ce cercle est de centre le pointé et de rayon la distance temporelle de référence R. D'après GRIGOLI, SCARABELLO et al., 2018.

Chaque nouvelle origine est reliée à un score (somme pondérée du nombre de pointés P et S associés, du nombre de pointés P et S non associés, de la profondeur et des résidus temporels) qui est comparé aux scores des autres origines appartenant au même événement. L'origine qui est envoyée au module de gestion des événements est celle qui a le score le plus élevé.

Les algorithmes d'association utilisés (méthode type recherche sur grille et méthode de clustering) se basent principalement sur les temps d'arrivée relatifs des différents pointés émis ainsi que la valeur de leurs résidus temporels pour associer les pointés entre eux. Or, les résidus temporels sont évalués en comparant les temps d'arrivée observés aux temps d'arrivée théoriques qui dépendent des temps de trajet calculés. Seulement, le calcul des temps de trajet théoriques est fonction de la géométrie du réseau de stations et de la structure du modèle de vitesse. Ces derniers facteurs vont donc être essentiels à prendre en compte pour améliorer le processus d'association. Cette amélioration vise à limiter les associations hybrides des faux pointés avec de vrais pointés. Ces faux pointés peuvent être de deux natures : ou bien ce sont des pointés qui ont été émis autour de vrais pointés dans une fenêtre temporelle très proche, comme il a été vu précédemment, ou bien ce sont des pointés qui sont reliés purement à du bruit. La sélection du premier type de faux pointé dans le processus d'association est fortement dépendant de la structure du modèle de vitesse et le deuxième type est plus lié à la géométrie du réseau.

4.2.2 Tenir compte de la configuration du réseau de stations

Quelque soit l'algorithme d'association utilisé, une distance maximale autorisée pour opérer le processus d'association est définie. La valeur qui est choisie est 250 km. Sachant que cette procédure s'attarde à détecter les événements de faible magnitude, le rayon de recherche est donc limité à une échelle régionale, car au-delà les chances de détecter des signaux de faible amplitude sont très petites.

De plus, la distance de référence utilisée pour accomplir le procédé de clustering est un facteur indispensable, voire déterminant, à définir pour mener à bien l'association basée sur cette méthode. Comme il a été écrit précédemment, l'estimation de cette distance dépend de deux paramètres fondamentaux : les différences de temps d'arrivée entre les pointés et les différences de temps de trajet entre les stations. Ce dernier paramètre dépend donc de la distance spatiale entre les stations. La prise en compte de la configuration du réseau de stations est donc capitale pour évaluer une distance de référence optimale pour clusteriser, donc détecter efficacement les événements.

Pour évaluer cette distance de référence optimale en tenant compte de la configuration du réseau de stations (Pestourie et al., 2017), j'ai d'abord généré automatiquement une grille de localisations épicentrales et hypocentrales de 45000 séismes synthétiques. Les localisations épicentrales sont comprises entre les intervalles de latitude [46°N-52°N] et de longitude [3°E-12°E] et les localisations hypocentrales sont comprises entre 2 km et 15 km. L'intervalle de profondeurs choisies est en lien avec les profondeurs qui sont majoritairement retrouvées dans les catalogue de séismes de la zone d'étude.

A partir du réseau de stations qui est utilisé pour cette étude, les temps d'arrivée des différentes phases sismiques P et S ont été simulés pour chacun des séismes synthétiques à partir du logiciel NonLinLoc (Lomax, Virieux et al., 2000). Le modèle de vitesse qui a été choisi pour générer ces temps d'arrivée théoriques a été le modèle régional Haslach utilisé majoritairement pour la détection des événements (cf Figure 3.39).

Pour chaque événement synthétique, une matrice des distances temporelles (somme vectorielle des différences de temps d'arrivée entre les pointés et des différences de temps de trajet entre les stations) est calculée à partir des 10 stations les plus proches de l'événement. Pour déterminer les temps de trajets entre les stations (rapport entre la distance spatiale entre les stations et la vitesse moyenne apparente horizontale des ondes P dans la croûte), la vitesse moyenne qui a été choisie correspond à la vitesse moyenne des ondes P dans la croûte pour le modèle Haslach, à savoir 6 km/s. A partir de cette matrice des distances, un algorithme de clustering DBSCAN, contenu dans le package Python Scikit-Learn, a été utilisé pour rechercher la valeur minimale de distance temporelle nécessaire pour former un cluster à partir de la matrice des distances pour chaque événement synthétique. Pour trouver cette valeur minimale, une gamme de valeurs de distance temporelle comprises entre 0.1 et 200 s avec un pas de 0.1 s a été testée. De plus, la valeur minimale de pointés nécessaires pour former un cluster a été paramétrée à 6.

A l'issue de cette recherche, chaque événement synthétique est donc caractérisé par une distance temporelle minimale nécessaire pour former des clusters. De ce fait, la valeur optimale de cette distance temporelle pour la totalité de la zone d'étude a été déduite de l'ensemble des événements synthétiques. Cette valeur optimale correspond à une valeur qui induirait des clusters sur toute la zone d'étude, quelque soit les localisations épicentrales et hypocentrales des événements. C'est donc la valeur qui couvre à 100% la zone d'étude, et dans notre cas, elle correspond à la valeur maximale rencontrée sur l'ensemble du jeu synthétique, à savoir 21.5 s (Figures 4.42 et 4.43).

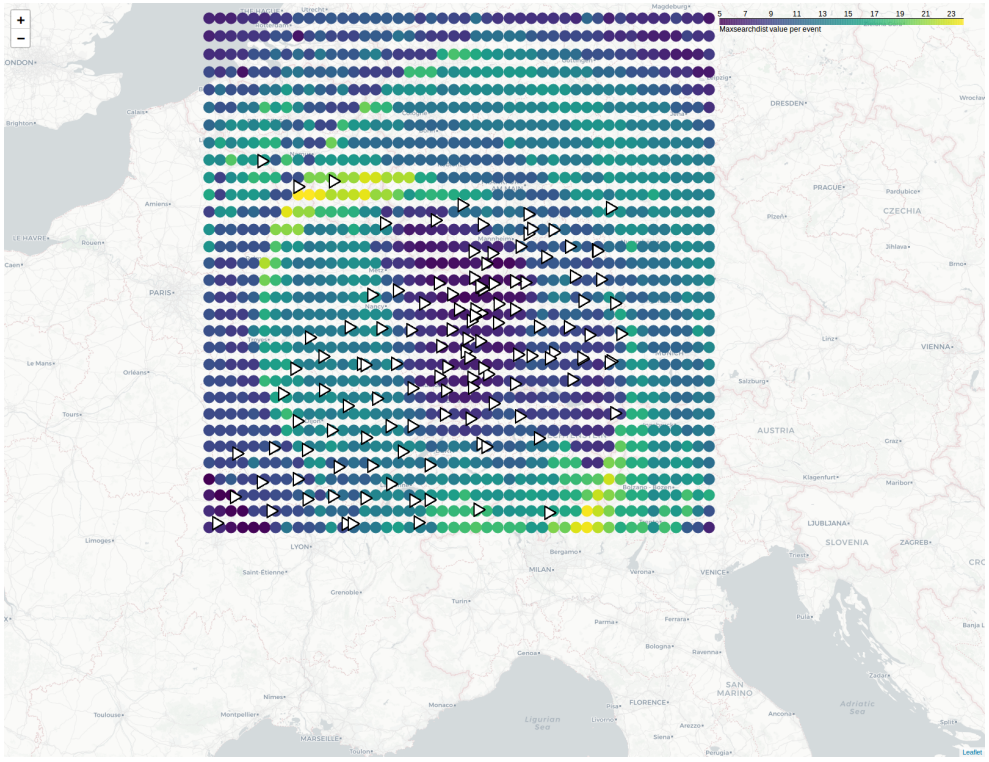


FIGURE 4.42: Représentation des valeurs de distance temporelle minimale pour former un cluster de pointés pour l'ensemble des événements synthétiques de la zone d'étude situés à une profondeur de 5 km. Chaque point correspond à la localisation épicentrale d'un événement synthétique. La couleur de ce point identifie la valeur de la distance temporelle pour cette événement (cf légende de la couleur sur la carte). Les triangles blancs correspondent aux stations qui ont été utilisées pour générer les temps d'arrivée.

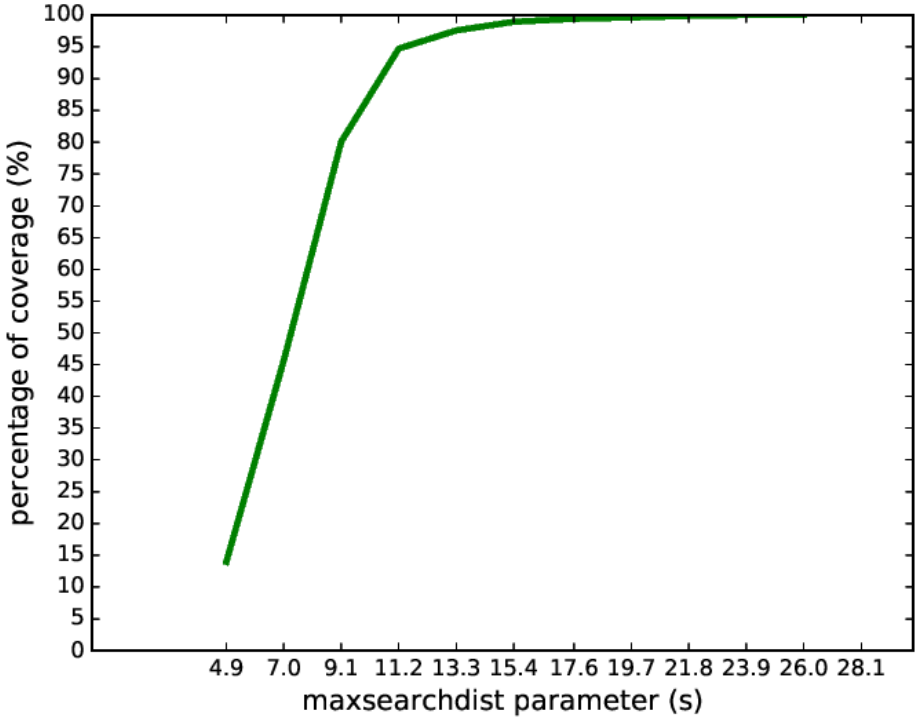


FIGURE 4.43: Graphique représentant le pourcentage de couverture de la zone d'étude en fonction de la valeur de la distance temporelle. Cette couverture correspond à la proportion d'événements qui sont effectivement générés à partir d'un cluster d'au moins 6 pointés pour une distance temporelle donnée.

4.2.3 Tenir compte du milieu de propagation des ondes sismiques

•Adapter le calcul des temps de trajet entre les stations

La nature des terrains affleurant dans cette zone d'étude est très variable : granitoïdes, roches volcaniques et roches métamorphiques des Massifs cristallins, couverture sédimentaire d'épaisseur, de nature et d'origine variables des bassins et des reliefs jurassiens (calcaires, dolomies, grès, argiles, bancs gypseux, etc.). La structure lithologique verticale et latérale de la croûte continentale est en conséquence très complexe. Affilier une seule vitesse moyenne apparente des ondes P pour toute la zone d'étude est donc très restrictive.

La vitesse des ondes sismiques étant fortement dépendante de la nature du milieu qu'elles traversent, une vitesse moyenne de 6 km/s pour la croûte continentale, comme exprimée par le modèle d'Haslach, peut facilement surestimer les temps trajets des ondes qui sont émises dans un milieu où la couche sédimentaire superficielle est régionalement plus épaisse de quelques kilomètres. En effet, le modèle d'Haslach a été établi à partir de l'étude des signaux émis par une explosion qui a eu lieu dans une carrière souterraine située à 2 km de profondeur dans le Massif de la Forêt Noire. La région est essentiellement constituée de gneiss (ortho- et paragneiss) traversés par des filons amphibolitiques. Cette étude a conduit à considérer un modèle de vitesse élaboré à partir d'une structure crustale constituée essentiellement d'une première couche granito-gneissique, d'une couche de granites plus profonds et d'une couche de basaltes et de gabbros lenticulaires. Seulement, les vitesses des ondes sismiques P sont généralement plus lentes dans des matériaux comme les calcaires (3.0-4.0 km/s) ou les grès (4.0-4.5 km/s) que dans des granites (5.5-6.0 km/s), des gneiss (de l'ordre de 5.5 km/s) ou des gabbros (6.5-7.0 km/s).

Or, la qualité de l'assemblage des pointés par clustering dépend de la distance temporelle qui est utilisée pour former les clusters. Et cette distance temporelle est conditionnée par la vitesse moyenne apparente des ondes P qui est choisie pour calculer les temps de trajet entre les stations. Par conséquent, une vitesse moyenne plus élevée diminue les temps de trajets et vice-versa. Ainsi, pour une même différence de temps d'arrivée entre les pointés, la valeur de la distance temporelle de référence diminue alors inversement à l'augmentation de la vitesse moyenne des ondes P utilisée.

De ce fait, si cette vitesse moyenne est surestimée pour une zone donnée, il y a un risque accru que des pointés ne soient pas associés au cluster en formation. Dans ce cas, leurs temps d'arrivée seraient en effet trop tardifs (puisque traversant un milieu de vitesse moyenne plus faible) relativement aux temps de trajets calculés entre les stations, augmentant artificiellement la distance temporelle qui les sépare des autres pointés inclus dans le cluster, et diminuant la probabilité que ces pointés tardifs soient finalement contenus dans le cluster.

Ainsi, dans cette configuration, si de multiples pointés ont été émis consécutivement dans le temps autour du signal à détecter, même si dans ce lot de pointés multiples il y a un vrai pointé qui correspond à la première arrivée des ondes P, un faux pointé a plus de chance d'être associé au cluster parce qu'à une distance temporelle suffisante pour y être inclus.

Afin de combler les lacunes de ce processus d'association, plusieurs vitesses moyennes des ondes P dans la croûte continentale ont alors été considérées : 4 km/s, 5 km/s et 6 km/s. De cette façon, trois instances ont été utilisées en parallèle pour générer le processus d'association par clustering : une instance avec pour vitesse moyenne des ondes P dans la croûte continentale de 4 km/s, une instance avec une vitesse moyenne de 5 km/s et une instance avec une moyenne de 6 km/s.

- **Adapter les modèles de vitesse pour générer des origines optimales**

Les deux procédés d'association (méthode recherche sur grille puis méthode basée sur le clustering des pointés) produisent des assemblages de pointés qui sont produits soit par rétro-projection à un hypocentre optimal, soit par clustering basé sur le calcul de distances temporelles. Dans tous les cas, des pointés supplémentaires peuvent y être ajoutés, en considérant la valeur de leurs résidus temporels notamment. Chaque assemblage de pointés génère finalement une origine qui est localisée avec un score optimal. Or, quelque soit le procédé d'association, ce score tient compte à la fois de la valeur des résidus et de la RMS.

Les valeurs maximales des résidus qui ont donc été autorisées pour maximaliser les détections ont été respectivement 2.5 s pour le premier procédé d'association (méthode recherche sur grille) et 2.8 s pour le deuxième procédé d'association (méthode basée sur le clustering). Ces valeurs ont été testées empiriquement et offrent un seuil maximal qui permet à la fois l'élimination des pointés avec des résidus excessifs et l'inclusion de pointés avec des résidus un peu plus élevés, augmentant alors le nombre de pointés P et S possibles tout en compensant un peu les grandes incertitudes liées au modèle de vitesse.

De plus, les valeurs maximales de RMS ont été plafonnées à 5 s pour le premier procédé d'association et à 6 s pour le deuxième procédé d'association. Ces valeurs ont été placées assez hautes mais elles ont permis d'évaluer la performance de l'association en révélant par exemple des événements dont les origines sélectionnées avaient systématiquement des RMS élevées, mais correspondaient pourtant à un vrai événement (séisme ou tir de carrière). Par conséquent, ces valeurs constituent un garant pour continuellement déceler des défaillances des procédés d'association qui peuvent être plus facilement corrigées. Ce qui augmente les chances de récupérer des événements qui sinon auraient été perdus.

Le modèle de vitesse est un paramètre fondamental qui conditionne la qualité des associations et des futures origines sélectionnées. Plusieurs modèles de vitesse ont alors été empiriquement testés de façon à évaluer si ces derniers pouvaient aboutir à des détections plus nombreuses et de meilleure qualité (moins de faux pointés). Les modèles testés empiriquement ont été ceux qui ont été générés automatiquement pour évaluer l'impact des incertitudes des pointés en fonction des localisations épicentrales et hypocentrales établies à partir d'une centaine de modèles de vitesse (50 modèles à 3 couches et 50 modèles multicouches).

L'objectif est de détecter ici, non pas d'obtenir des solutions hypocentrales de qualité exacte mais d'obtenir des solutions hypocentrales plus précises, c'est-à-dire qui minimisent les différences entre les temps d'arrivée des ondes P et S observés et les temps d'arrivée de ces ondes théoriques calculés pour les différentes stations. Ce n'est donc pas l'exactitude des modèles de vitesse au regard de la structure latérale et verticale réelle de la croûte continentale que nous cherchons, mais leur précision évaluée par la valeur de la RMS obtenue. Ceux qui ont donc été testés ont été ceux qui ont minimisé la RMS pour les événements localisés en juillet 2016, à savoir les modèles à 3 couches notifiés 11, 25, 31 et 38 et les modèles multicouches notifiés 10, 24, 27 et 35 (Figure 4.44 et 4.45). L'ensemble des 8 modèles sélectionnés présentent des vitesses moyennes crustales des ondes P et S équivalentes c'est-à-dire des valeurs comprises entre 4.5 km/s et 5.5 km/s pour les ondes P et des valeurs comprises entre 2.6 km/s et 3 km/s pour les ondes S (cf Annexe E pour le détails des différents modèles de vitesse). Ces modèles choisis ont par la suite été testés sur la détection automatique des événements au cours du mois de juillet 2016.

Ajouté à ces modèles de vitesse, un dernier modèle de vitesse a été testé. Il s'agit d'un modèle multicouche 1D minimum obtenu à partir du modèle 1D d'Haslach grâce à la procédure d'inversion des paramètres hypocentraux et des paramètres de vitesse proposée par Kissling et al., 1995, à travers le programme VELEST. L'approche proposée par Kissling et al., 1995 consiste en une série d'inversions simultanées des paramètres hypocentraux et des modèles de vitesses (Vp et Vs) de telle façon à approcher des solutions minimales, c'est-à-dire la RMS minimale. Les solutions sont à la fois la localisation des hypocentres, le calcul d'un modèle 1D en couches et la correction apportée aux stations (correction liée aux différences de temps de trajet sous chaque station). Le modèle 1D minimum est présenté dans l'annexe F.1.

Les tables des temps de trajet ont été calculées pour chaque modèle sélectionné à partir des outils TauP (Crotwell et al., 1999). Ces tables regroupent les temps de trajets des phases sismiques P et S pour des profondeurs comprises entre 0 et 35 km et des distances épicentrales comprises entre 0 km et 1400 km. Le programme de localisation LocSAT (Bratt et al., 1988) utilisent ces tables pour localiser les différentes origines détectées.

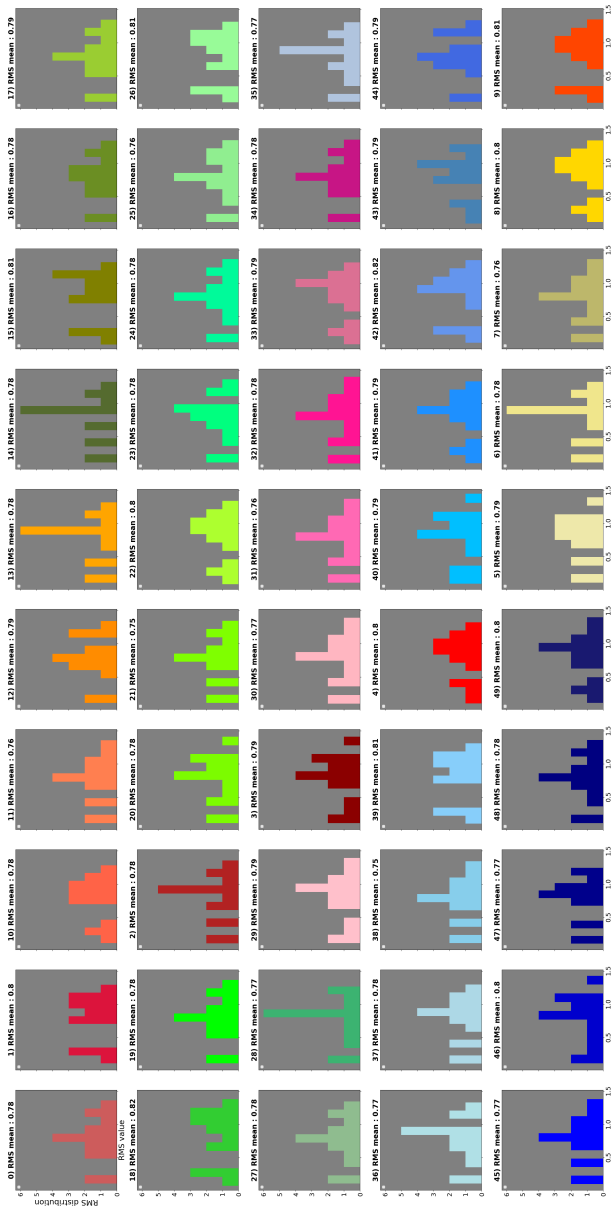


FIGURE 4.44: Distribution de la RMS évaluée sur des événements détectés en juillet 2016 et localisés avec 50 modèles de vitesse à 3 couches générés automatiquement (voir paragraphe 4.1.3). Chaque modèle de vitesse est numéroté et la valeur de la RMS moyenne obtenue pour chacun des modèles de vitesse est spécifiée en haut de chaque encart.

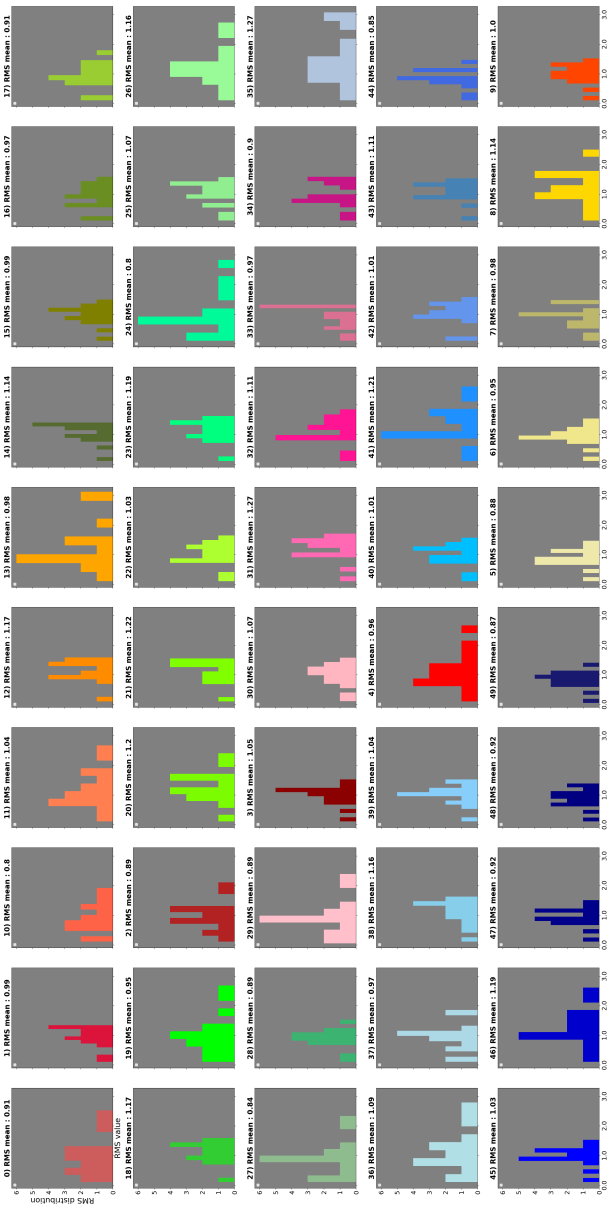


FIGURE 4.45: Distribution de la RMS évaluée sur des événements détectés en juillet 2016 et localisés avec 50 modèles de vitesse à multicouches générés automatiquement (voir paragraphe 4.1.3). Chaque modèle de vitesse est numéroté et la valeur de la RMS moyenne obtenue pour chacun des modèles de vitesse est spécifiée en haut de chaque encart.

Les différents modèles ont été testés sur les détections et localisations des événements pour le mois de juillet 2016. Les résultats ont été comparés avec ceux obtenus avec deux modèles régionaux 1D de référence qui sont classiquement utilisés par les analystes pour localiser les événements de la zone d'étude. Ces deux modèles sont celui d'Haslach et un autre modèle traditionnellement utilisé pour localiser les événements dans la région des Alpes (Figure 4.46, Frechet, 1978 ; Thouvenot et al., 2003).

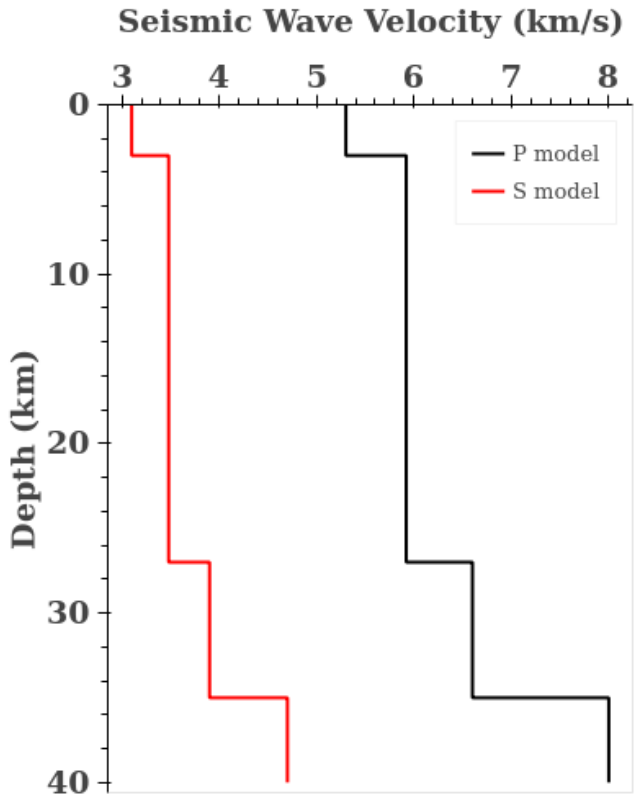


Figure 4.46: Modèle de vitesse des Alpes utilisé pour la détection, en combinaison avec le modèle de Haslach. Le modèle des Alpes a été élaboré à partir de l'analyse de profils sismiques et de l'hodochrone des ondes P déduit d'un tir d'une carrière identifié le 29 septembre 1977, près de la ville de Guillestre (Frechet, 1978 ; Thouvenot et al., 2003).

Les différents tests effectués n'ont pas permis de mettre en relief une solution de modèle de vitesse optimale pour toute la zone d'étude. En effet, si un modèle peut localement améliorer le processus final d'association et de localisation des origines (nombre de pointés supérieurs et diminution des faux pointés sélectionnés), il peut à l'inverse fortement dégrader la détection ailleurs, jusqu'à ne plus détecter les événements qui étaient précédemment présents dans le catalogue de référence automatique établi avec le modèle d'Haslach.

Le choix de modèles plus spécifiques, notamment les modèles multicouches, n'a donc pas apporté de plus-value sur les résultats de détection finaux, mais a au contraire souligné des hétérogénéités spatiales dans les détections plus marquées. Face à la variabilité lithologique du milieu de propagation, plusieurs de ces modèles de vitesse seraient en fait nécessaires. Seulement, déterminer les critères de détection (géographiques, pétrologiques, stratigraphiques, lithologiques, structuraux, sismologiques, etc) qui puissent permettre de découper efficacement et significativement la zone d'étude en plusieurs régions de détection, chacune affiliée par exemple à un modèle 1D minimum spécifique, devient éminemment complexe et demande une connaissance plus approfondie du comportement de la sismicité dans cette zone d'étude.

Par conséquent, la sélection combinée des deux modèles de référence, celui d'Haslach et celui des Alpes, a finalement produit les meilleurs résultats, avec plus d'homogénéité. Néanmoins, afin de récupérer le maximum d'événements, le choix d'en fixer automatiquement certains à une profondeur arbitraire a été privilégié.

En effet, en cas d'incertitudes trop fortes des modèles de vitesse choisis, en particulier vis-à-vis des couches superficielles, cela a permis à l'algorithme de localisation LocSAT de faire converger les solutions hypocentrales vers un minimum local, notamment pour les événements localisés plus superficiellement. La profondeur minimale arbitraire qui a été choisie est 2 km. Celle-ci correspond à la profondeur qui a été majoritairement retrouvée lorsque j'ai relocalisé l'ensemble des événements de la période 2016 en incluant les stations temporaires AlpArray pour l'année 2016 (cf Figure 3.38).

•Créer plusieurs instances pour optimiser le processus d'association sur toute la zone d'étude

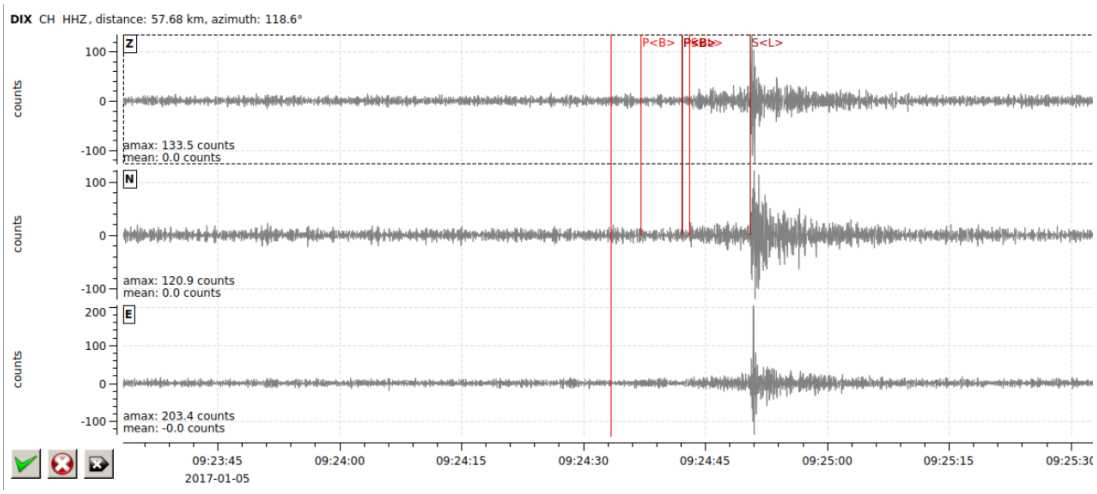
Six instances du processus d'association ont alors été introduites : deux pour le premier processus d'association (rétro-projection de pointés) et quatre pour le second processus d'association (clustering de pointés). Pour le premier processus d'association, chaque instance localise les origines avec un modèle de vitesse différent (la première instance utilise le modèle Haslach et la deuxième le modèle des Alpes). Pour le second processus d'association, chaque instance est définie par une combinaison d'une vitesse moyenne des ondes P dans la croûte continentale pour le clustering et d'un modèle de vitesse pour localiser (Table 4.1).

TABLE 4.1: Critères différenciant les différentes instances déployées du même processus d'association basé sur le clustering par la méthode DBSCAN (ESTER et al., 1996).

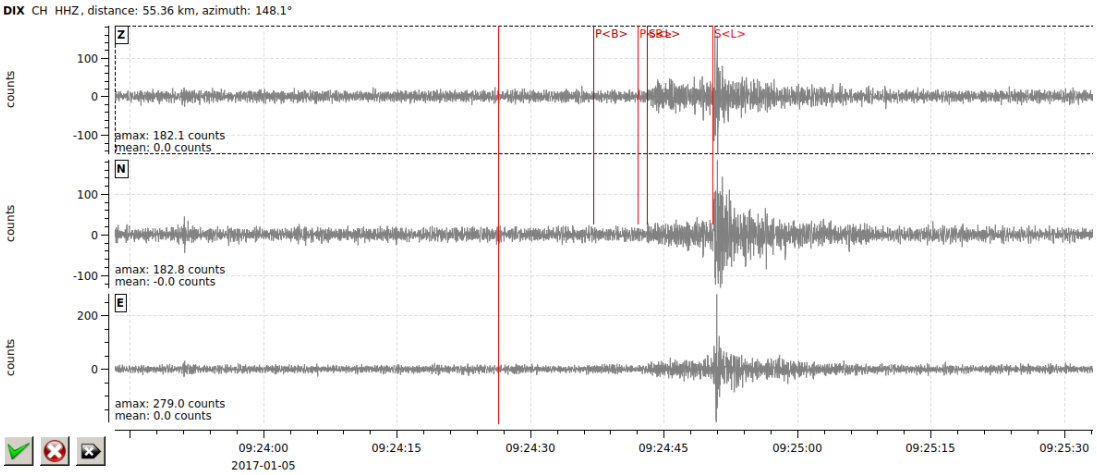
	Mean P-wave Velocity	Velocity Model
First Instance	4 km/s	Haslach Model
Second Instance	5 km/s	Haslach Model
Third Instance	5 km/s	Alps Model
Fourth Instance	6 km/s	Haslach Model

En plus des deux instances fournies par le premier processus d'association, l'utilisation en parallèle de ces 4 dernières instances améliorent nettement le processus d'association final. Par exemple, pour le séisme du 05 janvier 2017 à 09h24 qui est localisé au coeur du Massif du Chablais alpin (MLv 1.3), l'utilisation d'une vitesse moyenne des ondes P de 6 km/s, couplée au modèle Haslach, a permis d'associer correctement un pointé P à la station OGSI, située à l'extrême Sud de la nappe alpine de la Dent-Blanche (Figure 4.47a), et les deux pointés P et S à la station DIX, située au coeur du Valais Suisse (Figure 4.47c). En revanche, avec une vitesse moyenne de 4 km/s, et le même modèle d'Haslach, cette association échoue avec les deux stations : pour OGSI, c'est un faux pointé P précoce qui est choisi alors que pour la station DIX, ce sont deux faux pointés anticipés P et S qui ont été associés avec le reste des autres pointés (Figure 4.47b et d).

Pour ce séisme, des vitesses moyennes inférieures à 6 km/s induisent des distances temporelles entre les pointés supérieures à celles calculées pour des vitesses égales à 6 km/s, et donc des temps de trajet proportionnellement plus longs. Par conséquent, les pointés émis aux stations OGSI et DIX qui vont être à des distances temporelles suffisantes pour être assemblés avec les autres pointés émis au stations A173A, AIGLE, GIMEL, OGMY, OG35, A164A et A181A vont être ceux qui ont des temps d'arrivée anticipés, c'est-à-dire en l'occurrence des faux pointés. Ces dernières associations aboutissent à des résidus temporels négatifs élevés pour ces deux stations OGSI et DIX, augmentant alors la RMS de l'origine localisée : 4.7 (instance avec vitesse moyenne de 4 km/s), 5.4 (instance avec vitesse moyenne de 5 km/s) ou 4.6 secondes (instance avec vitesse moyenne de 5 km/s mais modèles des Alpes) au lieu de 1.4 (instance avec vitesse moyenne de 6 km/s) pour un même nombre de phases (18).



a) Pointés P et S sélectionnés (trait vertical rouge foncé) à la station DIX



b) Faux pointés P et S sélectionnés (trait vertical rouge foncé) à la station DIX

FIGURE 4.47: Comparaison de la performance de l'association produite avec deux instances du processus d'association considérant une vitesse moyenne des ondes P de 6 km/s (a) et 4 km/s (b).

De cette façon, pour améliorer le processus d'association des pointés pour cet événement avec les instances qui prennent en compte une vitesse moyenne des ondes P inférieure à 6 km/s pour calculer les temps de trajet entre les stations, il faudrait sélectionner une origine qui n'intègre pas les stations les plus éloignées (comme OG35, OGMY, A164A ou A181A). Cependant cela impliquerait de sélectionner des origines avec moins de phases (10 en l'occurrence) pour obtenir une RMS de moins de 2 s. Par conséquent, l'instance qui considère une vitesse moyenne de 6 km/s est l'instance optimale pour détecter ce séisme.

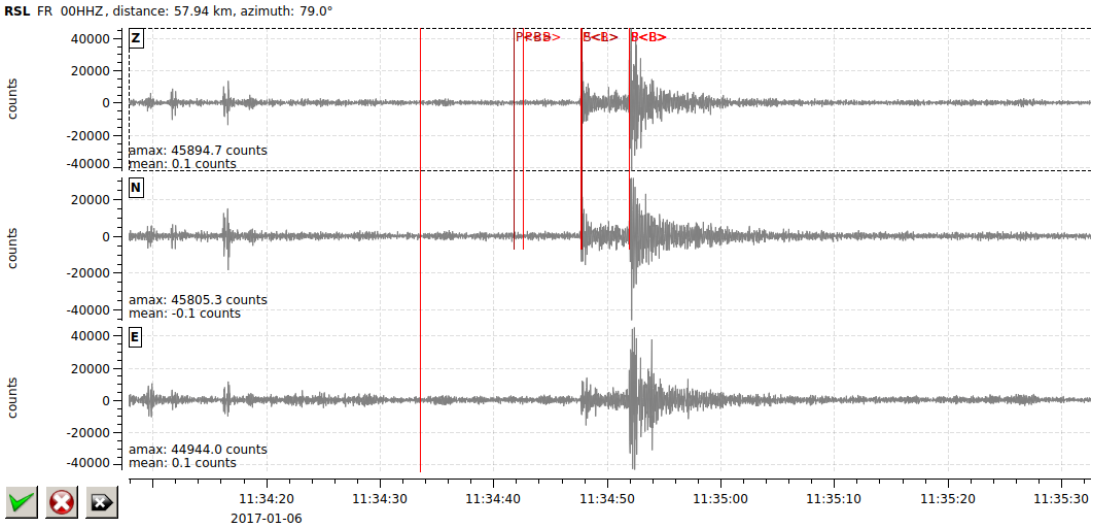
En revanche, pour le séisme du 06 janvier 2017 ayant eu lieu à 11h34 dans la région de Chambéry (MLv 0.9), c'est l'instance avec une vitesse moyenne des ondes P de 5 km/s, combinée au modèle des Alpes, qui va générer l'origine optimale : 9 phases et RMS de 1.4 secondes. En effet, si l'on prend l'exemple de la station RSL, de faux pointés P et S émis avec des temps d'arrivée anticipés sont sélectionnés dans le processus d'association opéré par les trois autres instances.

L'instance considérant une vitesse moyenne de 6 km/s calcule des distances temporelles globalement plus petites pour clusteriser, autorisant un plus grand nombre de phases (13 phases mais RMS de 5.2), donc l'inclusion de stations plus éloignées. Seulement, cela signifie également que la probabilité d'inclure de faux pointés est accrue, comme cela a été effectivement le cas dans cet exemple. Ainsi, l'inclusion d'un faux pointé dans le processus d'association émis à la station A215A a perturbé le procédé de clustering. La station RSL étant éloignée de la station A215A, pour que celle-ci soit à des distances temporelles suffisantes pour être associé au cluster, l'instance du processus d'association a alors sélectionné les faux pointés P et S émis de façon anticipée à cette station RSL, diminuant alors la différence des temps d'arrivée entre les stations, donc les distances temporelles utilisées pour clusteriser (Figure 4.48a).

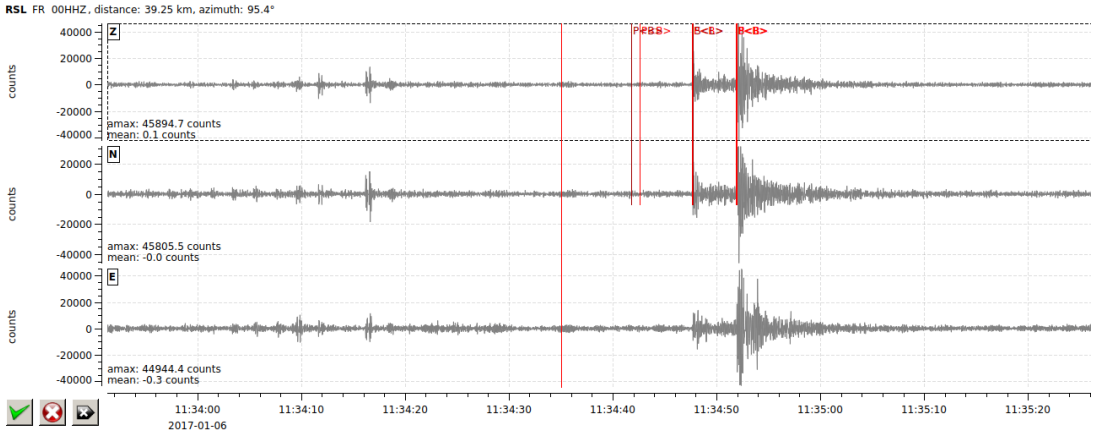
De même, l'instance considérant une vitesse moyenne de 4 km/s à tendance à calculer des distances temporelles plus grandes, diminuant la possibilité d'inclure des phases dans l'association (7 phases, RMS 1.4). Seulement, dans ce cas-ci, cette instance a sélectionné les faux pointés P et S émis de façon anticipée à la station RSL car cela compense des temps de trajet calculés entre les stations trop élevés (Figure 4.48b).

Enfin, l'instance considérant une vitesse moyenne de 5 km/s, mais combinée avec le modèle Haslach, tend à calculer des temps d'arrivée théoriques plus précoces qu'avec le modèle des Alpes, car les temps de trajets calculés pour ce modèle sont plus rapides. Dans cette configuration, la différence des temps d'arrivée théoriques et observés est élevée. Sachant que la valeur du résidu intervient dans le calcul du score qui va sélectionner la meilleure future origine, et qu'un poids élevé est donné à cette valeur, cette instance a sélectionné également les faux pointés P et S émis de façon anticipée à la station RSL car ils minimisent la valeur des résidus temporels (Figure 4.49c). L'instance avec une vitesse moyenne de 5 km/s, combinée au modèle des Alpes, est l'instance qui va

donc améliorer le mieux le processus d'association, sélectionnant notamment les pointés P et S corrects à la station RSL (Figure 4.49d)

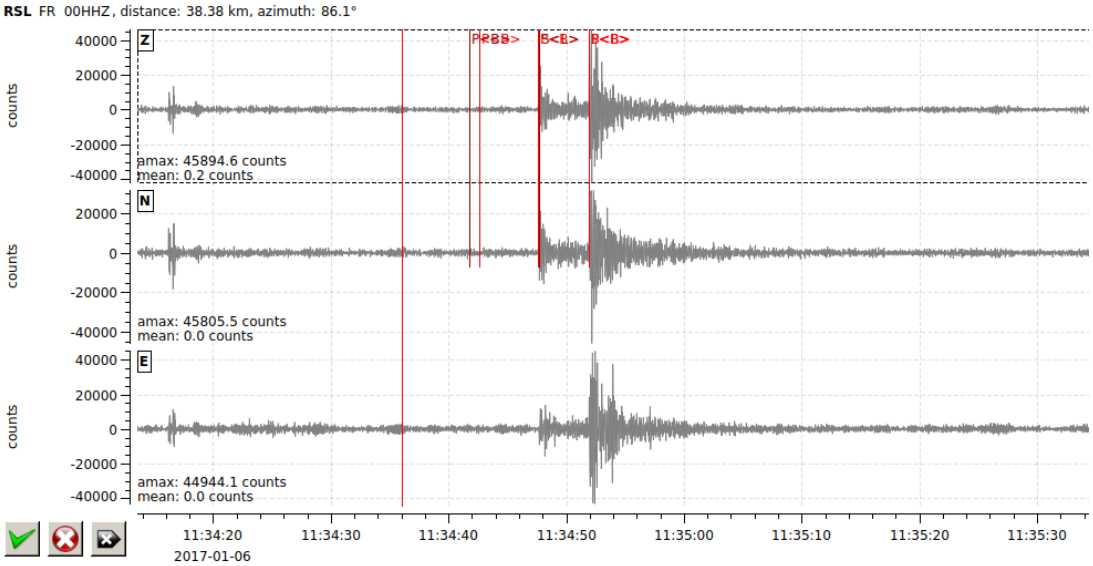


a) Faux pointés P et S sélectionnés (trait vertical rouge foncé) avec l'instance considérant une vitesse moyenne de 6 km/s et le modèle Haslach

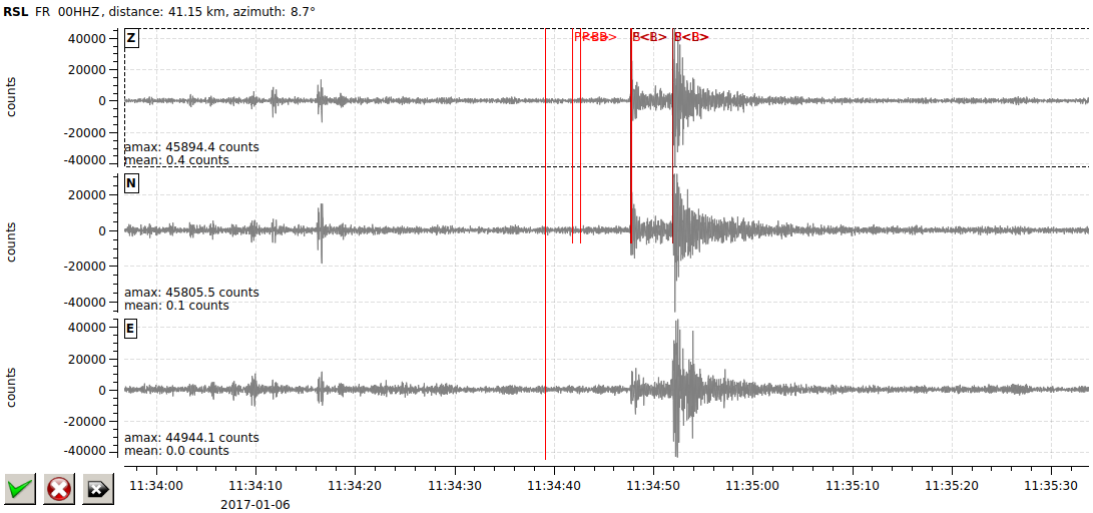


b) Faux pointés P et S sélectionnés (trait vertical rouge foncé) avec l'instance considérant une vitesse moyenne de 4 km/s et le modèle Haslach

FIGURE 4.48: Comparaison de la performance de l'association produite avec les quatre instances du processus d'association basé sur la méthode de clustering de DBSCAN (Ester et al., 1996) à partir de l'exemple de la station RSL).



c) Faux pointés P et S sélectionnés (trait vertical rouge foncé) avec l'instance considérant une vitesse moyenne de 5 km/s et le modèle Haslach



d) Faux pointés P et S sélectionnés (trait vertical rouge foncé) avec l'instance considérant une vitesse moyenne de 5 km/s et le modèle des Alpes

FIGURE 4.49: Comparaison de la performance de l'association produite avec les quatre instances du processus d'association basé sur la méthode de clustering de DBSCAN (Ester et al., 1996) à partir de l'exemple de la station RSL).

La prise en compte de la variabilité latérale et verticale du milieu de propagation est décisive pour améliorer le processus d'association des pointés P puis S entre eux. En effet, une évaluation erronée des temps de trajet au sein du réseau de stations peut faire perdre un vrai pointé au détriment d'un faux pointé émis dans une fenêtre temporelle très restreinte autour du signal à détecter, parce que ce dernier est à une distance temporelle compatible avec les autres pointés inclus dans le cluster ou est associé à un résidu temporel plus petit. Or, l'intégration de tels faux pointés déstabilise fortement la procédure d'association qui peut, en s'éloignant de la solution, plus facilement inclure d'autres faux pointés, dont ceux correspondant à du bruit. Par ailleurs, la prise en compte de la géométrie du réseau de stations dans la zone d'étude, c'est-à-dire de la distance entre les stations, est un autre garant pour limiter l'inclusion de faux pointés aberrants émis au sein du réseau.

4.3 Améliorer l'origine préférentielle pour chaque événement

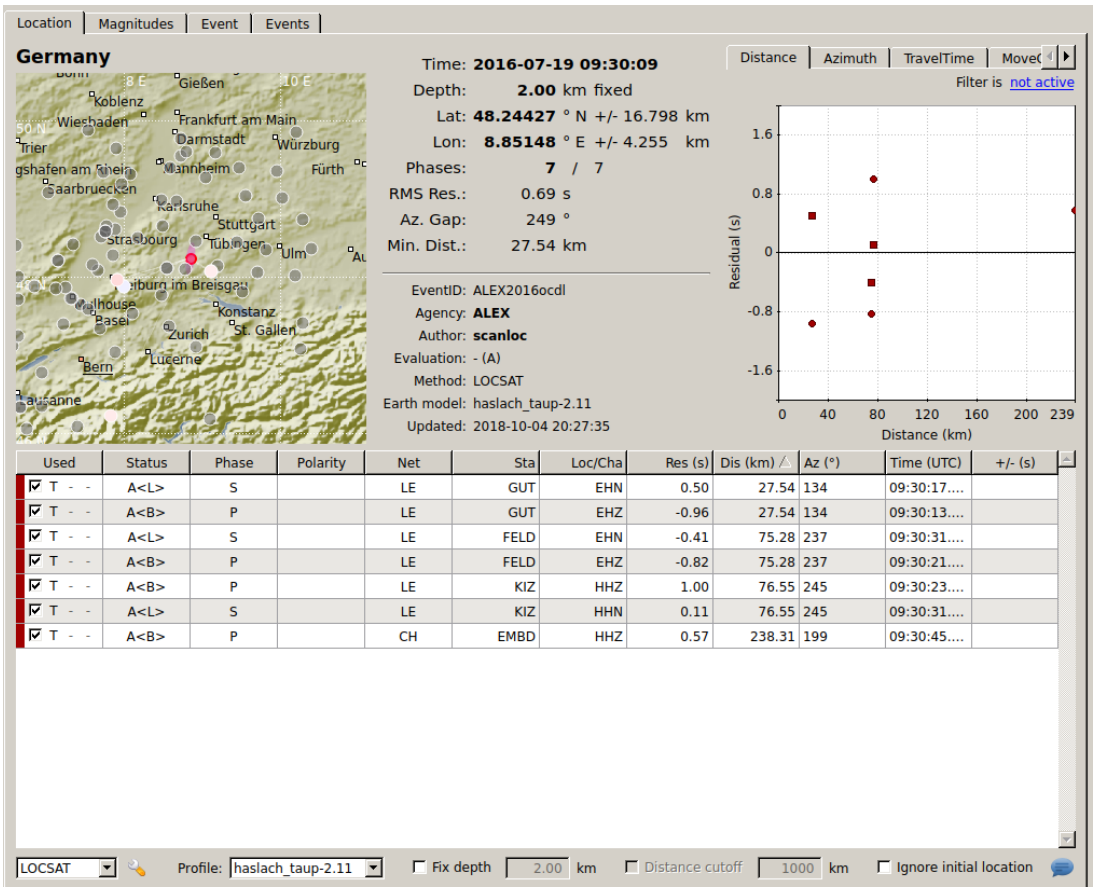
4.3.1 Combler les défaillances du protocole par défaut de sélection de l'origine préférentielle

Comme décrit précédemment dans la section 2, le système de détection de SeisComP3, qui est utilisé la plupart du temps en temps réel, crée plusieurs origines par événement. En effet, au fur et à mesure du temps qui passe, plus de phases sismiques vont être disponibles pour déclencher une nouvelle association de pointés et donc une nouvelle origine, y compris pour un même événement. Les origines sont sélectionnées sur la base d'un score, qui tient compte de plusieurs critères comme la RMS, la valeur des résidus, voire le nombre de pointés P et S associés ou non associés comme c'est le cas du deuxième processus d'association qui se base sur le clustering de pointés.

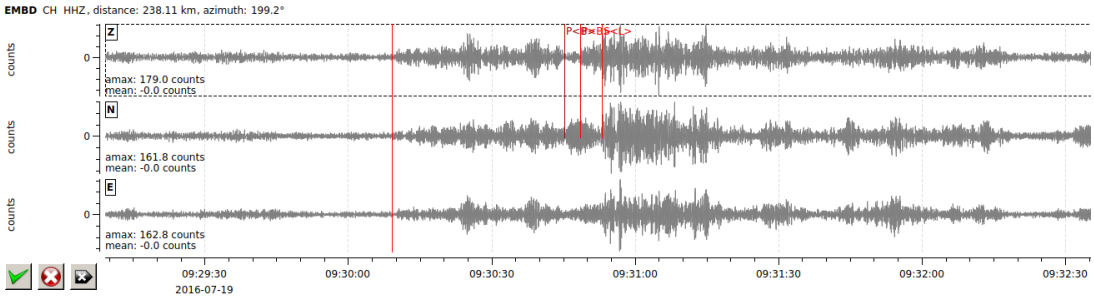
Pour chaque événement, le système de détection sélectionne parmi l'ensemble des origines une seule origine préférentielle en se basant principalement sur des critères comme la valeur de la RMS la plus basse ou le nombre maximal de phases. Ceci signifie qu'une origine contenant le plus grande nombre de phases et la plus petite valeur de RMS est considérée comme l'origine préférentielle, c'est-à-dire l'origine localisée avec la meilleure précision.

Cependant, ces deux critères ne sont pas suffisants pour sélectionner avec robustesse la meilleure origine. En effet, par exemple, le tir de la carrière de Dotternhausen, située dans la région d'Albstadt en Allemagne, qui a eu lieu le 19 juillet 2016 à 09h30, présente une origine préférentielle automatique estimée à 7 phases avec une RMS de 0.69 s (Figure 4.50a). Cette origine correspond effectivement à la meilleure combinaison nombre de phases maximal et RMS minimale.

Seulement, cette origine a été déclenchée suite à une association de pointés qui contient un faux pointé émis à la station EMBD, située au coeur du Valais suisse (Figure 4.50b). L'analyse des signaux montre que ce faux pointé correspond à du bruit et détériore la qualité de la localisation épicentrale (plus particulièrement latitudinale), hypocentrale (qui est fixée à la valeur par défaut de 2 km) et du calcul de la magnitude locale sur la composante verticale (MLv), qui est de 1.40. Sans prendre en compte ce faux pointé à cette station, l'origine préférentielle est estimée à partir de 6 phases avec une RMS égale à 0.69 s (Figure 4.51). Alors que les incertitudes des localisations épicentrales et hypocentrales restent élevées du fait du faible nombre de phases et des incertitudes liées au modèle de vitesse, la magnitude diminue, passant de 1.40 à 1.20. De ce fait, choisir cette dernière origine comme préférentielle éviterait de contaminer cet événement par un faux pointé, et améliorerait non seulement sa localisation mais aussi l'estimation de sa magnitude.

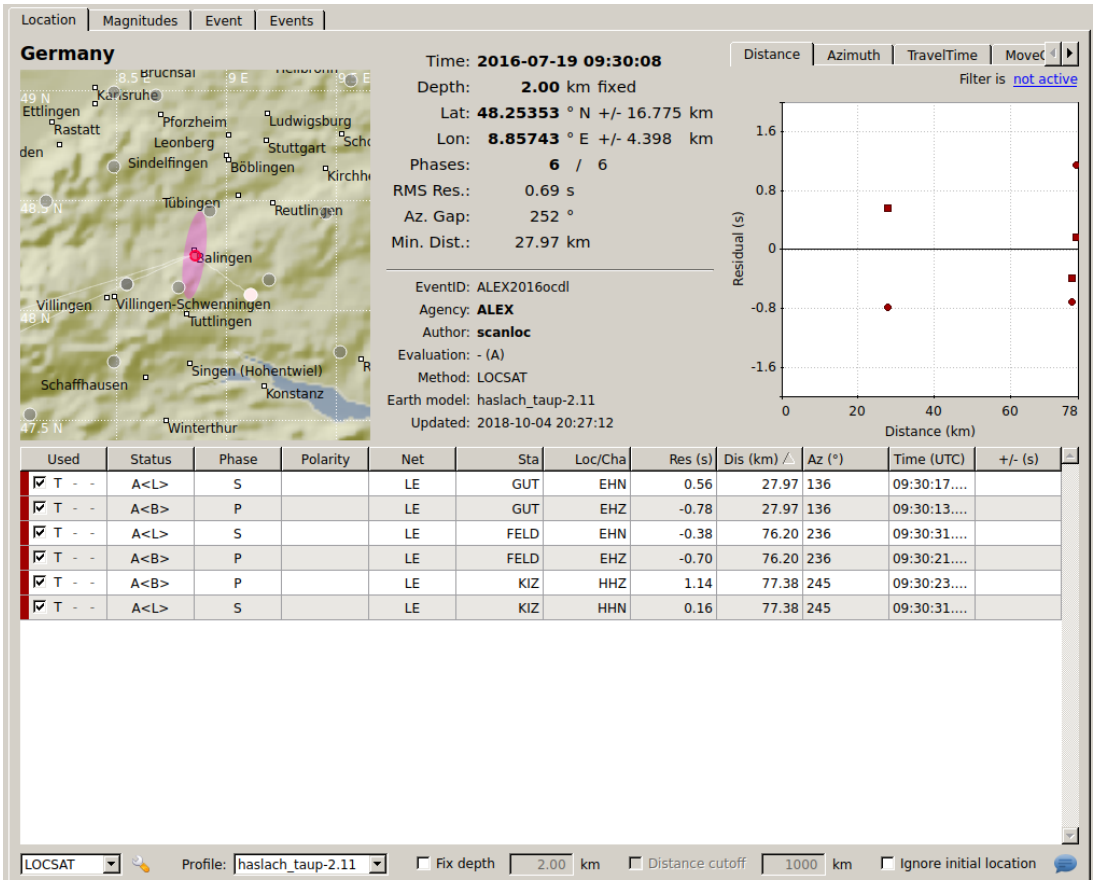


a) Origine préférentielle sélectionnée par le système de détection SeisComP3



b) Faux pointé émis à la station EMBD située à une distance épicentrale de 238 km et intégré à l'origine préférentielle définie par le système de détection SeisComP3

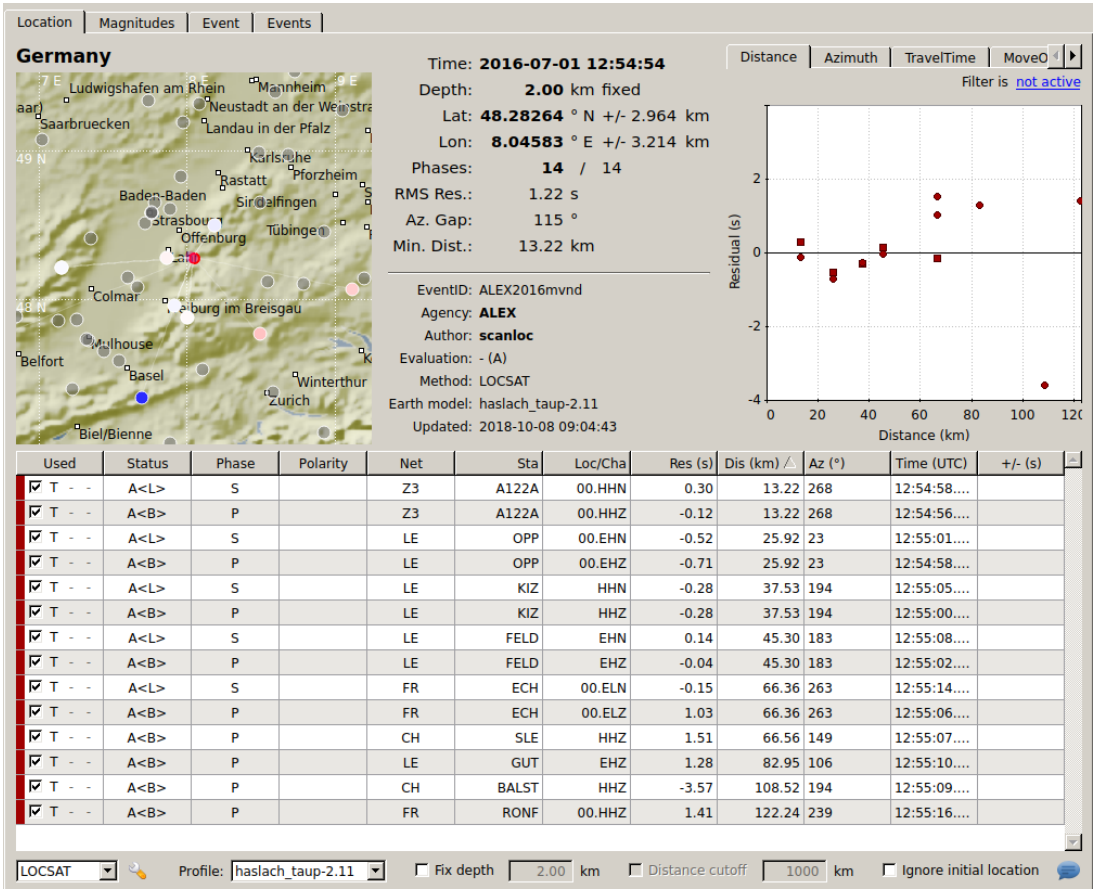
FIGURE 4.50: Exemple de défaillance de la procédure de sélection de l'origine préférentielle établie par le système de détection de SeisComP3. La figure présente deux origines pour un même événement correspondant à un tir de la carrière de Dotternhausen identifié le 19 juillet 2016 à 09h30 environ.



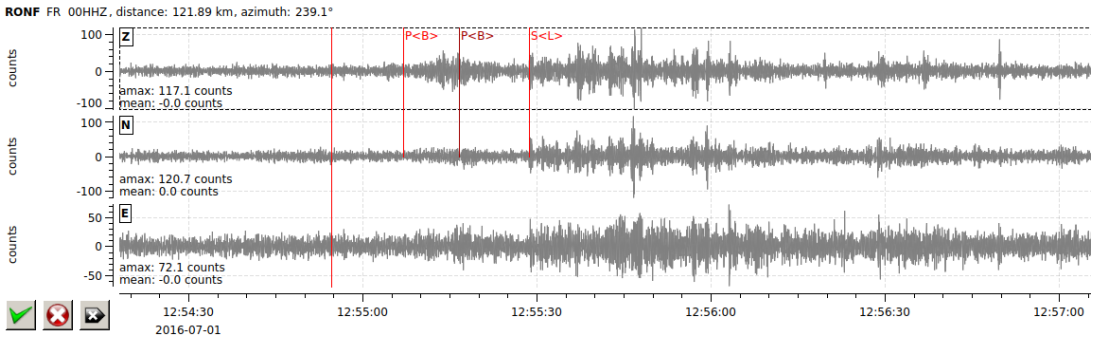
Origine préférentielle sans la prise en compte du faux pointé émis à la station EMBD située à une distance épicentrale de 238 km

FIGURE 4.51: Exemple de défaillance de la procédure de sélection de l'origine préférentielle établie par le système de détection de SeisComP3. La figure présente deux origines pour un même événement correspondant à un tir de la carrière de Dotternhausen identifié le 19 juillet 2016 à 09h30 environ.

De même, pour le tir de la carrière de Schuttertal, situé au coeur du Massif de la Forêt Noire, identifié le 01 juillet 2016 à 12h54 (MLv 0.9), le système de détection de SeisComP3 identifie une origine préférentielle pour cet événement à partir de 14 phases avec une RMS de 1.22 s (Figure 4.52). Seulement, là encore, cette origine a intégré un faux pointé émis à la station RONF (distance épicentrale 122 km). Ce faux pointé correspond également à du bruit impulsif pointé (Figure 4.52b). La meilleure origine pour cet événement n'intégrant pas le pointé à la station RONF serait alors une origine à 13 phases avec une RMS de 0.82 s (Figure 4.53).

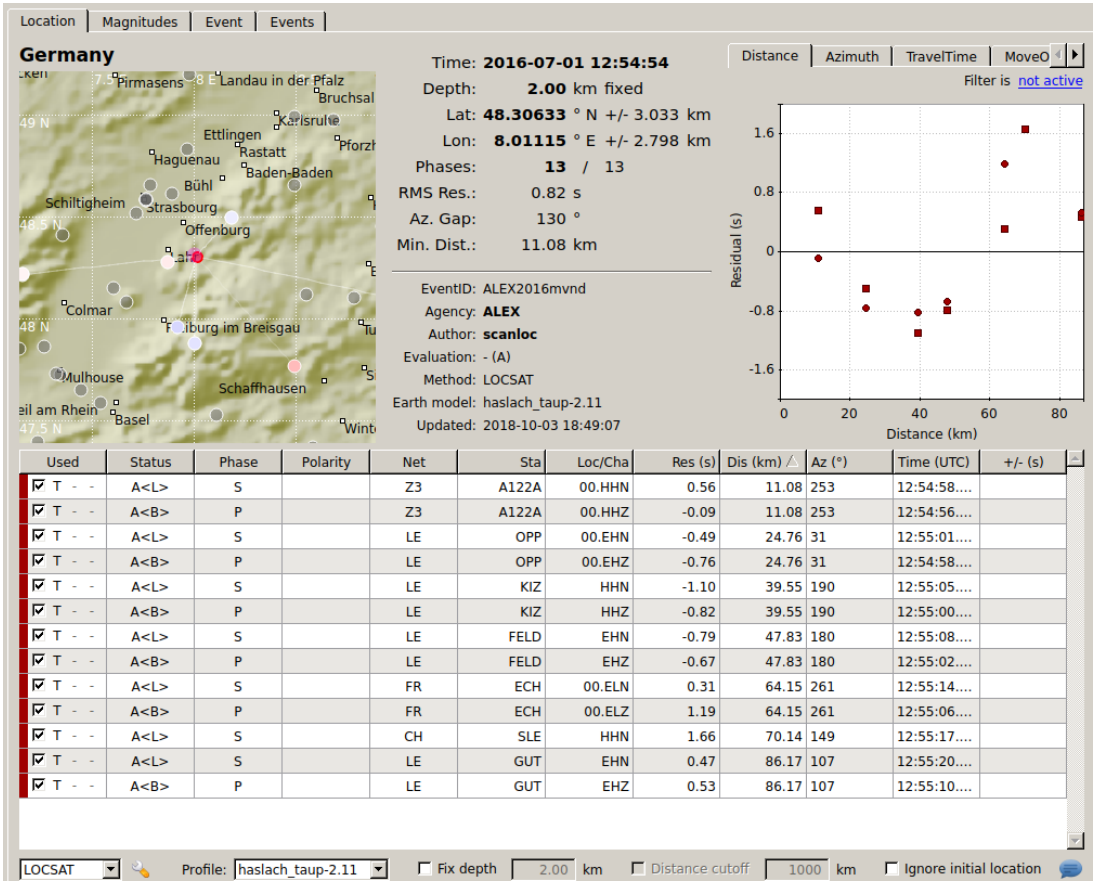


a)Origine préférentielle sélectionnée par le système de détection SeisComP3



b) Faux pointé émis à la station RONF située à une distance épicentrale de 122 km et intégré à l'origine préférentielle définie par le système de détection SeisComP3

FIGURE 4.52: Exemple de défaillance de la procédure de sélection de l'origine préférentielle établie par le système de détection de SeisComP3. La figure présente deux origines pour un même événement correspondant à un tir de la carrière de Schuttertal identifié le 01 juillet 2016 à 12h54 environ.



Origine préférentielle sans la prise en compte du faux pointé émis à la station RONF

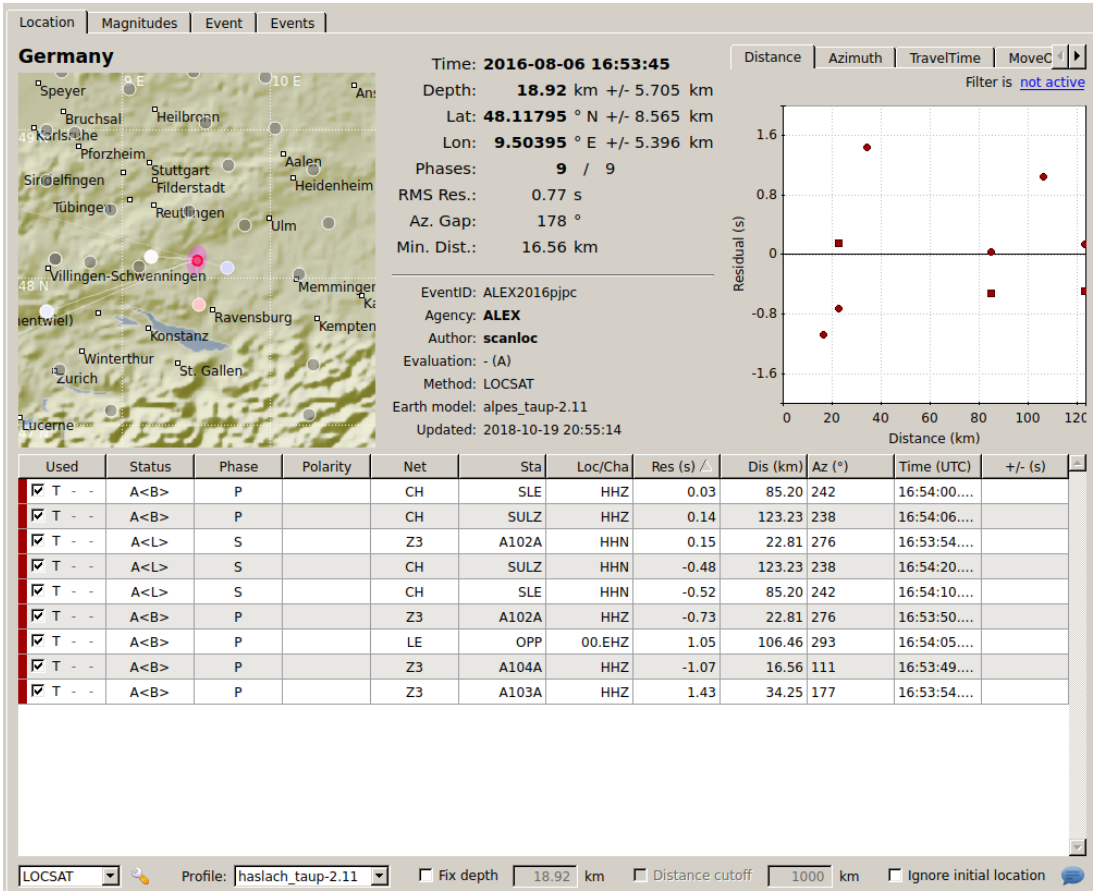
FIGURE 4.53: Exemple de défaillance de la procédure de sélection de l'origine préférentielle établie par le système de détection de SeisComP3. La figure présente deux origines pour un même événement correspondant à un tir de la carrière de Schuttertal identifié le 01 juillet 2016 à 12h54 environ.

Pour le séisme qui a eu lieu le 06 août 2016 au Nord du Lac Konztanz en Allemagne à 16h53 (MLv 0.64), l'origine préférentielle qui est sélectionnée par le système de détection de SeisComP3 est une origine déterminée à partir de 9 phases et une RMS de 0.77 s (Figure 4.54). Celle-ci a été localisée à partir d'une association de pointés qui contient un faux pointé émis à la station A103A, augmentant les incertitudes épicentrales et hypocentrales. Seulement, ici il ne s'agit pas d'un pointé relié à du bruit mais à un faux pointé du temps d'arrivée des ondes P (Figure 4.54b). Même si le vrai pointé automatique P a bien été produit, c'est le faux pointé P retardé qui a finalement été inclus dans le processus d'association, probablement parce que l'instance du processus d'association considère une vitesse moyenne apparente des ondes P de la croûte continentale (6 km/s) et un modèle de vitesse (Haslach) évaluant des temps de trajet trop rapides par rapport aux vrais temps d'arrivées observés.

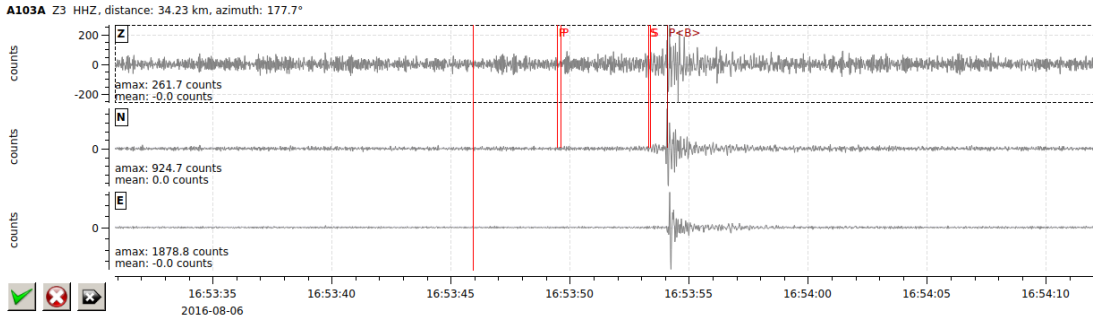
Dans le cas de ce séisme, la vraie origine préférentielle est alors une origine à 7 phases avec une RMS de 0.47 seconde (Figure 4.55. La station A103A n'est alors pas intégrée au processus de localisation de cette origine.

Sachant que les données acquises durent les mois de juillet et août 2016 ont servi à développer la procédure de détection proposée dans cette étude, la combinaison des différentes instances des processus d'association avec une procédure optimisée du choix de l'origine préférentielle apparaît bien indispensable pour détecter des origines préférentielles robustes, nettoyées de tout faux pointé.

Enfin, pour le tir de la carrière d'Arcey, situé dans la zone pré-jurassienne française, ayant eu lieu le 11 juillet 2016 à 15h12 (MLv 1.6), l'origine préférentielle définie par le système de détection de SeisComP3 est une origine à 15 phases avec une RMS de 1.14 s alors que la véritable origine préférentielle est une autre origine à 15 phases mais avec une RMS plus petite (1.02 s). Seulement, étant donné que cette dernière origine a été émise plus précocement, le système de détection privilégie, pour un même nombre de phases et une RMS équivalente, l'origine la plus tardive. Par conséquent, ce système gardera comme origine préférentielle une origine contenant deux faux pointés reliés à du bruit, émis à la station A122A (distance épicentrale de 121 km) et A103A (distance épicentrale de 215 km), et localisée avec de très fortes incertitudes épicentrales et hypocentrales (Figure 4.56).

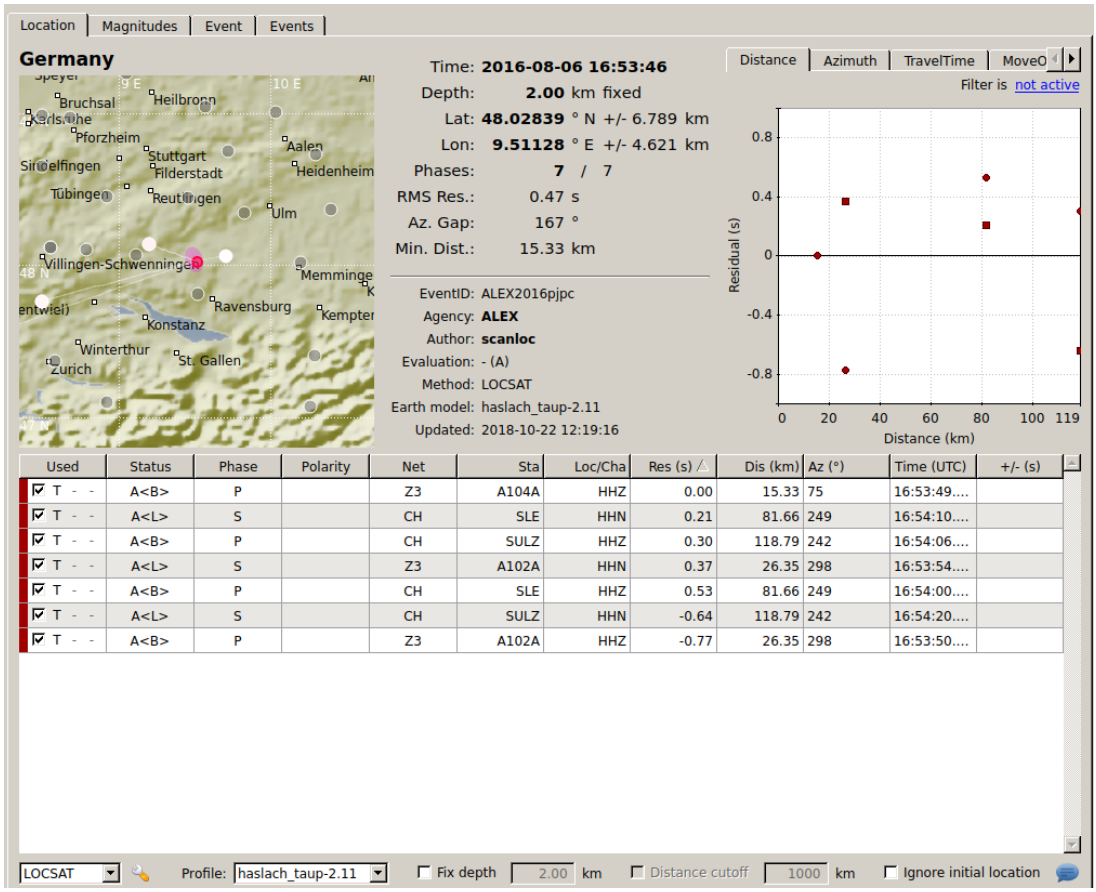


a) Origine préférentielle sélectionnée par le système de détection SeisComP3



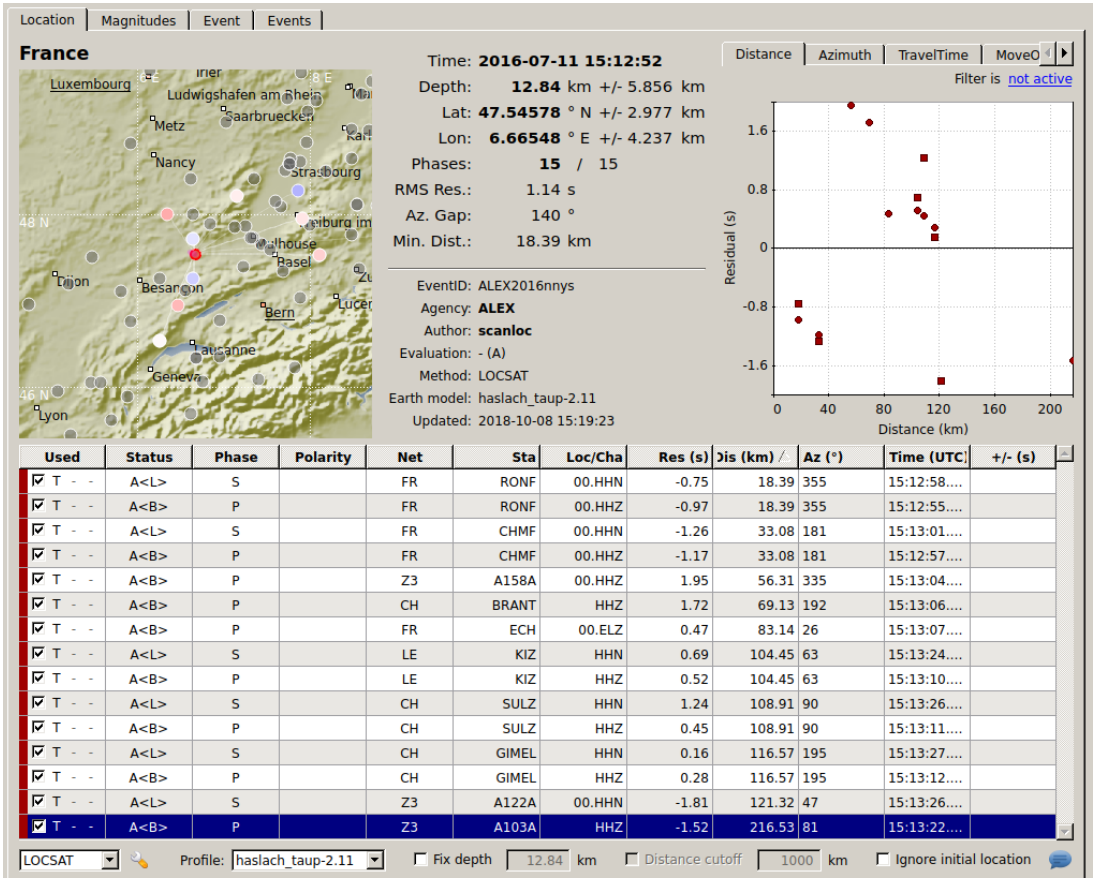
b) Faux pointé P émis à la station A103A située à une distance épicentrale de 34.25 km et intégré à l'origine préférentielle définie par le système de détection SeisComP3

FIGURE 4.54: Exemple de défaillance de la procédure de sélection de l'origine préférentielle établie par le système de détection de SeisComP3. La figure présente deux origines pour un même événement correspondant à un séisme identifié le 06 août 2016 à 16h53 environ.



Origine préférentielle sans la prise en compte du faux pointé P émis à la station A103A

FIGURE 4.55: Exemple de défaillance de la procédure de sélection de l'origine préférentielle établie par le système de détection de SeisComP3. La figure présente deux origines pour un même événement correspondant à un séisme identifié le 06 août 2016 à 16h53 environ.



a)Origine préférentielle sélectionnée par le système de détection SeisComP3

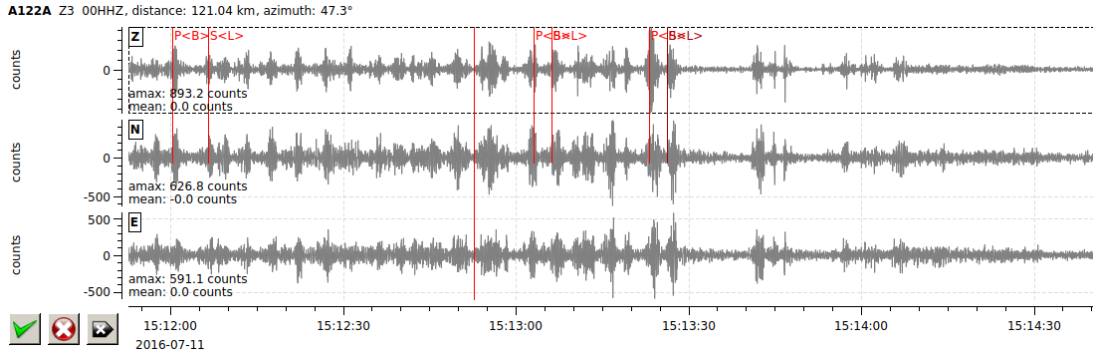
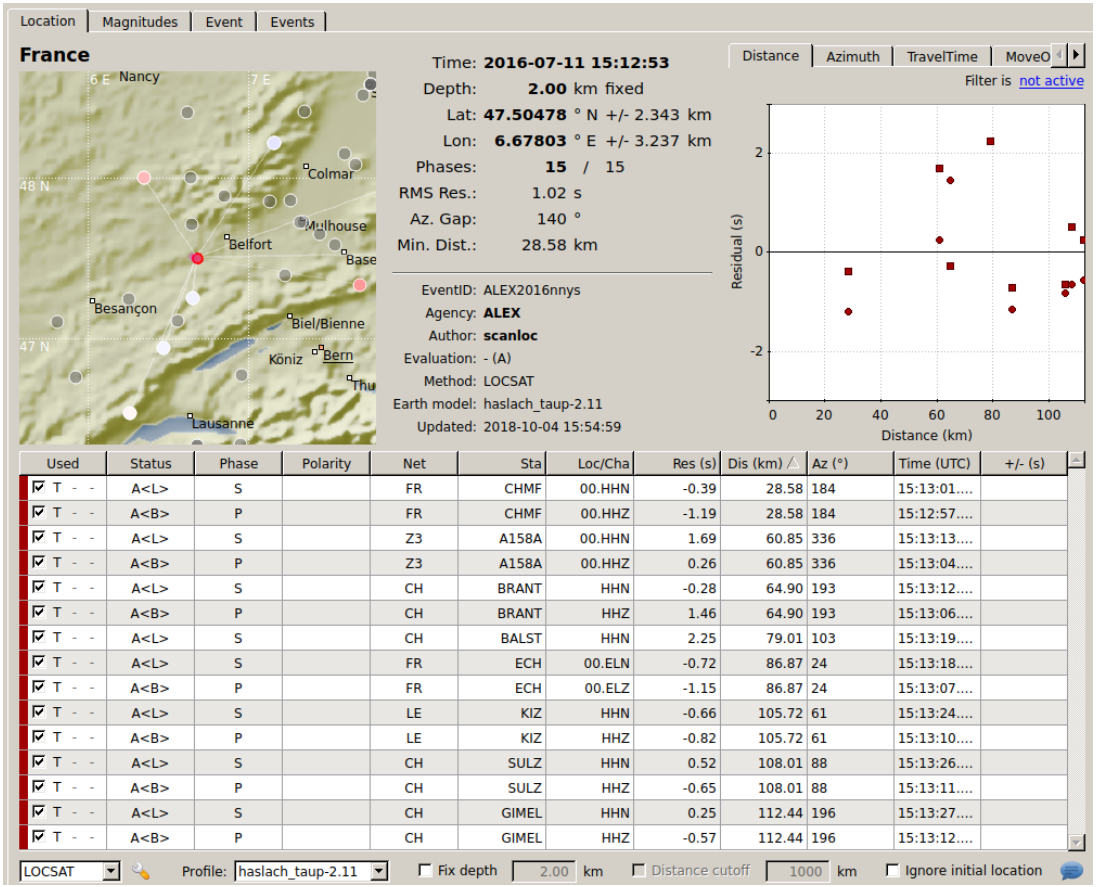


FIGURE 4.56: Exemple de défaillance de la procédure de sélection de l'origine préférentielle établie par le système de détection de SeisComP3. La figure présente deux origines pour un même événement correspondant à un tir de la carrière d'Arcey identifié le 11 juillet 2016 à 15h12 environ.



Used	Status	Phase	Polarity	Net	Sta	Loc/Cha	Res (s)	Dis (km)	Az (°)	Time (UTC)	+/- (s)
T	A<L>	S		FR	CHMF	00.HHN	-0.39	28.58	184	15:13:01....	
T	A<B>	P		FR	CHMF	00.HHZ	-1.19	28.58	184	15:12:57....	
T	A<L>	S		Z3	A158A	00.HHN	1.69	60.85	336	15:13:13....	
T	A<B>	P		Z3	A158A	00.HHZ	0.26	60.85	336	15:13:04....	
T	A<L>	S		CH	BRANT	HHN	-0.28	64.90	193	15:13:12....	
T	A<B>	P		CH	BRANT	HHZ	1.46	64.90	193	15:13:06....	
T	A<L>	S		CH	BALST	HHN	2.25	79.01	103	15:13:19....	
T	A<L>	S		FR	ECH	00.ELN	-0.72	86.87	24	15:13:18....	
T	A<B>	P		FR	ECH	00.ELZ	-1.15	86.87	24	15:13:07....	
T	A<L>	S		LE	KIZ	HHN	-0.66	105.72	61	15:13:24....	
T	A<B>	P		LE	KIZ	HHZ	-0.82	105.72	61	15:13:10....	
T	A<L>	S		CH	SULZ	HHN	0.52	108.01	88	15:13:26....	
T	A<B>	P		CH	SULZ	HHZ	-0.65	108.01	88	15:13:11....	
T	A<L>	S		CH	GIMEL	HHN	0.25	112.44	196	15:13:27....	
T	A<B>	P		CH	GIMEL	HHZ	-0.57	112.44	196	15:13:12....	

Origine préférentielle créée sans les faux pointés émis aux stations A122A et A103A

FIGURE 4.57: Exemple de défaillance de la procédure de sélection de l'origine préférentielle établie par le système de détection de SeisComP3. La figure présente deux origines pour un même événement correspondant à un tir de la carrière d'Arcey identifié le 11 juillet 2016 à 15h12 environ.

L'utilisation combinée du nombre maximal de phases et de la valeur minimale de la RMS ne suffit pas à définir l'origine préférentielle de manière robuste. Il y a donc nécessité de reconsidérer cette solution d'origine préférentielle en apportant d'autres critères (que ceux prédéfinis par SeisComP3) évaluant la qualité d'une origine qui puissent être facilement accessibles depuis la base de données des événements.

4.3.2 Définir des critères pour optimiser la sélection

Les incertitudes de localisation latitudinale et longitudinale apportent une information capitale indirecte pour évaluer si une origine a sa localisation polluée par d'éventuels faux pointés ou pour révéler quelle peut être la meilleure origine parmi celles qui présentent un même nombre de phases et une RMS équivalente.

De même, la valeur des résidus peut aider à reconnaître des origines contaminées par des faux pointés de part l'existence de résidus aberrants supérieurs à 3.5 secondes. Les valeurs des résidus peuvent également être très utiles pour repérer les origines avec les plus petites valeurs de résidus temporels (inférieures à 1.5 s).

Les distances épicentrales sont aussi des critères liés à la géométrie du réseau de stations à considérer (Bondar et al., 2004). La distance épicentrale minimale peut être utilisée pour départager des origines qui ont un nombre de phases, une valeur de RMS et des incertitudes de localisation épicentrales qui sont du même ordre de grandeur. La distance épicentrale maximale est aussi intéressante à prendre en compte car une valeur élevée peut indiquer l'existence d'un pointé aberrant émis à une station fortement éloignée géométriquement du reste des stations impliquées dans la localisation de l'origine, malgré des différences de temps d'arrivée et des résidus temporels qui peuvent être faibles.

La profondeur peut être également un paramètre intéressant pour exclure de la sélection des origines localisées à des profondeurs excessives (très supérieures à 30 km) du fait de l'existence d'un ou plusieurs faux pointé(s) associés au reste des vrais pointés.

Enfin, le nombre de phases S fournit une contrainte importante sur la profondeur focale des événements et sur les incertitudes de localisation (Gomberg et al., 1990 ; Husen et Hardebeck, 2010). Comme il est possible de l'observer dans les Figures 4.29 à 4.39, sans le pointé des temps d'arrivée des ondes S, la profondeur focale des événements est peu contrainte. Même si le pointé des temps d'arrivée des ondes S n'apporte pas nécessairement de l'exactitude (le pointé des ondes S est difficile du fait des conversions de phases notamment), le nombre de phases S inclus dans l'assemblage de pointés apporte de la précision

à la localisation et peut départager les origines qui ont un nombre de phases et une valeur de RMS similaires.

Le nombre de phases et la valeur de la RMS restent des paramètres clefs dans la sélection de l'origine préférentielle, en complément et en appui des autres critères décrits. Ces deux paramètres clefs servent également à affiner la sélection de l'origine préférentielle à travers une recherche itérative de la meilleure origine basée sur un intervalle de valeurs autour du nombre maximal de phases qui minimise à la fois la RMS et les imprécisions de localisation qu'évaluent les autres critères décrits ci-dessus.

4.3.3 Créer un module SeisComP3 qui détermine une meilleure origine préférentielle

L'analyse de l'ensemble des 708 événements détectés pour la période juillet-août 2016 a conduit à l'élaboration d'un arbre décisionnel qui initie la sélection de l'origine préférentielle à travers deux seuils de référence : le nombre maximal de phases (plusieurs valeurs seuils possibles) et la valeur minimale de RMS (un unique seuil égal à 2 s). Les autres critères (nombre de phases S, nombre de résidus temporels supérieurs à 1.5 s, nombre de résidus temporels $>$ 3.5 s, distance épicentrale minimale, distance épicentrale maximale, profondeur, incertitude latitudinale, incertitude longitudinale) viennent en appui pour affiner la sélection.

L'élaboration de l'arbre décisionnel a donc été effectuée manuellement après une étude précise de l'ensemble des critères potentiels pouvant influencer le choix de l'origine préférentielle pour tous les vrais événements détectés au cours des mois de juillet et août 2016. Cet arbre construit a été intégré dans un module SeisComP3 que j'ai codé et a été testé sur un jeu d'événements détectés par le BCSF-RéNaSS entre janvier et juillet 2016 puis un jeu d'événements détectés uniquement automatiquement selon la procédure développée dans ce travail de thèse pour les mois compris entre septembre et décembre 2016.

De cette façon, pour chaque événement détecté au cours du mois de juillet et août, j'ai vérifié son origine préférentielle. Si celle-ci ne correspondait pas à la véritable origine préférentielle (intégration d'un faux pointé par exemple dans l'association qui a conduit à l'origine préférentielle déclarée par SeisComP3), j'ai alors recherché d'autres critères, différents de ceux utilisés par défaut (RMS et nombre de phases), qui pourraient, s'ils étaient utilisés, faire basculer la sélection vers la véritable origine préférentielle.

L'arbre décisionnel construit manuellement répond donc à plusieurs chemins décisionnels possibles. Ces chemins décisionnels représentent l'ensemble des choix disponibles élaborés à partir de la valeur des critères utilisés pour sélectionner chaque origine (nombre de phases, RMS, nombre de phases S, nombre de résidus temporels supérieurs à 1.5 s, nombre de résidus temporels $>$ 3.5 s, distance épicentrale minimale, distance épicentrale maximale, profondeur, incertitude latitudinale, incertitude longitudinale), et ce, pour un large spectre de configurations possibles. Le facteur qui est finalement hautement

considéré pour optimiser le choix de l'origine préférentielle est la qualité de sa localisation.

Cet arbre décisionnel est donc d'abord construit sur la base du nombre maximal de phases qui est identifié pour chaque événement. Lorsque le nombre maximal de phases est supérieur à 20 ou bien inférieur 6, le choix de l'origine préférentielle est plus rapide et se base uniquement sur une recherche de l'origine préférentielle à travers la combinaison simultanée de la valeur minimale de la RMS et le nombre maximal de phases possible (Figure 4.58).

En revanche, le choix se complexifie pour un nombre de phases compris entre 6 et 19. En effet, en plus d'une recherche itérative de la meilleure combinaison possible entre la valeur de RMS minimale et le nombre de phases maximal, des critères supplémentaires tels que le nombre de phases S, la distance épicentrale minimale et les incertitudes de localisation épicentrale vont aider à affiner le diagnostic de cette nouvelle origine préférentielle (Figures 4.59, 4.60 et 4.61). Les autres critères comme la profondeur, les valeurs des résidus et la distance épicentrale maximale sont principalement utilisés pour rejeter les origines aberrantes quand c'est possible (Figure 4.58). Néanmoins, ces derniers critères sont très précieux lorsqu'il s'agit d'affiner la sélection d'une origine préférentielle pour les événements dont les origines présentent très peu de phases (généralement inférieur ou égal à 7, Figures 4.60 et 4.62).

Cet arbre décisionnel, présenté dans les Figures 4.58, 4.59, 4.60, 4.61 et 4.62, a été implémenté dans un module SeisComP3 que j'ai développé. Ce module, écrit en Python, extrait les différentes critères (RMS, nombre de phases, nombre de phases S, nombre de résidus temporels supérieurs à 1.5 s, nombre de résidus temporels $>$ 3.5 s, distance épicentrale minimale, distance épicentrale maximale, profondeur, incertitude latitudinale, incertitude longitudinale) pour chaque origine de chaque événement. En fonction de la valeur du nombre maximal de phases identifié pour chaque événement, relativement à la plus faible valeur de RMS, ce module évalue l'origine préférentielle en utilisant l'arbre décisionnel empiriquement construit. Celui-ci récupère finalement l'identifiant (ID) de l'origine préférentielle sélectionnée. Il notifie ensuite le système de messagerie du changement en activant le protocole de mise-à-jour de l'origine préférentielle. Le module de gestion des événements modifie alors dans la base de données des événements l'origine préférentielle par défaut en validant la nouvelle origine préférentielle.

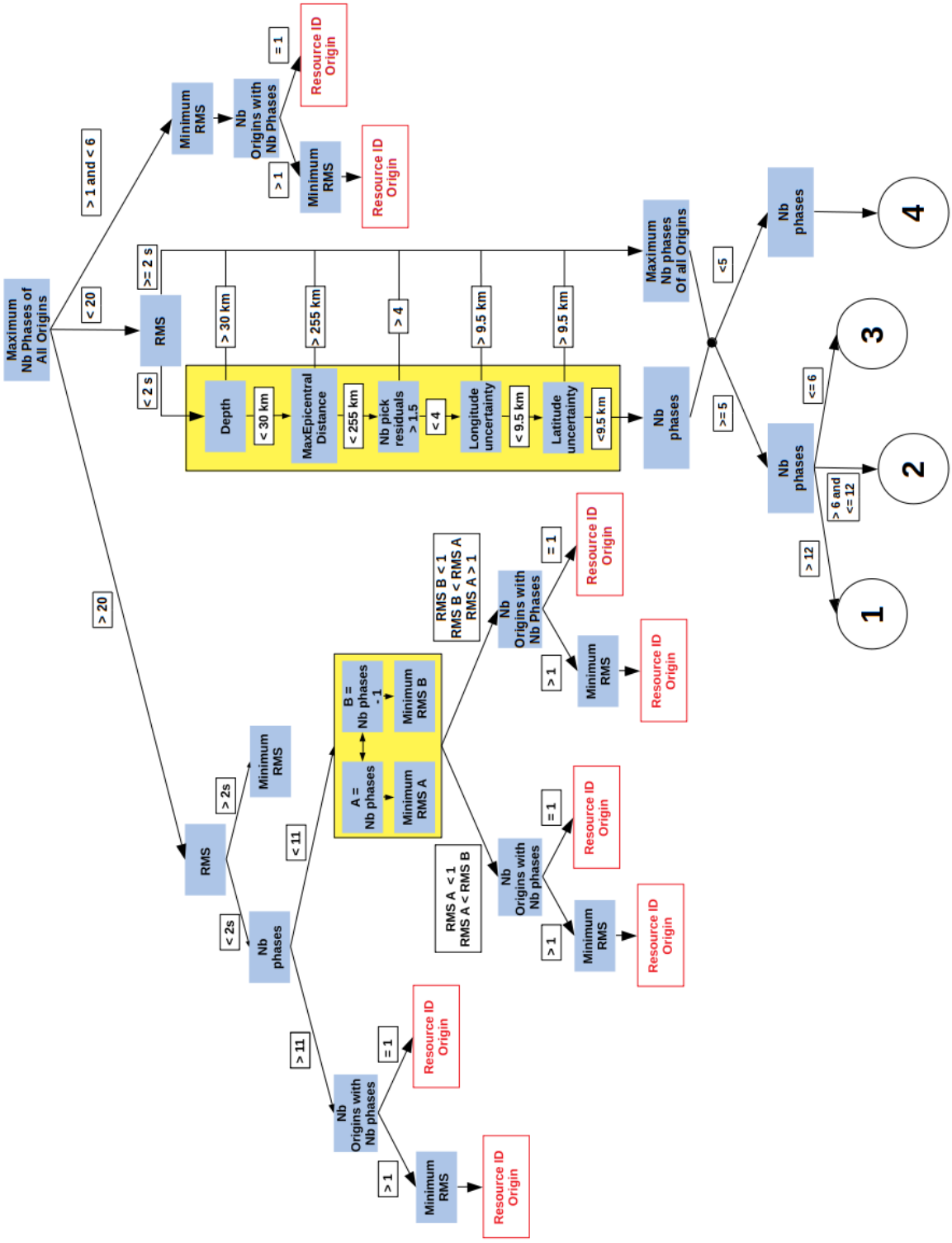


FIGURE 4.58: Architecture générale de l'arbre décisionnel qui sert à sélectionner la nouvelle origine préférentielle en se basant d'abord sur le nombre maximal de phases qui sont présentes pour chaque événement. A l'issue de la recherche de la nouvelle origine préférentielle, c'est l'identifiant de l'origine sélectionnée (ressource ID) qui est récupéré.

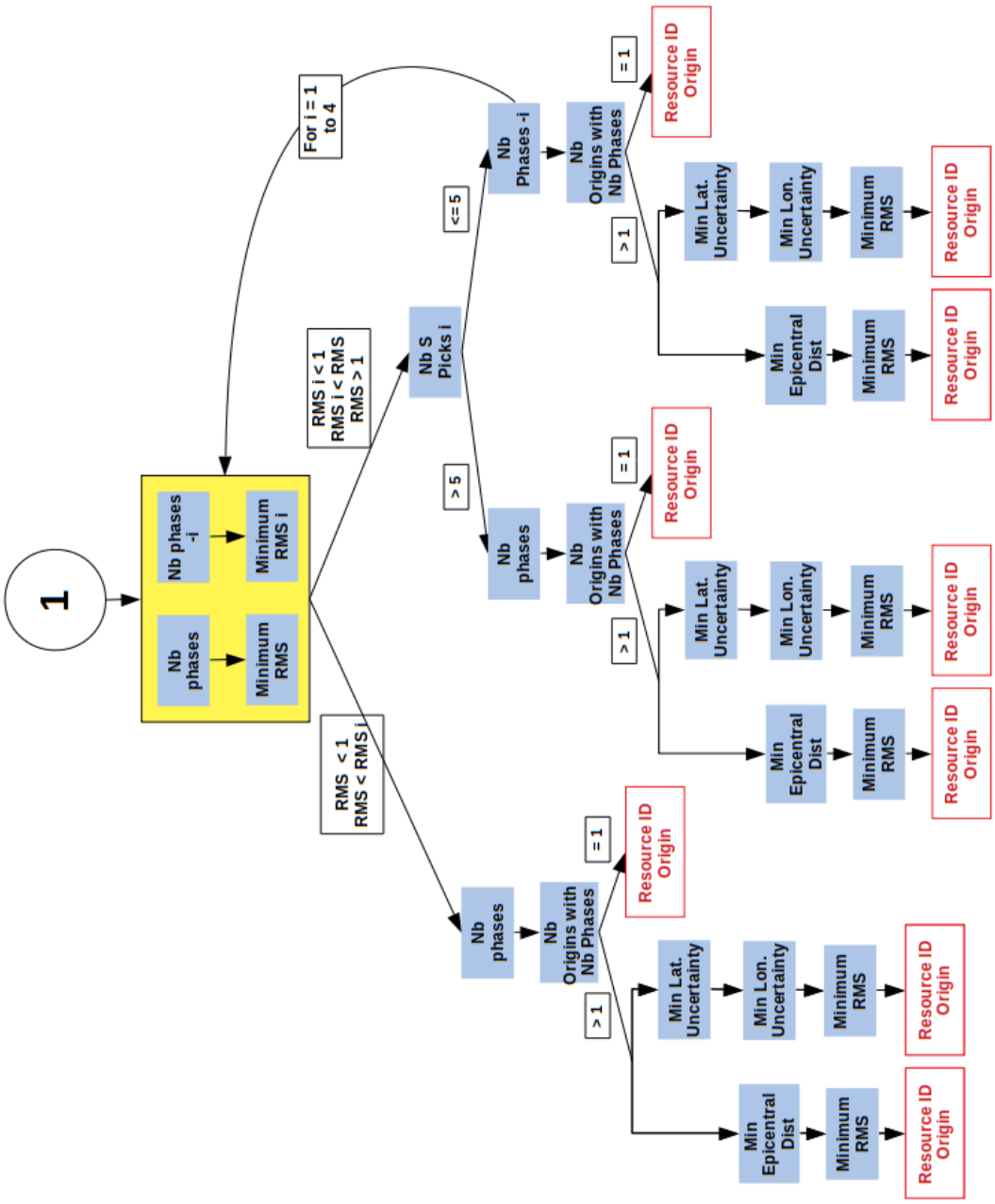


FIGURE 4.59: Composante plus détaillée (cercle numéroté 1 sur la figure 4.58) de l'arbre décisionnel. Cette composante sert à sélectionner la nouvelle origine préférentielle pour des événements dont le nombre maximal de phases, pour la valeur de RMS la plus faible, est compris entre 13 et 19 phases. A l'issue de la recherche de la nouvelle origine préférentielle, c'est l'identifiant de l'origine sélectionnée (ressource ID) qui est récupéré.

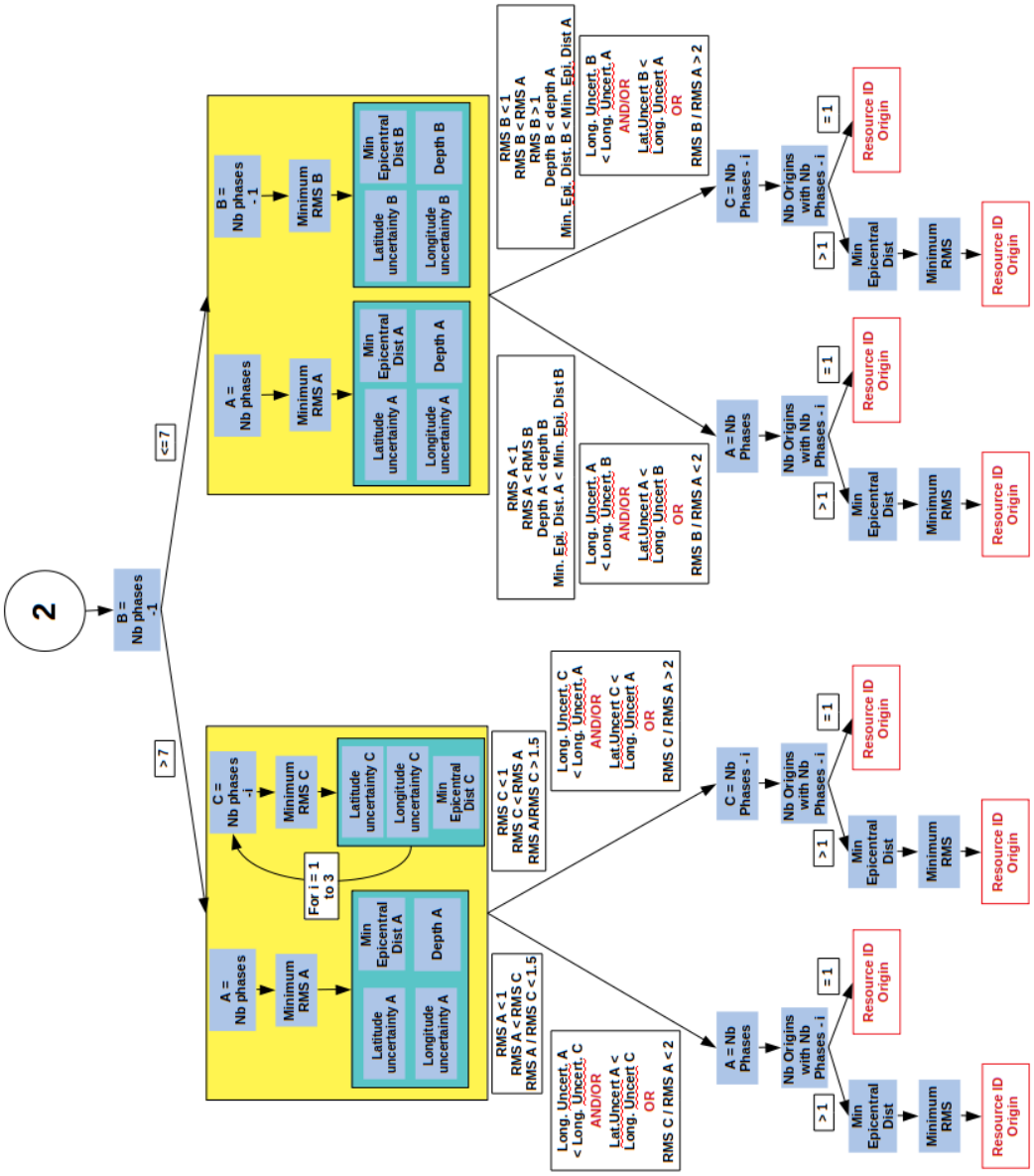


FIGURE 4.60: Composante plus détaillée (cercle numéroté 2 sur la figure 4.58) de l'arbre décisionnel. Cette composante sert à affiner la sélection de la nouvelle origine préférentielle pour des événements dont le nombre maximal de phases, pour la valeur de RMS la plus faible, est compris entre 7 et 12 phases. A l'issue de la recherche de la nouvelle origine préférentielle, c'est l'identifiant de l'origine sélectionnée (ressource ID) qui est récupéré.

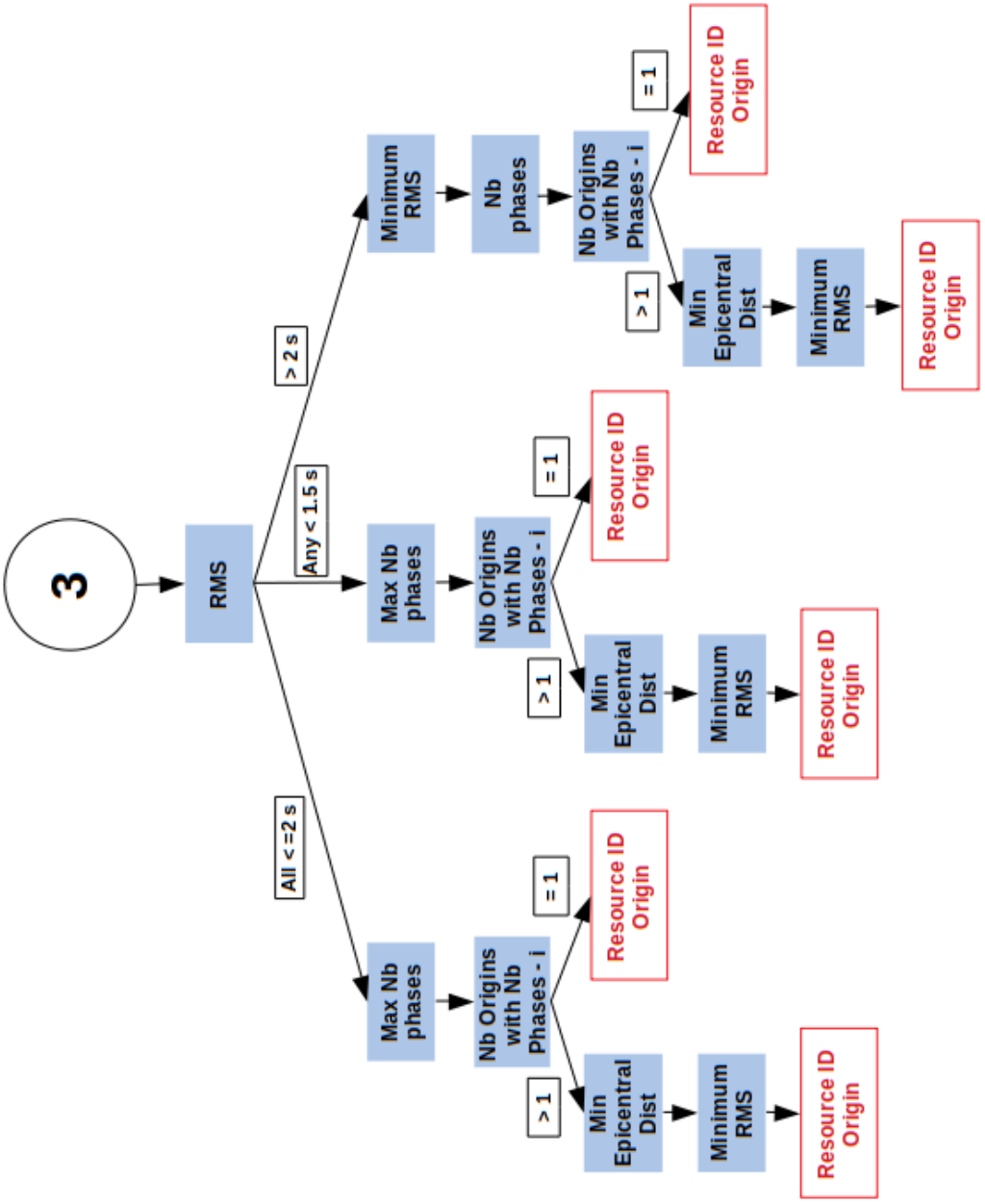


FIGURE 4.61: Composante plus détaillée (cercle numéroté 3 sur la figure 4.58) de l'arbre décisionnel. Cette composante sert à affiner la sélection de la nouvelle origine préférentielle pour des événements dont le nombre maximal de phases, pour la valeur de RMS la plus faible, est compris entre 5 et 6 phases. A l'issue de la recherche de la nouvelle origine préférentielle, c'est l'identifiant de l'origine sélectionnée (ressource ID) qui est récupéré.

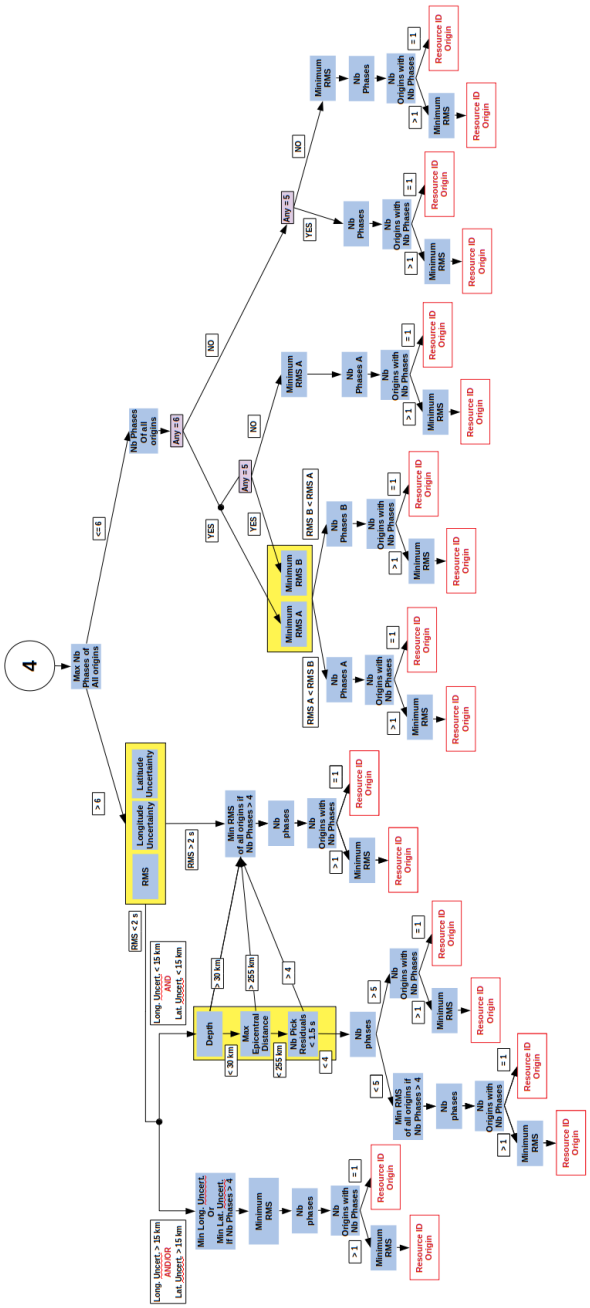


FIGURE 4.62: Composante plus détaillée (cercle numéroté 4 sur la figure 4.58) de l'arbre décisionnel. Cette composante sert à affiner la sélection de la nouvelle origine préférentielle pour des événements dont le nombre maximal de phases, pour la valeur de RMS la plus faible, est inférieur à 5 phases. A l'issue de la recherche de la nouvelle origine préférentielle, c'est l'identifiant de l'origine sélectionnée (ressource ID) qui est récupéré.

4.4 Récapitulatif

L'amélioration de la qualité des pointés automatiques P et S, du processus d'association ainsi que du processus de sélection de l'origine préférentielle conditionne fortement la réduction du nombre de séismes détectés avec un ou plusieurs faux pointés.

La prise en compte de la configuration du réseau de stations (géométrie, localisation, site d'implantation) est d'abord déterminante pour obtenir des pointés automatiques P et S, de meilleure qualité.

De plus, la considération du milieu de propagation, à travers la vitesse des ondes sismiques, est ensuite capitale pour améliorer le processus d'association, et limiter l'association de faux pointés avec de vrais pointés. En effet, ce milieu impacte fortement les temps de trajet qui sont calculés au sein du réseau de stations, que ce soit pour évaluer les résidus temporels associés aux différents pointés, ou pour évaluer les distances temporelles nécessaires pour former les clusters de pointés.

Enfin, la gestion de critères supplémentaires qui vont évaluer la qualité d'une origine localisée est décisive pour améliorer le processus de sélection de l'origine préférentielle parmi les origines qui constituent chaque événement. Si le nombre maximal de phases et la valeur minimale de la RMS sont suffisants pour estimer des événements qui sont générés avec beaucoup de phases (supérieurs à 20), ceci est beaucoup moins évident pour ceux qui en possèdent moins.

Or, la procédure de détection est ici établie pour détecter des signaux de faible amplitude correspondant à des séismes qui sont enregistrés à un faible nombre de stations. Les critères comme le nombre de phases S, les distances épicentrales maximale et minimale, la profondeur, le nombre de pointés avec des résidus supérieurs à 1.5 s et 3.5 s, les incertitudes de localisation latitudinales et longitudinales viennent donc affiner le processus de sélection de l'origine préférentielle. L'origine qui est finalement choisie est l'origine qui a certes le nombre maximal de phases et la plus petite valeur de RMS, mais qui est aussi localisée avec la plus grande précision.

La connaissance des caractéristiques du bruit enregistré aux stations et du milieu de propagation des ondes sismiques, couplée à la prise en compte des distances épicentrales et des facteurs évaluant la qualité des localisations des événements, conditionnent donc la performance de la détection finale des séismes. Cette détection dépend donc fortement de la nature du signal enregistré, qui reflète l'influence combinée des effets de la source, souvent très atténués, de la propagation des ondes dans le milieu et du bruit enregistré aux stations.

Face à une performance de détection des petits séismes multifactorielle, le nouveau système de détection contient :

— plusieurs instances de pointés P et S (dans notre cas 2) qui ont été introduites pour répondre à l'évolution spatio-temporelle des conditions de bruit aux stations et des distances épicentrales (Figure 4.63a) ;
— plusieurs instances du processus d'association qui ont été implémentées en parallèle (dans notre cas deux instances pour le processus d'association basé sur la rétro-projection des pointés à un hypocentre optimal, quatre instances pour le processus d'association basé sur le clustering de pointés) pour prendre en compte de façon optimale les variations verticales et latérales du milieu de propagation des ondes sismiques (Figure 4.63b) ;
— un module SeisComP3 qui a été développé pour sélectionner plus solidement l'origine préférentielle de chaque événement puis introduit dans le système de détection final (Figure 4.63c).

Afin d'assurer une synchronisation de l'ensemble des étapes de la nouvelle procédure, j'ai également développé un autre module SeisComP3 qui est intégré à la procédure de détection, directement après toutes les instances des processus d'association. Ce module vérifie en continu le statut de ces différentes instances et autorise la poursuite de la procédure de détection une fois que toutes ces dernières ont terminé leur action, à savoir lorsque toutes les origines ont été créées et localisées.

Si ces développements diminuent fortement la détection de séismes contaminés par de faux pointés, ils ne permettent en revanche pas d'annuler la détection des faux événements. Le chapitre suivant est donc dédié à expliquer comment à partir de l'apprentissage machine supervisé, il est possible de réduire la quantité de faux événements détectés, tout en veillant à identifier également automatiquement les vrais événements restants, à savoir les séismes et les tirs de carrière.

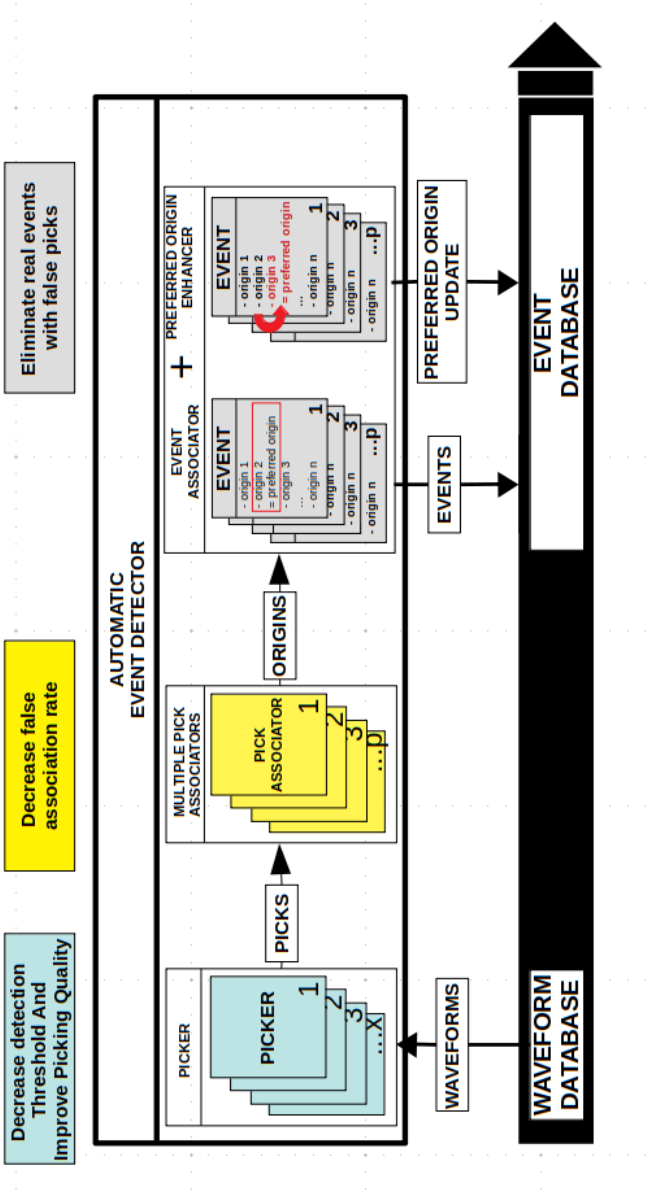


FIGURE 4.63: Procédure de détection nouvellement développée, qui vise à réduire le taux de séismes détectés avec de faux pointés tout en diminuant le seuil avec lequel ces derniers sont détectés. (a) L'instance de pointé automatique des ondes P et S est remplacée par plusieurs instances qui viennent améliorer la qualité des pointés émis en prenant en compte la variabilité spatio-temporelle des caractéristiques du bruit enregistré aux stations et des distances épicentrales. (b) Chaque instance des deux processus d'association (rétro-projection des pointés et clustering des pointés) est remplacée par plusieurs instances qui considèrent plus spécifiquement les variations latérales et verticales du milieu de propagation (modèle de vitesse pour les deux processus d'association et vitesse moyenne apparente des ondes P dans la croûte continentale pour le deuxième processus d'association). (c) Un nouveau module SeisComP3 que j'ai développé est introduit dans le système de détection pour sélectionner la véritable origine préférentielle basée sur d'autres critères (que ceux proposés par SeisComP3) qui évaluent la précision de localisation de cette origine.

Chapitre 5

Comment réduire la détection des faux événements et comment efficacement discriminer les séismes des tirs de carrière ?

Sommaire

5.1 Classer les événements avec l'apprentissage machine supervisé

5.1.1 Trouver une fonction de prédiction qui minimise l'erreur de généralisation

L'énoncé du problème de classification se base sur la supposition qu'il existe une fonction cible inconnue, une fonction d'étiquetage (nommée F), qui va permettre de labéliser en sortie (affilier une étiquette de classe) un ensemble d'événements (ici les faux événements, les séismes et les tirs de carrière) en données d'entrée (Figure 5.6a).

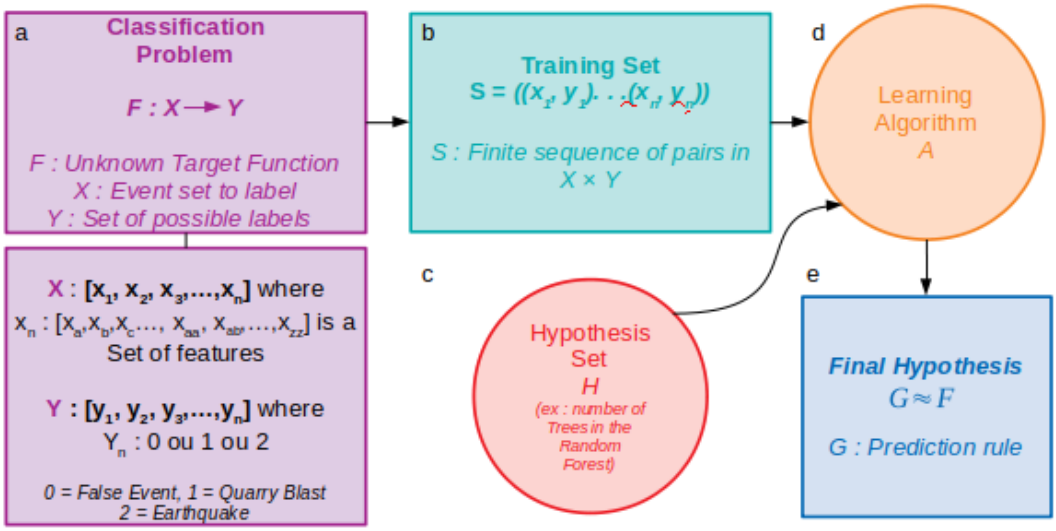


FIGURE 5.1: Cadre théorique de l'apprentissage machine supervisé. (a) Un ensemble d'observations (x_1 à x_n) appartenant au domaine X, avec chaque observation décrite par un vecteur d'attributs (x_a à x_{zz}), est mappé en un ensemble de labels, appartenant au domaine Y, par une fonction cible inconnue F. (b) Une base d'entraînement, échantillonnée à partir des données disponibles, est constituée d'un ensemble de couples (x,y) appartenant à X $\times$ Y, et sert de base d'entrée pour un algorithme d'apprentissage A (c) qui va générer une fonction de prédiction G qui va approximer la fonction cible inconnue (e), en se basant sur un espace initial d'hypothèses restreintes H (d). Cette fonction de prédiction doit être capable de prédire les labels de nouveaux événements en minimisant l'erreur de généralisation.

La résolution de ce problème par l'apprentissage machine supervisé va se baser sur la construction inductive d'une fonction de prédiction généralisable G, qui va approximer la fonction cible inconnue F, en commettant une erreur de prédiction (ou erreur empirique) la plus faible possible (Figure 5.6e). Dans le cadre de la classification, cette fonction de prédiction G est appelée classifieur.

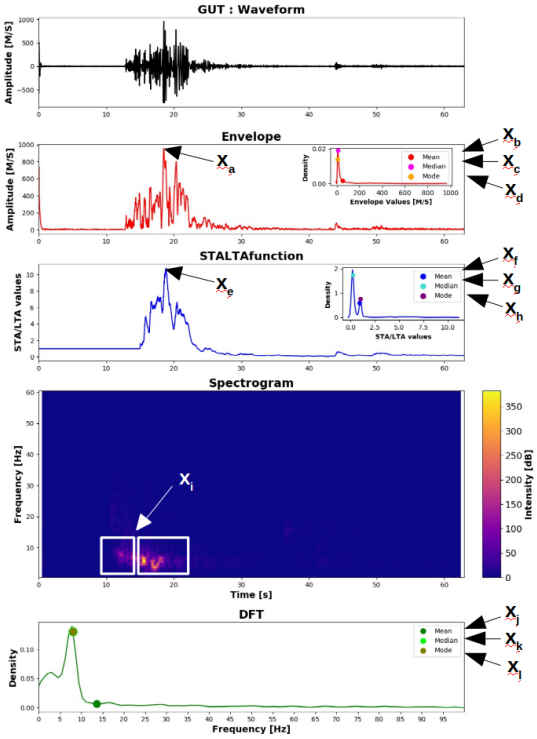
Ce classifieur est construit à partir d'un ensemble fini d'exemples, appelé base d'entraînement, dans lequel chaque exemple est une paire constituée du vecteur représentatif d'une observation, c'est-à-dire ici un vecteur d'attributs décrivant un événement, et d'une réponse associée, à savoir l'étiquette de classe de chaque événement (Figure 5.6b). Les Figures 5.2, 5.3 et 5.4 présentent quelques exemples d'attributs qui peuvent être utilisés pour décrire les événements qui sont à classer, à savoir les faux événements, les tirs de carrière et les séismes.

La base d'entraînement qui est utilisée (nommée S) correspond donc à un échantillonnage de l'ensemble des observations possibles (nommé X) et de leurs réponses associées (nommé Y) dont la distribution D est inconnue (Figure 5.6a, b). Seulement, l'hypothèse de base sous-jacente à l'apprentissage est que les données sont stationnaires, c'est-à-dire que les exemples de la base d'entraînement, sur laquelle la fonction de prédiction est apprise, sont représentatifs du problème général que l'on souhaite résoudre.

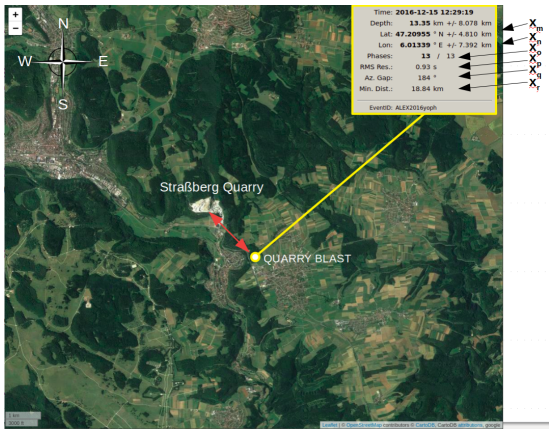
L'objectif de l'apprentissage supervisé est donc de rechercher la fonction de prédiction G qui aura de bonnes performances de généralisation. Autrement dit, la fonction de prédiction G trouvée réalisera une erreur de généralisation très faible. Seulement, cette erreur de généralisation est en fait difficile à estimer car elle est exprimée en fonction de deux paramètres inconnus : la distribution de l'ensemble de toutes les observations possibles D et la fonction d'étiquetage cible F. La seule information disponible est en fait contenue dans la base d'entraînement.

En suivant le principe inductif de la minimisation du risque empirique (MRE, Vapnik, 1999), cette erreur de généralisation sera donc approximée à travers le calcul de l'erreur empirique. En effet, ce principe suppose que la fonction de prédiction qui minimise l'erreur empirique, aboutit à une erreur de généralisation qui est proche de son minimum, et donc offre une borne supérieure à cette erreur de généralisation. Il s'agit alors dans ce nouveau cadre de trouver la fonction de prédiction G qui a l'erreur empirique la plus faible, c'est-à-dire qui minimise l'écart entre la réponse réelle y (l'étiquette de classe) et la réponse prédite G(x) par la fonction de prédiction G pour une observation donnée de la base d'entraînement S.

Seulement, la recherche de la fonction de prédiction G optimale sera restreinte par le choix de l'algorithme d'apprentissage et la configuration de ce dernier, qui vont définir l'espace des hypothèses possibles H pour trouver cette fonction G (Figure 5.6c). Cet espace d'hypothèses possibles H constitue l'ensemble des fonctions de prédiction candidates qui vont être considérées dans l'apprentissage pour prédire les réponses associées aux données d'observation. La recherche de cette fonction de prédiction optimale est donc biaisée vers un jeu particulier de règles de prédiction, ce qui est nommé biais inductif (Haussler, 1988 ; Mitchell, 1997). La fonction de prédiction optimale G, soit l'hypothèse finale qui appartient à l'ensemble H, est celle qui va se rapprocher le mieux de la fonction cible, et donc de minimiser l'erreur sur S.

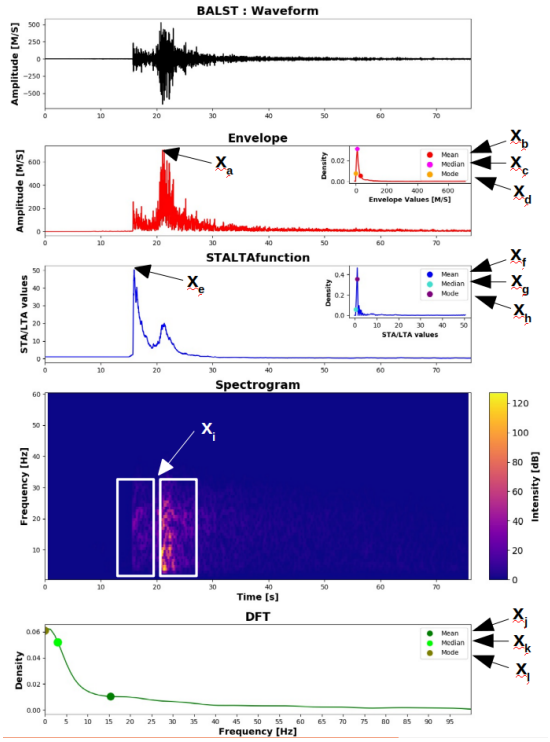


(a)

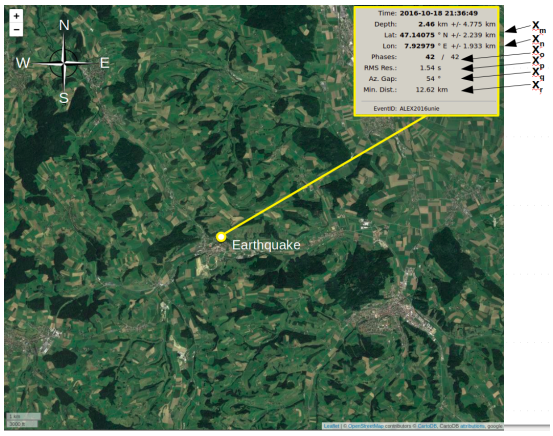


(b)

FIGURE 5.2: Aperçu de la variabilité des attributs qui peuvent être utilisés pour décrire un événement. L'événement qui est présenté ici est un tir de la carrière de Straßberg, identifié en Allemagne le 15 décembre 2016 à 12h29 (MLv 1.6). Par exemple, les attributs nommés x_a à x_d sont des paramètres statistiques (valeur maximale, moyenne, médiane, mode) qui décrivent l'enveloppe du signal enregistré sur la composante verticale de la station GUT pour cet événement. Les attributs nommés x_e à x_h sont les mêmes paramètres statistiques (valeur maximale, moyenne, médiane, mode) qui décrivent quant à eux la fonction STA/LTA, c'est-à-dire l'évolution des valeurs du rapport STA/LTA en fonction du temps, pour la même station. De même, pour x_j à x_l qui décrivent statistiquement la représentation spectrale discrète du signal échantillonné, obtenue par transformée discrète de Fourier (DFT). L'attribut x_i définit le rapport spectral de l'intensité du signal sur deux fenêtres temporelles adjacentes. Les attributs x_m à x_r apportent des informations sur l'origine préférentielle de l'événement localisé, à savoir par exemple sa localisation épicentrale et les incertitudes associées, le nombre de phases inclus dans l'association, la RMS des résidus, le gap azimutal ou la distance épicentrale minimale.



(a)



(b)

FIGURE 5.3: Aperçu de la variabilité des attributs qui peuvent être utilisés pour décrire un événement. L'événement qui est présenté ici est un séisme, identifié dans le canton de Zürich en Suisse le 18 octobre 2016 à 21h36 (MLv 1.4). Les attributs proposés sont les mêmes que que dans la Figure 5.2. Pour les attributs liés au signal, c'est la composante verticale de la première station (BALST) qui est ici présentée.

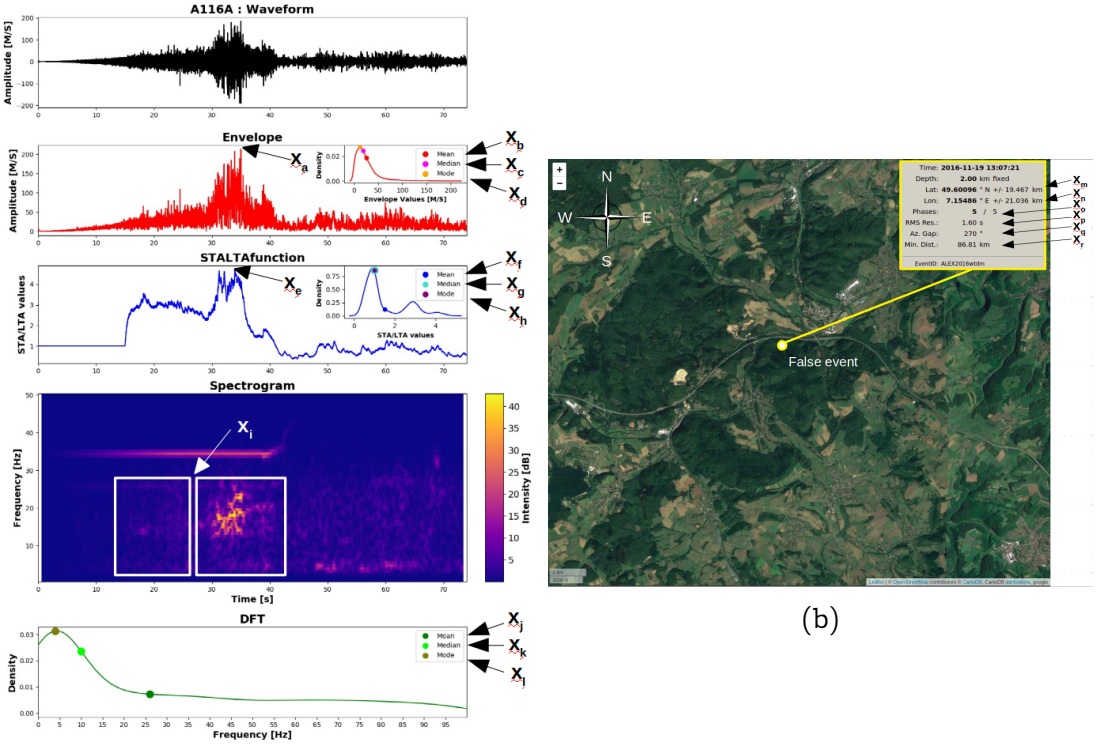


FIGURE 5.4: Aperçu de la variabilité des attributs qui peuvent être utilisés pour décrire un événement. L'événement qui est présenté ici est un faux événement, identifié au Sud-Ouest de Frankfurt, non loin de deux carrières, le 19 novembre 2016 à 13h07. Les attributs proposés sont aussi les mêmes que que dans la Figure 5.2. Pour les attributs liés au signal, c'est la composante verticale de la première station (A116A) qui est également présentée.

Si l'algorithme d'apprentissage renvoie une fonction de prédiction dont l'erreur empirique reflète son erreur de généralisation lorsque la taille de la base d'entraînement tend vers l'infini, et, si l'algorithme permet de trouver une fonction de prédiction qui minimise l'erreur de généralisation dans la classe d'hypothèses considérée, alors l'erreur empirique de cette fonction de prédiction sur la base d'entraînement S converge en probabilité vers son erreur de généralisation.

La borne supérieure de l'erreur de généralisation s'exprime donc effectivement en fonction de l'erreur empirique de la fonction de prédiction apprise sur une base d'entraînement, mais aussi en fonction de la complexité de la classe d'hypothèses utilisées (nombre de nœuds dans un réseau neuronal, profondeur d'un arbre décisionnel, etc.). Cette complexité traduit la capacité de la classe d'hypothèses à résoudre le problème de prédiction. Plus cette capacité est grande, plus l'erreur empirique sur S est faible, mais plus il y a un risque que l'erreur de généralisation soit en revanche élevée (Figure 5.5). Cette borne exhibe ainsi le compromis qui existe entre l'erreur empirique à minimiser et la capacité de la classe d'hypothèses à contrôler.

L'erreur qui est estimée sur la base d'entraînement n'est donc pas forcément représentative de la performance de la fonction de prédiction sur de nouvelles observations. Il est alors nécessaire de disposer d'un second ensemble d'exemples étiquetés, appelé base de test, auquel l'algorithme d'apprentissage n'avait pas accès, pour estimer l'erreur moyenne de la fonction produite, qui sera cette fois plus représentative de son erreur de généralisation. L'objectif pour l'algorithme d'apprentissage est de trouver une fonction ayant de bonnes performances de généralisation et non celle qui sera capable de reproduire parfaitement les réponses associées aux exemples d'entraînement.

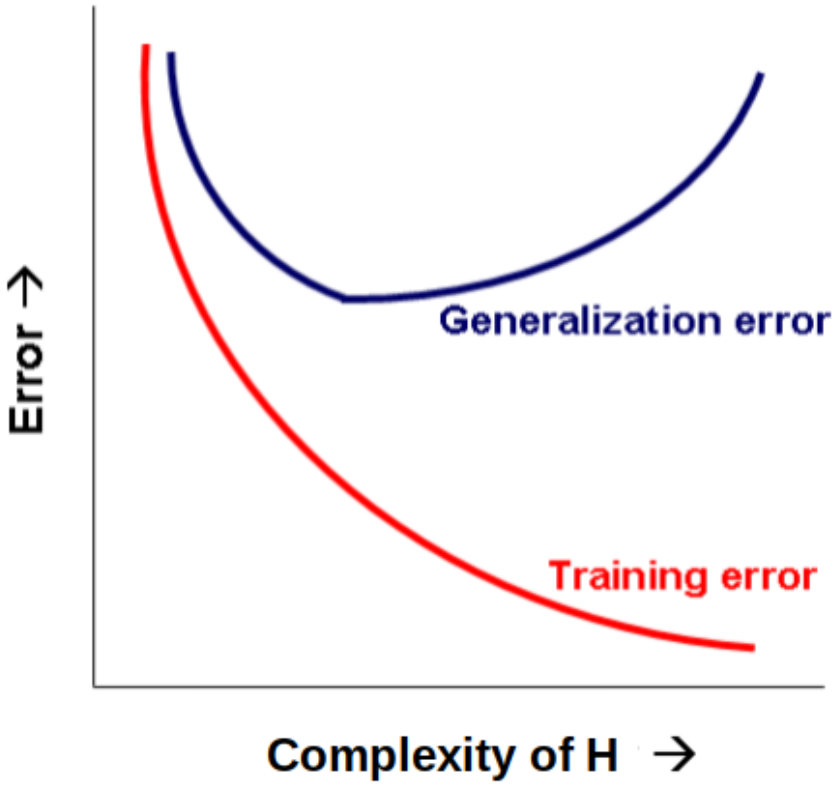


FIGURE 5.5: Évolution de l'erreur empirique (ou erreur d'entraînement) et de l'erreur de généralisation en fonction de la complexité de la classe d'hypothèses utilisées (H). L'erreur empirique diminue avec la complexité de la classe d'hypothèses utilisées alors que l'erreur de généralisation est d'abord élevée pour de faibles niveaux de complexité, diminue jusqu'à ce que la complexité de la classe d'hypothèses utilisées corresponde à la distribution inconnue des observations, puis s'élève de nouveau pour des classes d'hypothèse de plus haute complexité. Modifié d'après BELKIN et al., 2019.

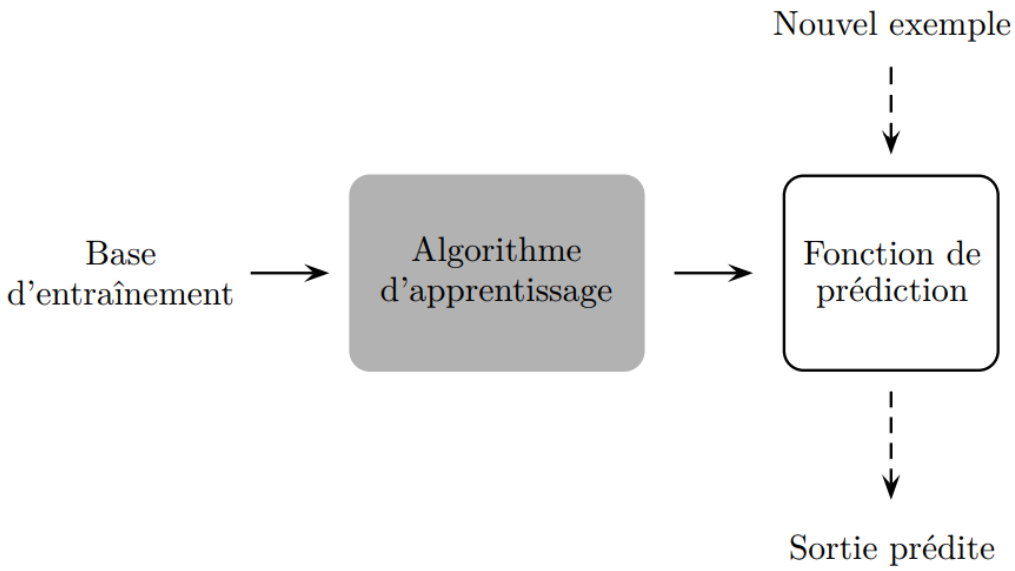


FIGURE 5.6: Illustration des deux phases d'un problème d'apprentissage supervisé. Dans la phase d'apprentissage (schématisée par les traits pleins), une fonction minimisant l'erreur empirique sur une base d'entraînement est trouvée parmi une classe de fonctions hypothétiques prédéfinies. Dans la phase de test (schématisée par les traits pointillés), les sorties de nouveaux exemples sont prédites par la fonction de prédiction. D'après AMINI, 2015.

5.1.2 Définir les contraintes de l'espace d'apprentissage qui élèvent l'erreur de généralisation

•Une taille d'échantillon petite

Dans le cadre du problème de la classification, afin d'évaluer la performance des prédictions du classifieur, il est nécessaire d'avoir une base d'entraînement qui ait des étiquettes de classe correctement affiliées aux différents événements. Or, dans la zone d'étude, c'est seulement à partir de 2012 que les tirs de carrière ont été inclus dans le catalogue maintenu par le BCSF-RéNaSS. De plus, comme il a été déjà décrit dans la section 3, ce n'est qu'à partir de 2016 que cette base de données a pu être réellement soigneusement discriminée. Par conséquent, la seule base de données vraiment robuste dont je dispose pour résoudre ce problème de classification des faux événements, des séismes et des tirs de carrière, est une base relativement petite d'environ 10389 événements (728 faux événements, 5537 séismes, et 4124 tirs de carrière) qui ont été détectés entre 2016 et 2019. Cette taille est effectivement relativement petite si on la compare aux bases de données qui peuvent être produites actuellement dans d'autres domaines appliqués comme la biomédecine avec les données omiques (ex : données de séquençage entier du génome, plusieurs dizaines de milliards de nucléotides répartis sur plus de 20 000 génomes différents, Wainberg et al., 2018) ou plus théoriques comme la reconnaissance d'images à partir de la base des 300 millions d'images du monde réel éditée par Google (ImageNet et JFT-300M, Sun et al., 2017).

•Des données de grande dimension

Chaque événement de la base de données que je possède peut être décrit à travers un espace d'attributs qui est grand. En effet, quel que soit l'événement (faux événement, séisme ou tir de carrière), celui-ci peut être d'abord décrit à travers les paramètres qui vont définir son origine, à savoir sa localisation (coordonnées géographiques, profondeur, distance à un site anthropique comme une carrière, etc.) et les incertitudes associées (ellipsoïdes de confiance), son temps d'occurrence (heure de la journée, jour de la semaine, etc.), sa qualité (distances épicentrales, RMS, résidus temporels, nombre et type de phases, gap azimutal, etc) et sa magnitude (magnitude locale, magnitude de coda, magnitude de surface, etc.) par exemple.

Un événement peut ensuite être également défini grâce aux signaux qui ont servi à l'identifier. Ces derniers fournissent des informations indirectes précieuses sur la source qui a émis ces signaux, mais aussi sur les effets du milieu de propagation et les caractéristiques du bruit enregistré aux différentes stations. Cet ensemble d'informations peut être extrait à partir d'une description complète des signaux dans le domaine temporel (formes d'onde, enveloppe), fréquentiel (spectre) et tempo-fréquentiel (spectrogrammes, sonogrammes, décomposition en ondelettes, transformation de Wigner-Ville, transformation d'Hilbert-Huang, etc.).

Quelques exemples d'attributs sont présentés dans la Figure 5.3 ou 5.2 ou 5.4. Dans ce travail de thèse, un total de 361 attributs ont été identifiés pour décrire chaque événement. Le détail de ces attributs est adressé dans le tableau S1 qui se trouve dans le supplément de l'article qui est présenté ultérieurement.

Seulement, la précision des algorithmes de classification a tendance à se détériorer à mesure que la dimensionnalité des attributs augmente, en raison d'un phénomène appelé la "malédiction de la dimensionnalité" (Bellman, 1961 ; Trunk, 1979 ; Koppen, 2000). En effet, si la distance euclidienne est choisie pour comparer relativement chaque observation de la base d'entraînement (à savoir chaque vecteur d'événement) dans l'espace euclidien correspondant, la distance qui sépare chaque point d'observation augmente avec le nombre d'attributs (Figure 5.7).

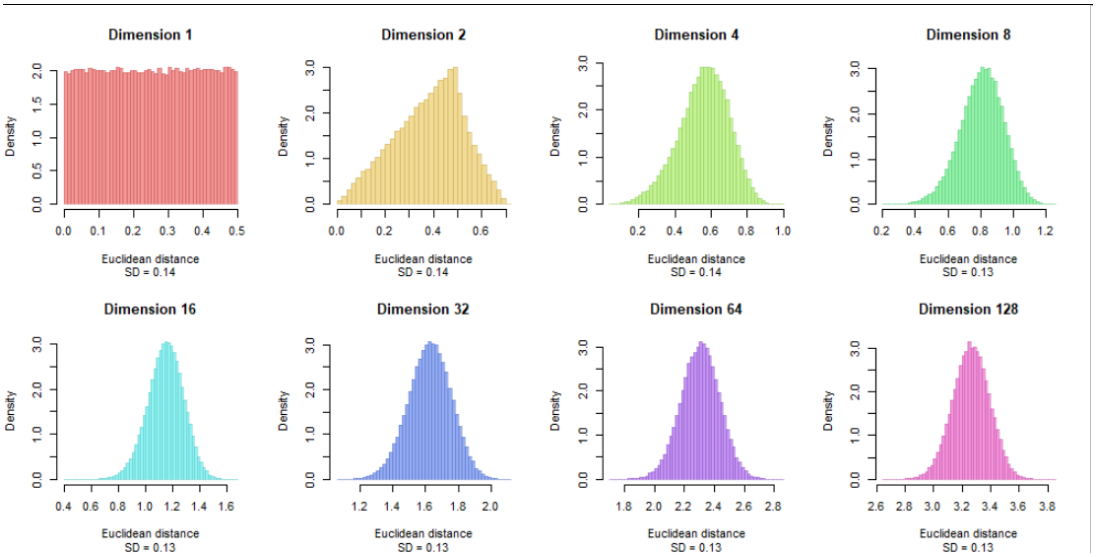


Figure 5.7: Évolution de la distribution de la densité de 500 points d'observation en fonction de la distance euclidienne et de la dimensionnalité de l'espace d'attributs. Chaque histogramme représente une dimensionnalité (de 1 à 128). Au fur et à mesure que la dimensionnalité des attributs augmente, les distributions tendent vers une forme gaussienne avec une distance moyenne entre chaque point d'observation de l'espace euclidien qui s'accroît. Cette accroissement de la distance qui sépare chaque point d'observation limite les possibilités de regroupement des observations en des classes bien identifiées. Une comparaison relative des distances (normalisation par rapport à la distance maximale) montre qu'à mesure que la dimension augmente, les distances se concentrent autour d'une valeur centrale, soulignant le fait que les points d'observation tendent à être environ tous à la même distance. SD = écart-type.

Avec l'augmentation de la dimensionnalité de l'espace d'attributs, les possibilités de combinaisons uniques d'attributs se multiplient donc, éloignant les points d'observation les uns des autres (Figure 5.8). Cet accroissement exponentiel du nombre attributs est d'ailleurs nécessairement associé à une augmentation de la redondance ou la non significativité de plusieurs entre eux, au regard du problème posé.

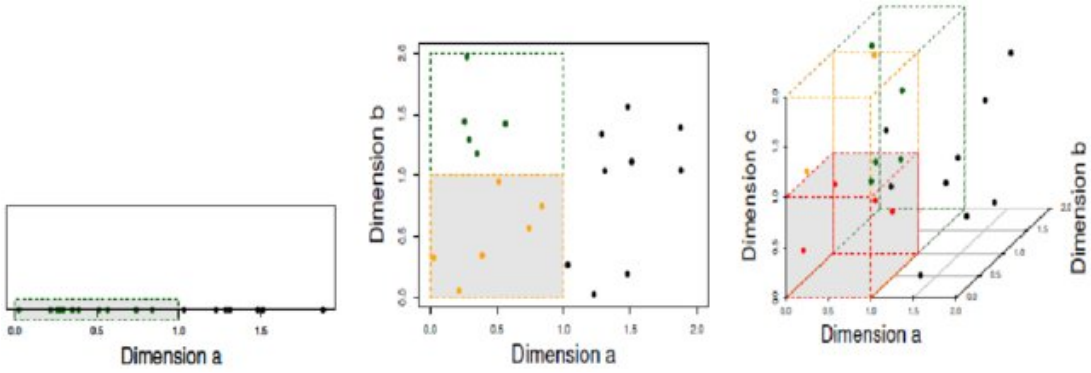
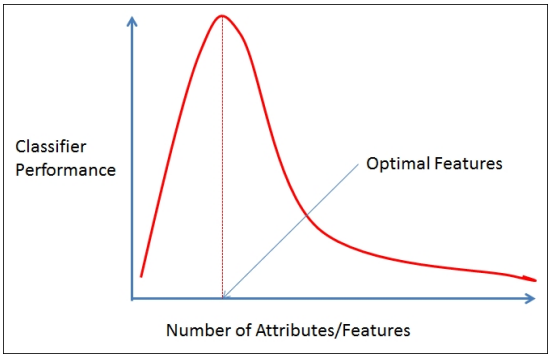
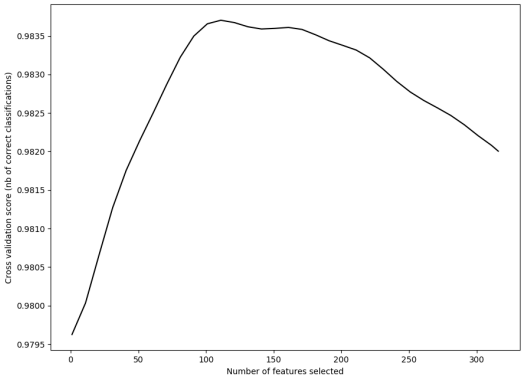


FIGURE 5.8: Effet de l'augmentation de la dimensionnalité des attributs sur la densité des points d'observation (chaque point d'observation correspond à un vecteur de n attributs, n=1, 2, 3). Les données dans une seule dimension sont relativement compactes. En ajoutant une dimension, les points d'observation s'écartent. Des dimensions supplémentaires éparpillent largement les points d'observation, diminuant fortement leur densité au sein d'un espace d'attributs à plus grande dimension. D'après PARSONS et al., 2004.

Par conséquent, au fur et à mesure que la dimensionnalité des attributs s'étend, il y a un risque accru de sur-adapter la fonction de prédiction à des cas particuliers. Ceci génère donc des classifieurs avec de mauvaises performances de généralisation et rend donc plus difficile les prédictions correctes sur de nouvelles observations (Figure 5.9). Lorsque l'espace d'attributs progresse en dimensionnalité, les données perdent en densité mais le classifieur généré se complexifie aussi.



(a) Représentation théorique du phénomène de Hughes



(b) Estimation grossière de la performance d'un classifieur obtenu à partir d'un jeu d'entraînement constitué de séismes et de tirs de carrières extraits du catalogue BCSF-RéNaSS, pour la période 2017-2019. Chaque donnée d'observation (séisme ou tir de carrière) est représentée par un vecteur allant de 10 à 350 attributs aléatoirement choisis parmi la banque d'attributs définie initialement (cf supplément de l'article qui va suivre). L'algorithme d'apprentissage utilisé pour générer le classifieur est Random Forest.

FIGURE 5.9: Représentation graphique du phénomène de Hughes. Dans le cadre de l'apprentissage machine supervisé, la performance d'un classifieur est fortement dépendante de la dimensionnalité de l'espace d'attributs qui est donnée en entrée. Si la performance du classifieur augmente d'abord proportionnellement avec le nombre d'attributs utilisés jusqu'à atteindre une performance optimale, celle-ci chute rapidement du fait de l'accroissement possible du nombre de combinaisons uniques d'attributs, éloignant les observations les unes des autres. Pour de nouveau optimiser la performance du classifieur, il faudrait augmenter la taille de l'échantillon du jeu d'entraînement afin de diminuer les distances qui séparent chaque point d'observation dans un espace à grande dimension d'attributs.

•Une répartition déséquilibrée des classes d'événement

Par ailleurs, en plus d'une dimensionalité élevée, ce jeu de données présente un fort déséquilibre des classes d'événement en direction des faux événements. En effet, les faux événements ne représentent que 7% du total des événements pour la période 2016-2019. De façon moins marquée, les tirs de carrière représentent 40% des événements détectés pour la même période contre 53% pour les séismes.

Dans le cadre d'un problème de classification binaire, s'il s'agit d'identifier les faux événements de l'ensemble des vrais événements, alors les vrais événements constituent 93% de la base d'entraînement. Or, dans la réalité, lorsque les petits séismes sont détectés, ce sont les faux événements qui sont majoritaires pour plus de 95% du total des événements détectés. Ce jeu de données collectés n'est donc pas représentatif du profil de détection qui est généré lorsque les seuils de détection sont fortement diminués.

Or, dans un cas ou dans l'autre, des données très déséquilibrées posent des difficultés supplémentaires. La plupart des fonctions de prédiction apprises à partir d'une base d'entraînement fortement déséquilibrée présentent effectivement un biais en faveur de la classe majoritaire (ici les vrais événements) et, dans des cas extrêmes, peuvent ignorer complètement la classe minoritaire dans leurs prédictions (J. M. Johnson et al., 2019). Les algorithmes d'apprentissage présentent donc des difficultés à généraliser le comportement de la classe minoritaire et la capacité prédictive de la fonction de prédiction apprise est faible.

En effet, les probabilités ou les scores prédits par de nombreux algorithmes d'apprentissage ne sont pas calibrés. Ceci signifie que la distribution et le comportement des probabilités prédites peuvent ne pas correspondre à la distribution attendue des probabilités observées dans les données d'apprentissage. Ceci est particulièrement courant avec les algorithmes d'apprentissage automatiques non linéaires complexes qui ne font pas directement des prédictions probabilistes, mais utilisent plutôt des approximations. Par exemple, les algorithme d'apprentissage basés sur les forêt aléatoires comme Random Forest (Breiman, 2001) estiment leur prédiction sous la forme d'un score qui évalue le nombre d'arbres décisionnels qui a prédit correctement le label d'un événement par rapport au nombre total d'arbres utilisés pour construire la fonction de prédiction.

•Des données bruitées et hétérogènes

Comme il a été écrit dans les sections précédentes, les signaux associés aux faux événements et aux vrais événements (séismes, tirs de carrière) qui sont détectés dans la zone d'étude ont des contenus fréquentiels, des amplitudes et des durées qui peuvent être très fortement similaires (Inbal et al., 2018 ; Poli et al., 2020). De nombreux vrais événements sont détectés avec de très faibles rapport signal/bruit. D'ailleurs, le succès de l'opération de pointé des temps d'arrivée des différentes phases sismiques a été fortement dépendant des conditions de bruit enregistré aux stations. La prise en compte de ces conditions de bruit est effectivement un critère décisif pour obtenir un pointé automatique de qualité, nous l'avons vu. L'ensemble des signaux sismiques qui sont détectés sont donc de faible amplitude, et fortement contaminés par du bruit stationnaire, voire non stationnaire.

De plus, la diversité des formes d'ondes associées aux tirs de carrière rend souvent difficile la tâche de discrimination des vrais événements entre eux, comme je l'ai déjà évoqué précédemment dans les chapitres 1 et 2. L'analyse de la forme d'onde peut alors parfois brouiller la qualité de la discrimination.

Par ailleurs, les solutions épicentrales et hypocentrales des événements détectés peuvent apporter beaucoup d'incertitudes à l'identification des événements, en se basant uniquement sur ces paramètres. Par exemple, la profondeur est souvent mal contrainte, les incertitudes latitudinales et longitudinales des épicentres des vrais événements peuvent être aussi fortes que celles des épicentres des faux événements, car ce sont souvent des petits événements qui sont détectés et localisés avec très peu de phases.

Or, tous ces paramètres évoqués juste au-dessus font partie des attributs que j'ai sélectionnés pour former la banque des 361 attributs qui vont servir à classer les événements (faux événement, séisme, tir de carrière). Cela signifie que des attributs statistiques décrivant le contenu fréquentiel absolu des signaux ou bien le contenu fréquentiel de ces signaux relativement au temps (spectrogramme par exemple) ou bien l'enveloppe du signal peuvent être utilisés. Or, si ces attributs peuvent apporter des informations précieuses sur la nature de chaque événement, ils peuvent également introduire beaucoup de confusion lorsque par exemple, d'une classe d'événement à une autre, les contenus fréquentiels se chevauchent ou les formes d'ondes tendent à être similaires.

Face à la complexité et la diversité intrinsèques du jeu d'événements (i.e. diversité des signaux, effets du milieu de propagation, contenu en bruit élevé, localisations des origines peu contraintes), il y a un fort risque que l'algorithme d'apprentissage apprenne sur des corrélations parasitaires dans les données d'apprentissage.

Lorsque l'algorithme d'apprentissage exploite des artefacts ou des informations parasites dans le jeu de données pour en déduire une fonction de prédiction erronée, ce comportement est nommé métaphoriquement "Clever-Hans" (Pfungst, 1911 ; Lapuschkin et al., 2019 ; Schramowski et al., 2020, Figure 5.10). Clever-Hans était le nom d'un cheval intelligent qui semblait avoir appris à répondre à des questions arithmétiques, mais qui n'avait en fait appris à lire que les indices sociaux qui lui permettaient de donner la bonne réponse. Dans des environnements contrôlés où il ne pouvait ni voir les visages des gens, ni recevoir d'autres commentaires, ce cheval intelligent n'a en fait pas pu répondre à ces questions.

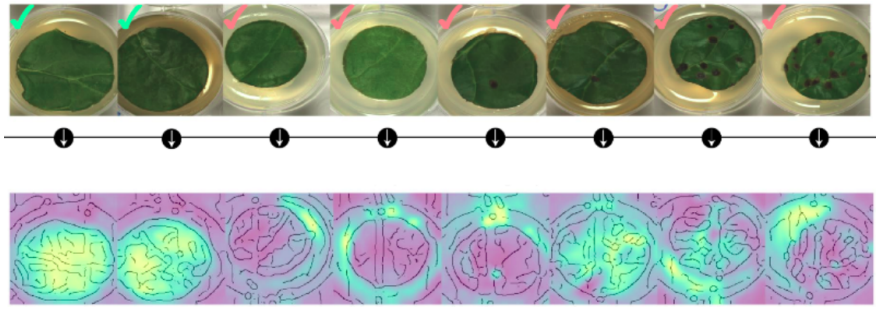


Figure 5.10: Exemple de comportement type "Clever-Hans". Des feuilles de Betterave sont soumises à un stress biotique (inoculation des feuilles avec un pathogène fongique *Cercospora beticola*, typique de la famille des *Chenopodiaceae* à laquelle appartient la Betterave). Un réseau de neurones convolutif (CNNs) a été utilisé pour classer des images RGB, produites par imagerie hyperspectrale, de feuilles de Betterave infectées et saines. Les disques de tissu foliaire ont été placés dans des boîte de pétri contenant une solution d'agar (figures du haut). A chaque photographie correspond une coche de couleur (vert pour feuille saine et rose pour feuille infectée). Les résultats de la classification par le réseau neuronal convolutif sont représentés graphiquement pour chaque échantillon (figures du bas). Les couleurs jaune-vert correspondent aux régions qui ont été utilisées pour obtenir le diagnostic de classification alors que les couleurs bleu-violet correspondent aux régions non utilisées. Pour une meilleure lisibilité, les régions colorées ont été superposées à l'image originale filtrée. Il est possible alors d'observer que le réseau neuronal convolutif profond a correctement classifié les feuilles infectées sur la base d'arguments artéfactuels, en l'occurrence ici la solution d'agar entourant les disques foliaires, illustrant ainsi un comportement de type "Clever-Hans". Modifié d'après Schramowski et al., 2020.

•Conséquence : un fort risque de sur-apprentissage

L'ensemble des paramètres décrits, une taille d'échantillon petite avec une dimensionnalité des données élevée, un déséquilibre des classes d'événement et des informations possiblement parasitaires, limitent l'aptitude des algorithmes d'apprentissage à générer des fonctions de prédictions généralisables. En effet, tous ces paramètres qui décrivent la base de données que je possède peuvent facilement induire en erreur l'algorithme d'apprentissage, et aboutir à un sur-apprentissage des données collectées, donc une erreur de généralisation élevée.

L'erreur de généralisation peut se décomposer de la manière suivante : une erreur bayésienne intrinsèque irréductible associée à tous les classifieurs, une erreur d'approximation et une erreur d'estimation. L'erreur d'approximation désigne l'erreur minimale réalisable par une fonction de prédiction G au sein de l'espace d'hypothèses H. Ce terme mesure le risque encouru lorsque l'on se restreint à une certaine classe d'hypothèses, à savoir le niveau de biais inductif atteint (Shalev-Shwartz et al., 2014).

Le fort taux d'informations parasitaires contenues dans les données d'apprentissage peut facilement aboutir à un fort biais inductif, et donc à une erreur d'approximation élevée. Seulement, la haute dimensionalité de l'espace d'attributs peut véhiculer une richesse d'informations telle que des motifs multiples de classification possibles des événements peuvent se révéler, complexifiant grandement l'espace d'hypothèses qui sert à générer la fonction de prédiction. Dans ce cas-ci, l'erreur d'approximation devient beaucoup plus faible.

L'erreur d'estimation représente la différence entre l'erreur d'approximation et l'erreur globalement réalisée par la fonction de prédiction, dans le cadre du principe de minimisation du risque empirique. Cette erreur d'estimation mesure l'éloignement de la fonction de prédiction, apprise par l'algorithme d'apprentissage, de la meilleure fonction de prédiction disponible au sein de la classe d'hypothèses H. Étant une propriété incompressible de l'algorithme d'apprentissage, cette erreur dépend fortement de la taille du jeu d'entraînement, mais aussi de la taille de la classe d'hypothèses sélectionnée H. Ainsi, pour une classe d'hypothèses finie, l'erreur d'estimation augmente logarithmiquement avec la taille (et donc la complexité) de la classe d'hypothèses et décroît avec la taille de l'échantillon d'entraînement. Cette erreur d'estimation existe parce que l'erreur empirique est seulement une estimation de l'erreur globale de généralisation.

L'échantillon que je possède étant relativement petit et assez peu représentatif de l'ensemble des classes d'événements à identifier (ici les faux événements), l'erreur d'estimation à l'issue de l'apprentissage a de fortes chances d'être élevée.

Le jeu de données que je détiens apporte donc de fortes contraintes sur son apprentissage. La fonction de prédiction générée présente un risque fort d'erreur d'approximation et d'erreur d'estimation. Or, l'objectif d'un apprentissage automatique optimal, guidé par un algorithme d'apprentissage, est de minimiser l'erreur totale de généralisation. Je suis donc nécessairement confrontée à un compromis entre le biais et la complexité de la fonction de prédiction à générer.

D'une part, choisir une classe d'hypothèses très riche peut diminuer l'erreur d'approximation mais peut en même temps augmenter l'erreur d'estimation, puisqu'un espace d'hypothèses riche peut conduire à un sur-apprentissage (Figure 5.11). D'un autre côté, choisir un petit ensemble d'hypothèses réduit l'erreur d'estimation mais peut augmenter l'erreur d'approximation ou, en d'autres termes, peut conduire à un sous-apprentissage (Figure 5.11). Bien sûr, le choix optimal pour l'espace d'hypothèses est un espace réduit qui contient un seul classifieur, le classifieur optimal de Bayes. Seulement, ce classifieur optimal dépend de la distribution sous-jacente D de l'ensemble des observations (des événements) ayant lieu dans la nature, qui est totalement inconnue. De tout manière, l'apprentissage aurait été inutile si nous avions connu cette distribution D.

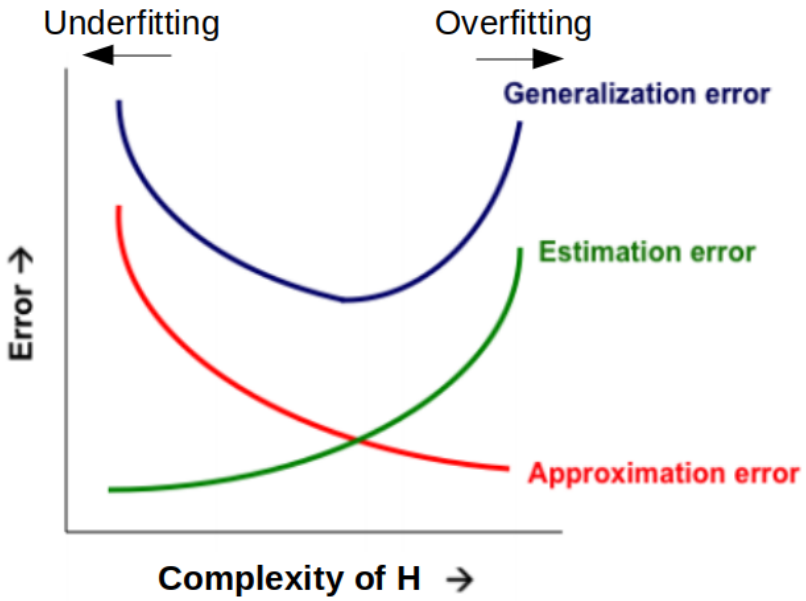


FIGURE 5.11: Représentation théorique de l'erreur d'approximation et de l'erreur d'estimation en fonction de la complexité de l'espace d'hypothèses. Pour une taille d'échantillon fixe, à mesure que la complexité de l'espace d'hypothèses augmente, l'erreur d'approximation diminue, tandis que l'erreur d'estimation augmente. Une valeur élevée de l'un ou de l'autre contribue à une erreur de généralisation élevée. L'erreur d'approximation élevée est associée à un sous-apprentissage alors qu'une erreur d'estimation élevée est associée à un sur-apprentissage. AGARWAL, 2018.

La sous-section suivante s'attache alors à définir les réponses que je peux apporter aux contraintes inhérentes au jeu de données que je possède. L'objectif est clairement de générer un apprentissage automatique qui puisse étudier un espace d'hypothèses suffisamment riche, tout en conservant une erreur d'estimation raisonnable pour obtenir la "meilleure" fonction de prédiction.

5.1.3 Réduire les contraintes pour optimiser l'apprentissage

- **Réajuster le déséquilibre des classes d'événement**

Une première réponse aux contraintes inhérentes au jeu de données disponibles serait d'augmenter la taille de ce dernier. En effet, ce jeu de données est très déséquilibré puisque les faux événements ne représentent que 7% du nombre total d'événements. Afin de compenser ce déséquilibre, j'ai utilisé les faux événements détectés au cours du mois de juillet et août 2016. Ces deux mois font partie d'un catalogue test automatique qui a servi de base pour améliorer la détection automatique des petits séismes. Environ 24000 événements sont alors disponibles pour combler le déséquilibre de classe entre les faux et les vrais événements. Tous ces faux événements ont été revus manuellement.

De plus, afin d'égaliser les proportions des séismes et des tirs de carrière dans ce jeu de données (40% de tirs de carrière et 53% de séismes), un sous-échantillonnage de ces événements peut être réaliser. Ce procédé autorise plus facilement un ré-échantillonnage ultérieur de la base d'entraînement générée à partir de ce jeu de données pour tester la performance de l'apprentissage automatique à partir de bases d'entraînement différentes.

Toutefois, en fonction de l'algorithme d'apprentissage choisi, l'entraînement sur des bases d'entraînement différentes, même légèrement variables, peut conduire à des résultats de prédiction très instables du fait d'une adaptabilité trop forte de l'algorithme aux variations de la base d'entraînement, diminuant alors l'erreur de généralisation.

De ce fait, pour obtenir un classifieur qui puisse prédire efficacement les différents types d'événement en dehors du jeu d'entraînement (faux événement, séisme, tir de carrière), il est nécessaire de délimiter clairement un espace d'hypothèses qui puisse solidement réduire les effets des contraintes inhérentes au jeu de données disponible. Le choix de l'algorithme d'apprentissage ainsi que sa configuration, en sont les éléments fondateurs.

•**Définir un espace d'hypothèses optimal**

- Choisir un algorithme d'apprentissage stable. La stabilité d'un algorithme d'apprentissage peut être reliée à la notion de variance. En effet, un algorithme d'apprentissage stable est un algorithme qui réalise une erreur moyenne de généralisation faible s'il est entraîné sur plusieurs bases d'entraînement différentes. Cette erreur de généralisation moyenne est formalisée comme étant la somme de plusieurs erreurs : la variance, le biais (élevé au carré) et une erreur irréductible ou bruit intrinsèque associée à la distribution inconnue de l'ensemble des observations possibles. Cette erreur de généralisation moyenne est évaluée à partir de la moyenne des prédictions émises par l'ensemble des fonctions de prédiction apprises sur les différents échantillons de jeu d'entraînement.

Le biais (élevé au carré) évalue l'écart entre les prédictions émises par la fonction de prédiction moyenne et les prédictions attendues émises par la fonction de prédiction optimale. La variance évalue de combien une fonction de prédiction apprise à partir d'un échantillon d'entraînement particulier, s'éloigne de la fonction de prédiction moyenne (Agarwal, 2018). La variance traduit donc le degré de flexibilité de l'algorithme d'apprentissage utilisé, c'est-à-dire la capacité de ce dernier à changer sa fonction de prédiction lorsqu'un jeu d'entraînement différent est utilisé. Un algorithme avec une variance élevée aura ainsi une faible stabilité et sera donc sujet au sur-apprentissage (Figure 5.12).

En revanche, ce dernier sera caractérisé par un faible biais puisque son adaptabilité (il peut produire une fonction de prédiction différente lorsque l'échantillon d'entraînement change) tend à diminuer les écarts entre les prédictions produites par chaque fonction de prédiction et celles attendues.

Un algorithme d'apprentissage optimal est donc un algorithme capable de réduire la variance tout en conservant un faible biais. L'algorithme qui a été choisi dans ce travail de thèse a été l'algorithme de Random Forest (forêt aléatoire, Breiman, 2001). Cet algorithme base son apprentissage sur la construction d'un ensemble d'arbres décisionnels qui forme alors une forêt.

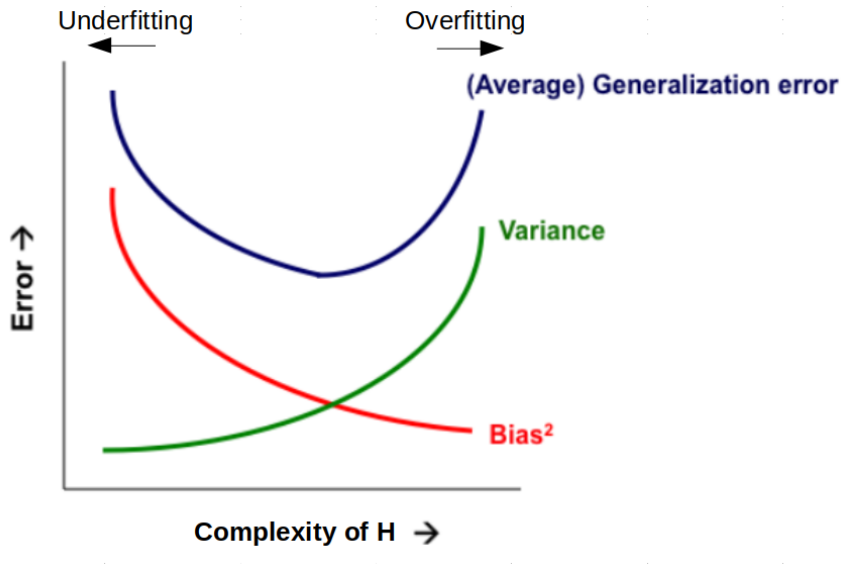


FIGURE 5.12: Évolution du biais et de la variance en fonction du degré de complexité de l'espace d'hypothèses. Pour une taille d'échantillon fixe, à mesure que la complexité de l'espace d'hypothèses augmente, le biais diminue, tandis que la variance augmente. Une valeur élevée de chacun contribue à une erreur de généralisation moyenne élevée. Un biais élevé est associé à un sous-apprentissage alors qu'une variance élevée est associée à un sur-apprentissage. Modifié d'après AGARWAL, 2018.

Cependant, les arbres de décision sont généralement très sensibles aux données sur lesquelles ils sont entraînés. Chaque arbre a effectivement la potentialité de capturer des interactions complexes entre les attributs. Un petit changement dans les données d'apprentissage peut provoquer facilement une modification des chemins décisionnels, rendant le processus de prédiction très instable (forte variance). Pour gagner en stabilité (et donc diminuer la variance), l'algorithme de Random Forest effectue un double échantillonnage : celui de la base d'entraînement et celui des attributs qui constituent cette base d'entraînement (Figure 5.13).

Le tirage aléatoire effectué sur la base d'entraînement est réalisé par une méthode d'échantillonnage avec remplacement (Bootstrap). Au sein de la forêt, chaque arbre décisionnel est donc construit indépendamment à partir d'un échantillon aléatoire de la base d'entraînement. Par conséquent, les échantillons qui ont été utilisés pour élaborer chaque arbre individuel sont de même longueur et issus de la même population, celle de l'échantillon original (la base d'entraînement originelle). Les données sont alors identiquement distribuées, et cela signifie que le biais de l'algorithme qui sera évalué sur la totalité de la forêt aléatoire sera le même que celui qui aurait été évalué à partir d'un seul arbre au sein de la forêt.

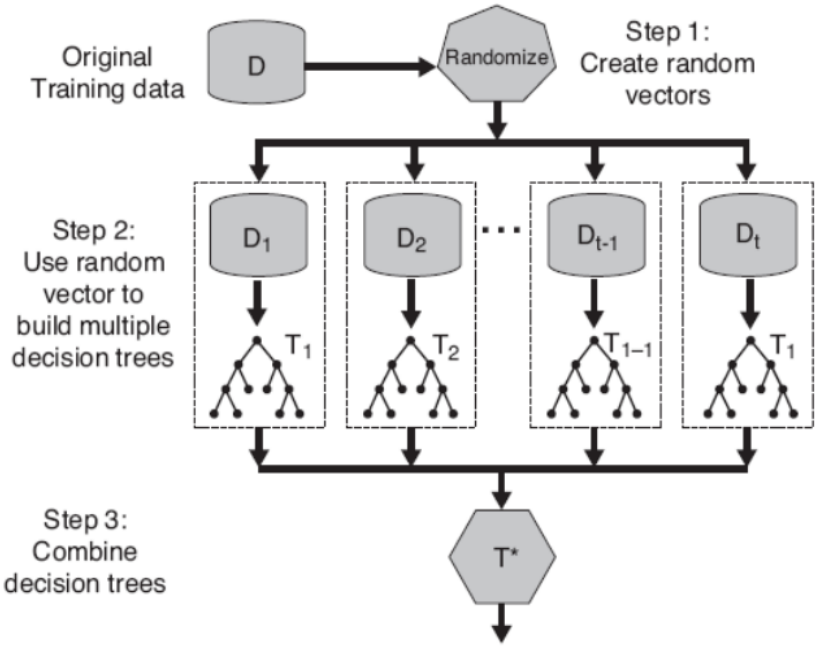


FIGURE 5.13: Méthode d'agrégation avec bootstrap (bagging) utilisée par Random Forest pour effectuer ses prédictions. La base d'entraînement est d'abord échantillonnée aléatoirement par bootstrap. Chaque échantillon aléatoire généré est utilisé pour construire un arbre décisionnel. Chaque embranchement de l'arbre décisionnel est élaboré à partir d'un deuxième échantillonnage aléatoire de l'espace d'attributs. La prédiction finale est définie en agrégeant les prédictions de l'ensemble des arbres décisionnels. Dans le cas de la classification, cette prédiction finale correspond à un vote majoritaire. Modifié d'après Cao et al. (2020)

.

Par ailleurs, en plus de l'échantillonnage de la base d'entraînement, un deuxième échantillonnage aléatoire est réalisé à partir de l'espace d'attributs qui définit chaque observation. Ce second échantillonnage vise à réduire la variance de l'algorithme d'apprentissage en décorrélant les arbres décisionnels entre eux. En effet, chaque arbre est une construction récursive de séries de fractionnements binaires qui séparent les différentes observations en sous-groupes successifs (Figure 5.14).

Chaque fois qu'une partition est considérée dans la construction de l'arbre, au lieu de la totalité des p attributs, un échantillon aléatoire de n attributs est tiré parmi l'ensemble du jeu complet des attributs décrivant chaque événement (observation). Parmi ces n attributs candidats potentiels pour générer un noeud de fractionnement, un seul attribut est sélectionné : c'est celui qui minimise l'erreur de classification des différentes observations au noeud formé. Ainsi, à chaque nouvel embranchement de l'arbre décisionnel, un nouvel échantillonnage de n attributs est prélevé (généralement $n = \sqrt{p}$). Cet échantillonnage aléatoire des attributs réduit donc la possibilité de générer des arbres similaires. En effet, si le même jeu de p attributs était toujours considéré à chaque noeud et pour chaque arbre, chaque arbre décisionnel serait systématiquement construit à partir de la même sélection hiérarchique des attributs les plus discriminants, en particulier au sommet de l'arbre.

Les prédictions des différents arbres décisionnels au sein de la forêt aléatoire sont ensuite agrégées pour aboutir à une prédiction finale (la prédiction finale correspond au vote majoritaire dans le cas de la classification). Les prédictions effectuées par les différents arbres étant faiblement corrélées du fait du sous-échantillonnage aléatoire récursif des attributs, leur agrégation diminue fortement la variance globale de l'algorithme d'apprentissage. Seulement, en présence de ce sous-échantillonnage, la solution prédictive étant recherchée dans un sous-espace restreint, la complexité du modèle est certes plus faible mais le biais est également plus grand. Si Random Forest réduit la variance, il y a tout de même un réglage biais-variance à réaliser. C'est ce que permet le réglage des hyperparamètres.

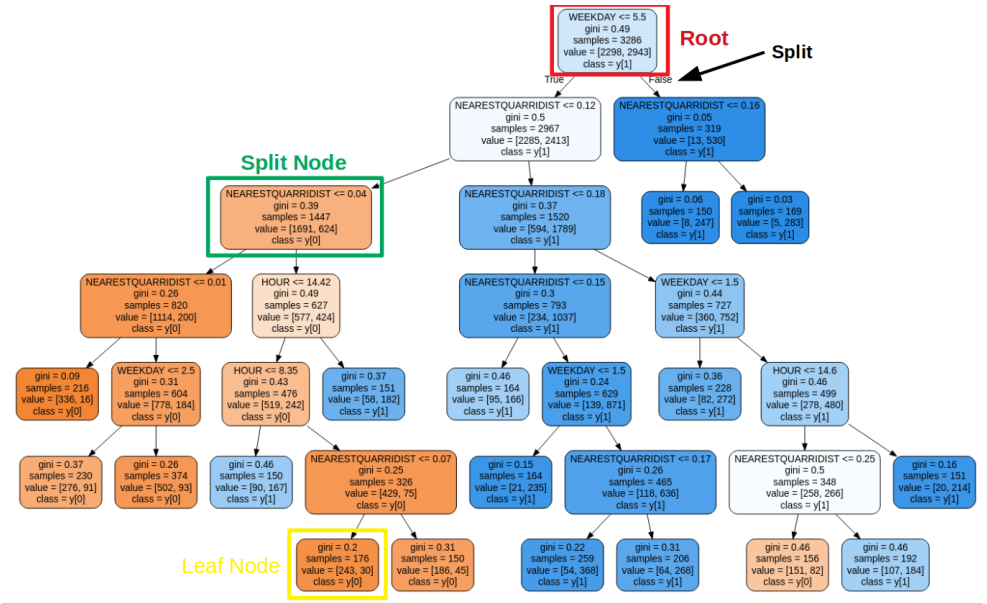


FIGURE 5.14: Exemple d'arbre décisionnel généré par l'algorithme d'apprentissage Random Forest. L'arbre décisionnel présenté a été établi à partir d'un jeu de données contenant l'ensemble des séismes et des tirs de carrière détectés entre 2017 et 2019 par le BCSF-RéNaSS (une part de 70 % est réservée au jeu d'entraînement et une part de 30% au jeu test). Chaque événement est représenté par un vecteur de trois attributs : l'heure de la journée, le jour de la semaine et la proximité de l'événement à la carrière. L'espace des hypothèses qui a été utilisé pour générer la fonction de prédiction (c'est-à-dire le classifieur) a été restreint afin de visualiser plus facilement la solution de classification. La sélection de l'attribut à chaque noeud de l'arbre a été établie à partir du calcul de l'impureté de Gini qui donne accès à l'importance de chaque attribut dans le processus de classification, soit son pouvoir discriminant. Chaque noeud apporte 5 informations. La première est la question posée au sujet des données, basée sur la valeur de l'attribut (par exemple à la racine de l'arbre : est-ce que le jour de la semaine est inférieur à 5.5, c'est-à-dire est-ce le jour de la semaine n'est pas samedi ?). Chaque question a soit une réponse "vrai", soit une réponse "fausse", qui va séparer le noeud en deux sous-groupes (le groupe "vrai" à gauche et le groupe "faux" à droite), et ainsi de suite en descendant dans l'arbre. La deuxième information de chaque noeud est la valeur de l'impureté de Gini. La troisième information est le nombre d'observations (événements) dans le noeud. La quatrième information est le nombre d'échantillons de ces observations dans chaque classe (ici classe 1 séisme ou classe 0 tir de carrière). Par exemple, la racine a 2298 échantillons dans la classe 0 et 2943 dans la classe 1. La dernière information est la classe majoritaire pour les observations à ce noeud. Par exemple, à la racine, la classe majoritaire est la classe 1 (séisme). Les prédictions finales se font aux noeuds terminaux c'est-à-dire au niveau des feuilles.

- Configurer des hyperparamètres de façon optimale. Les hyperparamètres sont des paramètres inhérents à l'algorithme d'apprentissage lui-même (Figure 5.15). Dans le cas de Random Forest, ces paramètres sont clairement reliés à l'architecture des arbres décisionnels à construire. Le nombre d'arbres fait partie des hyperparamètres à définir. En effet, un choix optimal du nombre d'arbres à inclure dans la forêt aura une influence direct sur le biais : augmenter le nombre diminuera le biais (Arlot et al., 2014). De même, augmenter la profondeur de l'arbre décisionnel diminue le biais mais augmente la variance.

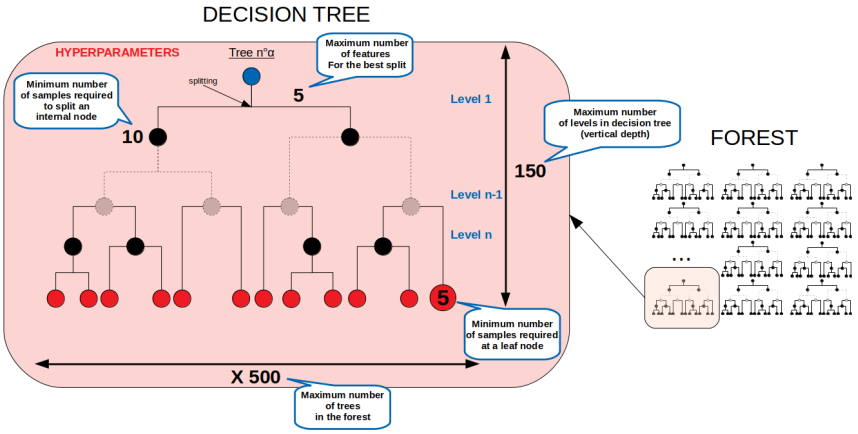


Figure 5.15: Principaux hyperparamètres associés à la configuration interne des arbres décisionnels constituant l'armature de l'apprentissage de l'algorithme de Random Forest. Ces hyperparamètres délimitent l'espace des hypothèses possibles pour rechercher la fonction de prédiction optimale (soit le meilleur classifieur).

Afin d'établir le choix optimal des hyperparamètres qui vont délimiter l'espace des hypothèses possibles, il est nécessaire d'explorer la plus grand étendue d'hyperparamètres possible afin de minimiser l'erreur de généralisation de la future fonction de prédiction apprise, et donc obtenir le juste équilibre entre le biais et la variance.

Plusieurs méthodes existent pour effectuer cette recherche multiple comme la recherche sur grille, la recherche aléatoire ou l'optimisation bayésienne. Dans ce travail de recherche, c'est la combinaison des deux méthodes, recherche aléatoire puis recherche sur grille, qui est utilisée. La recherche sur grille effectue une recherche exhaustive (explore toutes les combinaisons possibles) sur un ensemble de valeurs des hyperparamètres préalablement spécifié. C'est un algorithme de recherche très simple qui conduit aux prédictions les plus précises tant que des combinaisons suffisantes et pertinentes sont données (Bergstra et al., 2012). La recherche aléatoire (Bergstra et al., 2012) est une amélioration fondamentale de la recherche sur grille. Cette recherche s'effectue sur un échantillonnage aléatoire des valeurs d'hyperparamètres à partir des distributions statistiques préalablement définies. L'utilisation de cette méthode est souvent suggérée au début de la procédure d'optimisation des hyperparamètres pour réduire rapidement l'espace de recherche, avant d'utiliser un autre algorithme guidé pour obtenir un résultat plus fin (passage d'un schéma d'échantillonnage grossier à fin, Yu et al., 2020). L'échantillonnage plus fin est établi grâce à une recherche sur grille dans cette étude.

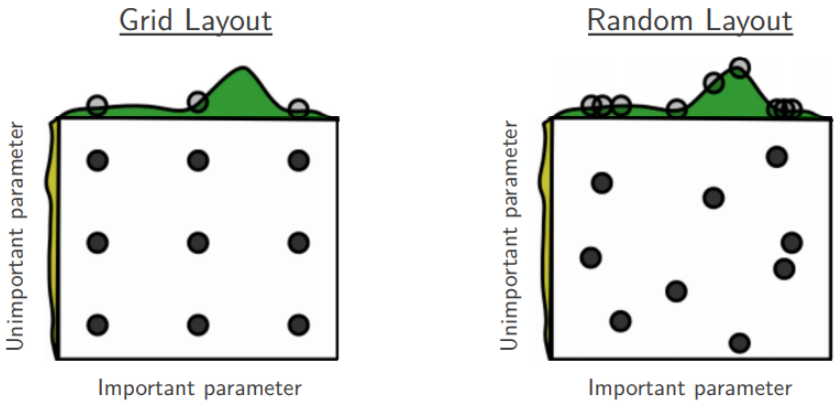


Figure 5.16: Recherche sur grille versus recherche aléatoire dans le cas de deux hyperparamètres et neuf combinaisons testées. Points noirs : combinaisons testées des hyperparamètres. Courbes jaune et verte : fonction objective de chaque hyperparamètre testé (fonction qui sert de critère pour déterminer la meilleure solution au problème d'optimisation des hyperparamètres et qui évalue l'erreur de validation). Sur l'axe y, la courbe est presque plate, signifiant que cet hyperparamètre a un faible impact sur la fonction objective totale. Cependant, sur l'axe x, un minimum clair apparaît, correspondant à la valeur optimale de cet hyperparamètre. Points gris : projection des combinaisons d'hyperparamètres testés sur la courbe verte. Le nombre de points gris est plus élevé pour la recherche aléatoire que pour la recherche sur grille, ce qui signifie que plus de valeurs ont été testées. D'après Bergstra et al., 2012.

L'estimation de l'erreur de généralisation est établie pour les différentes combinaisons d'hyperparamètres testées à l'aide de la stratégie de la validation croisée (Figure 5.17). Celle-ci se fonde sur le principe suivant : le jeu de données (ici la période 2017-2019) est partitionné en k sous-ensembles indépendants. Chaque sous-ensemble sert successivement d'échantillon de validation et le reste d'échantillon d'entraînement. L'échantillon d'entraînement est utilisé pour entraîner l'algorithme d'apprentissage (Random Forest) qui va sélectionner la fonction de prédiction, puis l'erreur commise est évaluée avec les données de validation. La performance de la validation croisée est estimée comme étant la moyenne arithmétique sur les k estimations de performance des ensembles de validation. La principale idée derrière la validation croisée est que chaque échantillon de l'ensemble de données disponible a la possibilité d'être testé (Raschka, 2018). La figure 5.17 illustre le processus de validation croisée à partir d'un partitionnement des données en 5 sous-ensembles. Dans ce cas précis, cinq fonctions de prédiction sont générées à partir des 5 itérations sur un jeu d'entraînement différent mais de longueur identique.

Si la validation croisée pour la recherche des hyperparamètres optimaux est intégrée dans la démarche globale d'apprentissage supervisée, le jeu de données est d'abord divisé en deux parties : une partie réservée à l'apprentissage proprement dit (période 2017-2019) et une partie destinée à constituer le jeu test (ici janvier-août 2016, Figure 5.18 étape 1). La recherche des hyperparamètres optimaux s'applique au jeu servant à l'apprentissage qui est lui-même subdivisé en jeu d'entraînement et jeu de validation.

La méthode de validation croisée type k-fold est utilisée pour chaque combinaison d'hyperparamètres testés. Plusieurs fonctions de prédiction sont produites avec, pour chacune, l'estimation de leur performance de prédiction (Figure 5.18 étape 2). Ce sont les valeurs des hyperparamètres qui ont produit les meilleurs résultats lors de la procédure de validation croisée qui sont par la suite utilisés pour dimensionner l'espace d'hypothèses et sélectionner la fonction de prédiction optimale (Figure 5.18 étape 3). Le jeu test indépendant (période janvier-août 2016) est ensuite utilisé pour évaluer la performance de cette fonction de prédiction (Figure 5.18 étape 4). Enfin, cette fonction de prédiction validée par le jeu test est déployé et utilisé sur un nouveau jeu de données (ici événements détectés au cours de la procédure de détection automatique développée dans ce travail de thèse pour la période septembre-décembre 2016, (Figure 5.18 étape 5)).

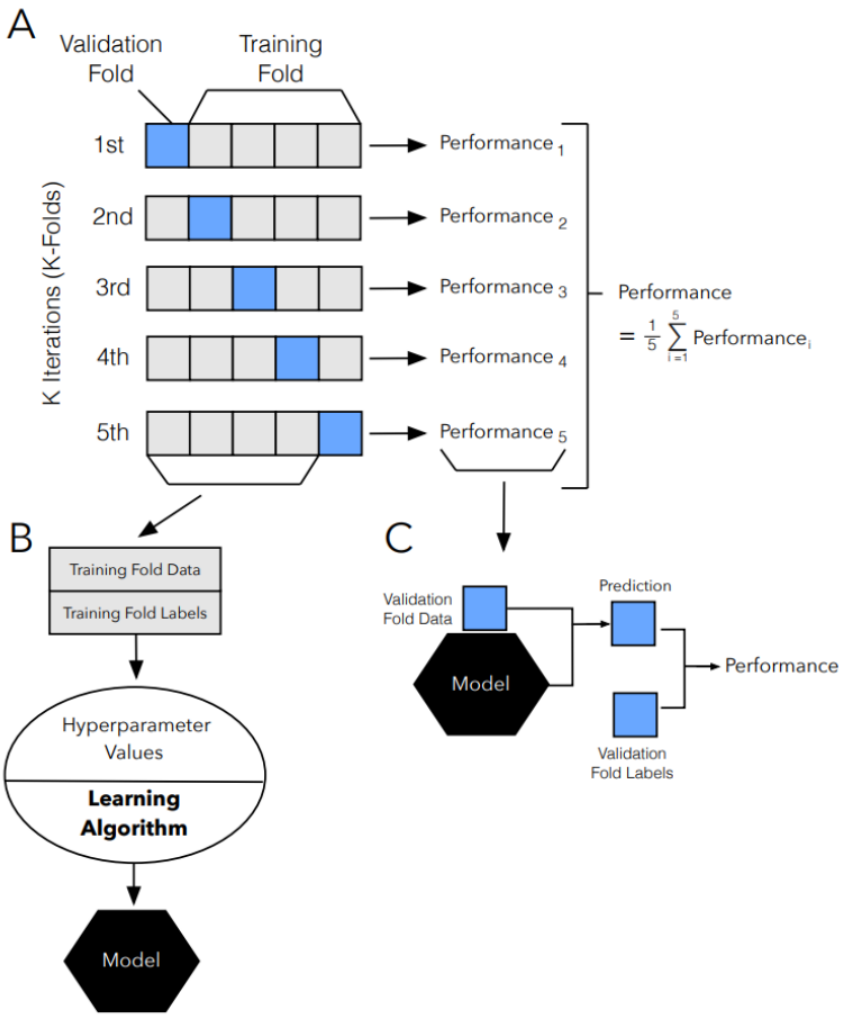


FIGURE 5.17: Principe de la validation croisée. Le processus d'apprentissage est itéré k fois (ici 5 fois). (a) A chaque itération le jeu de données est découpé en k parties (ici 5 parties) : une partie est utilisée pour la validation (c) et les $k-1$ (4) parties restantes sont fusionnées en un sous-ensemble d'entraînement pour l'apprentissage à partir d'une combinaison d'hyperparamètres donnée (b). La fonction de prédiction apprise est testée avec le jeu de validation (c). La performance globale de la validation croisée (par exemple le calcul de la précision) correspond à la moyenne arithmétique des k (ici 5) estimations de performance de la fonction de prédiction sur les ensembles de validation (a). D'après RASCHKA, 2018

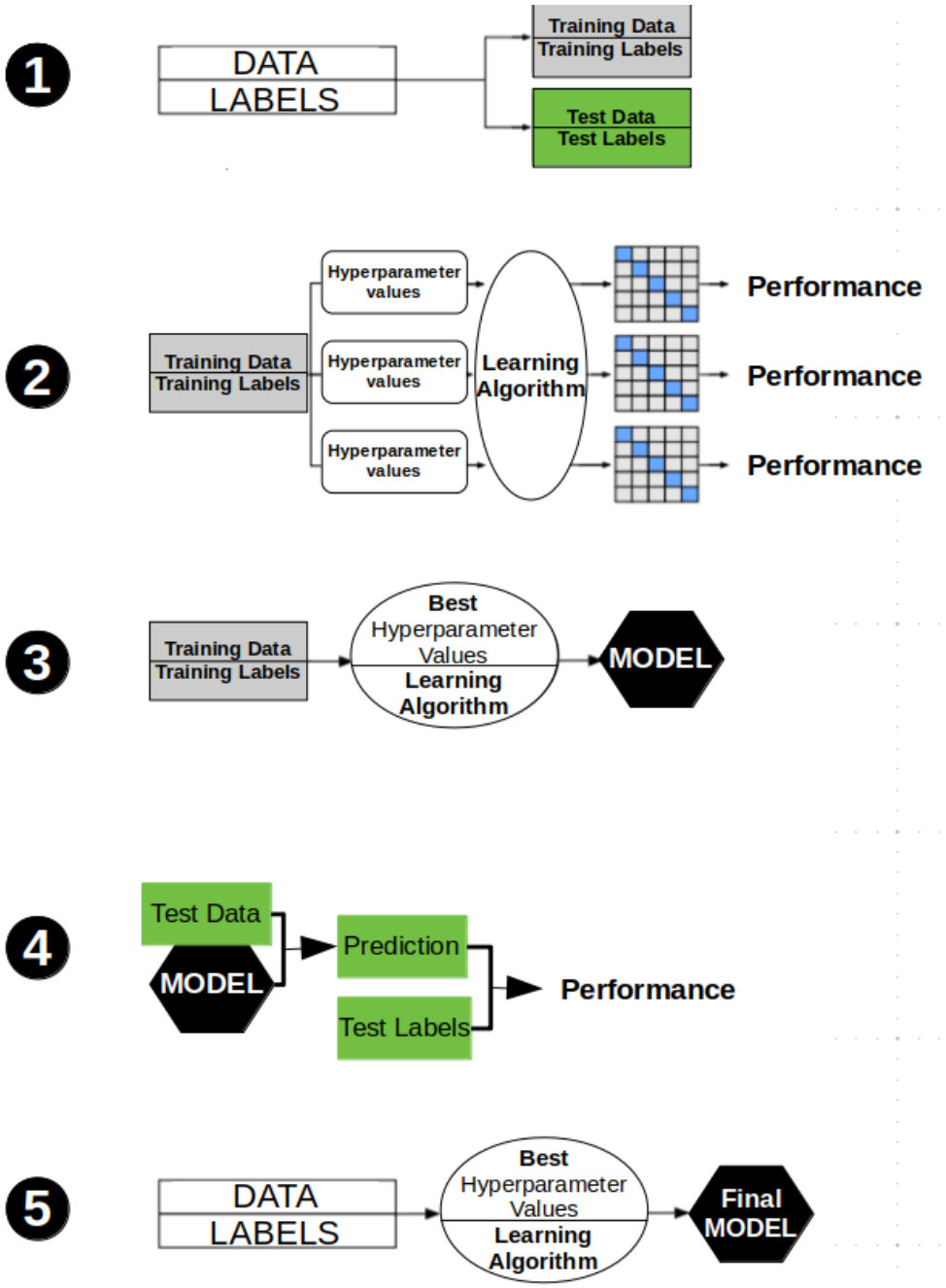


FIGURE 5.18: Intégration de la procédure de recherche des hyperparamètres optimaux par validation croisée dans la procédure d'apprentissage dans le but de trouver la fonction de prédiction qui minimise l'erreur de généralisation. D'après RASCHKA, 2018.

L'algorithme d'apprentissage Random Forest permet de générer une fonction de prédiction moyenne stable grâce à l'agrégation d'un ensemble d'arbres décisionnels complètement indépendants. La configuration de ces hyperparamètres (i.e nombre d'arbres dans la forêt aléatoire, profondeur d'un arbre décisionnel, etc.) aide à établir un cadre de recherche optimal de cette fonction de prédiction à partir d'une classe d'hypothèses restreintes. Cette classe d'hypothèses restreintes permet à la fois de diminuer le biais associé à l'algorithme lui-même tout en veillant à maintenir une variance qui puisse être acceptable. De cette façon, le choix de Random Forest, couplé à la stratégie de recherche des hyper-paramètres optimaux par validation croisée type k-fold, est un bon compromis pour limiter le sur-apprentissage à partir d'une base d'entraînement de taille petite.

•Privilégier l'interactivité avec l'algorithme d'apprentissage

L'idée primordiale est qu'il y a un niveau de connaissance préalable du problème spécifique en question (ici la classification des faux événements, des séismes et des tirs de carrière) qui puisse permettre de concevoir des espaces d"hypothèses pour lesquels l'erreur d'approximation et l'erreur d'estimation ne soient pas trop grandes.

Plusieurs sources de connaissances préalables sont possibles et peuvent être intégrées dans le pipeline de l'apprentissage (Rueden et al., 2019). Elles peuvent provenir d'un groupe individuel de personnes ayant une expérience significative sur un domaine de connaissances donné. Il s'agit dans ce cas d'une connaissance d'expertise que peuvent avoir par exemple les analystes sur l'identification des tirs de carrière grâce aux formes d'onde. Ces connaissances préalables peuvent découler directement du savoir scientifique disciplinaire (propriétés intrinsèques du bruit par exemple) ou d'un savoir formalisé par une communauté scientifique particulière (comme le rapport d'amplitude maximale entre les ondes P et S en tant que facteur discriminant des tirs de carrière et des séismes). Ces connaissances peuvent enfin provenir d'un savoir intuitif partagé et validé implicitement par le raisonnement humain (comme par exemple le repérage des tirs de carrière à travers les épicentres qui sont situés très proches des carrières).

Lorsque la performance de l'apprentissage automatique est évaluée, celle-ci peut être effectivement exprimée à travers de nombreuses métriques dédiées telles que l'exactitude (proportion d'événements bien classé), la sensitivité (proportion de vrais événements ou de séismes correctement prédits par exemple) ou la spécificité (proportion de faux événements ou de tirs de carrière correctement prédits par exemple). Ces métriques évaluent donc la capacité prédictive des fonctions de prédiction apprises et soulèvent les taux d'erreurs de classification si le problème posé est de ce type. Néanmoins, ces dernières ne donnent pas accès au processus de décision qui a conduit la fonction de prédiction à opérer tel ou tel choix. Pourtant, cette transparence du chemin décisionnel est nécessaire pour juger si la fonction de prédiction est bien valide et généralisable ou si elle a fondé l'ensemble de ses décisions sur des corrélations erronées dans les données d'apprentissage (Lapuschkin et al., 2019).

La transparence du processus de décision peut être obtenue par l'intervention de l'être humain. En effet, à travers ses connaissances et son expertise, ce dernier est capable de révéler les corrélations parasitaires ou artefactuelles qui ont conduit l'algorithme d'apprentissage à générer une solution erronée et les corriger pour obtenir des stratégies de décision plus fiables (Gilpin et al., 2018). L'interaction entre le système d'apprentissage et l'utilisateur humain est donc une clef indispensable pour minimiser l'erreur de généralisation véhiculée par les fonctions apprises (S. Lundberg et al., 2017 ; Schramowski et al., 2020). L'interactivité est la voie que j'ai alors choisie dans ce travail de thèse. En effet, face à un jeu de données d'une grande dimensionnalité (nombre d'attributs mais également nature et quantité d'informations contenues dans chacun très élevées), l'inclusion des connaissances préalables dans le processus d'apprentissage va permettre de délimiter plus efficacement encore l'espace des hypothèses qui va servir à générer la fonction de prédiction optimale.

Cette interactivité peut se dessiner à plusieurs niveaux au cours du processus d'apprentissage (Figure 5.19). Cette interaction est effectivement nécessaire pour déceler des informations parasitaires ou artefactuelles ayant été incluses dans le processus d'apprentissage, pour vérifier la significativité des attributs sélectionnés et leur impact sur la construction des règles de classification, et pour valider la généralisabilité et la plausibilité de la fonction de prédiction apprise en étudiant son comportement sur plusieurs instances ou sur différents jeux de données (Figure 5.19).

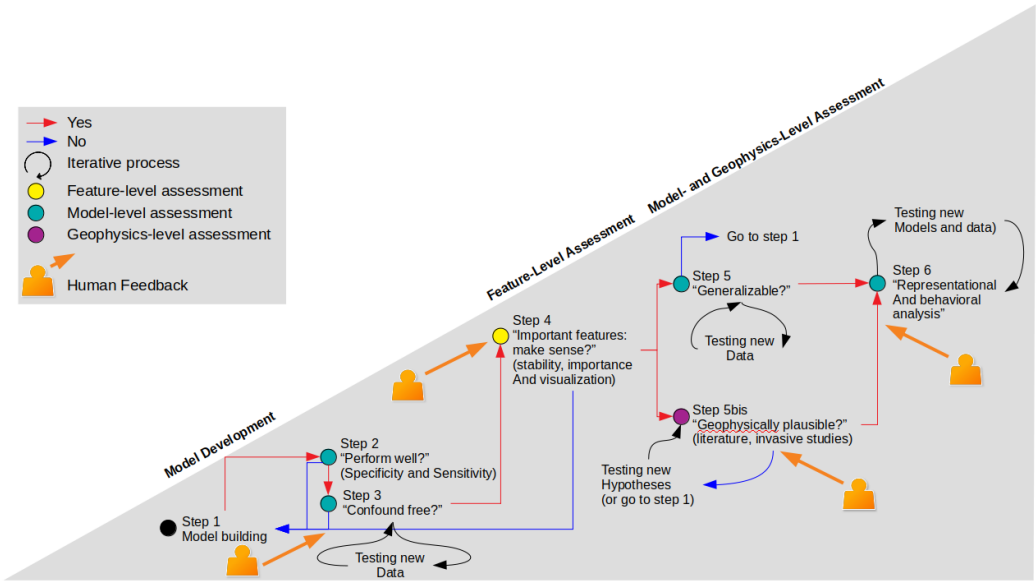


FIGURE 5.19: Procédure d'apprentissage qui implémente un cadre unifié d'interprétation de la fonction de prédiction (modèle). La première étape (étape 1) est la construction du modèle. Cette étape est une condition pré-requise qui reste en dehors du cadre d'interprétation. Les propriétés basiques du modèle sont évaluées à travers son pouvoir prédictif (étape 2) et sa capacité à inclure dans sa prédiction des informations parasitaires ou artefactuelles (étape 3). Dans le cas où chacune des étapes ne montre pas un modèle de qualité suffisante, le modèle et/ou la qualité des données sont revus (retour à l'étape 1). Si le modèle passe ce contrôle qualité, la prochaine étape est celle de l'estimation de ce modèle à l'échelle des attributs (étape 4). Plusieurs options sont possibles pour identifier les attributs significatifs (ex : tests de bootstrap, sélection des attributs basée sur leur importance relative). Si les attributs significatifs identifiés ne fournissent pas des résultats sensibles, la construction du modèle est révisée (étape 1). Dans le cas contraire, la généralisabilité et la plausibilité du modèle sont testés. La généralisabilité est testée à travers de nouvelles données (étape 5). La validité géophysique du modèle est examinée à travers les résultats de la littérature par exemple (étape 5bis). Cette étape peut également être effectuée plusieurs fois au cas où le modèle suggère de nouvelles théories qui devraient être évaluées. La dernière étape est l'analyse représentationnelle et comportementale du modèle (étape 6) et contribue à mieux comprendre les processus de décision du modèle, en examinant son comportement sur plusieurs instances ou sur plusieurs jeux de données. Cette dernière étape peut nécessiter souvent d'autres modèles pour permettre une comparaison. Cependant, si d'autres modèles sont déjà disponibles, cette étape peut être effectuée plus tôt. Enfin, les résultats de l'étape 6 pourraient fournir des preuves convergentes pour l'étape 5bis. Modifié d'après KOHOUTOVA et al., 2020.

Seulement, pour rendre opérable cette interactivité, il faut pouvoir visualiser les processus de décision de la fonction de prédiction apprise, avoir accès aux attributs sélectionnés ainsi que leur importance relative et avoir une connaissance de la valeur physique de ces attributs que la littérature peut nous fournir.

Choisir un algorithme d'apprentissage interprétable est donc indispensable. La notion d'algorithme interprétable est ici à relier à la capacité humaine de comprendre comment l'algorithme d'apprentissage utilise les attributs en entrée pour choisir ses prédictions (S. Lundberg et al., 2017). L'algorithme de Random Forest a donc été aussi choisi car ce dernier se base sur l'édification d'un arrangement hiérarchique de règles de classification qui sont assez facilement interprétables (Doshi-Velez et al., 2017 ; Drouin et al., 2019). En effet, la visualisation des chemins décisionnels à travers un arbre suffit à comprendre comment et pourquoi la fonction de prédiction peut arriver à sa prédiction (Samek, 2020, Figure 5.20).

De plus, l'analyse de l'arbre apporte une information capitale donnée par l'impureté de Gini (affichée dans chaque noeud de l'arbre). En effet, cette impureté de Gini donne accès à l'importance de chaque attribut dans le processus de classification, soit son pouvoir discriminant.

L'impureté de Gini est une métrique utilisée pour déterminer quel est l'attribut qui doit être utilisé et avec quel seuil pour pouvoir fractionner les données en des groupes plus petits (passage d'un noeud parent à deux noeuds fils dans l'arbre). Ce critère mesure la fréquence à laquelle une observation aléatoirement choisie dans la base d'entraînement serait incorrectement labélisée si elle était aléatoirement labélisée selon la distribution des labels dans l'échantillon formé au noeud (c'est-à-dire si la moitié des observations dans l'échantillon est "A" et l'autre moitié est "B", une observation aléatoirement labélisée en se basant sur la composition de cet échantillon a 50% de chance d'être labélisée incorrectement). L'impureté de Gini atteint 0 quand toutes les observations dans l'échantillon tombent dans une seule catégorie (c'est-à-dire s'il y a seulement un label possible dans l'échantillon, une observation sera identifiée avec ce label 100% du temps). Cette mesure est donc essentiellement la probabilité qu'une nouvelle observation soit incorrectement classifiée à un noeud donné dans un arbre décisionnel, en se basant sur le jeu d'entraînement.

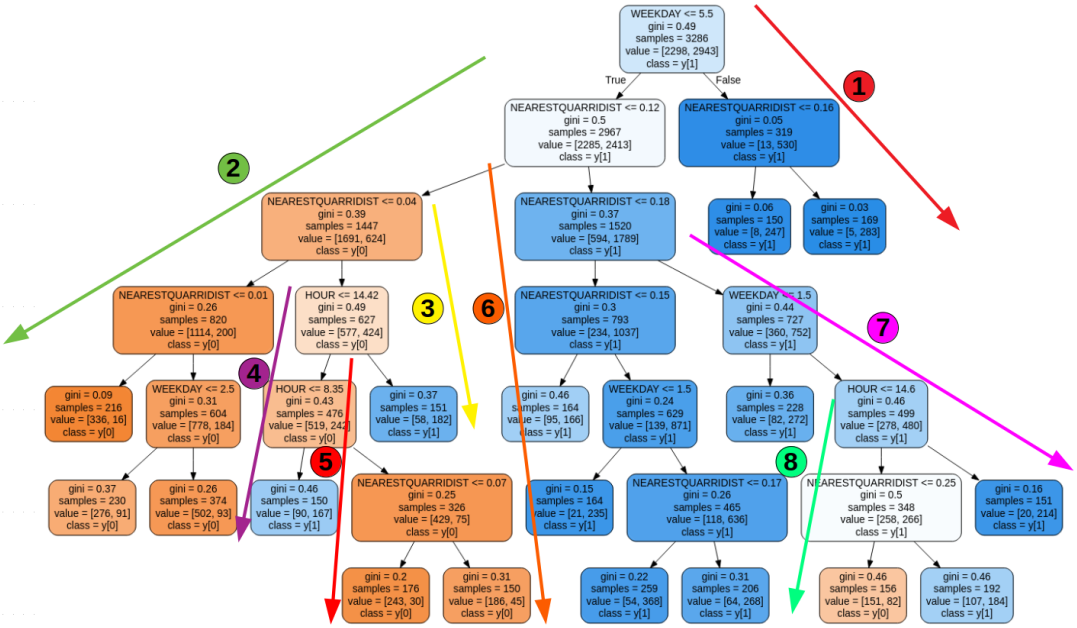


FIGURE 5.20: Exemple d'arbre décisionnel généré par l'algorithme d'apprentissage Random Forest pour classer les séismes et tirs de carrière détectés au cours de la période 2017-2019 (cf Figure 5.14 pour plus de détails). L'analyse des différents chemins décisionnels de cet arbre montre d'abord que lorsque le jour de la semaine est samedi (> 5.5), quelque soit la proximité de l'événement à une carrière, les événements répondant à ces critères sont classés comme séismes (classe 1, chemin 1). Si maintenant le jour de la semaine est différent de samedi, plusieurs configurations sont possibles. Si la proximité de l'événement à la carrière la plus proche est inférieure ou égale à 0.04 degré (soit $\leq 4.45km$) alors tous les événements répondant à ce dernier critère sont classés comme étant des tirs de carrière (classe 0, chemin 2). Si en revanche, la distance de l'événement à la carrière la plus proche est comprise entre 0.04 degré (4.45 km) et 0.12 degré (13.34 km) et que l'heure d'occurrence de cet événement est avant 8.35 h ou après 14.42 h, alors les événements qui possèdent ces derniers critères sont classés comme séismes (classe 1, chemin 4 et chemin 3), sinon si l'heure d'occurrence est comprise entre 8.35 h et 14.42 h, ces événements sont identifiés comme des tirs de carrière (classe 0, chemin 5). De plus, si la distance de l'événement à la carrière la plus proche est comprise entre 0.12 degré (13.34 km) et 0.18 degré (20 km), et que le jour de la semaine n'est toujours pas samedi, alors les événements qui tombent dans cette catégorie de valeurs sont classés comme des séismes (classe 1, chemin 6). Enfin, si la distance de l'événement à la carrière la plus proche est supérieure ou égale à 0.18 degré (20 km), que le jour de la semaine est compris entre mardi et vendredi, et que l'heure d'occurrence de l'événement se situe après 14.6h, alors les événements qui répondent à ces critères sont identifiés comme des séismes (classe 1, chemin 7). Néanmoins, si l'heure d'occurrence de l'événement se situe avant 14.6 h et que la distance de l'événement à la carrière la plus proche est comprise entre 0.18 degré (20 km) et 0.25 degré (27.79 km), alors les événements sont plutôt classés comme des tirs des carrière (classe 0, chemin 8).

Parce que les forêts aléatoires sont un ensemble d'arbres décisionnels individuels, l'impureté de Gini peut être mise à profit pour estimer l'importance des attributs en calculant la diminution moyenne de l'impureté de Gini entre les noeuds parent et fils que l'attribut divise. En effet, après chaque fractionnement binaire à partir d'un noeud parent, les noeuds fils générés doivent avoir un coefficient de Gini inférieur, car le but des fractionnements est de rendre les distributions des observations dans les nœuds fils aussi pures que possible (c'est-à-dire une impureté de 0). En effet, toutes les observations dans un nœud doivent tendre vers un maximum de similarité pour pouvoir progressivement atteindre une prédiction finale à un noeud terminal pour une unique classe d'observation. Par conséquent, l'attribut qui a été utilisé pour scinder le noeud parent en deux noeuds fils a diminué l'impureté de Gini.

Ainsi, si la diminution moyenne de l'impureté de Gini est calculée pour chaque attribut utilisé dans les arbres de la forêt, il est alors possible de déduire le degré d'importance de chacun. Ce calcul correspond à la somme moyennée des diminutions de l'impureté pour tous les noeuds où l'attribut est utilisé, pondérée par la proportion des échantillons qui atteignent ce noeud dans chaque arbre décisionnel de la forêt aléatoire (Louppe et al., 2013). C'est donc un calcul qui peut évaluer le degré d'importance d'un attribut à travers tous les arbres qui forment la forêt. Une valeur de diminution moyenne de l'impureté de Gini élevée indiquera une importance élevée de l'attribut.

Si j'évalue l'importance relative des attributs pour la forêt aléatoire auquel l'arbre décisionnel de la Figure 5.20 appartient, il est possible de constater que l'heure d'occurrence des événements et la distance de l'événement à la carrière la plus proche sont les attributs les plus discriminants (Figure 5.21).

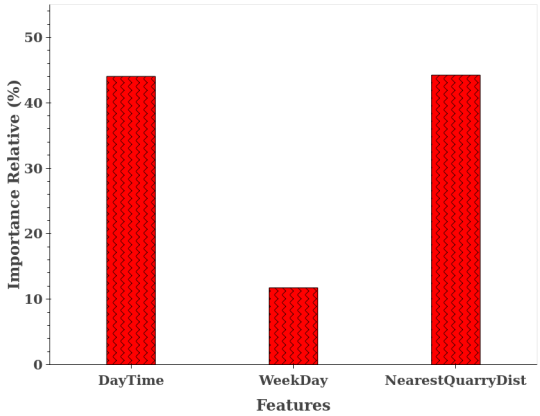


Figure 5.21: Importance relative des attributs calculée à partir de la forêt aléatoire contenant l'arbre décisionnel présenté dans la Figure 5.20 pour la prédiction des labels des séismes et des tirs de carrière du catalogue BCSF-RéNaSS pour la période 2017-2019.

L'étude de l'arbre décisionnel (élaborée dans la Figure 5.20) le confirme. En effet, d'après cet arbre décisionnel, tous les événements qui sont situés à moins de 4.5 km d'une carrière la plus proche sont considérés comme des tirs de carrière. Lorsque cette distance augmente, c'est l'heure d'occurrence de l'événement qui va déterminer son label. En l'occurrence, pour des distances supérieures à 4.5 km mais inférieures à 13 km, si l'heure est comprise entre 8.35 h et 14.42 h, c'est-à-dire dans la période d'activité maximale classique des carrières dans la zone d'étude mais aussi ailleurs dans le monde (Voyles et al., 2019), les événements qui tombent dans cette intervalle de valeurs sont identifiés comme des tirs de carrière. Dès que la distance s'éloigne de 13 km, la plupart des événements sont classés comme des séismes à l'exception d'un échantillon.

Si cette classification apparaît globalement pertinente (un tir de carrière est un événement qui est situé très proche d'une carrière et a lieu aux heures traditionnelles d'activité des carrières), elle reste tout de même fragile. En effet, un échantillon d'événements est d'abord classé comme tir de carrière alors que ces derniers sont localisés à une distance comprise entre 20 et 28 km d'une carrière la plus proche. Plusieurs hypothèses sont possibles pour expliquer ce résultats : soit ces événements sont en fait mal identifiés, soit ils sont bien classés mais très mal localisés et/ou la carrière qui leur est associée n'a pas été répertoriée dans la base de données des carrières.

Par ailleurs, les attributs sélectionnés vont conduire à classer les séismes par défaut : ce sont des événements qui sont très éloignés des carrières, ou lorsqu'ils sont plus proches, n'ont pas lieu aux heures d'activité maximale des carrières. Ce qui est alors très restrictif.

Or, pourtant, si la performance prédictive de ce classifieur est évaluée, il est possible d'observer que 80% des tirs de carrière et 90% de séismes ont été bien classés. Ce qui n'est, au premier abord, pas si mauvais. Seulement, l'analyse des arbres décisionnels amène à penser à un sur-apprentissage.

En effet, une proportion non négligeable de tirs de carrière ont bien lieu en dehors des pics traditionnels d'activité des carrières, de nombreux tirs de carrière sont également mal localisés et la base de données des carrières, même si riche, n'est certainement pas exhaustive. De même, les séismes sont également bien détectés pendant les pics d'activité des carrières et peuvent même être localisés non loin des sites de ces carrières. Un taux élevé d'exceptions à cette règle de classification souligne la forte instabilité de cette fonction de prédiction, et son incapacité à généraliser. Même si ces résultats offrent une base intéressante, d'autres attributs doivent être considérés pour affiner cette classification primordiale. On observe là la nécessité absolue d'apporter une expertise humaine pour estimer la validité d'une fonction de prédiction apprise.

Seulement, j'ajouterai que si l'analyse d'un arbre décisionnel est à portée de l'Homme, cette analyse devient plus difficile lorsqu'il s'agit d'évaluer l'ensemble des arbres (de l'ordre de plusieurs centaines) qui composent la forêt aléatoire. L'apport d'outils automatisés pour analyser ce flux d'arbres est donc d'un grand apport et reste une perspective intéressante à approfondir (Lapuschkin et al., 2019 ; Samek, 2020, Figure 5.22).

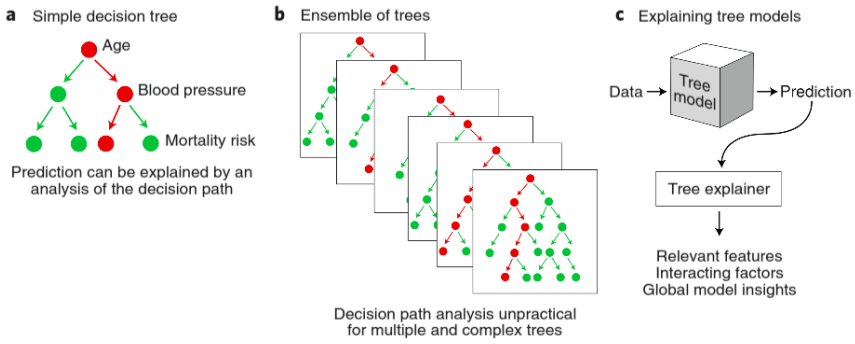


Figure 5.22: (a) Des arbres décisionnels simples peuvent être facilement compris en visualisant le chemin de décision. (b) Du fait de leur complexité, des modèles de pointe basés sur un ensemble d'arbres deviennent extrêmement difficiles à interpréter dans leur totalité. (c) Des outils automatisés (par exemple TreeExplainer, S. M. Lundberg et al., 2020) sont nécessaires pour extraire les attributs pertinents et trouver les effets d'interaction dans les modèles basés sur les arbres. D'après Samek, 2020.

L'être humain peut donc vérifier en partie le choix pertinent des attributs (qui sont préalablement sélectionnés automatiquement) et peut valider les règles de classification élaborées par l'algorithme de Random Forest (seuil utilisé pour générer un noeud, organisation hiérarchique des noeuds, pertinence de la prédiction au regard de la règle de classification choisie, etc.), réduisant l'espace d'hypothèses possibles à un espace plus riche et pertinent.

Choix de l'algorithme, sélection des hyperparamètres optimaux et forte interactivité avec les connaissances préalables humaines sont les trois grands facteurs qui vont aider à délimiter un espace d'hypothèses riche et pertinent dans le but de sélectionner une fonction de prédiction qui minimise l'erreur de généralisation malgré les contraintes apportées par mon jeu de données disponible.

•Optimiser la résolution du problème de classification posé

Si une seule fonction de prédiction est utilisée pour simultanément prédire les labels des faux événements, des tirs de carrière et des séismes, la performance de classification est diminuée. En effet, comme il a été décrit précédemment dans le chapitre 2, le bruit non-stationnaire à l'origine de la détection des faux événements est souvent de même ordre d'amplitude, de contenu fréquentiel et de durée que les séismes ou les tirs de carrière. De plus, étant donné que la procédure de détection détecte un taux élevé d'événements de faible magnitude, les signaux associés à ces événements se détachent très souvent à peine du bruit.

En considérant une seule fonction de prédiction, une sélection automatique d'un sous-ensemble optimal d'attributs par élimination récursive à partir du pool initial des 361 attributs (cf tableau S1 du supplément de l'article) révèle que les attributs qui possèdent l'importance la plus forte ne sont pas ceux qui décrivent les caractéristiques du signal associé à l'événement à classer. A la place, ce sont des attributs comme le nombre de phases total, le nombre de phases S, le facteur de corrélation entre la différence des temps d'arrivée P-S et la distance épicentrale, le nombre de stations, la RMS des résidus temporels ou bien l'heure d'occurrence des événements qui ressortent principalement (Figure 5.23).

Si effectivement un faux événement peut être globalement plus facilement classé par rapport à un séisme avec ces critères : ce dernier possède un nombre plus faible de phases, un nombre presque négligeable de phases S, un facteur de corrélation entre la différence des temps d'arrivée P-S et la distance épicentrale plus faible, un nombre de stations impliquées dans la détection de l'événement plus petit, une RMS des résidus temporels plus élevée, et une heure d'occurrence concentrée à la période d'activité anthropique la plus élevée (c'est-à-dire entre 7 heures et 18 heures) ; le constat est plus difficile pour les tirs de carrière.

En effet, ces derniers ont souvent lieu aux périodes de la journée où le niveau de bruit d'origine anthropique est le plus élevé, leurs signaux sont donc fortement contaminés par du bruit. La détection des phases est alors plus difficile et beaucoup de tirs de carrière sont détectés avec très peu de stations. Ajouté à cela, les ondes S sont souvent de plus faible amplitude pour les tirs de carrière et il est couramment difficile d'identifier et de pointer les temps d'arrivée de ces ondes. Par ailleurs, ces événements sont très superficiels et peuvent avoir lieu dans des terrains sédimentaires, leurs solutions hypocentrales peuvent donc être localisées avec une plus grande incertitude puisque les modèles de vitesse qui sont utilisés pour détecter ne tiennent pas compte des variations latérales et verticales de l'épaisseur de la couche sédimentaire. Par conséquent, ces tirs de carrière peuvent partager les mêmes caractéristiques que les faux événements et être classés en tant que tels.

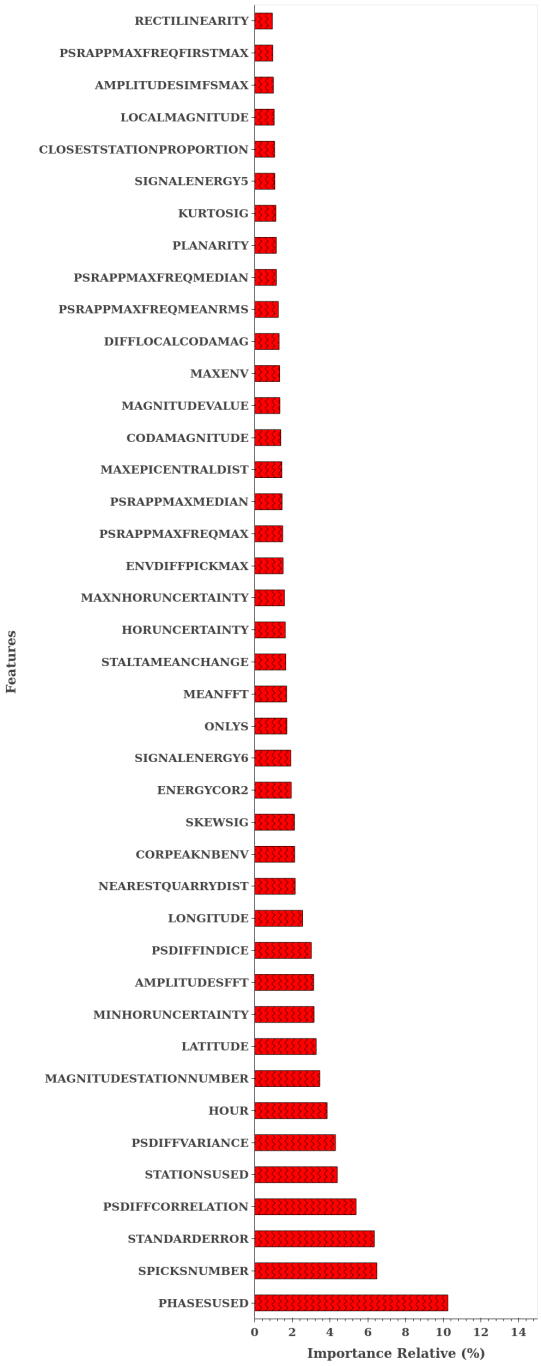


FIGURE 5.23: Importance relative des attributs pour la prédiction des labels des faux événements, des séismes et des tirs de carrière avec un seul classifieur. La sélection automatique des attributs a été effectuée par élimination récursive pour obtenir un sous-ensemble optimal qui est visualisé ici (c'est une sélection primordiale). Cette combinaison a été évaluée par validation croisée sur 5 itérations (jeu de données de la période 2017-2019) puis testée sur un nouveau jeu test comprenant les événements détectés par le BCSF-RéNaSS entre janvier et août 2016. L'équilibre de classe a été respectée dans les jeux d'entraînement et de validation. Le jeu test a une taille qui correspond à 30% de celle du jeu d'entraînement. Le détail des attributs est présenté dans le tableau 1 du supplément de l'article qui va suivre.

Lorsque cette fonction de prédiction apprise à partir d'un jeu d'entraînement élaboré à partir de cette sélection primordiale d'attributs (toujours la période 2017-2019) est utilisée pour prédire le labels des événements d'un jeu test (période janvier 2016-août 2016), on constate que 25% des tirs de carrière sont effectivement identifiés comme des faux événements contre 7% pour les séismes.

En revanche, si deux fonctions de prédiction sont désormais apprises, une pour prédire les faux événements et les vrais événements et une pour prédire les séismes et les tirs de carrière parmi les vrais événements identifiés, on constate que 15% des tirs de carrière sont identifiés comme faux événements contre 5% pour les séismes.

Avec deux fonctions de prédiction, si les attributs liés à la configuration du réseau de stations ressortent également, le résultat de la sélection automatique des attributs par élimination récursive montre également que des attributs reliés au signal comme le degré de polarisation planaire ou bien le degré de complexité de la fonction STA/LTA prennent de 2 à 3 fois plus d'importance lorsque que le problème de la classification des faux événements est traité de manière binaire (c'est-à-dire faux événements versus vrais événements, Figure 5.24).

Par ailleurs, des attributs comme l'heure d'occurrence de l'événement et la variance des fréquences contenues dans le spectre du signal ne sont plus sélectionnés pour classifier les faux événements parmi les autres événements dans le cas d'une approche binaire, alors qu"ils le sont indéniablement dans une approche ternaire (faux événements versus tirs de carrière versus séismes). Or, si ces deux attributs peuvent être précieux pour distinguer un séisme d'un tir de carrière, ils amènent à des confusions lorsqu'ils sont utilisés pour identifier les faux événements parmi le reste des autres événements. En effet, ces faux événements présentent en moyenne statistiquement une variance spectrale qui se rapproche fortement de celle des tirs de carrière et leurs heures d'occurrence sont majoritairement comprises dans le même intervalle temporel que celui des tirs de carrière. Le problème se complexifie davantage s'il s'avère que des séismes répondent aussi à cet ensemble de critères, à savoir une variance spectrale plus faible, une heure d'occurrence identifiée dans le pic d'activité anthropique, un nombre inférieur de phases (dont les phases S) et une RMS résiduelle plus élevée.

Avec une approche binaire, la variance spectrale et l'heure d'occurrence de l'événement sont sélectionnés lorsqu'il s'agit uniquement de classer les tirs de carrière et les séismes. Leur importance relative est d'ailleurs très élevée (10% et 13% respectivement, Figure 5.25). Comparativement à la classification ternaire simultanée des tirs de carrière, des séismes et des faux événements, ces derniers sont donc utilisés plus fréquemment pour la construction des arbres décisionnels dans cette approche binaire.

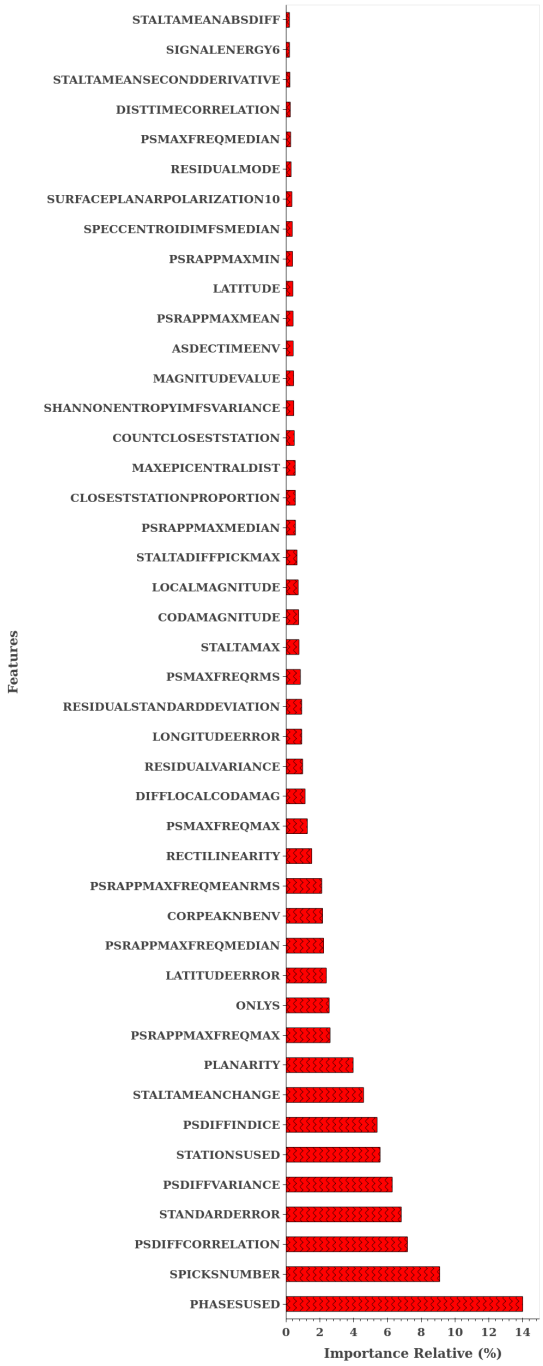


FIGURE 5.24: Importance relative des attributs pour la prédiction des labels des faux événements et des vrais événements (ensemble unitaire de séismes et de tirs de carrière) avec une approche binaire. La sélection des attributs a été effectuée automatique par élimination récursive pour combiner un sous-ensemble optimal qui est visualisé ici (c'est une sélection primordiale). Cette combinaison a été évaluée par validation croisée sur 5 itérations (jeu de données de la période 2017-2019) puis testée sur un jeu test comprenant les événements détectés par le BCSF-RéNaSS entre janvier et août 2016. L'équilibre de classe a été respectée dans le jeu d'entraînement. Le jeu test a une taille qui corresponde à 30% de celle du jeu d'entraînement. Voir tableau S1 du supplément de l'article qui va suivre pour plus de détails sur les attributs.

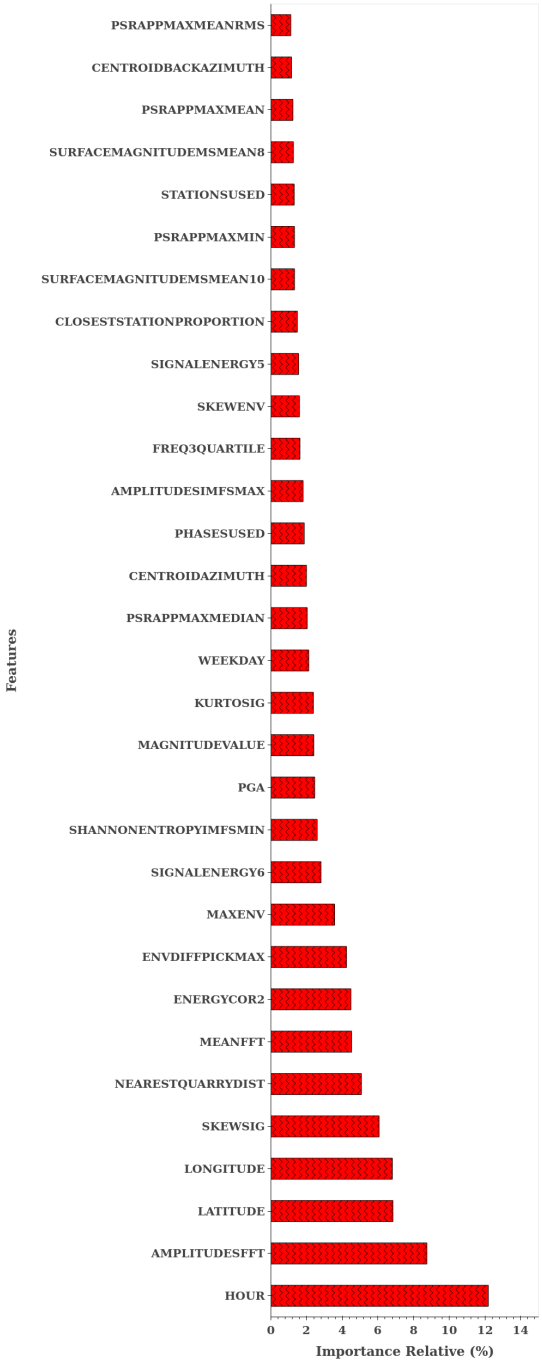


FIGURE 5.25: Importance relative des attributs pour la prédiction des labels des séismes et des tirs de carrière avec une approche binaire. La sélection des attributs a été effectuée automatique par élimination récursive pour combiner un sous-ensemble optimal qui est visualisé ici (c'est une sélection primordiale). Cette combinaison a été évaluée par validation croisée sur 5 itérations (jeu de données de la période 2017-2019) puis testée sur un jeu test comprenant les événements détectés par le BCSF-RéNaSS entre janvier et août 2016. L'équilibre de classe a été respectée dans le jeu d'entraînement. Le jeu test a une taille qui corresponde à 30% de celle du jeu d'entraînement. Voir tableau S1 du supplément de l'article qui va suivre pour le détail des attributs.

J'ai donc privilégié dans ce travail de thèse, une approche binaire séquentielle. Deux fonctions de prédiction sont effectivement apprises : une pour prédire les faux événements et les vrais événements et une autre pour prédire les séismes et les tirs de carrière parmi les vrais événements. En s'attardant plus spécifiquement sur les attributs qui vont d'abord définir ce qu'est un faux événement relativement à un vrai événement, puis un séisme naturel relativement à un tir de carrière, cette approche limite les effets parasitaires d'une approche ternaire en éliminant des corrélations artefactuelles comme celle de lier l'heure d'occurrence de l'événement avec son incertitude de localisation et la variance spectrale des signaux associés. Cette approche binaire permet alors de mieux solidement gérer l'hétérogénéité et la complexité des données.

En outre, l'utilisation de l'algorithme de Random Forest, combinée à une sélection optimale de ses hyperparamètres de configuration, permet de contrebalancer les effets liés à la taille, petite, et à la dimensionnalité, assez élevée, du jeu de données, en minimisant les erreurs d'approximation et d'estimation.

Par ailleurs, l'apport de connaissances préalables dans le processus d'apprentissage, transmises par l'interaction avec l'être humain, aide à déceler les informations parasitaires véhiculées par le jeu de données qui peuvent conduire à une sélection automatique d'attributs redondants et/ou non significatifs. Cette interactivité est aussi un garant pour estimer la validité et la plausibilité de la fonction de prédiction apprise (cohérence et significativité des règles de classification, pertinence des attributs, généralisabilité des règles apprises, etc.). Cette interactivité offre alors un cadre structurel à un jeu de données qui présente des informations très hétérogènes et diversifiées, et donc qui offre un espace de solutions prédictives multiples mais pas toutes vraisemblables.

Enfin, le déséquilibre des classes d'événements au sein du jeu de données mère est corrigé par l'inclusion de faux événements revus manuellement et un sous-échantillonnage des autres classes d'événement (séismes et tirs de carrière).

5.2 Choisir la fonction de prédiction optimale dans l'espace des hypothèses possibles

5.2.1 Rechercher la combinaison optimale d'attributs

Comme écrit précédemment, le jeu de données qui a servi pour l'entraînement et la validation croisée comprend les vrais événements détectés par le BCSF-RéNaSS entre janvier 2017 et décembre 2019. Ce jeu de données a été complété avec un lot de faux événements détectés automatiquement au cours des mois de juillet et août 2016 pour assurer la discrimination des vrais et des faux événements. Ces derniers ont d'ailleurs été revus manuellement. Les proportions des différentes classes d'événements ont été équilibrées dans le jeu d'entraînement, et leur représentativité a été estimée dans les différents jeux de validation.

Les valeurs des hyperparamètres utilisés pour contraindre l'espace des hypothèses possibles sont répertoriées dans le tableau suivant (Table 5.1).

TABLE 5.1: Hyperparamètres optimaux utilisés pour contraindre l'espace des hypothèses possibles avec l'algorithme d'apprentissage Random Forest.

	Hyperparameter Value
Tree Depth	150
Minimum number of samples required to split a node	5
Minimum number of samples required at a leaf node	5
Number of Trees	500

•**Rechercher les attributs pour classer les vrais et faux événements**

En ce qui concerne la discrimination des faux événements et des vrais événements, l'utilisation du sous-ensemble optimal d'attributs extrait de la procédure d'élimination récursive des attributs (cf Figure 5.24), après avoir testé sa performance via validation croisée sur 5 itérations, montre que celle-ci aboutit à un classifieur capable de prédire correctement 92% des vrais événements et 99 % des faux événements sur un jeu test contenant les vrais événements détectés par le BCSF-RéNaSS entre janvier 2016 et août 2016, complété par environ 2500 faux événements (Figure 5.26).

En revanche, lorsque ce classifieur est utilisé sur un nouveau jeu de données détecté automatiquement selon la procédure qui a été développée dans cette thèse (septembre 2016-décembre 2016), la performance prédictive de ce dernier se dégrade fortement pour les vrais événements : seulement 60% d'entre eux sont correctement prédits. En revanche, 98% des faux événements sont quant à eux bien prédits (Figure 5.26).

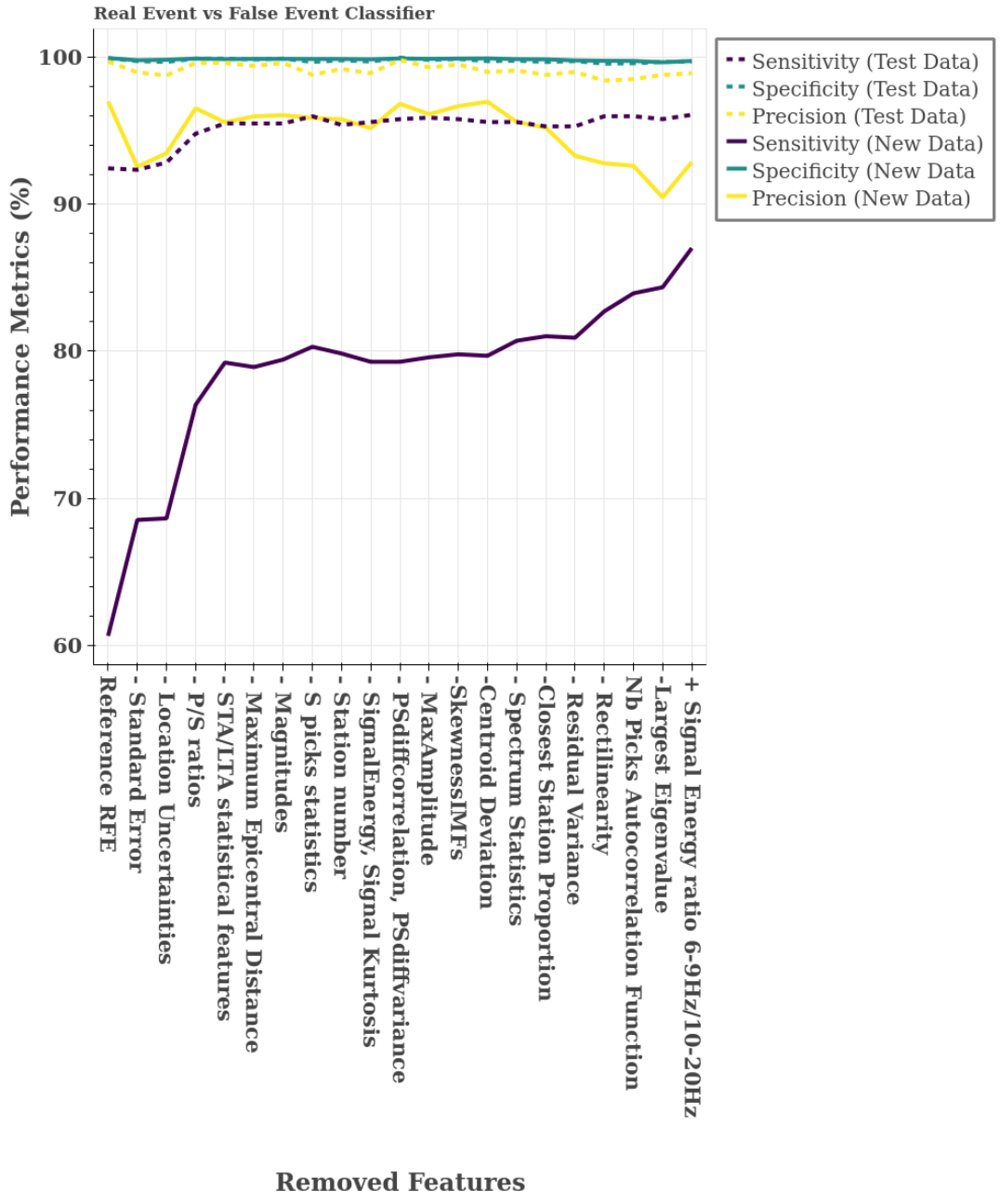


FIGURE 5.26: Effets du retrait (symbole -) et/ou de l'ajout (symbole +) d'attributs sur la capacité prédictive de classifieurs discriminant les vrais événements et les faux événements. La spécificité désigne le taux de faux événements correctement prédits (c'est-à-dire le rapport des vrais négatifs sur la somme des vrais négatifs et des faux positifs). La sensitivité désigne le taux de vrais événements correctement prédits (c'est-à-dire le rapport des vrais positifs sur la somme des vrais positifs et des faux négatifs). La précision désigne la proportion de vrais événements correctement prédits relativement à l'ensemble des événements prédits positivement (c'est-à-dire le rapport entres les vrais positifs et la somme des vrais positifs et des faux positifs). Référence RFE = sélection automatique des attributs effectuée par élimination récursive des attributs (cf Figure 5.24). Voir tableau S1 du supplément de l'article pour le détail des attributs.

Parmi les attributs automatiquement sélectionnés par la procédure d'élimination récursive des attributs, plusieurs attributs contribuent à diminuer la performance prédictive du classifieur vis-à-vis du jeu d'événements produits automatiquement. L'erreur standard (RMS des résidus temporels) est d'abord en moyenne plus élevée pour ce jeu automatique : $1.32 \pm 1.17s$ contre $0.37 \pm 0.10s$ pour les jeux d'entraînement, de validation et de test. De nombreux vrais événements compris dans le jeu automatique ont donc des erreurs standards qui peuvent se rapprocher des erreurs standards des faux événements, qui sont en moyenne égales à $2.56 \pm 2.04s$.

De plus, les épicentres des vrais événements du jeu automatique ont des incertitudes longitudinales et latitudinales élevées, respectivement $9.46 \pm 9.41km$ et $9.51 \pm 9.60km$. Ces incertitudes épicentrales sont effectivement 2 à 3 fois plus élevées que celles estimées pour les vrais événements compris dans les jeux d'entraînement, de validation et de test. Là encore, les incertitudes épicentrales de nombreux vrais événements tendent à s'approcher des incertitudes estimées pour les épicentres des faux événements (incertitudes latitudinales et longitudinales moyennes égales à $13.96 \pm 13.39km$ et $14.44 \pm 14.26km$). Par ailleurs, les écarts-types calculés sur les incertitudes latitudinales et longitudinales montrent une grande dispersion des valeurs pour les vrais événements du jeu automatique et la totalité des faux événements.

Or, les attributs décrivant l'erreur standard et les incertitudes épicentrales font partie des attributs qui ont une importance relative forte (de 3 % pour les incertitudes longitudinales à 7 % pour la RMS des résidus). Par conséquent, étant donné qu'un quart des vrais événements présentent des incertitudes latitudinales et longitudinales supérieures à 10 km et que environ 13% ont une erreur standard supérieure à 2 s, ces attributs ont été retirés.

Le retrait de ces attributs n'a pas beaucoup d'effet sur la performance prédictive du nouveau classifieur vis-à-vis du jeu test (celui-ci prédit correctement les vrais et faux événements avec la même performance prédictive que le classifieur précédent). En revanche, la nouvelle fonction de prédiction générée à partir de la nouvelle combinaison d'attributs améliore la qualité de prédiction des vrais événements contenus dans le jeu automatique, en prédisant correctement environ 70% d'entre eux, tout en maintenant un taux élevé de prédictions correctes des faux événements (Figure 5.26).

De même, les attributs décrivant les rapports d'amplitude et spectraux entre les ondes P et S ont été retirés de la combinaison des attributs. Les valeurs de ces différents rapports sont en moyenne plus élevées pour les faux événements. Seulement, ces valeurs expriment plus une conséquence du processus de génération des pointés qu'une propriété physique à relier aux faux événements eux-mêmes. En effet, les variations d'amplitude, souvent très impulsives, du bruit non-stationnaire haute fréquence conduit à assimiler ces variations à une arrivée d'ondes P de forte amplitude et à haute fréquence.

De plus les temps d'arrivée des ondes S étant détectés une fois que les pointés P sont émis, les pointés "S" dans le cas des faux événements sont émis dans le sillage des faux pointés "P".

Afin de ne pas introduire des corrélations artefactuelles (des rapports d'amplitude et spectraux P/S élevés associés à des faux événements alors que ce sont des rapports artefactuels), les attributs reliés à la description complète de ces rapports P/S dans les domaines temporel et fréquentiel ont donc été retirés. Ceci permet d'éviter les erreurs de prédiction pour des vrais événements dont les rapports d'amplitude et spectraux P/S sont élevés. Le retrait de l'ensemble de ces attributs induit une performance prédictive supérieure pour les vrais événements, quel que soit le jeu de données : le nouveau classifieur détecte respectivement 95% et 76% des vrais événements pour le jeu test et le jeu automatique, tout en maintenant de haut niveaux de prédiction correcte des faux événements.

Par ailleurs, la plupart des attributs dépeignant la fonction STA/LTA (évolution du rapport STA/LTA au cours du temps), ont été supprimés, à l'exception de la valeur maximale du rapport STA/LTA. L'estimation du degré de complexité de cette fonction STA/LTA traduit les fortes fluctuations liées au bruit enregistré. Les attributs décrivant statistiquement cette fonction STA/LTA constituent donc de forts discriminants pour identifier les faux événements. Néanmoins, de nombreux vrais événements détectés avec de faibles rapports signal/bruit peuvent être également associés à une fonction STA/LTA complexe, sensible au niveau de bruit contaminant le signal sismique détecté. Cette sensibilité est fortement exacerbée puisqu'une fenêtre temporelle STA de durée courte (0.5 s) a été initialement choisie. Par conséquent, ces attributs pouvant introduire facilement de la confusion, ils ont alors été retirés. Ce retrait a conduit à une amélioration de la capacité prédictive du classifieur résultant qui est en mesure de prédire correctement près de 96% des vrais événements pour le jeu test et 79% pour le jeu automatique (Figure 5.26).

De la même manière, des attributs tels que la distance épicentrale maximale, les magnitudes des événements (magnitude locale et magnitude de coda), la proportion de pointés S ou bien la déviation de l'événement par rapport au centroïde des stations impliquées dans la détection de cet événement, ont été supprimés de la combinaison d'attributs. Les faux événements étant détectés à partir d'une association de faux pointés décorrélés entre eux, ces derniers sont alors détectés avec des distances épicentrales élevées, leurs magnitudes sont alors généralement élevées et la déviation de l'événement par rapport au centroïde des stations est forte. De plus, la proportion des phases S dans l'association est faible puisqu'il n'y a pas à proprement parler des arrivées d'ondes S détectables dans le cas de ces faux événements. Seulement, les vrais événements peuvent aussi partager toutes ces caractéristiques, en particulier pour les événements détectés aux confins du réseau de stations et pour les événements dont le contenu en bruit est élevé et qui rend difficile la détection des phases S.

Quelques attributs redondants ont aussi été retranchés de la sélection des attributs comme le nombre de stations utilisées, la variance des résidus temporels ou la proportion de stations du réseau qui sont les plus proches de l'événement et qui sont impliquées dans sa détection. Le nombre de stations utilisées, qui a un poids élevé dans les attributs, est effectivement très indirectement corrélé au nombre de phases utilisées, qui est l'attribut avec l'importance relative maximale. L'utilisation de ces deux attributs, nombre de stations et nombre de phases utilisées, augmente la probabilité de classer les vrais événements détectés avec très peu de stations et très peu de phases comme faux événements. De plus, la variance des résidus temporels est un indice de la dispersion des valeurs des résidus assez superflue puisque l'écart-type des résidus (racine carrée de la variance), inclus dans la sélection, apporte déjà cette information. De même, la proportion des stations les plus proches de l'événement est assez liée à la distance épicentrale minimale, également incluse dans la sélection des attributs.

D'autres attributs ont également été retirés alors qu'ils contribuent très activement à la discrimination des vrais et faux événements. Ce sont le degré de rectilinéarité du signal, le nombre de pics dans la fonction d'auto-corrélation ainsi que, en moindre mesure, la plus grande valeur propre initiale de la matrice de covariance calculée à partir du signal sur les trois composantes.

La rectilinéarité mesure la polarisation linéaire du champ d'onde : une valeur élevée de cette rectilinéarité représente un champ d'onde linéairement polarisé comme c'est le cas par exemple des ondes P longitudinales, des ondes S transversales et des ondes de Love (Greenhalgh et al., 2018). Quand un signal rectilinéairement polarisé est contaminé par du bruit, même si le bruit est par exemple polarisé de façon sphérique, sa direction de polarisation va changer nettement (Zheng et al., 1992). Par conséquent, la variabilité des phases sismiques qui peuvent être repérées dans le signal, associée à un fort contenu en bruit, amène à une trajectoire des particules 3D très complexe, s'éloignant d'une polarisation purement linéaire (Cliet et al., 1987). Si l'apport de la rectilinéarité dans la combinaison d'attributs a tendance à améliorer la prédiction des faux événements, la complexité du signal associé aux vrais événements rend plus difficile leur prédiction.

Le degré de rectilinéarité étant formulé en fonction de la l'ordre de grandeur de la valeur propre maximale de la matrice de covariance, le même constat peut être établi avec l'attribut qui exprime cette valeur propre maximale. Cet attribut donne des informations sur la cohérence spatiale du champ d'onde observé, et est classiquement utilisé pour détecter les signaux sismiques (Wagner et al., 1996 ; Seydoux et al., 2016). Seulement, lorsque ces signaux sismiques sont détectés avec de faibles rapports signal/bruit, leurs valeurs propres maximales diminuent, rendant plus difficile la prédiction correcte des vrais événements qui sont associés à ces types de signaux (Saenger et al., 2009).

Il en est de même pour l'estimation du degré de périodicité du signal, traduite par l'estimation du nombre de pics dans la fonction d'autocorrélation. En effet, le bruit d'origine anthropique étant généralement non-stationnaire et non-Gaussien (Groos et al., 2009 ; Steim, 2015) et les signaux sismiques détectés étant souvent de faible amplitude et contaminés par du bruit, les profils de ces fonctions d'autocorrélation peuvent être très similaires, accentuant la difficulté d'utiliser un tel attribut pour correctement prédire les vrais et faux événements.

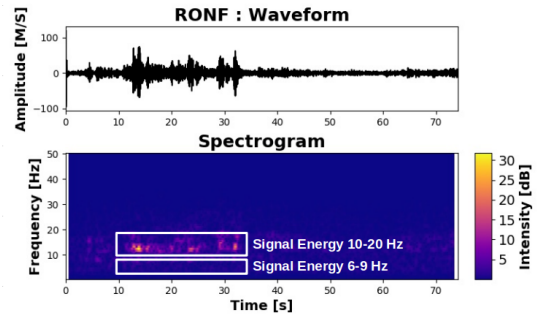
Si le retrait de ces trois précédents attributs aboutit à une amélioration de la capacité prédictive du classifieur vis-à-vis des vrais événements (respectivement 96 % pour le jeu test et 82-84% pour le jeu automatique), la qualité de prédiction des faux événements est quant à elle légèrement dégradée, comme en témoigne la diminution de la valeur de la précision (rapport entre le nombre de vrais événements correctement prédits et la somme du nombre de faux événements incorrectement prédits plus le nombre de vrais événements correctement prédits). Cependant, au regard du grand nombre de faux événements détectés (près de 45 000), cette dégradation de la prédiction des faux événements ne se manifeste presque pas sur la valeur du taux de faux événements correctement prédits. Par conséquent, le taux de faux événements incorrectement détectés reste donc acceptable.

Enfin, les attributs qui sont reliés à l'énergie du signal dans les différentes gammes fréquentielles testées (1-3 Hz, 3-6 Hz, 6-9 Hz, 1-5 Hz, 5-10 Hz, 10-20 Hz, 20-50 Hz) et à la description du spectre issu de la transformation discrète de Fourier (médiane, énergie dans les gammes fréquentielles 0-12.5 Hz, 12.5-25 Hz, 25-37.5 Hz et 37.5-50 Hz, nombre de pics) n'ont pas été gardés non plus. Ces attributs sont effectivement peu discriminants (importance relative inférieure à 1%) et complexifient la tâche de discrimination. Comme il a déjà été évoqué, les signaux associés aux différents événements (faux événements, tirs de carrière, séismes) présentent des contenus fréquentiels et des amplitudes assez équivalentes. De plus, au sein d'une même classe d'événements, les signaux peuvent présenter des amplitudes et des intensités très variables pour une même gamme fréquentielle, relatant la taille différentielle des sources de ces signaux.

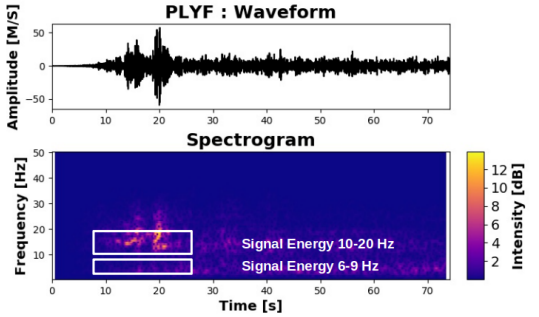
En revanche, si ce sont plutôt des rapports d'énergie du signal sur des gammes de fréquence différentes qui sont considérés, l'effet discriminant s'amplifie. En effet, si l'attribut définissant le rapport de l'énergie du signal entre les gammes de fréquence 6-9 Hz et 10-20 Hz est ajouté à la combinaison des attributs optimaux, celui-ci, ayant une importance relative non négligeable (environ 5 %), améliore de façon notoire la performance prédictive du classifieur résultant. Celui-ci est désormais capable de prédire correctement environ 87 % des vrais événements du jeu automatique et 96 % des vrais événements du jeu test, tout en maintenant un fort taux de prédiction des faux événements (99.7% de prédictions correctes).

Pris individuellement, le signal filtré aux gammes de fréquences comprises entre 10 et 20 Hz peut être d'une intensité comparable pour les séismes, les faux événements et parfois les tirs de carrière. De la même manière, en fonction de la nature du bruit, le signal filtré aux gammes de fréquences comprises entre 6 et 9 Hz peut être en moyenne de même intensité pour les différents types d'événement. En revanche, la prise en compte du rapport combiné de l'énergie du signal entre les gammes de fréquence 6-9 Hz et 10-20 Hz permet de distinguer plus facilement les événements entre eux. L'énergie du signal dans la gamme de fréquence 6-9 Hz est effectivement relativement plus élevée pour les vrais événements (tirs de carrière et séismes) que l'énergie du signal dans la gamme de fréquence 10-20 Hz, comparativement au rapport de l'énergie du signal pour ces gammes de fréquence estimé pour les faux événements (Figure 5.27).

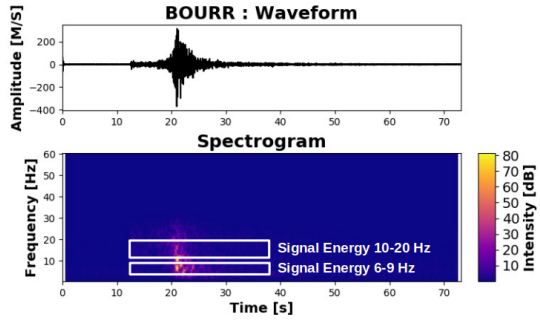
Le classifieur final qui prédit les vrais et faux événements est présenté dans l'article de la sous-section suivante.



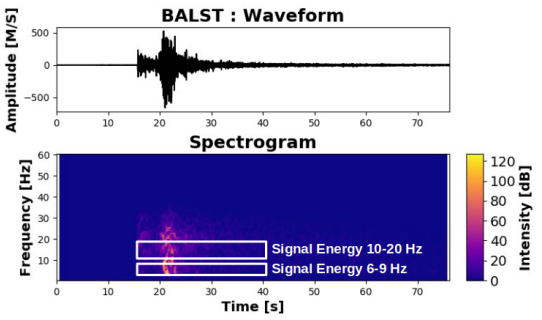
(a) Spectrogramme correspondant au signal détecté à la station RONF et qui est relié à un faux événement créé le 03 décembre 2016 à 06h23



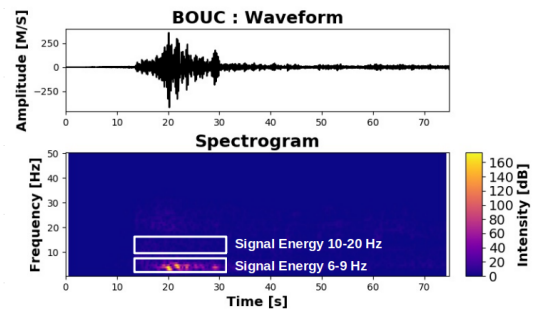
(b) Spectrogramme correspondant au signal détecté à la station PLYF et qui est relié à un faux événement créé le 19 décembre 2016 à 07h05



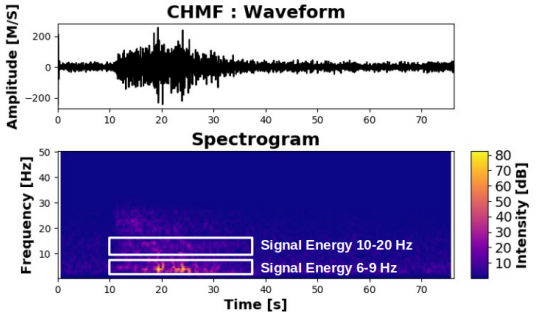
(c) Spectrogramme correspondant au signal détecté à la station BOURR et qui est relié à un séisme identifié le 18 octobre 2016 à 21h36 dans le canton de Zürich (MLv 1.4)



(d) Spectrogramme correspondant au signal détecté à la station BALST et qui est relié à un séisme identifié le 18 octobre 2016 à 21h36 dans le canton de Zürich (MLv 1.4)



(e) Spectrogramme correspondant au signal détecté à la station BOUC et qui est relié à un tir de carrière Bonnefoy identifié le 15 décembre 2016 à 12h29 dans la région de Besançon (MLv 1.6)



(f) Spectrogramme correspondant au signal détecté à la station CHMF et qui est relié à un tir de la carrière Bonnefoy identifié le 15 décembre 2016 à 12h29 dans le canton de Besançon (MLv 1.6)

FIGURE 5.27: Comparaison de l'intensité du signal pour les gammes fréquentielles 6-9 Hz et 10-20 Hz entre les deux grands types d'événements : (a), (b) faux événements et vrais événements dont (c) (d) les séismes et (e) (f) les tirs de carrière.

•Rechercher les attributs pour classer les séismes et les tirs

En ce qui concerne la classification des séismes et des tirs de carrière, l'utilisation du sous-ensemble optimal des attributs sélectionnés par élimination récursive conduit à une fonction de prédiction capable de prédire correctement 82% des tirs de carrière et 97% des séismes contenus dans le jeu test (janvier 2016-août 2016). Lorsque cette fonction de prédiction est utilisée sur le jeu automatique (septembre 2016-décembre 2016), le classifieur est capable de prédire correctement 88% des séismes et 89% des tirs de carrière (Figure 5.28).

L'exclusion des attributs véhiculant des informations sur la distance de l'événement à la carrière la plus proche, le nombre de phases utilisées, le nombre de phases S, les magnitudes locales (MLv et ML) amènent à mieux prédire les séismes du jeu automatique avec une amélioration de 2% de vraies prédictions. En revanche, cela a peu d'effet sur les séismes contenus dans le jeu test et cela dégrade assez fortement la qualité de la prédiction des tirs de carrière dans les deux jeux de données.

En effet, un tir de carrière est situé proche d'une carrière, contient peu de phases S dans l'association qui l'a détecté et donc possède un nombre de phases moins grand, mais la réciproque n'est pas forcément vraie pour les séismes. De plus, les magnitudes estimées pour les tirs de carrière sont en moyenne plus élevées comme décrit dans le chapitre 2 (1.58 contre 1.40 pour les séismes). Ces attributs apparaissent donc significatifs pour classer les tirs de carrière alors qu'ils le sont beaucoup moins pour les séismes (Figure 5.28).

Toutefois, cet effet négatif tend à s'annuler lorsque ce sont les attributs statistiques qui décrivent l'enveloppe du signal qui sont retirés de la combinaison d'attributs (la tendance s'inverse, la prédiction correcte des tirs de carrière s'améliore). Si ces derniers ressortent fortement de la sélection automatique des attributs, ils apportent de la confusion à la fonction de prédiction générée. Comme il a été illustré précédemment, il est souvent difficile de distinguer un séisme d'un tir de carrière basé uniquement sur sa forme d'onde, du fait notamment des effets liés au milieu de propagation. Or, ici, lorsque la combinaison d'attributs est dénuée des informations liées à l'enveloppe (donc indirectement la forme d'onde) et à la distance de l'événement à la plus proche carrière, la qualité de la prédiction des séismes contenus dans le jeu test reste élevée (97% de séismes correctement prédits). Celle-ci s'améliore également nettement pour les séismes contenus dans le jeu automatique (passage de 90 à 93% de séismes correctement prédits, Figure 5.28). Ces attributs, pourtant utilisés pour discriminer les événements par les analystes, ne sont pas ceux qui vont fondamentalement aider à classer les séismes et les tirs de carrière entre eux.

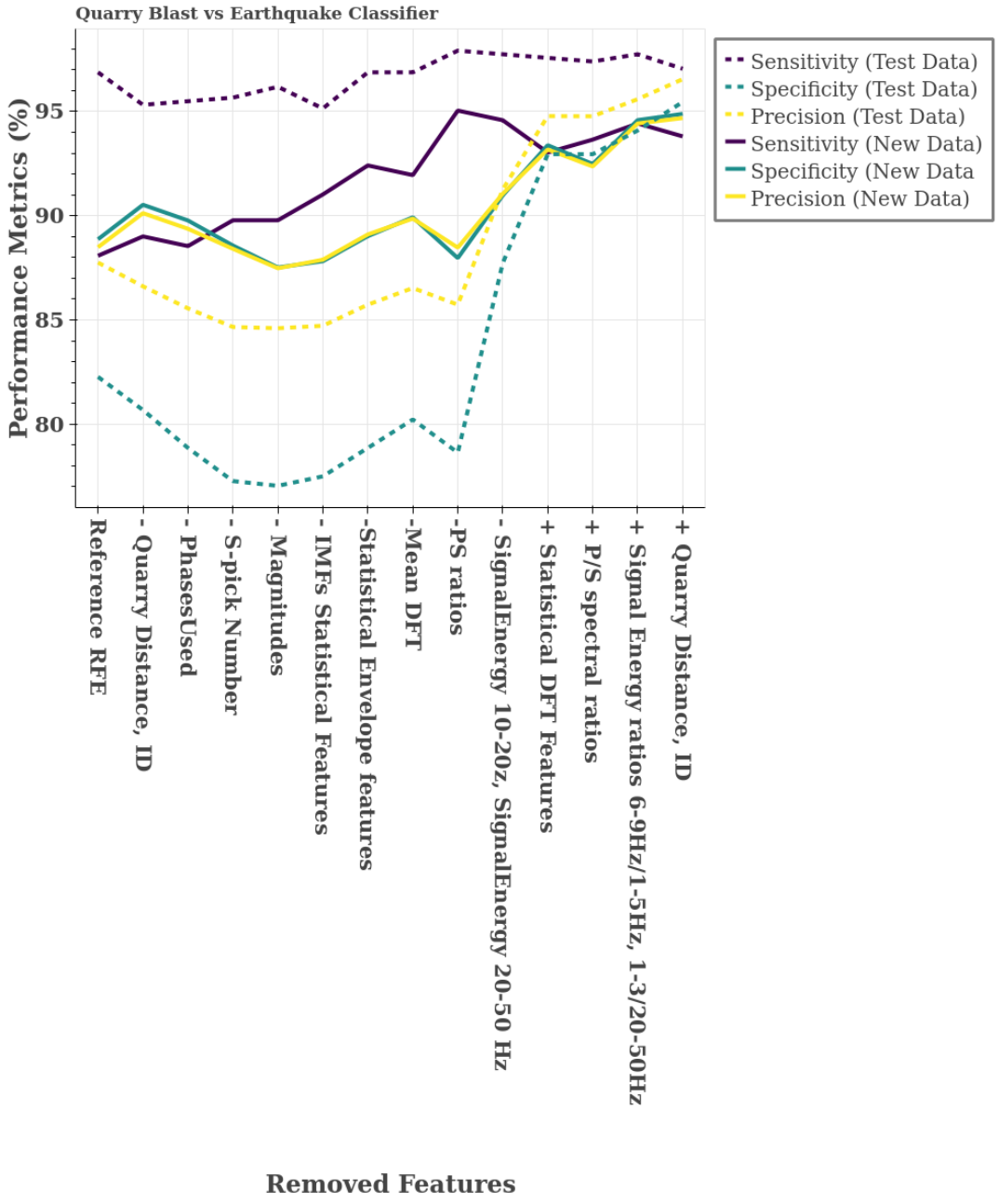


FIGURE 5.28: Effets du retrait (symbole -) et/ou de l'ajout (symbole +) d'attributs à partir d'une sélection initiale automatique d'attributs, effectuée par élimination récursive, sur la capacité prédictive de classifieurs discriminant les séismes et les tirs de carrières. La spécificité désigne le taux de tirs de carrière correctement prédits (c'est-à-dire le rapport entres les vrais négatifs et la somme des vrais négatifs et des faux positifs). La sensitivité désigne le taux de séismes correctement prédits (c'est-à-dire le rapport entres les vrais positifs et la somme des vrais positifs et des faux négatifs). La précision désigne la proportion de séismes correctement prédits relativement à l'ensemble des événements prédits positivement (c'est-à-dire le rapport entres les vrais positifs et la somme des vrais positifs et des faux positifs). Référence RFE = sélection automatique des attributs effectuée par élimination récursive des attributs. Voir e tableau S1 du supplément de l'article pour le détail des attributs.

Le même constat peut être fait pour les attributs statistiques décrivant les valeurs des rapports des amplitudes maximales entre les ondes P et S pour chaque événement (Figure 5.28). Si ces attributs semblent être importants pour la prédiction des tirs de carrière, ils ont un effet plutôt négatifs sur la prédiction des séismes (leur retrait de la combinaison des attributs amène à une augmentation du taux de prédiction correcte des séismes pour les deux jeux test et automatique).

De même, l'attribut exprimant la moyenne des valeurs du spectre du signal ainsi que les attributs représentant l'énergie du signal dans les gammes fréquentielles 10-20 Hz et 20-50 Hz ont été retirés de la sélection automatique. Si les tirs de carrière sont associés à des signaux qui ont un contenu fréquentiel globalement plus basse-fréquence que ceux associés aux séismes, les motifs fréquentiels peuvent être en fait très variables d'un événement à l'autre, et d'une station à l'autre. Ce sont donc les rapports d'énergie du signal relatifs qui sont donc considérés : le rapport de l'énergie du signal entre les gammes fréquentielles 6-9 Hz et 1-5 Hz ainsi que le rapport de l'énergie du signal entre les gammes fréquentielles 3-6 Hz et 20-50 Hz. En effet, les signaux reliés aux tirs de carrière présentent une énergie plus élevée dans la gamme de fréquences 1-5 Hz ou 3-6 Hz relativement à la gamme de fréquences 6-9 Hz ou 20-50 Hz, et inversement pour les séismes. Ceci souligne notamment la superficialité générale des tirs de carrière qui génèrent beaucoup d'ondes de surface de faible fréquence (Gitterman et al., 1998).

Enfin, l'apport supplémentaire d'attributs reliés à la description du signal dans le domaine fréquentiel apporte plus de contraintes à la discrimination des séismes et des tirs de carrière. Il semble que les attributs reliés à des informations contenues dans le spectrogramme du signal dans le domaine tempo-fréquentiel (variance des valeurs du spectre du signal, nombre de pics contenus dans le spectre, rapports spectraux entre les ondes P et S, fréquence cumulée de 25%, fréquence cumulée de 75%) expriment davantage ce qu'est un séisme relativement à un tir de carrière, et vice versa. En présence de ces attributs, le nouveau classifieur converge vers une capacité équivalente à prédire correctement les séismes et les tirs de carrière, quel que soit le jeu de données utilisé (Figure 5.28).

A l'inverse, si ce sont les attributs qui définissent les rapports d'amplitudes maximales entre les ondes S et P ainsi que les caractéristiques de l'enveloppe qui sont choisis, cela engendre un déséquilibre dans la capacité prédictive du classifieur résultant : les tirs de carrière sont en conséquence plus correctement prédits que les séismes, et inversement si ces attributs sont retirés sans ajout d'informations sur les spectrogrammes des signaux. Cette instabilité de prédiction en présence de ces attributs soulignent bien l'incapacité du classifieur à généraliser.

Si en revanche, on décide d'ajouter de nouveau les attributs qui relatent les informations liées à la distance de l'événement à la carrière la plus proche, la performance prédictive vis-à-vis des séismes se dégrade derechef. Par conséquent, si ces informations sont très utiles pour affiner la prédiction des tirs de carrière, elles apportent de la confusion et de l'instabilité à la fonction de prédiction, qui aura plus de difficulté à prédire certains séismes (surtout s'ils sont véritablement situés près d'une carrière). Ces informations sont alors définitivement retirées de la sélection des attributs. La fonction de prédiction finale est présentée dans l'article qui va suivre.

5.2.2 Comprendre les erreurs de classification

L'interactivité Homme-machine peut se traduire aussi à travers l'analyse des erreurs de classification réalisées par le classifieur automatique. En effet, l'analyse de ces erreurs amène à estimer la réelle performance prédictive des classifieurs en évaluant la pertinence des règles de classification émises, de sorte à déceler les défaillances de la fonction de prédiction apprise.

- **Erreurs de classification pour les vrais et faux événements**

Erreurs de classification des faux événements. Parmi les faux événements incorrectement prédits par le classifieur, 24 % d'entre eux sont issus d'une association de faux pointés avec 1 ou 2 vrais pointés émis au moins à une station qui a enregistré un signal sismique isolé. De ce fait, les attributs qui sont calculés pour ces faux événements vont véhiculer une information supplémentaire à relier avec cet apport de signal cohérent non-stationnaire.

Cela peut avoir pour effet d'augmenter le nombre de phases utilisées pour l'association, qui est l'attribut qui a une importance relative la plus élevée. Cela peut aussi modifier la valeur des attributs tels que l'entropie de Shannon, qui traduit le caractère aléatoire du signal, ou la différence absolue moyenne d'ordre 1 de l'enveloppe du signal, qui exprime le degré de non-stationnarité du signal.

Par exemple, le faux événement détecté le 11 décembre 2016 à 15h10 présente un signal sismique isolée enregistré à la station A119A où deux pointés (P et S) sont émis (Figure 5.29). Cet événement présente alors 8 phases. De plus, le signal détecté à la station A119A apportant de la cohérence, la corrélation entre les premières arrivées détectées aux stations (les pointés P) et la distance épicentrale devient forte (0.99). Ce faux événement n'est pas correctement prédit par le classifieur, en particulier parce que l'apport du signal sismique "parasite" éloigne les valeurs de certains attributs des caractéristiques généralement rencontrées chez les faux événements comme le faible nombre de phases ou le fort caractère aléatoire.

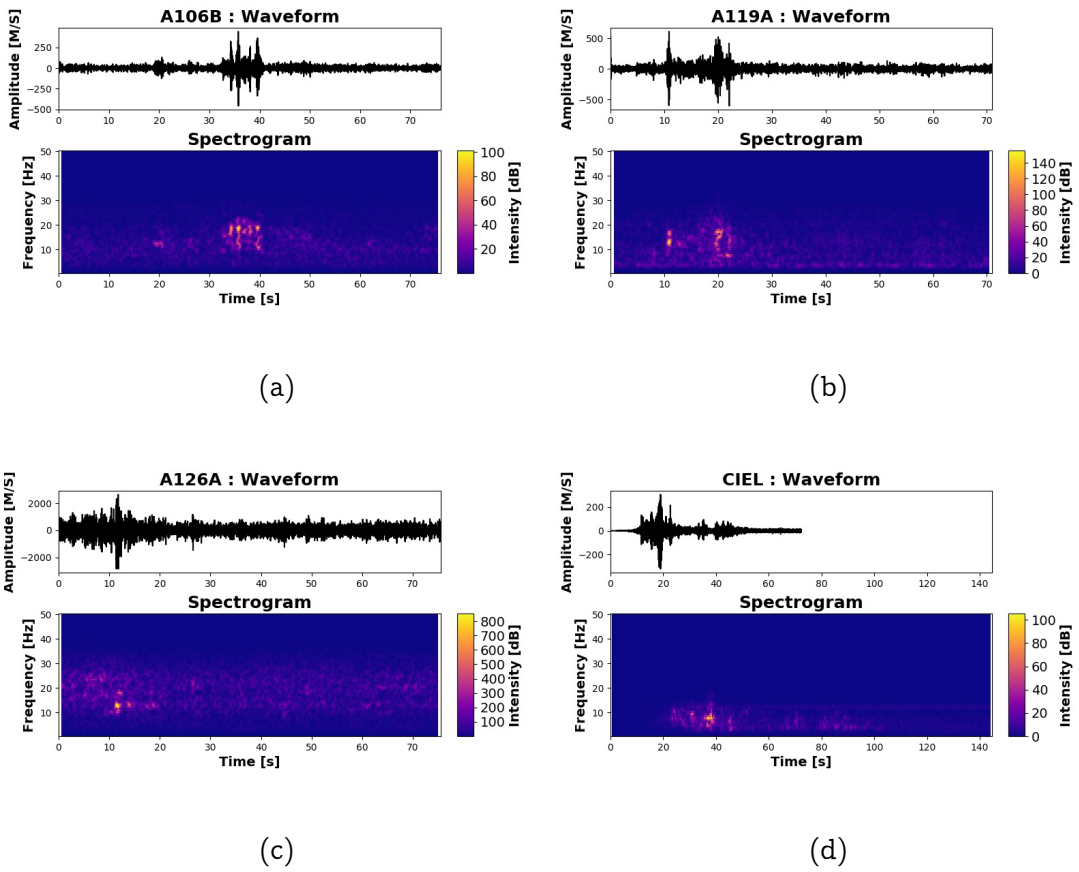


FIGURE 5.29: Formes d'onde et spectrogrammes des signaux associés à un faux événement détecté le 11 décembre 2016 à 15h10 et incluant un signal sismique isolé à la station A119A (composante verticale).

De plus, ce faux événement est prédit comme vrai événement par le classifieur avec une probabilité de 0.55 (soit 275 arbres sur 500, le vote majoritaire étant à partir de 250 arbres).

Parmi les arbres décisionnels de la forêt aléatoire qui vont correctement prédire cet événement, l'attribut déterminant qui va contribuer de façon notoire à orienter la prédiction finale vers celle de faux événement est le degré de polarisation planaire du signal qui est élevé pour ce dernier (0.79). Le degré de polarisation planaire est fortement corrélé à la profondeur de la source : la valeur est élevée pour les signaux associés au bruit d'origine anthropique puisqu'ils se propagent principalement sous forme d'ondes de surface de Rayleigh (Havskov et Alguacil, 2004).

Erreurs de classification des vrais événements. A l'inverse, les vrais événements qui sont incorrectement prédits sont caractérisés par un nombre de phases plus petites, et sont souvent détectés avec des distances épicentrales minimales supérieures. Ces derniers peuvent être fortement contaminés par du bruit stationnaire (rapports signal/bruit faibles), diminuant le rapport de l'énergie du signal entre les gammes fréquentielles 6-9 Hz et 10-20 Hz.

En fonction du degré de certitude de la prédiction, les vrais événements incorrectement prédits avec une forte probabilité peuvent détenir encore de faux pointés non éliminés par les développements exposés précédemment. Dans ces cas extrêmes, l'inclusion de faux pointés induit une augmentation de la valeur des résidus temporels et une diminution du facteur de corrélation entre les premières arrivées des ondes P et la distance épicentrale. Cet effet souligne l'importance de développer une procédure de détection adaptée des petits séismes, sans quoi, le taux de perte des événements serait conséquent.

Parmi les arbres décisionnels qui contribuent à prédire correctement ces vrais événements, qui sont finalement systématiquement mal classés par vote majoritaire, la valeur de l'entropie de Shannon calculée à partir du signal dans le domaine tempo-fréquentiel (voir détails des attributs dans le tableau S1 du supplément de l'article qui va suivre) est un critère décisif pour orienter le choix final de prédiction vers la prédiction correcte, à savoir l'étiquette "vrai événement".

•Erreurs de classification pour les séismes et les tirs de carrière

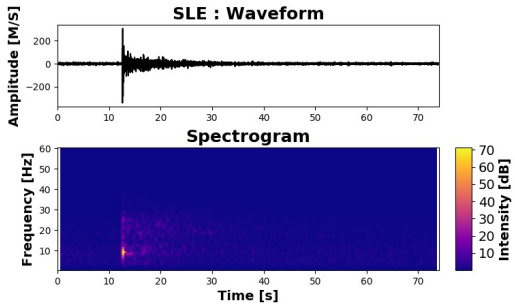
Erreurs de classification des séismes. Si je prends l'exemple de séismes appartenant à un essaim régulièrement observé au Nord du Lac Konstanz en Allemagne, plusieurs d'entre eux sont systématiquement mal classés. Ces séismes sont classés par le classifieur comme étant des tirs de carrière avec des probabilités de prédiction comprises entre 0.53 et 0.69. Ceci signifie que, pour ces événements, 265 à 345 arbres décisionnels, sur un total de 500 arbres inclus dans la forêt, aboutissent à la prédiction finale de tir de carrière. Le vote majoritaire est estimé à au moins 250 arbres.

Ces séismes mal classés sont caractérisés par des signaux dont l'intensité se concentre à des gammes fréquentielles particulièrement basses ($<$ 10 Hz, Figure 5.30). De plus, la variance du spectre est inférieure à la moyenne des variances estimées pour l'ensemble des séismes du jeu de données (période 2016-2019). Le nombre de pics estimé dans le spectre du signal est aussi particulièrement bas.

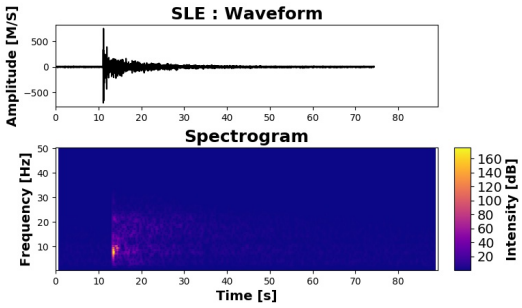
Aussi, pour les séismes prédits comme étant des tirs de carrière avec les plus fortes probabilités, la moyenne des magnitudes de surface (à la période 10 s) est plus élevée (de l'ordre de 1.30) que la moyenne des magnitudes de surface estimées pour l'ensemble des séismes du jeu de données (qui est équivalent à 0.78). Ces séismes partagent donc les mêmes caractéristiques que les tirs de carrière, traduisant probablement la superficialité de leurs sources.

L'analyse de quelques arbres décisionnels montre que le séisme détecté le 10 novembre à 09h28 (MLv 1.34), et appartenant à l'essaim de séismes identifié au Nord du Lac Konstanz en Allemagne, est correctement prédit par le classifieur si les attributs traduisant la forme de la distribution des valeurs du signal associé à l'événement sont impliqués dans le chemin décisionnel (Figure 5.31). Ces attributs sont le coefficient d'asymétrie de la distribution (en anglais *skewness*) et le coefficient d'aplatissement de cette distribution (en anglais *kurtosis*).

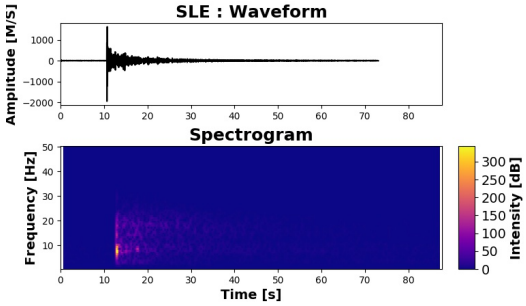
En revanche, lorsque ce même événement est prédit comme tir de carrière par un autre arbre décisionnel, ce sont les attributs reliés aux rapports de l'énergie du signal entre les gammes fréquentielles 3-6 Hz et 20-50 Hz puis 1-5 Hz et 6-9 Hz qui vont guider la prédiction finale du chemin décisionnel, combinés avec la valeur maximale du rapport spectral entres les ondes P et S (Figure 5.32). En effet, les signaux associés à cet événement présentent une intensité maximale relativement plus élevée aux faibles gammes fréquentielles, comme c'est le cas de nombreux tirs de carrière.



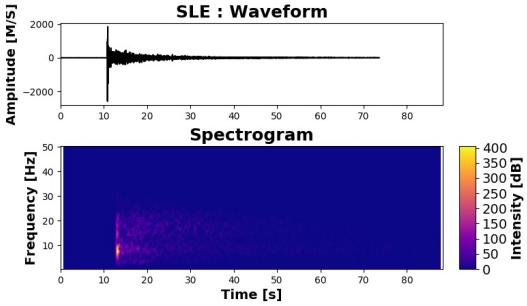
(a) Séisme détecté le 03 novembre 2016 à 6h34 (MLv 1.22)



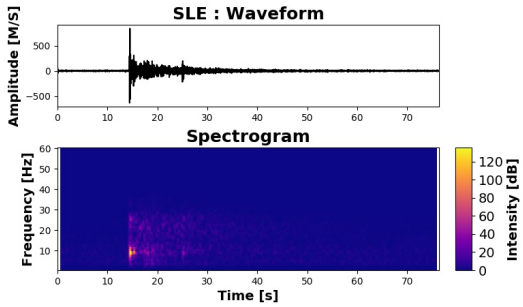
(b) Séisme détecté le 10 novembre à 09h28 (MLv 1.34)



(c) Séisme détecté le 21 novembre à 16h14 (MLv 1.71)



(d) Séisme détecté le 27 novembre à 11h00 (MLv 1.67)



(e) Séisme détecté le 01 décembre à 13h58 (MLv 1.25)

FIGURE 5.30: Signaux associés à 6 séismes incorrectement classifiés par le classifieur automatique des séismes et des tirs de carrière. Les signaux sont enregistrés à la première station SLE (distance épicentrale moyenne = 22 km) et présentent une intensité maximale à des fréquences relativement basses pour des séismes (< 10 Hz). Les séismes ont été détectés au Nord du Lac Konstanz en Allemagne.

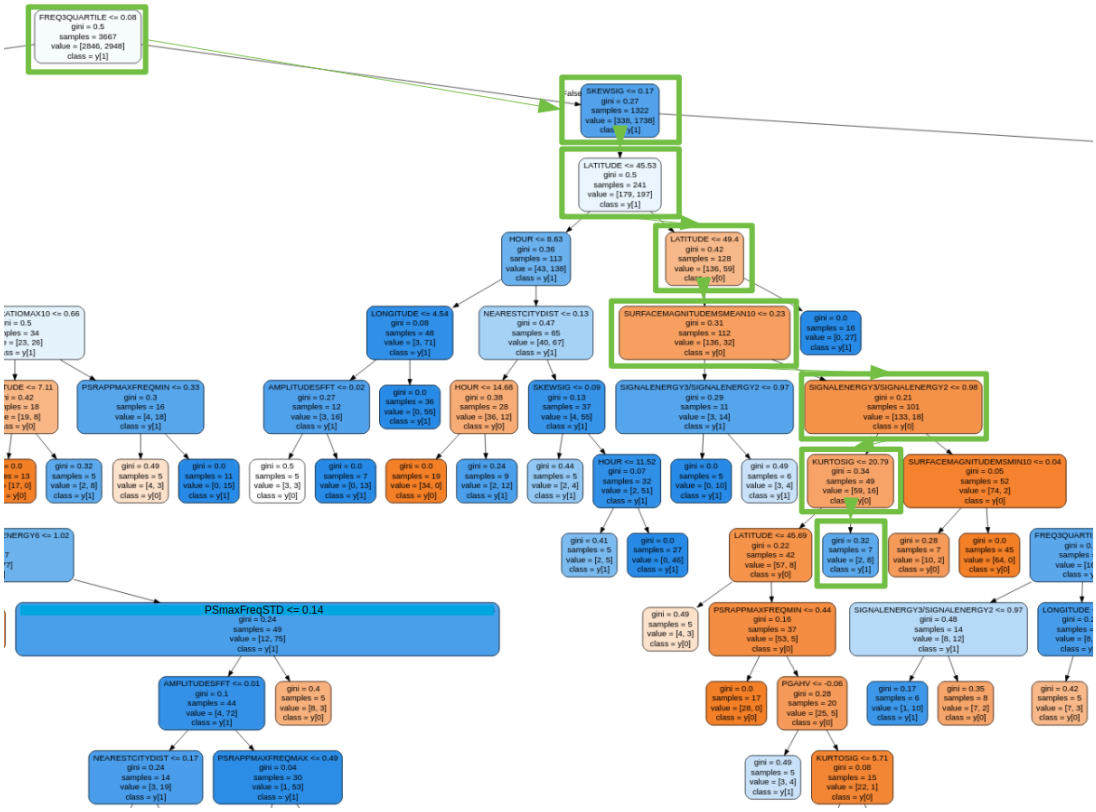


FIGURE 5.31: Extrait d'un arbre décisionnel tiré aléatoirement de la forêt aléatoire, aboutissant à la prédiction correcte du séisme détecté le 10 novembre à 09h28 (MLv 1.34) au Nord du Lac Konstanz en Allemagne. Le chemin décisionnel est représenté en vert. La classe y[1] représente la classe des séismes et y[0] la classe des tirs de carrière.

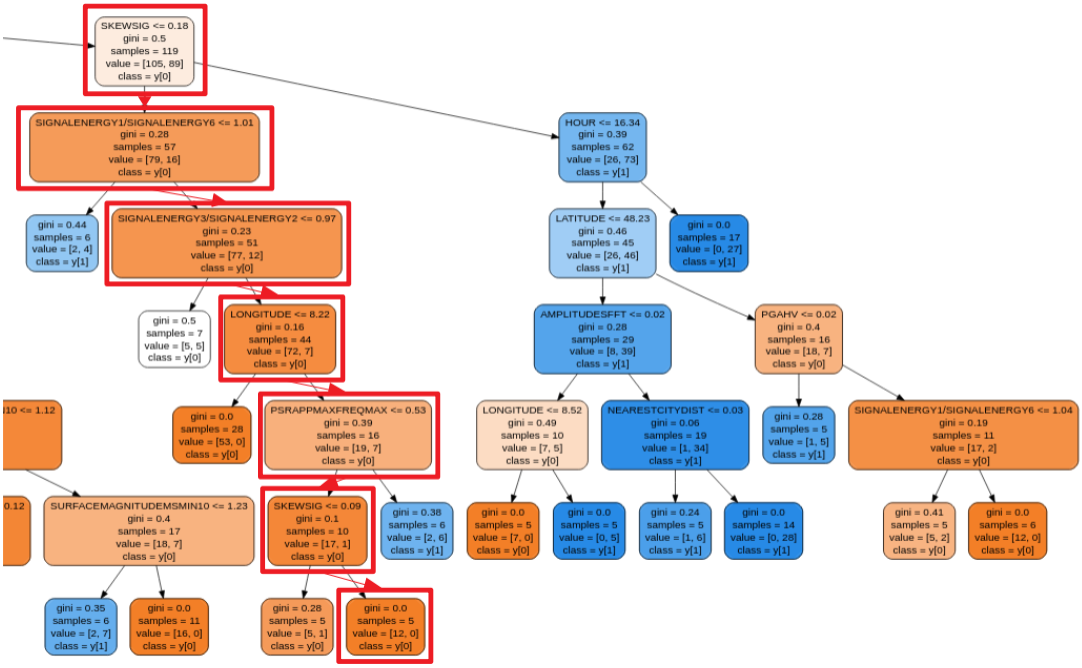
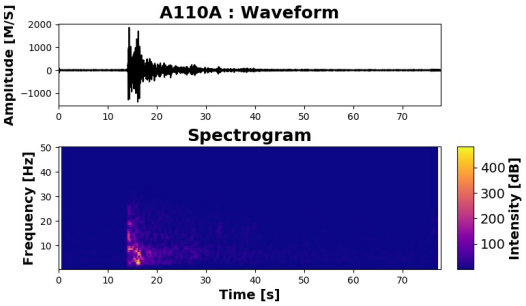
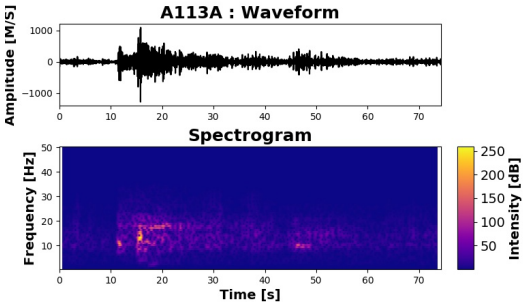


FIGURE 5.32: Extrait d'un arbre décisionnel tiré aléatoirement de la forêt aléatoire, aboutissant à la prédiction incorrecte du séisme détecté le 10 novembre à 09h28 (MLv 1.34) au Nord du Lac Konstanz en Allemagne. Le chemin décisionnel est représenté en rouge. La classe y[1] représente la classe des séismes et y[0] la classe des tirs de carrière.

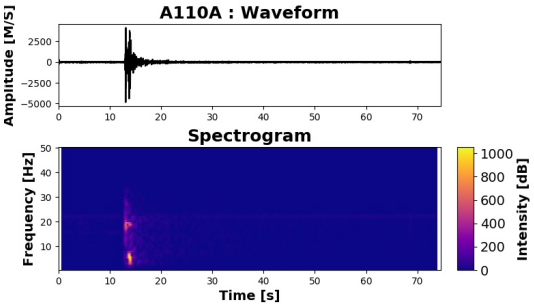
Erreurs de classification des tirs de carrière. De la même façon, les tirs de carrière incorrectement classés sont reliés à des signaux dont l'intensité reste élevée à des gammes de fréquences supérieures à 10 Hz et dont le spectre possède une valeur de variance plus élevée que la moyenne des variances estimées pour l'ensemble des tirs de carrière du jeu de données (Figure 5.33).



(a) Tir de la carrière de Groß-Bieberau identifié le 09 novembre 2016 à 13h44 (MLv 2.11, première station A110A, distance épicentrale 33 km)



(b) Tir de la carrière de Mühltal détecté en Allemagne le 20 septembre 2016 à 11h25 (MLv 2.08, première station A110A, distance épicentrale 24 km)



(c) Tir de la carrière de Mühltal détecté en Allemagne le 20 septembre 2016 à 11h25 (MLv 2.08, seconde station A113A, distance épicentrale 33 km)

FIGURE 5.33: Signaux associés à 2 tirs de carrière incorrectement classifiés par le classifieur automatique des séismes et des tirs de carrière. Les signaux présentent une intensité maximale à des fréquences relativement élevées pour des tirs de carrière ($>$ 10 Hz).

L'analyse de quelques arbres décisionnels montre par exemple que le tir de la carrière de Groß-Bieberau identifié le 09 novembre 2016 à 13h44 (MLv 2.11) est incorrectement prédit par le classifieur lorsque les attributs décrivant le rapport de l'énergie du signal entre 1-5 Hz et 6-9 Hz, la variance du spectre et le rapport minimum spectral entre les ondes P et S sont impliqués dans l'élaboration du chemin décisionnel (Figure 5.34).

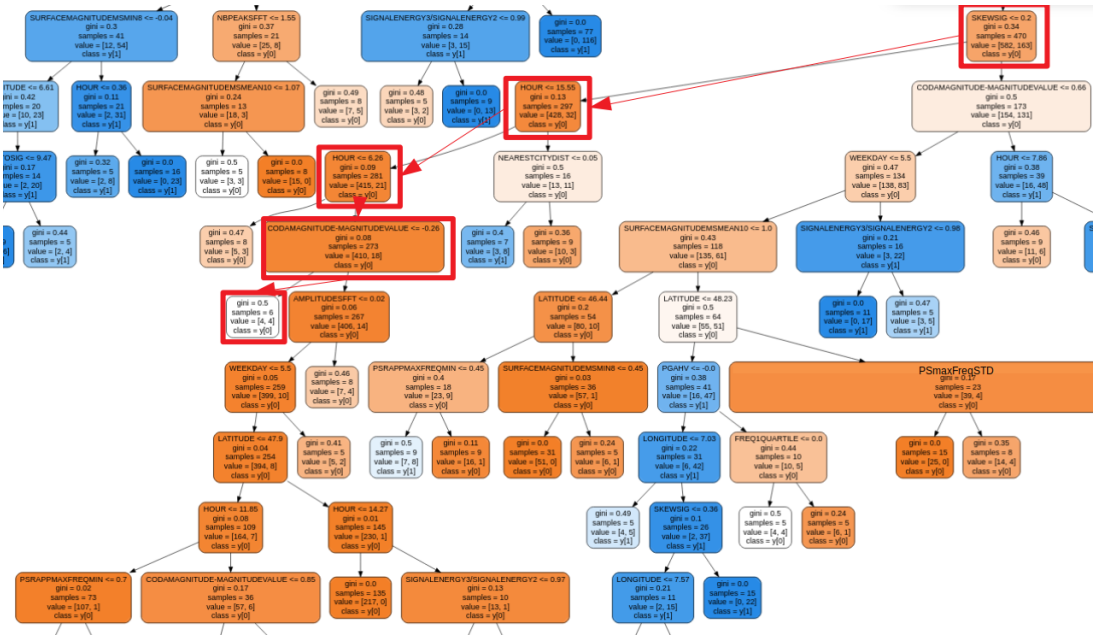


FIGURE 5.34: Extrait d'un arbre décisionnel tiré aléatoirement de la forêt aléatoire, aboutissant à la prédiction correcte du tir de la carrière de Groß-Bieberau détecté le 09 novembre 2016 à 12h44 (MLv 2.11) en Allemagne. Le chemin décisionnel est représenté en rouge. La classe y[1] représente la classe des séismes et y[0] la classe des tirs de carrière.

Comme il a été décrit dans le chapitre 2, les carrières exploitent une diversité de matériaux, qui va des roches sédimentaires aux roches métamorphiques, en passant par les roches magmatiques. La carrière de Groß-Bieberau exploite du gabbro qui est une roche compétente qui véhicule des signaux de haute fréquence faiblement atténués. Les caractéristiques du signal enregistré apportant des confusions, ces tirs de carrière peuvent se rapprocher des caractéristiques des séismes.

Le tir de carrière décrit précédemment est en revanche bien classé si le chemin décisionnel qui conduit à la prédiction finale présente à ses embranchements les attributs tels que le coefficient d'asymétrie moyen de la distribution des amplitudes des signaux associés à cet événement, l'heure de l'événement ainsi que la différence moyenne entre la magnitude locale et la magnitude de coda, qui une fonction sensible de la profondeur de la source (KOPER et al., 2016 ; HOLT et al., 2019). D'autres informations que l'analyse pure du spectrogramme apparaissent alors nécessaires pour assurer une couverture de prédiction plus large (Figure 5.35.

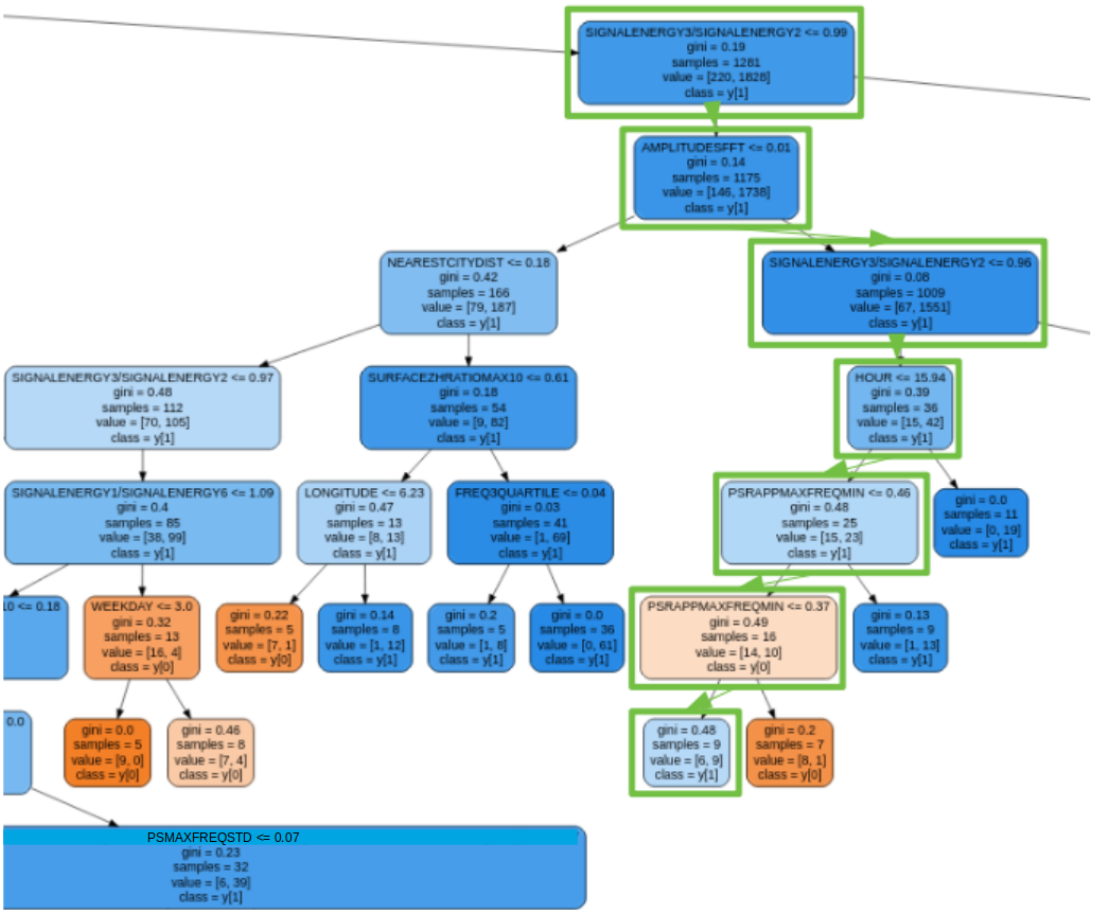


FIGURE 5.35: Extrait d'un arbre décisionnel tiré aléatoirement de la forêt aléatoire, aboutissant à la prédiction incorrecte du tir de la carrière de Groß-Bieberau détecté le 09 novembre à 13h44 (MLv 2.11) en Allemagne. Le chemin décisionnel est représenté en vert. La classe y[1] représente la classe des séismes et y[0] la classe des tirs de carrière.

L'injection des connaissances préalables dans le système d'apprentissage amène à détecter les corrélations parasitaires sur lesquelles l'algorithme peut fonder son apprentissage, limitant la minimisation de l'erreur de généralisation. A travers l'analyse des attributs, des arbres décisionnels ainsi que les résultats de prédiction des différents classifieurs, la performance prédictive de ces derniers a pu être estimée. Reste à connaître le potentiel des fonctions de prédiction optimales qui en découlent.

5.3 Utiliser la fonction de prédiction optimale et évaluer sa performance finale

L'article ci-dessous replace le problème de classification dans le contexte de ce travail de recherche. Il brosse de manière synthétique la méthodologie adoptée et redéfinit la procédure d'apprentissage dans le cadre de l'interactivité. Il présente ensuite les résultats obtenus avec les deux classifieurs optimaux sélectionnés puis les discute. Le premier classifieur identifie les faux événements et les vrais événements. Le second classifieur discrimine les vrais événements préalablement identifiés en les étiquetant comme séismes ou tirs de carrière. La performance des classifieurs est évaluée à travers deux modes : un mode dit test et un mode dit opérationnel. Plusieurs métriques sont utilisées pour évaluer cette performance (sensitivité, spécificité, précision mais également probabilité de prédiction). Cet article donne enfin des éléments qui révèlent le degré de validité et de plausibilité des classifieurs sélectionnés à travers l'analyse des attributs, des arbres décisionnels et des informations tirées de la littérature sur ce sujet de classification.

5.3.1 Article : Monitoring Regional Seismicity Using Hybrid Intelligence

Monitoring Regional Seismicity using Hybrid Intelligence

Alexandra Renouard[1], Alessia Maggi[1], Marc Grunberg[2], Cécile Doubre[1], Clément Hibert[1]

[1]Université de Strasbourg, CNRS, IPGS/EOST, UMR7516, 5 rue René Descartes, 67100 Strasbourg, France
[2]Université de Strasbourg, CNRS, EOST, UMS830, F-67000 Strasbourg, France

Corresponding author: Alexandra Renouard (alexandra.renouard@unistra.fr)

Mailing Address: Université de Strasbourg, CNRS, IPGS/EOST, 5 rue René Descartes, 67100 Strasbourg, France.

Key Words: Machine-learning Discrimination, False Alarms, Quarry Blasts, Earthquake detection

Abstract

Small-magnitude earthquakes shed light on the distribution and occurrence of earthquakes, especially in stable continental regions where natural seismicity remains difficult to explain under slow strain rate conditions. However, capturing them in catalogs is strongly hindered by signal-to-noise issues, resulting in high rates of false as well as man-made events also being detected. Accurate and robust classification of all these events is then critical for optimally detecting small earthquakes. This requires uncovering recurrent salient features that can firstly rapidly identify false events from real events, then accurately recognize earthquakes form man-made events (mainly quarry blasts) despite high signal variability and noise content. In this study, we combine the complementary strengths of human and interpretable rule-based machine-learning algorithms for solving this classification problem. We use human expert knowledge to co-create two reliable machine-learning classifiers through human-assisted selection of classification features and review of events for which the classifier predictions are uncertain. The two classifiers are integrated into the SeiscomP3 operational monitoring system. The first one discards false events from the set of events obtained with a low STA-LTA threshold; the second one labels the remaining events as either earthquakes or quarry blasts. When run in an operational setting, the first classifier correctly detected more than 99% of false events and just over 93% of earthquakes; the second classifier correctly labeled 95% of quarry blasts and 96% of earthquakes. After a manual review of only the second classifier low-confidence outputs, the final catalog contained fewer than 2% of misclassified events. These results confirm that machine-learning strengthens the quality of earthquake catalogs and that the performance of machine-learning classifiers can be improved through human expertise. Our study promotes a broader implication of hybrid intelligence monitoring within seismological observatories.

1 INTRODUCTION

Even if small earthquakes rarely make the news, the benefits of their study are real (Brodsky, 2019; Ross, Trugman, et al., 2019). Due to their high frequency of occurrence, small earthquakes can bring statistical robustness to the observed seismic processes such as recurrent earthquakes triggered by local perturbations in the regional stress field. This is particularly important in continental plate interiors hosting low-to-moderate seismicity, such as the northeastern European Upper Rhine Graben area, where no mechanism is universally accepted to explain earthquake occurrence under very slow strain rate conditions (Gallen & Thigpen, 2018; Bezada & Smale, 2019; Leclère & Calais, 2019).

Recent worldwide deployment of seismic networks provides high-quality volumes of recorded seismograms, hiding a gold mine of information on small earthquakes (Levandowki et al., 2018). However, capturing them in catalogs is strongly hindered by signal-to-noise issues. The automated detection approaches used by most seismological observatories worldwide are based on arrival time differences (Lindenbaum et al., 2017). They use standard amplitude threshold algorithms, such the ratio of the short-term to the long-term average signal energy (STA/LTA), to automatically pick seismic wave arrival times, then associate them in coherent groups to infer earthquake locations. If observatories lower the detection threshold to recover lower-amplitude earthquake signals, they will also detect many more data-glitches, transient noise or man-made signals related to human activities (Arrowsmith et al., 2014; Díaz et al., 2017; Ross, Meier, & Hauksson, 2019). Consequently, high rates of false events and man-made events contaminate the automated earthquake catalogs, and fewer than 10% of automated detections remain in the final analyst-reviewed catalogs (Draelos et al., 2018). Decreasing the minimum detection magnitude also increases operational cost, since analysts have to screen a huge number of false events and risk missing real events in the process (Draelos et al., 2012). Some authors have suggested labelling arrival times before associating them (Yeck et al., 2019), however small events have lower signal-to-noise ratio arrivals that can easily be mistaken for false arrivals (McBrearty et al., 2019) and be therefore excluded from the catalogs.

Yet another challenge in building high-quality earthquake catalogs is discriminating between natural earthquakes and anthropogenic events such as explosions or quarry blasts. Although experienced analysts are able to distinguish these events by taking into account both waveform characteristics and source parameters of detected events (origin-time, location, polarities etc.), as catalogs grow in size thanks to improved seismic networks, the discrimination step becomes more complex and less repeatable (Onagawa et al., 2019). Machine-learning tools have been proposed to help analysts classify seismological signals since the 1990s (Dowla et al., 1990; Wang & Teng, 1995; Tiira, 1999; Maggi et al., 2017; Perol et al., 2018; Linville et al., 2019; Rouet-Leduc et al., 2019; Zhu & Beroza, 2019). They have low operational cost, and can analyse large volumes of real-time data (Meier et al., 2019), but are not yet implemented routinely outside of volcanic observatories.

In this paper, we describe how we trained machine-learning algorithms to classify events resulting from operational seismic monitoring of the Upper Rhine Graben area, using a hybrid approach that combines the advantages of machine-learning algorithms and human expertise while overcoming their respective limitations (Patel et al., 2019; Gennatas et al., 2020). We chose to classify events after the association step, as proposed by Draelos et al. (2012) and Z. Li et al. (2018).

We developed our classifiers within the SeisComP3 framework, one of the main earthquake monitoring systems used for detecting local, regional, and global seismicity in many countries across the world (Olivieri & Clinton, 2012), thereby also addressing the gap that has been observed between research developments in earthquake classification and their implementation on an operational level (Sparks et al., 2012).

2 DATA AND METHODS

Figure 1a describes the data flow we implemented. We first optimised the SeiscomP3 automated event detector already operational at the French National Seismological Service (BCSF-RéNaSS) to increase the number of small events detected, which also greatly increased the number of false detections. We then extracted waveform and event-based features for each detected event, and fed them through two successive rule-based classifiers: one to discriminate between false detections and real seismic events (classifier 1), and one to discriminate between seismic events of natural or anthropogenic origin (classifier 2). The machine-learning algorithm we selected to build these classifiers was supervised Random Forest (Breiman, 2001), which produces robust classifications through human-interpretable learning mechanisms (Drouin et al., 2019; Lundberg et al., 2020). It grows multiple independent base learners (decisions trees) to build a classification model, outputs prediction results from all of them, and combines these results to form a final prediction with a probability estimate. In order to co-construct more trusted classifiers, we included human feedback when refining the classifier, as suggested by Schaumberg et al. (2020). We obtained the final classifiers by proceeding in stages, as illustrated in Figure 1b and described below.

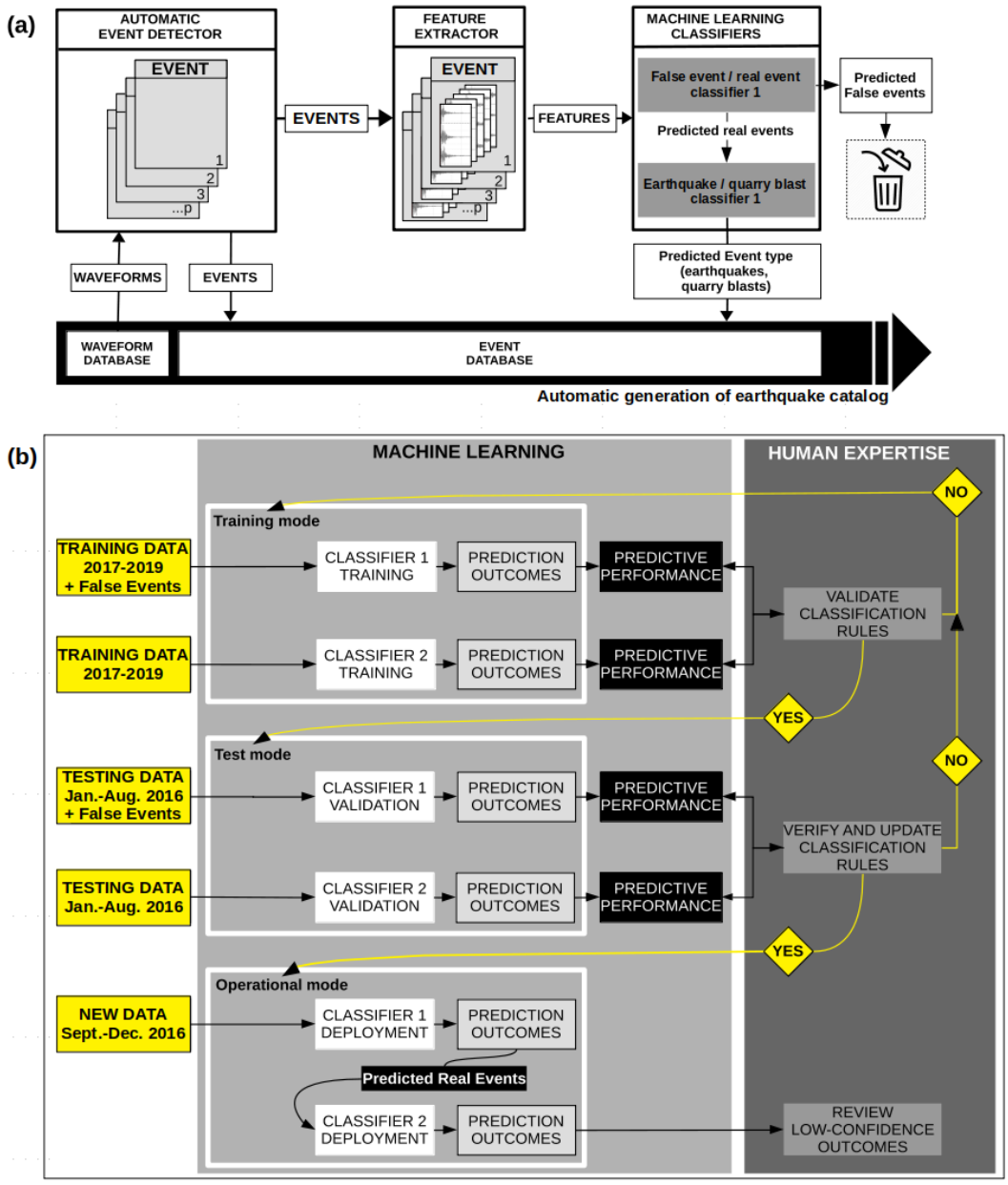


Figure 1. (a) Operational classifiers incorporated within the automatic detection procedure in near real-time (for details of the optimised event detector processing, see Figure S2 in the electronic supplement to this article). (b) Human-In-the-Loop Machine Learning architecture used to create the final operational classifiers (classifier 1 and 2).

2.1 Data

We conducted this study using 100 Hz seismic waveform data recorded between 2016 and 2019 by 226 seismic stations in the northeastern European Upper Rhine Graben area (see figure S1 in the electronic supplement to this article). Half of these were permanent stations (103 three-component broadband seismometers and 10 three-component strong motion sensors) and half were temporary stations from the AlpArray Seismic Network installed around the Alpine arc from 2015 to 2020 (Hetényi et al., 2018).

In addition to the raw waveform data, we needed a set of manually classified seismic events to train and test our machine-learning classifiers. We retrieved 10389 manually reviewed seismic event solutions (728 false alarms, 5537 earthquakes, and 4124 quarry blasts) from the database maintained by BCSF-RéNaSS between 2016 and 2019 (see figure S1). To test our machine-learning trained classifiers in a fully operational mode, we ran the full system presented in Figure 1a on four months of continuous data (09/2016-12/2016).

2.2 Feature Extraction

Each event used to perform our machine-learning procedure was coded into a vector of 361 features (see Table S1 in the electronic supplement to this article). More than half represent time and frequency domain characteristics of the vertical component seismograms from stations close to each event; a quarter represent characteristics of the 3-component seismograms from these same stations; and about one-fifth represent characteristics of the preferred origins themselves (e.g. event magnitudes, origin times and locations, uncertainties, and quality scores).

Following O'Rourke et al. (2016), we calculated our waveform features using data from the five closest stations starting from 10 s before the P-wave arrival times. We defined the duration of the data window using the STA/LTA (short-term average / long-term average) functions used for picking: we cut the signals when the value of the STALTA function after the first S arrivals (observed or inferred) descended towards its value before the first P arrivals. The waveform data were rotated to radial and transverse components, detrended and tapered before removing the instrument response.

2.3 Training mode

Our first step was to train the Random Forest machine algorithm to produce semi-automatic classification rules (Figure 1b). For this step, we used data retrieved from the BCSF-RéNaSS catalog between 2017 and 2019. Compared to many machine-learning datasets, whose number of samples run into the millions, those from seismic classification problems including ours are small, and may not fully represent the full spread of possible data. Since false events were underrepresented in the BCSF-RéNaSS dataset, we added 23747 manually reviewed false events from a supplementary un-labeled dataset from the same region. We populated each classifier's training set by randomly selecting 50% false events and 50% real events (earthquakes, quarry blasts) for classifier 1, and 50% earthquakes and 50% quarry blasts for classifier 2.

We optimised hyper-parameters linked to the Random Forest algorithm (e.g. number of trees, tree depth) using a five-fold cross-validated random-search and refined their values using a five-fold cross-validated grid-search (Bergstra & Bengio, 2012). We assessed the extracted features to find the most accurate and efficient feature representation for each classifier. We used a recursive feature elimination algorithm to select the best features (Gregorutti et al., 2017), combined with a five-fold cross-validated selection of the optimal number of features to retain. Human expertise was used to validate the rules proposed by each classifier.

2.4 Testing mode

We tested the performance of each trained classifier on representative testing sets (Figure 1b). For both classifiers, the real events in the testing set were extracted from the BCSF-RéNaSS catalog (Jan. to Aug. 2016). For classifier 1, we added previously unseen false events until the testing set reached 30% of the size of the training set. In order to estimate how our classifiers generalised, we performed 50 runs for each one, randomly resampling the training data at every run. We used human expertise to examine the rules proposed by each classifier for the first runs and for any subsequent runs that resulted in significantly different predictions.

2.5 Operational mode

We deployed the final trained classifiers in operational mode with only the best features computed for each incoming event (Figure 1b). The classifiers were run on four months (Sept. to Dec. 2016) of the automatic catalog generated by our optimised SeisComP3 detection procedure. All the incoming events predicted as real by classifier 1 were fed into classifier 2, labelled as either earthquakes or quarry blasts, and then tagged as such in the database. In order to correctly estimate operational performance and be able to manually check all prediction outcomes, the events identified as false by classifier 1 were not automatically removed from the database. Here also, we performed 50 runs for each classifier, randomly resampling the training data at every run. We used human expertise to review the final low-confidence outputs (those with low prediction probability) in order to analyse and remove the few remaining misclassifications.

3 RESULTS

We analyse each classifier in turn by first presenting the relative weight given by the classifier to each feature to check they are consistent with the physical process that generates the data, as suggested by Kohoutová et al. (2020) and J. Li et al. (2020), and then by presenting the standard classifier evaluation metrics and the distribution of classification probabilities.

3.1 Classifier 1: False Events vs Real Events

Given that false events are generated by incorrectly associating random local noise, the features that contributed most strongly to discriminating false from real events were related to location quality and pick statistics. We found the most important features to be the number of phases used, the standard deviation of the event-station distance, that of time residual distributions, and the maximum value of the function that depicts the time variations of the STA/LTA ratio (Figure 2). False events were often located with few phases because the random and transient character of impulsive anthropogenic noise made it unlikely that multiple stations in a network would reach STA/LTA pick-triggering values in a time-consecutive order (Coviello et al., 2019). The distributions of epicentral distances and time residuals had larger standard deviations for false events as they were influenced by the systematic mislocation of unrelated anthropogenic noise sources, induced by the fortuitous alignment of non-seismic phases (Arrowsmith et al., 2018). The maximum value of the STALTA function contributed also strongly because false events were generated by strong impulsive noise which triggered larger peak value in the STALTA function.

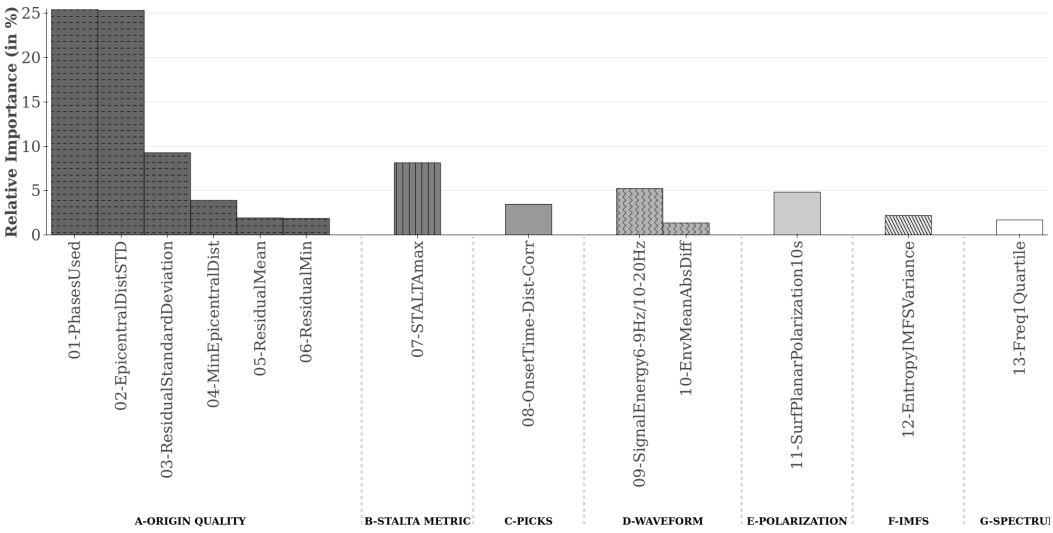


Figure 2. Relative importance of best features for classifier 1. The features belonging to group A estimate the event location quality, especially poor for false events. The features of groups B and C give information on the quality of the pick association process. Higher maximum values of the STALTA function (STALTAmax) usually indicate false picks and lower correlation coefficients between first time arrivals and epicentral distances (OnsetTime-Dist-Corr) underline false associations. Among the five signal-related features (groups D to G), the degree of planar polarization (SurfPlanarPolarization10) is correlated to the source depth: its value is higher for man-made signals since they propagate mainly as surface Rayleigh waves. The ratio of seismic energy in the low-frequency and high-frequency ranges (SignalEnergy6-9/10-20Hz) as well as the first quartile of the signal spectrum (Freq1Quartile) enhance real event prediction, especially for night-time events when seismic noise level decrease at low-frequency bands. The signal randomness and non-stationarity, higher for cultural noise, are described by the Shannon Entropy variance of the decomposed Intrinsic Mode Functions (EntropyIMFSVariance) and the mean absolute 1-order difference of the signal envelope (EnvMeanAbsDiff). See Table S1 for more detailed feature description.

The predictive performance of classifier 1 is shown in Table 1. The classifier achieved 99% precision in testing mode. Among the missed events, the majority were few poorly recorded quarry blasts or some earthquakes located out of the network. If we set aside teleseismic events, for which neither the seismic network nor the 4-20 Hz bandpass butterworth filter used to detect events were designed, the loss-rate for earthquakes was below 2% (Table 2). The classifier accurately predicted real events, especially earthquakes, with few false positives. The high predictive quality of the false event /real event classifier is underlined by its F-Measure score above 0.95. We also evaluated classifier 1 in operational mode: although the overall precision dropped to 92%, more than 99% of the false events were correctly identified. This is reassuring, as one of the side-effects of dropping the detection threshold in the automated picking phase is to increase the number of false events. In operational mode, the classifier missed fewer than 7% of earthquakes (Table 2). The majority of missed events were poorly recorded events such as quarry blasts or other events that analysts could not identify because of unclear signal signature and/or high minimum epicentral distance. The few misclassified false events were mostly located within the network and 24% of them incorporated isolated seismic signals in their association (for locations of the misclassified events in both modes, see figures S3 and S5 in the electronic supplement to this article).

Table 1. Confusion matrix and classification metrics[a] for the false event vs real event classifier

	Testing mode		**Operational mode**	
	Predicted false	**Predicted real**	**Predicted false**	**Predicted real**
Expected false	3466 ± 2	10 ± 2	46442 ± 5	117 ± 5
Expected real	40 ± 2	977 ± 2	242 ± 4	1395 ± 4
Specificity (%)	99.71 ± 0.06		99.74 ± 0.01	
Sensitivity (%)	96.07 ± 0.11		85.21 ± 0.22	
Precision (%)	98.99 ± 0.21		92.26 ± 0.33	
F-Measure	0.975 ± 0.001		0.886 ± 0.002	

[a] Specificity: the correctly predicted false event rate (i.e. the ratio of true negatives to true negatives plus false positives). Sensitivity: the correctly predicted real event rate (i.e. the ratio of true positives to true positives plus false negatives). Precision: the proportion of correctly predicted real events relative to all true positive detections (i.e. the ratio of true positives to true positives plus false positives. F-Measure: a summary statistic that combines precision and sensitivity ($2 \times \text{precision} \times \text{sensitivity}/(\text{precision} + \text{sensitivity})$).

Table 2. Description of the real events missed by the false event vs real event classifier

	Testing mode	**Operational mode**
Missed Earthquakes		
Proportion (%)	3.92 ± 0.28	6.95 ± 0.24
Number	$\frac{22 \pm 2}{576}$	$\frac{48 \pm 2}{694}$
	Teleseismic events: 11 ± 1	Teleseismic events: 0
Missed Quarry Blasts		
Proportion (%)	3.91 ± 0.40	19.32 ± 0.38
Number	$\frac{17 \pm 2}{441}$	$\frac{159 \pm 3}{822}$
Missed Unknown Events		
Proportion (%)	-	27.36 ± 0.84
Number	-	$\frac{33 \pm 1}{121}$

Figure 3a shows the prediction probability distribution for the classifier 1 in operational mode (for testing mode, see figure S6 in the electronic supplement to this article). Nearly 75% of the real events identified by the classifier were predicted with probabilities of over 0.8; after manual verification, 0.4% of these turned out to be false events. Conversely, nearly 90% of false events identified by the classifier were predicted with near certainty (probabilities of being real events under 0.1); after manual verification, fewer than 0.2% of these turned out to be real events. But what about the intermediate-level predictions? About 1% of the events in the operational mode catalog were predicted with probabilities ranging between 0.4 and 0.6; nearly 20% of these were incorrectly predicted. These incorrect predictions disproportionately involved false events being identified as real events: after manual verification, 40% of the events predicted to be real with probabilities close to 0.5 were in fact false events. In an operational setting, therefore, it would make sense to trust the near certainty predictions of this classifier, but we should probably ask operators to verify its intermediate-level predictions of real events.

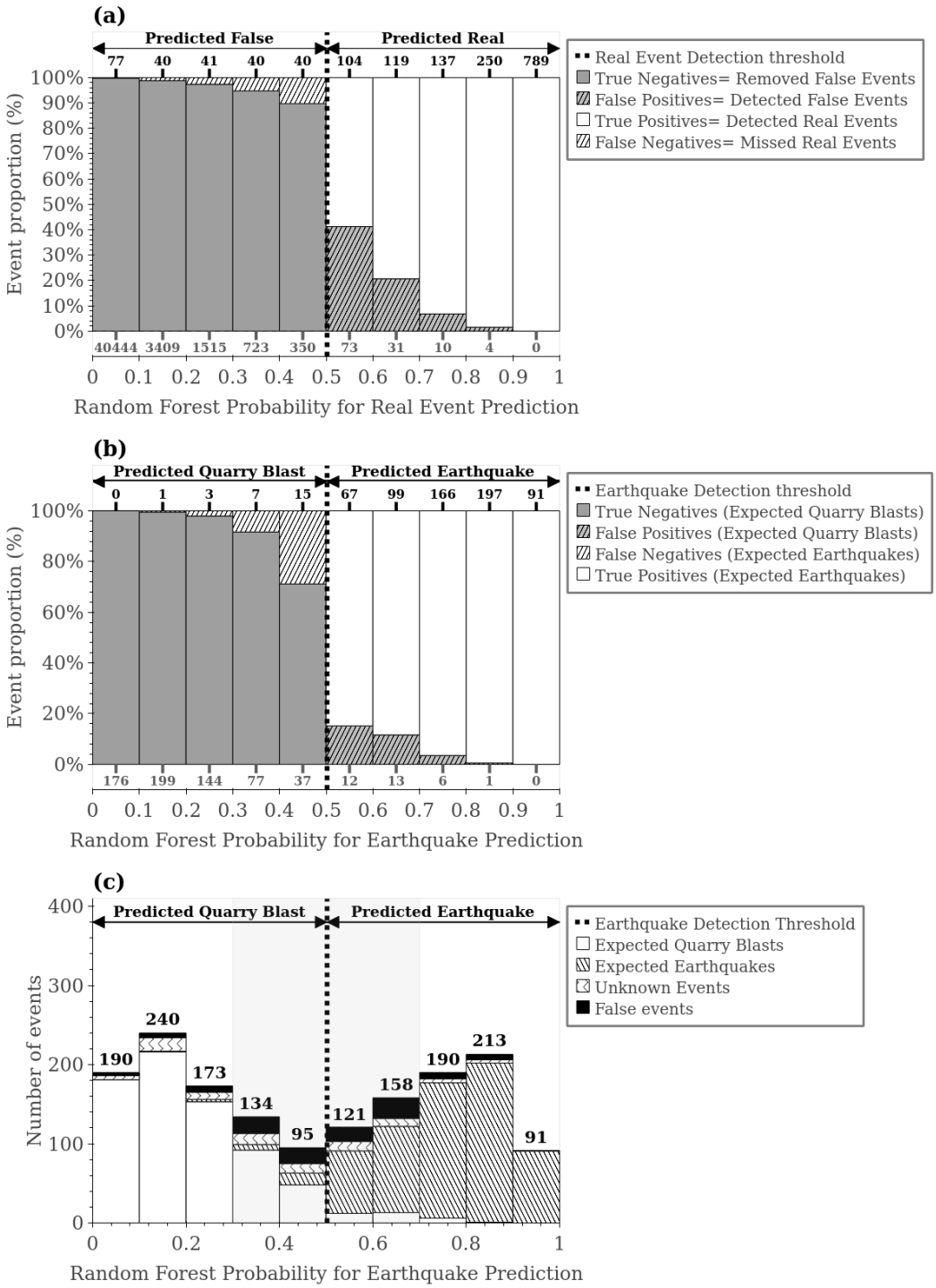


Figure 3. Distribution of prediction probabilities for (a) the trained false event vs real event and (b) the trained earthquake vs quarry blast classifiers in operational mode. (c) Full prediction outcomes for the earthquake vs quarry blast classifier.

3.2 Classifier 2: Earthquakes vs Quarry Blasts

Analysts use three main criteria to manually discriminate quarry blasts from earthquakes: their proximity to known blasting sites, their occurrence within daylight hours, and the similarity of their waveforms to those of previous quarry blasts in the area (Voyles et al., 2019). All three were among the most discriminant features for classifier 2 (Figure 4), but they were not alone. The epicentre's latitude, longitude, and distance to the nearest city allowed better quarry blast prediction accuracy than the distance to the nearest quarry alone. Instead of comparing each waveform to previous ones from known blast sites (too time consuming) we encapsulated waveform shape by calculating the skewness of the 4-20 Hz filtered seismogram. The three analyst criteria discussed here, though powerful, were insufficient by themselves, because natural earthquakes also occur near quarries during working hours and quarry-blast signals can vary with even slight ray-path changes (Dickey et al., 2019).

We improved matters by adding a feature that coded the variance of the discrete Fourier transform amplitudes, and therefore detected the narrower frequency spectrum and spectral scalloping typical of blast-related signals (Kortström et al., 2016). Compared to earthquakes of similar magnitude, quarry blasts generate longer duration coda waves (Koper et al., 2016) and have higher low-frequency surface-wave amplitudes (Musil & Plešinger, 1996) because of the shallowness of their source depth. Earthquakes that occur at shallow depths share these characteristics, but generate a higher proportion of high-frequency S-wave energy than quarry blasts, giving a small role to features that encode the P-to-S wave spectral ratios and the ratio of vertical to horizontal peak ground acceleration (Fereidoni & Atkinson, 2017).

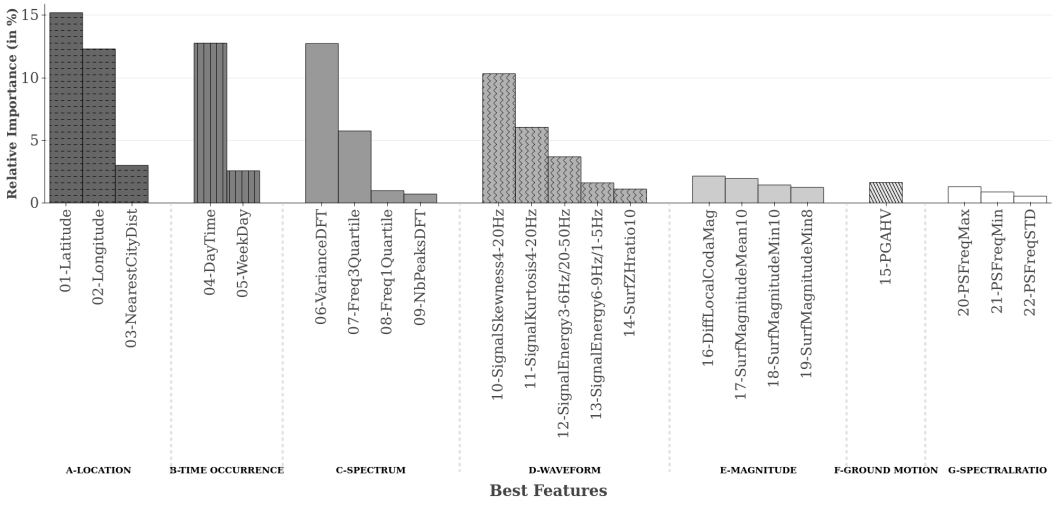


Figure 4. Relative importance of best features for classifier 2. The features belonging to group A and B give information on potential candidates for being quarry blasts (their epicenter position within or very near a blasting site, its occurrence on working days and daylight hours). The signal-related features (groups C to G) validate or invalidate the preceding diagnostic using information on the spectral frequency content (group C), the waveform shape (group D: SignalSkewness4-20Hz and SignalKurtosis4-20Hz) and the wavefield properties (surface waves: group D -SignalEnergy6-9Hz/1-5Hz, SignalEnergy3-6Hz/20-50Hz, SurfZHratioMax10- and group E; P- and S-waves: group F and group G). See Table S1 for more detailed feature description.

The performance of classifier 2 is shown in Table 3. The classifier obtained high scores in both testing mode and operational mode, with a precision between 94.7 and 96.6% and a F-Measure score between 0.95 and 0.96. It identified incorrectly just over 3-4% of earthquakes and 4-5% of quarry blasts. When attempting to classify the few false events incorrectly labeled by the first classifier (118 out of over 46 560), it split them equally between earthquakes and quarry blasts. We manually checked the misclassified earthquakes and quarry blasts. The misclassified earthquakes were located near blasting sites, occurred in working hours, and had low frequency variance. Most of their waveforms were similar to those associated with quarry blasts. Almost half the misclassified earthquakes had high P/S spectral ratios and some of them had high surface magnitude values, probably due to the shallowness of their sources. The misclassified quarry blasts had more high frequency content, potentially due to the nature of the extracted material (very competent rocks such as basalt, gabbro, or rhyodacite), and were difficult to classify manually. The locations of earthquakes and quarry blasts predicted by classifier 2 for both modes are shown in figures S4 and S5.

Table 3. Confusion matrix and classification metrics[a] for the earthquake vs quarry blast classifier

	Testing mode		**Operational mode**	
	Predicted quake	**Predicted blast**	**Predicted quake**	**Predicted blast**
Expected quake	558 ± 2	18 ± 2	620 ± 2	26 ± 2
Expected blast	20 ± 2	421 ± 2	35 ± 3	630 ± 3
Specificity (%)	95.54 ± 0.49		94.76 ± 0.34	
Sensitivity (%)	96.90 ± 0.36		96.04 ± 0.30	
Precision (%)	96.60 ± 0.36		94.68 ± 0.33	
F-Measure	0.966 ± 0.002		0.953 ± 0.002	

[a] Specificity: the correctly predicted quarry blast rate (i.e. the ratio of true negatives to true negatives plus false positives). Sensitivity: the correctly predicted earthquake rate (i.e. the ratio of true positives to true positives plus false negatives). Precision: the proportion of correctly predicted earthquakes relative to all true positive detections (i.e. the ratio of true positives to true positives plus false positives. F-Measure: a summary statistic that combines precision and sensitivity ($2 \times \text{precision} \times \text{sensitivity}/(\text{precision} + \text{sensitivity})$).

Figure 3b,c shows the prediction probability distributions for the expected earthquakes and quarry blasts in operational mode (see Figure S6 for testing mode). Confusion between earthquakes and quarry blasts was due in large part to very shallow earthquakes, earthquakes that occurred close to quarries or in urban environments, or the false events let through by classifier 1.

Probability values allow analysts to streamline their operational processes. In our 4-month operational dataset, almost 80% of misclassified quarry blasts, 85% of misclassified earthquakes, nearly 70% of false events, and more than half of the manually unclassifiable events had probabilities between 0.3 and 0.7 (Figure 3). If analysts concentrated on re-visiting events in this probability range, they would screen just over 30% of the events let through by classifier 1 and have to correct about one in three events screened. This works out to 6-7 events that need to be screened per day, compared to 582 events per day if no machine-learning classifier were available. At the end of the human-assisted machine-learning procedure (enhanced detection, followed by the two classifiers, followed by manual screening) only 1% of the events would be manually re-tagged and the final catalog would

contain only 1% of misclassified quarry blasts, 0.5% of misclassified earthquakes, and 0.08% of false events.

4 DISCUSSION

The hybrid approach we designed involved a strong partnership between humans and machine-learning algorithms that improved their respective performances. From an operational standpoint, machine-learning classifiers can reduce the number of events that require manual discrimination. This is particularly important where the number of expected false events and/or anthropogenic events is high (in our case because we lowered the network's detection threshold in order to lower the regional completeness magnitude), as classifiers can eliminate the cry wolf effect caused by false alarm fatigue and help reduce the number of missed real events (Heldt, 2015; Lim et al., 2019). Our procedure removed more than 99% of false events from the original 48 000 detected events in a few minutes, missing fewer than 7% of earthquakes out of 14% of missed real events, whereas manual review of the same data took several months and missed 30% of the real events. Machine-learning classifiers can also assist analysts in making diagnoses and resolve erroneous labeling. In training mode, classifier 2 uncovered 1.63% of mis-classified earthquakes and 2.04% of mis-classified quarry-blasts in the manually labeled training set, comparable to the proportion of mislabeled events in other catalogs (e.g Utah catalog, Linville et al., 2019). Adding manual review of the uncertain classifications using the probability information would then result in even cleaner catalogs.

However, human input should not be relegated simply to checking the output of the machine-learning classifiers. Recent studies have underscored the strong link between the validity of classifiers and their interpretability (Rudin, 2019; J. Li et al., 2020). To help detect and avoid biases in the classifiers, especially where the number of training samples is small, such as in seismic discrimination problems like ours, we need humans to use their domain expertise to assist in selecting features and validating models.

Some features that seem at first glance to be good candidates for driving event classification turn out to be irrelevant, and we need domain-level knowledge to understand why. We found that features related to absolute event locations should not be relied upon too heavily, as they can have large uncertainties due to poor knowledge of the seismic velocity structures. Furthermore, we confirmed that in a geological context ruled by heterogeneities, sharp lateral discontinuities, and path effects, features that code for the signal's envelope discriminate poorly between earthquakes and quarry blasts.

Without knowing why and how a classification model works, it is difficult to know when it will fail, to which seismic event subgroups it applies, and how it can advance our understanding of the mechanisms underlying event classification performance (Kohoutová et al., 2020). We chose to implement Random Forest classifiers because they give direct access to the sequences in which the features are taken into account by the decision trees they are made from. These sequences strongly influence the final outcome. For example, we found that classifier 1 trees that used location-related features (number and quality of picks, epicentral uncertainties etc.) to perform the bulk of the classification and waveform-related features to refine it performed better than the trees that used the same features in inverse order (see Figure S7 in the electronic supplement to this article). Man-made signals carry ample energy in the 1-10-Hz frequency band often used to observe regional seismic signals in urban environments (Inbal et al., 2018; Poli et al., 2020). This overlap in frequency content and amplitude makes it difficult to use signal-related features as primary predictors. Another example concerns classifier 2: its decision trees first split events into geographically dependent daylight vs non-daylight groups, then refined each group based on its waveform features (see Figures S8 and S9 in the electronic supplement to this article). This correlates with previously noted regional variabilities in the effectiveness of signal discriminants (Baumgardt & Young, 1991; Tibi et al., 2019).

We analysed several tens of trees out of the 500 in each classifier, which allowed us to validate or invalidate some choices made by the Random Forest algorithm's recursive feature selection, and remove certain features entirely. The refined classifiers generalised better than the original versions, and improved their predictions for data taken from an entirely different study area (the Pyrenees region). Even finer understanding and refinement of the classification rules would require analysing the entire forest using automated methods and tools (Lapuschkin et al., 2019; Lundberg et al., 2020; Samek, 2020).

Because no previous studies have combined a false vs real event classifier with a quake vs blast classifier, we compare our two classifiers separately to those documented in the literature. Additional validation for machine-learning classifiers can be provided by the geophysical plausibility of their classification rules (Kohoutová et al., 2020).

Our finding that the number of phases used for an event location is a strong discriminator between false and real events was previously noted by Draelos et al. (2012), who found that this single feature could be used to correctly classify 76% of the events at the International Data Center. We also agree with Draelos et al. (2012) that the lowest slowness residual, the lowest slowness uncertainty and the highest signal-to-noise ratio contribute significantly to false event vs real event discrimination. Some studies in earthquake early warning exploited signal impulsivity measured through kurtosis or skewness to distinguish large earthquakes from noise in the Western U.S. (Meier et al., 2019). However we found these features less useful in our moderate-seismicity context because many of our noise sources generated impulsive, transient signals with amplitudes similar to many earthquakes (Westfall, 2014). Instead, we found that polarisation features, such as the degree of planarity of the surface wave-field, helped improve classifier 1 because they are highly correlated with the source depth, as previously observed by Chouet et al. (1997) and Mousavi et al. (2016). Many of the best features retained by our quake vs blast classifier were also used in previous studies: the daytime hours discriminant was used in Switzerland, Alaska, and western United States by Wiemer and Baer (2000) and in South Africa by Zaliapin and Ben-Zion (2016); spectral parameters as well as spectral ratios were used in Southern California, USA by Allmann et al. (2008) and in Turkey by (Kuyuk et al., 2011); ground motion parameters were used in US western Alberta by Fereidoni and Atkinson (2017), surface-wave magnitude was used in Italy by Bonner et al. (2011), and the change in coda energy was used in Utah, USA by Koper et al. (2016).

5 CONCLUSION

We implemented two Random Forest classifiers that can be integrated into the SeiscomP3 workflow of the French seismic monitoring center BCSF-RéNaSS, allowing us to lower the detection threshold of the network without analysts being overwhelmed by the increase in the number of false detections. When run in an operational setting, our system detected more small earthquakes and quarry-blasts while requiring direct input from analysts for fewer than 1% of the events, and led to a final catalog containing only 1% of misclassified quarry blasts, 0.5% of misclassified earthquakes, and 0.08% of false events.

As suggested by many recent studies (Alber et al., 2019; Kong et al., 2019; Tibi et al., 2019; Kohoutová et al., 2020; J. Li et al., 2020; Lundberg et al., 2020), we have preferred a hybrid approach that integrates humans in the system at all levels of the machine-learning implementation, including feature selection, model refinement, and decision making on events for which the classifier predictions are uncertain. We believe such close human-machine integration is necessary to provide optimal classification results, especially in fields, such as seismic discrimination, where natural variability of events is high but sample sizes for training are low.

6 DATA AND RESOURCES

This work included data from the permanent seismic networks operated by the French seismological and geodetic network (RESIF), the Swiss Seismological Service (SED), the German Research Center for Geosciences in Potsdam (GFZ), the German State Office of Geology, Natural Resources and Mining of Freiburg (LGRB), and the Royal Observatory of Belgium (ROB) as well as AlpArray temporary seismic network Z3 (Hetényi et al., 2018; AlpArray Seismic Network, 2015). The waveform data are available through EIDA (http://www.orfeus-eu.org/eida, last accessed September 2019). The catalog used for training and testing phases is provided by the French National Service of Observation (BCSF-RéNaSS) and available using a FDSN protocol (http://renass.unistra.fr, last accessed July 2020). The catalog produced in operational mode is a currently unpublished catalog of the wide region surrounding the Upper Rhine Graben area, but is available upon request. The quarry database is also a currently unpublished database and available upon request. Some features are provided by the French geological survey (BRGM) and available via a web feature service (http://geoservices.brgm.fr/odmgm, last accessed July 2019). All data processing used in the study is made under the SeisComP3 framework. All SeisComP3 modules were written in Python (feature extraction, classification, event labeling and false event removal) and can be fully integrated in the SeisComP3 monitoring system. The codes can be available upon request.

7 AKNOWLEDGMENTS

We would like to acknowledge the High Performance Computing (HPC) center of the University of Strasbourg for supporting this work by providing scientific support and access to computing resources. Part of the computing resources were funded by the Equipex Equip@Meso project (Programme Investissements d'Avenir) and the CPER Alsacalcul/Big Data. We acknowledge the operation of the AlpArray temporary seismic network Z3 (Hetényi et al., 2018; AlpArray Seismic Network, 2015), which is part of the project AlpArray-FR funded by Agence Nationale de la Recherche (contract ANR-15-CE31-0015). We warmly thank Clément Grellier for his guidance in the world of SeisComP3 and HPC. We acknowledge funding from the LABEX ANR-11-LABX-0050-G-EAU-THERMIE-PROFONDE.

References

Alber, M., Buganza Tepole, A., Cannon, W. R., De, S., Dura-Bernal, S., Garikipati, K., ... Kuhl, E. (2019). Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. *npj Digital Medicine*, *115*(2). doi: 10.1038/s41746-019-0193-y

Allmann, B. P., Shearer, P. M., & Hauksson, E. (2008). Spectral discrimination between quarry blasts and earthquakes in southern california. *Bulletin of the Seismological Society of America*, *98*(4), 2073–2079. doi: 10.1785/0120070215

AlpArray Seismic Network. (2015). *AlpArray Seismic Network (AASN) Temporary Component, AlpArray Working Group. doi* : 10.12686/*alparray*/$z3_2015$.

Arrowsmith, S., Euler, G., Marcillo, O., Blom, P., Whitaker, R., & Randall, G. (2014). Development of a robust and automated infrasound event catalogue using the internationalmonitoring system. *Geophysical Journal International*, *200*(3), 1411–1422. doi: 10.1093/gji/ggu486

Arrowsmith, S., Young, C., & Pankow, K. (2018). Implementation of the waveform correlation event detection system (wceds) method for regional seismic event detection in utah. *Bulletin of the Seismological Society of America*, *108*(6), 3548–3561. doi: 10.1785/0120180097

Baumgardt, D. R., & Young, G. B. (1991). Regional seismic waveform discriminants and case-based event identification using regional arrays. *Bulletin - Seismological Society*

of America, *80*(6), 1874–1892.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*, 281–305.

Bezada, M. J., & Smale, J. (2019). Lateral variations in lithospheric mantle structure control the location of intracontinental seismicity in australia. *Geophysical Research Letters*, *46*(22). doi: 10.1029/2019GL084848

Bonner, J. L., Stroujkova, A., & Anderson, D. (2011). Determination of love-and rayleigh-wave magnitudes for earthquakes and explosions. *Bulletin of the Seismological Society of America*, *101*(6), 3096–3104. doi: 10.1785/0120110131

Breiman, L. (2001). Random forests. *Machine Learning*. doi: 10.1023/A:1010933404324

Brodsky, E. E. (2019). The importance of studying small earthquakes. *Science*, *364*(6442), 736–737. doi: 10.1126/science.aax2490

Chouet, B., Saccorotti, G., Martini, M., Dawson, P., De Luca, G., & Milana, G. (1997). Source and path effects in the wave fields of tremor and explosions at stromboli volcano, italy. *Journal of Geophysical Research: Solid Earth*, *102*(B7), 15129–15150. doi: 10.1029/97JB00953

Coviello, V., Arattano, M., Comiti, F., Macconi, P., & Marchi, L. (2019). Seismic characterization of debris flows: Insights into energy radiation and implications for warning. *Journal of Geophysical Research: Earth Surface*, *124*(6), 1440–1463. doi: 10.1029/2018JF004683

Díaz, J., Ruiz, M., Sánchez-Pastor, P. S., & Romero, P. (2017). Urban seismology: On the origin of earth vibrations within a city. *Scientific Reports*, *7*(15296). doi: 10.1038/s41598-017-15499-y

Dickey, J., Borghetti, B., Junek, W., & Martin, R. (2019). Beyond correlation: A path-invariant measure for seismogram similarity. *Seismological Research Letters*, *91*(1), 356–369. doi: 10.1785/0220190090

Dowla, F. U., Taylor, S. R., & Anderson, R. W. (1990). Seismic discrimination with artificial neural networks: preliminary results with regional spectral data. *Bulletin - Seismological Society of America*, *80*(5), 1346–1373.

Draelos, T. J., Peterson, M. G., Knox, H. A., Lawry, B. J., Phillips-Alonge, K. E., Ziegler, A. E., ... Faust, A. (2018). Dynamic tuning of seismic signal detector trigger levels for local networks. *Bulletin of the Seismological Society of America*, *108*(3), 1346–1354. doi: 10.1785/0120170200

Draelos, T. J., Procopio, M. J., Lewis, J. E., & Young, C. J. (2012). False event screening using data mining in historical archives. *Seismological Research Letters*, *83*(2), 267–274. doi: 10.1785/gssrl.83.2.267

Drouin, A., Letarte, G., Raymond, F., Marchand, M., Corbeil, J., & Laviolette, F. (2019). Interpretable genotype-to-phenotype classifiers with performance guarantees. *Scientific Reports*, *9*(4071). doi: 10.1038/s41598-019-40561-2

Fereidoni, A., & Atkinson, G. M. (2017). Discriminating earthquakes from quarry blasts based on shakemap ground-motion parameters. *Bulletin of the Seismological Society of America*, *107*(4), 1931–1939. doi: 10.1785/0120160308

Gallen, S. F., & Thigpen, J. R. (2018). Lithologic controls on focused erosion and intraplate earthquakes in the eastern tennessee seismic zone. *Geophysical Research Letters*, *45*, 9569–9578. doi: 10.1029/2018GL079157

Gennatas, E. D., Friedman, J. H., Ungar, L. H., Pirracchio, R., Eaton, E., Reichmann, L. G., ... Valdes, G. (2020). Expert-augmented machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(9), 4571–4577. doi: 10.1073/pnas.1906831117

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing*, *27*(3), 659–678.

Heldt, T. (2015). Beep, beeep, beeeep, beeeeeep. *Science Translational Medicine*, *310*(7). doi: 10.1126/scitranslmed.aad4451

Hetényi, G., Molinari, I., Clinton, J., Bokelmann, G., Bondár, I., Crawford, W. C., ... Zieke, T. (2018). The alparray seismic network: A large-scale european experiment

to image the alpine orogen. *Surveys in Geophysics*, *39*(5), 1009–1033. doi: 10.1007/s10712-018-9472-4

Inbal, A., Cristea-Platon, T., Ampuero, J. P., Hillers, G., Agnew, D., & Hough, S. E. (2018). Sources of long-range anthropogenic noise in southern california and implications for tectonic tremor detection. *Bulletin of the Seismological Society of America*, *108*(6), 3511–3527. doi: 10.1785/0120180130

Kohoutová, L., Heo, J., Cha, S., Lee, S., Moon, T., Wager, T. D., & Woo, C.-W. (2020). Toward a unified framework for interpreting machine-learning models in neuroimaging. *Nature Protocols*, *15*(4), 1399–1435. doi: 10.1038/s41596-019-0289-5

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. doi: 10.1785/0220180259

Koper, K. D., Pechmann, J. C., Burlacu, R., Pankow, K. L., Stein, J., Hale, J. M., ... McCarter, M. K. (2016). Magnitude-based discrimination of man-made seismic events from naturally occurring earthquakes in utah, usa. *Geophysical Research Letters*, *43*(20), 10638–10645. doi: 10.1002/2016GL070742

Kortström, J., Uski, M., & Tiira, T. (2016). Automatic classification of seismic events within a regional seismograph network. *Computers and Geosciences*, *87*, 22–30. doi: 10.1016/j.cageo.2015.11.006

Kuyuk, H. S., Yildirim, E., Dogan, E., & Horasan, G. (2011). An unsupervised learning algorithm: Application to the discrimination of seismic events and quarry blasts in the vicinity of istanbul. *Natural Hazards and Earth System Science*, *11*(1), 93–100. doi: 10.5194/nhess-11-93-2011

Lapuschkin, S., S., W., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, *10*(1096). doi: 10.1038/s41467-019-08987-4

Leclère, H., & Calais, E. (2019). A parametric analysis of fault reactivation in the new madrid seismic zone: The role of pore fluid overpressure. *Journal of Geophysical Research: Solid Earth*, *124*(10), 10630–10648. doi: 10.1029/2018JB017181

Levandowki, W., Herrmann, R. B., Briggs, R., O., B., & R., G. (2018). An updated stress map of the continental united states reveals heterogeneous intraplate stress. *Nature Geoscience*, *11*, 433–437. doi: 10.1038/s41561-018-0120-x

Li, J., Liu, L., Le, T. D., & Liu, J. (2020). Accurate data-driven prediction does not mean high reproducibility. *Nature Machine Intelligence*, *2*, 13–15. doi: 10.1038/s42256-019-0140-2

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. doi: 10.1029/2018GL077870

Lim, J. R., Liu, B. F., & Egnoto, M. (2019). Cry wolf effect? evaluating the impact of false alarms on public responses to tornado alerts in the southeastern united states. *Weather, Climate, and Society*, *11*(3), 549–563. doi: 10.1175/WCAS-D-18-0080.1

Lindenbaum, O., Rabin, N., Bregman, Y., & Averbuch, A. (2017). Multi-channel fusion for seismic event detection and classification. In *2016 ieee international conference on the science of electrical engineering, icsee 2016.* doi: 10.1109/ICSEE.2016.7806088

Linville, L., Pankow, K., & Draelos, T. (2019). Deep learning models augment analyst decisions for event discrimination. *Geophysical Research Letters*, *46*(7), 3643–3651. doi: 10.1029/2018GL081119

Lundberg, S., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... S.-I., L. (2020). From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, *2*, 56–67. doi: 10.1038/s42256-019-0138-9

Maggi, A., Ferrazzini, V., Hibert, C., Beauducel, F., Boissier, P., & Amemoutou, A. (2017, April). Implementation of a Multistation Approach for Automated Event Classification at Piton de la Fournaise Volcano. *Seismological Research Letters*, *88*(3), 878–891.

McBrearty, I. W., Delorey, A. A., & Johnson, P. A. (2019). Pairwise association of seismic arrivals with convolutional neural networks. *Seismological Research Letters*, *90*(2).

534 doi: 10.1785/0220180326
535 Meier, M. A., Ross, Z. E., Ramachandran, A., Balakrishna, A., Nair, S., Kundzicz, P., ...
536 Yue, Y. (2019). Reliable real-time seismic signal/noise discrimination with machine
537 learning. *Journal of Geophysical Research: Solid Earth*, *124*(1), 788–800. doi: 10
538 .1029/2018JB016661
539 Mousavi, S. M., Horton, S. P., Langston, C. A., & Samei, B. (2016). Seismic features and
540 automatic discrimination of deep and shallow induced-microearthquakes using neural
541 network and logistic regression. *Geophysical Journal International*, *207*(1), 29–46.
542 doi: 10.1093/gji/ggw258
543 Musil, M., & Plešinger, A. (1996). Discrimination between local microearthquakes and
544 quarry blasts by multi-layer perceptrons and kohonen maps. *Bulletin of the Seismo-*
545 *logical Society of America*, *86*(4), 1077–1090.
546 Olivieri, M., & Clinton, J. (2012). An almost fair comparison between earthworm and
547 seiscomp3. *Seismological Research Letters*, *83*(4), 720–727. doi: 10.1785/0220110111
548 Onagawa, R., Shinya, M., Ota, K., & Kudo, K. (2019). Risk aversion in the adjustment of
549 speed-accuracy tradeoff depending on time constraints. *Scientific Reports*, *9*(1), 1–12.
550 doi: 10.1038/s41598-019-48052-0
551 O'Rourke, C. T., Baker, G. E., & Sheehan, A. F. (2016). Using p/s amplitude ratios
552 for seismic discrimination at local distances. *Bulletin of the Seismological Society of*
553 *America*, *106*(5), 2320–2331. doi: 10.1785/0120160035
554 Patel, B. N., Rosenberg, L., Willcox, G., Baltaxe, D., Lyons, M., Irvin, J., ... Lungren, M. P.
555 (2019). Human–machine partnership with artificial intelligence for chest radiograph
556 diagnosis. *npj Digital Medicine*, *2*(1), 1–10. doi: 10.1038/s41746-019-0189-7
557 Perol, T., Gharbi, M., & Denolle, M. (2018). Convolutional neural network for earthquake
558 detection and location. *Science Advances*, *4*(2), 1–8. doi: 10.1126/sciadv.1700578
559 Poli, P., Boaga, J., Molinari, I., Cascone, V., & Boschi, L. (2020). The 2020 coronavirus
560 lockdown and seismic monitoring of anthropic activities in northern italy. *Scientific*
561 *Reports*, *10*(9404). doi: 10.101038/s41598-020-66368-0
562 Ross, Z. E., Meier, M.-A., & Hauksson, E. (2019). P wave arrival picking and first-motion
563 polarity determination with deep learning. *Journal of Geophysical Research: Solid*
564 *Earth*, *123*(6), 5120–5129. doi: 10.1029/2017JB015251
565 Ross, Z. E., Trugman, D. T., Hauksson, E., & Shearer, P. M. (2019). Searching for
566 hidden earthquakes in southern california. *Science*, *364*(6442), 767–771. doi: 10.1126/
567 science.aaw6888
568 Rouet-Leduc, B., Hulbert, C., & Johnson, P. A. (2019). Continuous chatter of the cascadia
569 subduction zone revealed by machine learning. *Nature Geoscience*, *12*(1), 75–79. doi:
570 10.1038/s41561-018-0274-6
571 Rudin, C. (2019). Stop explaining black box machine learning models for high stakes
572 decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*, 206–
573 215. doi: 10.1038/s42256-019-0048-x
574 Samek, W. (2020). Learning with explainable trees. *Nature Machine Intelligence*, *2*, 16–17.
575 doi: 10.1038/s42256-019-0142-0
576 Schaumberg, A. J., Juarez-Nicanor, W. C., Choudhury, S. J., Pastrián, L. G., Pritt, B. S.,
577 Prieto-Pozuelo, M., ... Fuchs, T. J. (2020). Interpretable multimodal deep learning for
578 real-time pan-tissue pan-disease pathology search on social media. *Modern Pathology*.
579 doi: 10.1038/s41379-020-0540-1
580 Sparks, R. S., Biggs, J., & Neuberg, J. W. (2012). Monitoring volcanoes. *Science*, *335*(6074),
581 1310–1311. doi: 10.1126/science.1219485
582 Tibi, R., Linville, L., Young, C., & Brogan, R. (2019). Classification of local seismic events
583 in the utah region: A comparison of amplitude ratio methods with a spectrogram-
584 based machine learning approach. *Bulletin of the Seismological Society of America*,
585 *109*(6), 2532–2544. doi: 10.1785/0120190150
586 Tiira, T. (1999). Detecting teleseismic events using artificial neural networks. *Computers*
587 *and Geosciences*, *25*(8), 929–939. doi: 10.1016/S0098-3004(99)00056-4
588 Voyles, J. R., Holt, M. M., Hale, J. M., Koper, K. D., Burlacu, R., & Chambers, D. J. (2019).

A new catalog of explosion source parameters in the utah region with application to ml–mc-based depth discrimination at local distances. *Seismological Research Letters*, *91*(1), 222–236. doi: 10.1785/0220190185

Wang, J., & Teng, T.-L. (1995). Artificial neural network-based seismic detector. *Bulletin - Seismological Society of America*, *85*(1), 308–319. doi: 10.1016/0148-9062(96)86904-x

Westfall, P. H. (2014). Kurtosis as peakedness, 1905–2014. r.i.p. *American Statistician*, *68*(3), 191–195. doi: 10.1080/00031305.2014.917055

Wiemer, S., & Baer, M. (2000). Mapping and removing quarry blast events from seismicity catalogs. *Bulletin of the Seismological Society of America*, *90*(2), 525–530. doi: 10.1785/0119990104

Yeck, W. L., Patton, J. M., Johnson, C. E., Kragness, D., Benz, H. M., Earle, P. S., ... Ambruz, N. B. (2019). Glass3: A standalone multiscale seismic detection associator. *Bulletin of the Seismological Society of America*, *109*(4), 1469–1478. doi: 10.1785/0120180308

Zaliapin, I., & Ben-Zion, Y. (2016). Discriminating characteristics of tectonic and human-induced seismicity. *Bulletin of the Seismological Society of America*, *106*(3), 846–859. doi: 10.1785/0120150211

Zhu, W., & Beroza, G. C. (2019). Phasenet: A deep-neural-network-based seismic arrival-time picking method. *Geophysical Journal International*, *216*(1), 261–273. doi: 10.1093/gji/ggy423

5.3.2 Supplément de l'article

SUPPLEMENTAL MATERIAL

Title: Monitoring Regional Seismicity using Hybrid Intelligence

Authors: Alexandra Renouard, Alessia Maggi, Marc Grunberg, Cécile Doubre, Clément Hibert

The supplement includes one table and 9 figures :

- 1) the description of the 361 features initially used to create the automatic machine-learning classification rules (Table S1); **page 2**
- 2) the study area and the distribution of the station network used (Figure S1); **page 4**
- 3) the optimised detection and post-detection procedure developed to detect small earthquakes (figure S2); **page 5**
- 4) the geographical locations of the training data, the testing data and the operational data (Figures S3, S4 and S5) **page 6**;
- 5) the distribution of prediction probabilities for classifiers 1 and 2 in testing mode (Figure S6) **page 9**;
- 6) the visualization of a simplified part of a decision tree for each classifier: classifier 1 (Figure S7) and classifier 2 (Figure S8 + its map projection Figure S9) **page 10**.

Table S1. Feature description. (**separate file (*TableS1.pdf*) containing the details of the 361 features initially used for classifier training)**. References cited for some features : surface magnitudes (Bonner et al., 2006, onner et al., 2006; Russel, 2006, ussel, 2006; Selby, 2001, elby, 2001), coda magnitudes (Holt et al., 2019, olt et al., 2019; Koper et al., 2016, oper et al., 2016), ratio of vertical to horizontal amplitudes in Rayleigh waves (Tanimoto and Rivera, 2008, animoto and Rivera, 2008), polarization analysis (Jurkevics, 1988, urkevics, 1988; Vidale, 1986, idale, 1986), complexity measure (Batista et al., 2014, atista et al., 2014), spectral centroid (Tzanetakis and Cook, 2002, zanetakis and Cook, 2002), spectrogram features (Provost et al., 2017, rovost et al., 2017), empirical mode decomposition and Hilbert spectrum (Huang et al., 1998, uang et al., 1998).

References

Batista, G. E. A. P. A., Keogh, E. J., Tataw, O. M., and de Souza, V. M. A. 2014. Cid: an efficient complexity-invariant distance for time series, *Data Min. Knowl. Disc.* **28** 634–669, doi: doi: 10.1007/s10618-013-0312-3.

Bonner, J. L., Russel, D. R., Harkrider, D. G., Reiter, D. T., and B., H. R. 2006. Development of a time-domain, variable-period surface-wave magnitude measurement procedure for application at regional and teleseismic distances, part ii: Application and ms–mb performance, *Bull. Seismol. Soc. Am.* **96**, no 2, 678–696, doi: 10.1785/0120050056.

Holt, M. M., Koper, K. D., Yeck, W., D'Amico, S., Li, Z., Hale, J. M., and Burlacu, R. 2019. On the portability of ml–mc as a depth discriminant for small seismic events recorded at local distances, *Bull. Seismol. Soc. Am.* **109**, no 5, 1661–1673. doi: 10.1785/0120190096.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., . . . Liu, H. H. 1998. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis, *Proc. Math. Phys. Eng. Sci.* **454**, no 1971, 903–995.

Jurkevics, A. 1988. Polarization analysis of three-component array data, *Bull. Seismol. Soc. Am.* **78**, no 5, 1725–1743.

K. D., Pechmann, J. C., Burlacu, R., Pankow, K. L., Stein, J., Hale, J. M., Roberson, P., and McCarter, M. K. 2016. Magnitude-based discrimination of man-

made seismic events from naturally occurring earthquakes in Utah, USA, *Geophys. Res. Lett.* **43**, no 20, 10638–10645, doi:10.1002/2016GL070742.

Provost, F., Hibert, C., and Malet, J.-P. 2017. Automatic classification of endogenous landslide seismicity using the random forest supervised classifier, *Geophys. Res. Lett.* **44**, no 1, 113-120, doi:10.1002/2016GL070709.

Russel, D. R. 2006. Development of a time-domain, variable-period surface-wave magnitude measurement procedure for application at regional and teleseismic distances, part i: Theory, *Bull. Seismol. Soc. Am.* **96**, no 2, 665–677, doi: 10.1785/0120050055

Selby, N. D. 2001. Association of rayleigh waves using backazimuth measurements: Application to test ban verification, *Bull. Seismol. Soc. Am.* **91**, no 3, 580–593, doi: 10.1785/0120000068.

Tanimoto, T., and Rivera, L. 2008. The ZH ratio method for long-period seismic data: sensitivity kernels and observational techniques, *Geophys. J. Int.* **172**, no 1, 187–198, doi: 10.1111/j.1365-246X.2007.03609.x

Tzanetakis, G., and Cook, P. 2002. Musical genre classification of audio signals. *IEEE Trans. Speech Audio Process.* **10**, no 5, 293–301, doi: 10.1109/TSA.2002.800560

Vidale, J. E. 1986. Complex polarization analysis of particle motion, *Bull. Seismol. Soc. Am.* **76**, no 5, 1393–1405.

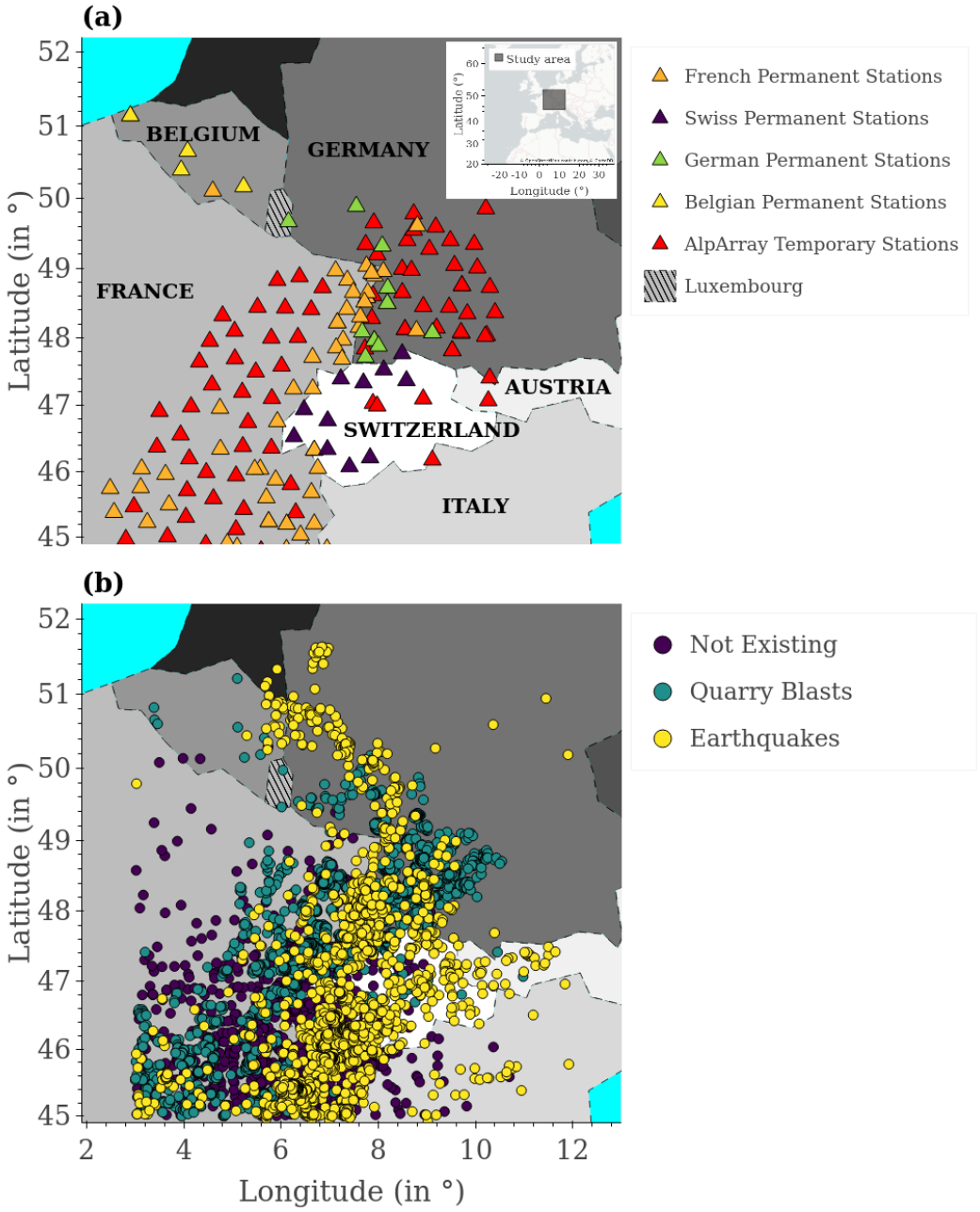


Figure S1. (a) Station network of the studied area and (b) events retrieved from the French national catalog (BCSF-RéNaSS)

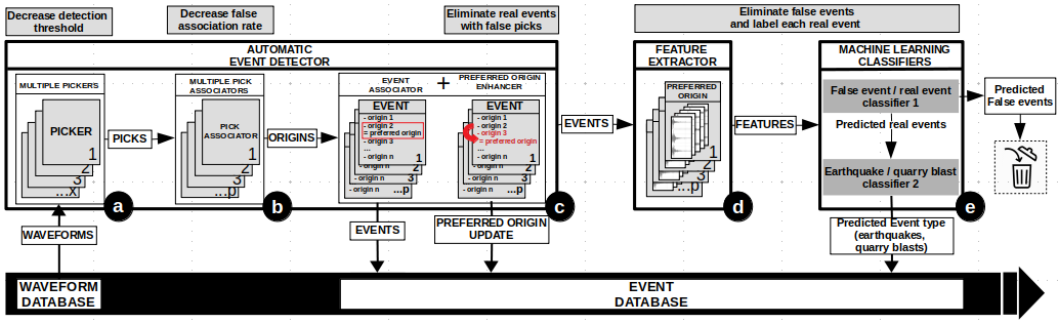


Figure S2. SeisComP3 optimised detection and post-detection procedure. (a) We improved the quality of the automatic P- and S-arrival time picking by implementing multiple picker instances that account for the space-time-varying noise characteristics of individual stations. We also increased the sensitivity of the STA/LTA pick triggers. (b) We enhanced the pick association process (process of grouping together phase arrival picks to create and locate an origin) by implementing multiple pick associator instances that account for the space-varying velocity characteristics of the seismic wave propagation medium. This helps to decrease misdetections (actual seismic origins including some non-seismic picks in their association). (c) The origins derived from the pick associators are fed into an event associator algorithm that grouped all the origins for each event and designated a preferred one. To address the remaining misdetections, we designed a first SeisComP3 post-detection module that automatically impedes a misdetection to be a preferred origin. If most of the misdetections are discarded, many false detections (pick association caused by noise or glitches) continue to be processed and overwhelm the event alert system. To discriminate between false detections and true detections, we implemented two machine-learning classifiers within a second SeisComP3 post-detection module. This modules extracts the best features (d) then discriminates between false detections, earthquakes, and quarry blasts (e). False detections are automatically removed from the database.

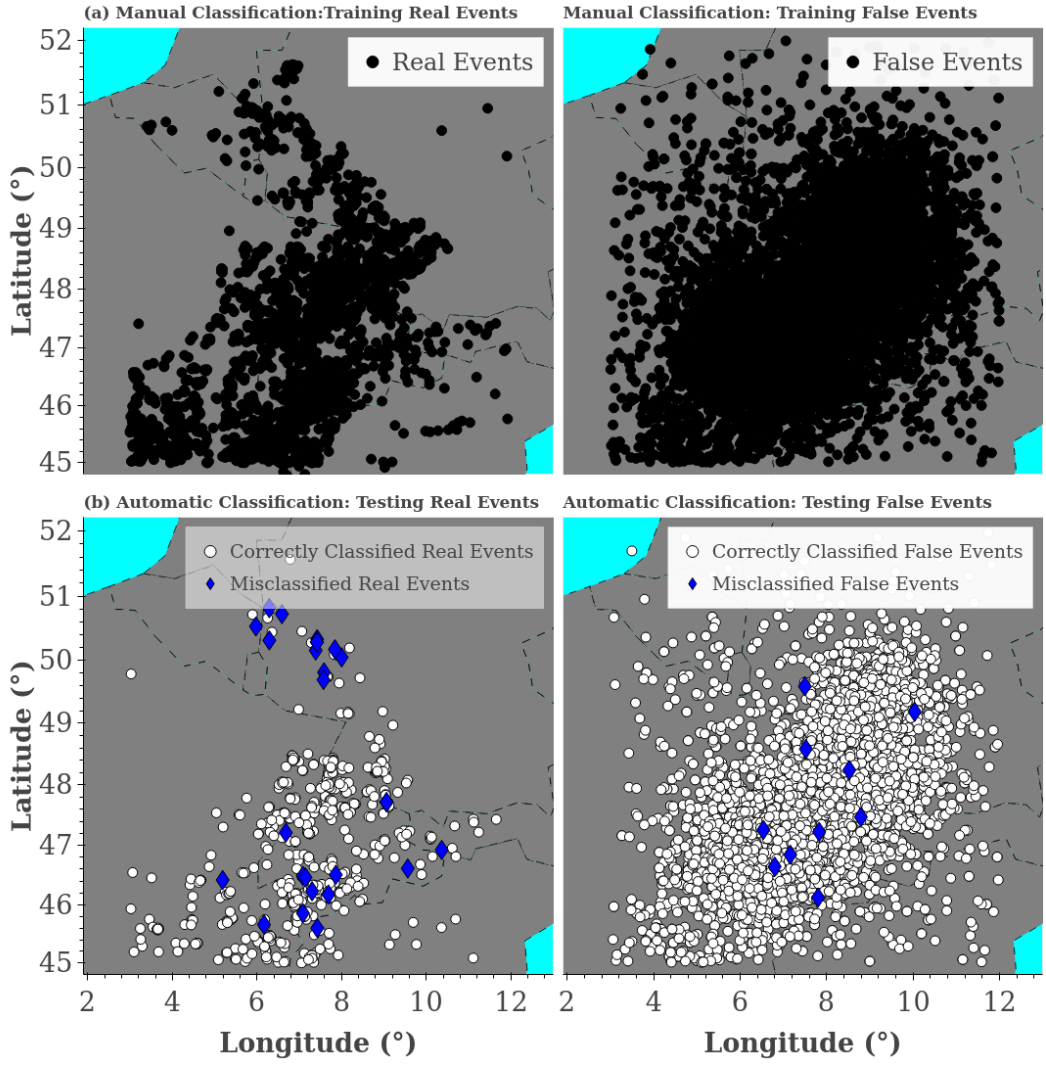


Figure S3. Locations of manually classified real and false events making up the training set (a) compared to the locations of automatically classified real and false events coming from the testing set (b). The events misclassified by classifier 1 are represented with blue diamonds whereas the correctly classified events appeared in white circles.

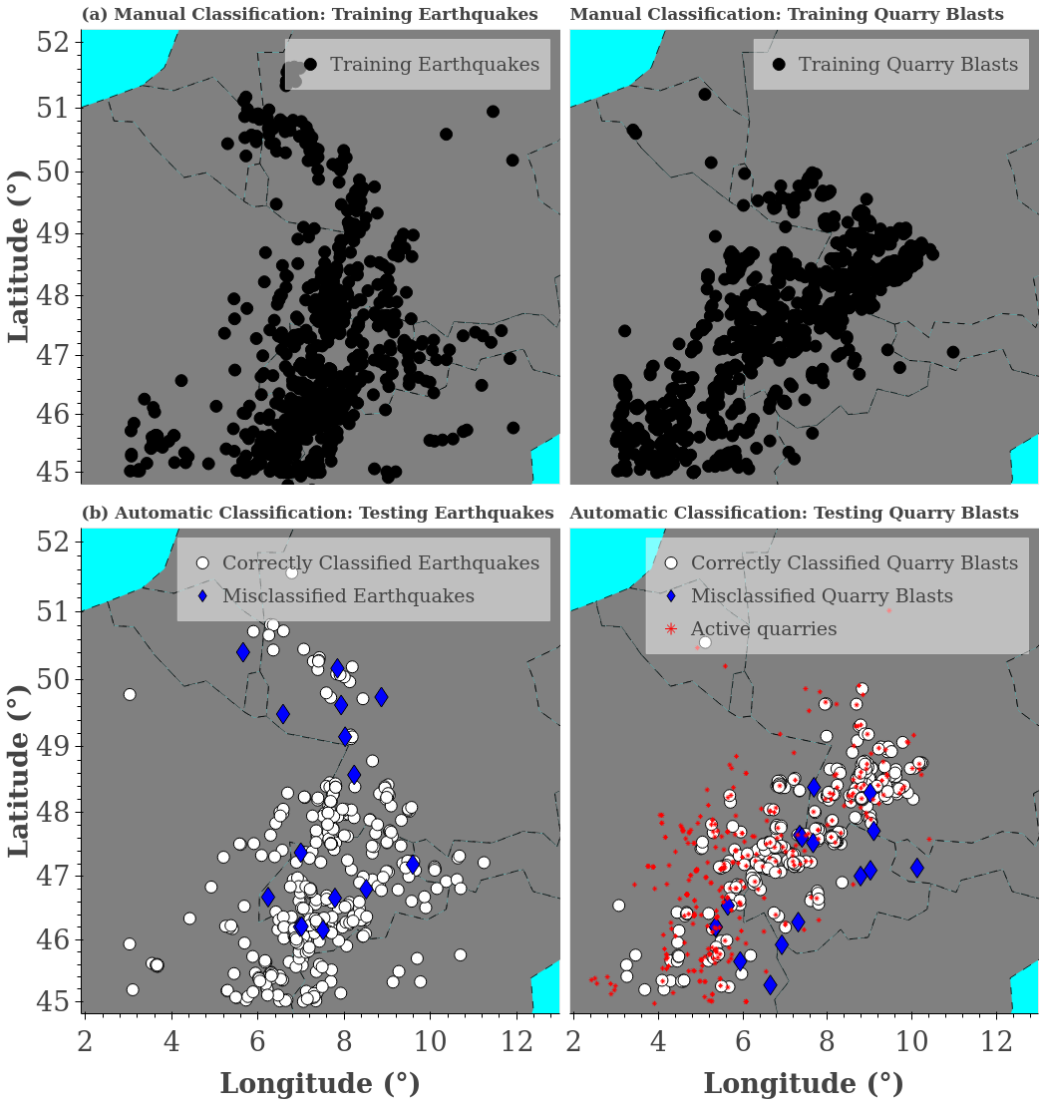


Figure S4. Locations of manually classified earthquakes and quarry blasts making up the training set (a) compared to the locations of automatically classified earthquakes and quarry blasts coming from the testing set (b). The events misclassified by classifier 2 are represented with blue diamonds whereas the correctly classified events appeared in white circles.

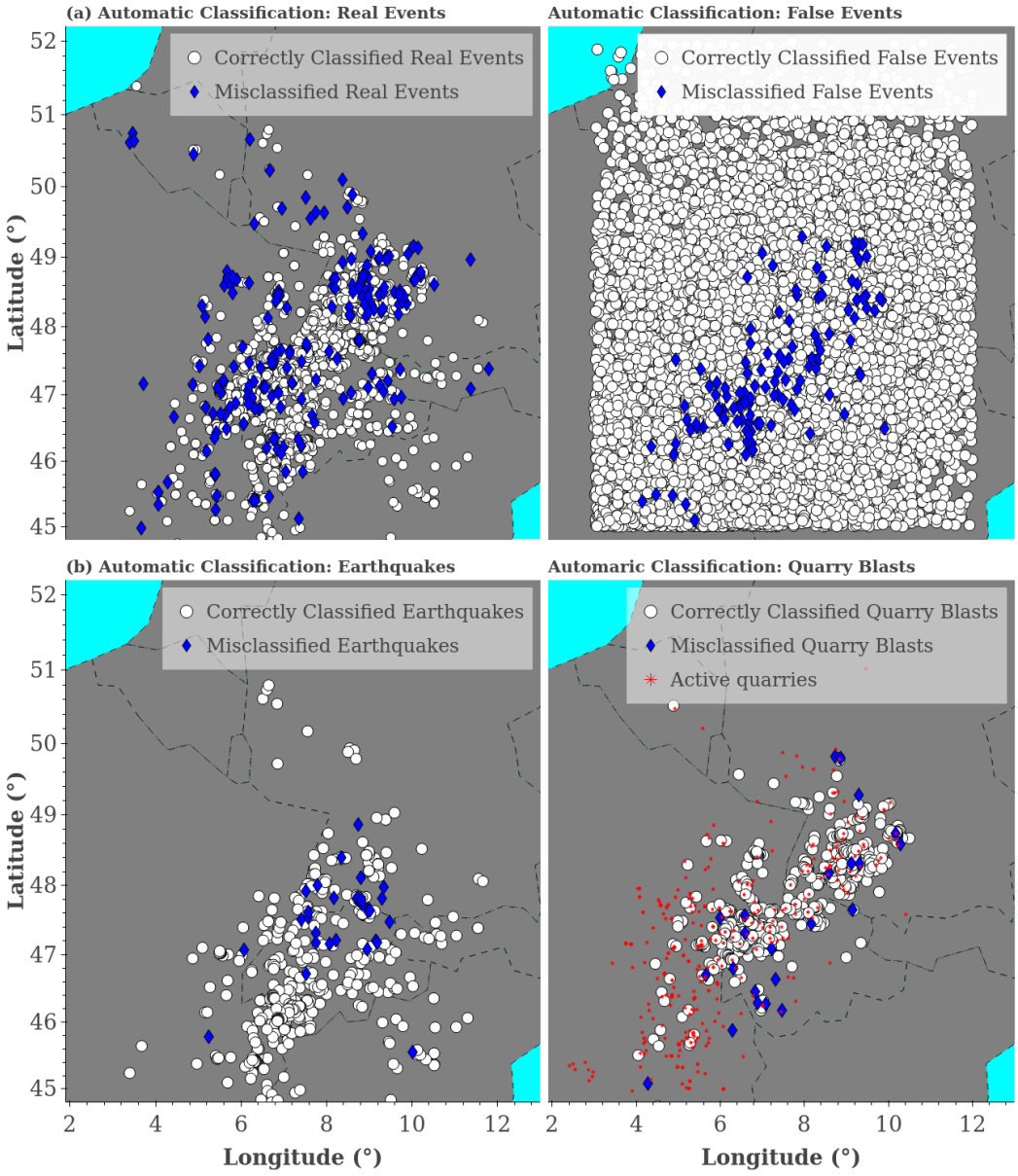


Figure S5. Locations of events predicted by classifier 1 (a) and classifier 2 (b) in operational mode.

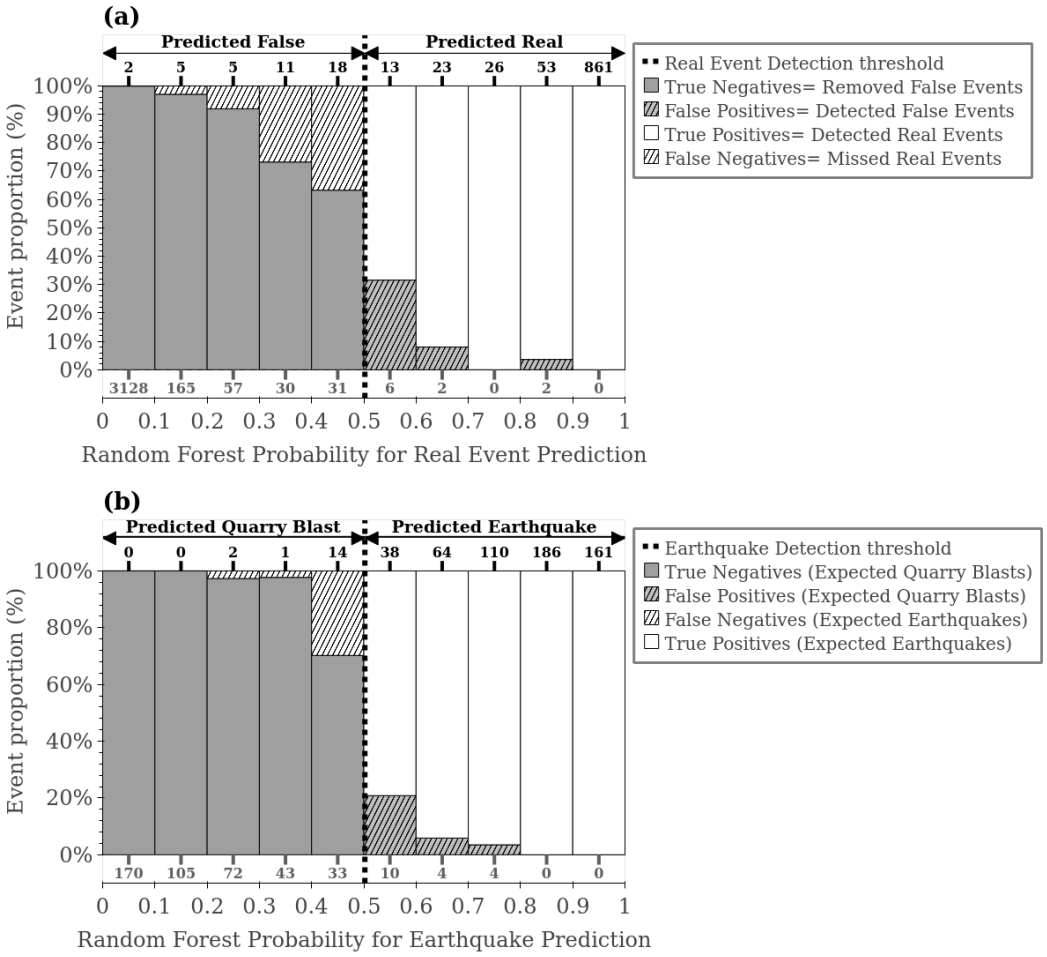


Figure S6. Distribution of prediction probabilities for (a) the trained false event vs real event and (b) the trained earthquake vs quarry blast classifiers in testing mode.

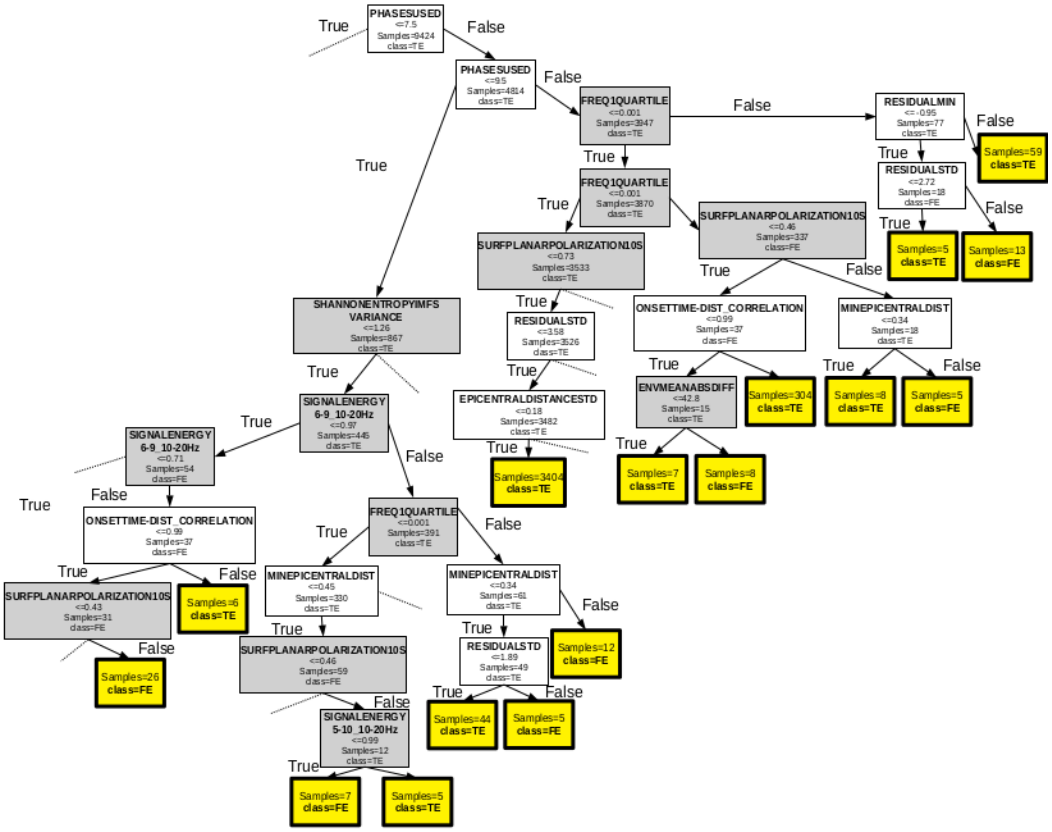


Figure S7. Simplified part of a decision tree randomly extracted from the classifier 1. The features linked to the event location and pick association quality, shaded in white, recursively alternate with the signal-related features, shaded in grey. Each partition in the tree (i.e. the tree node) is created from a feature value threshold. For instance, the first tree node corresponds to the question: "is the number of phases lower than 7.5?". The two answers (true, false) create two branches in the tree (a split). If the answer is true, the left branch of the tree is concerned; if false, it is the right branch. The procedure continues iteratively, until a decision tree hyper-parameter criteria is reached (i.e. maximum depth of the tree or minimum number of samples reached at a partition). The final partitions shaded in yellow (i.e. the leaves) make the final predictions (RE= real event, FE=false event).

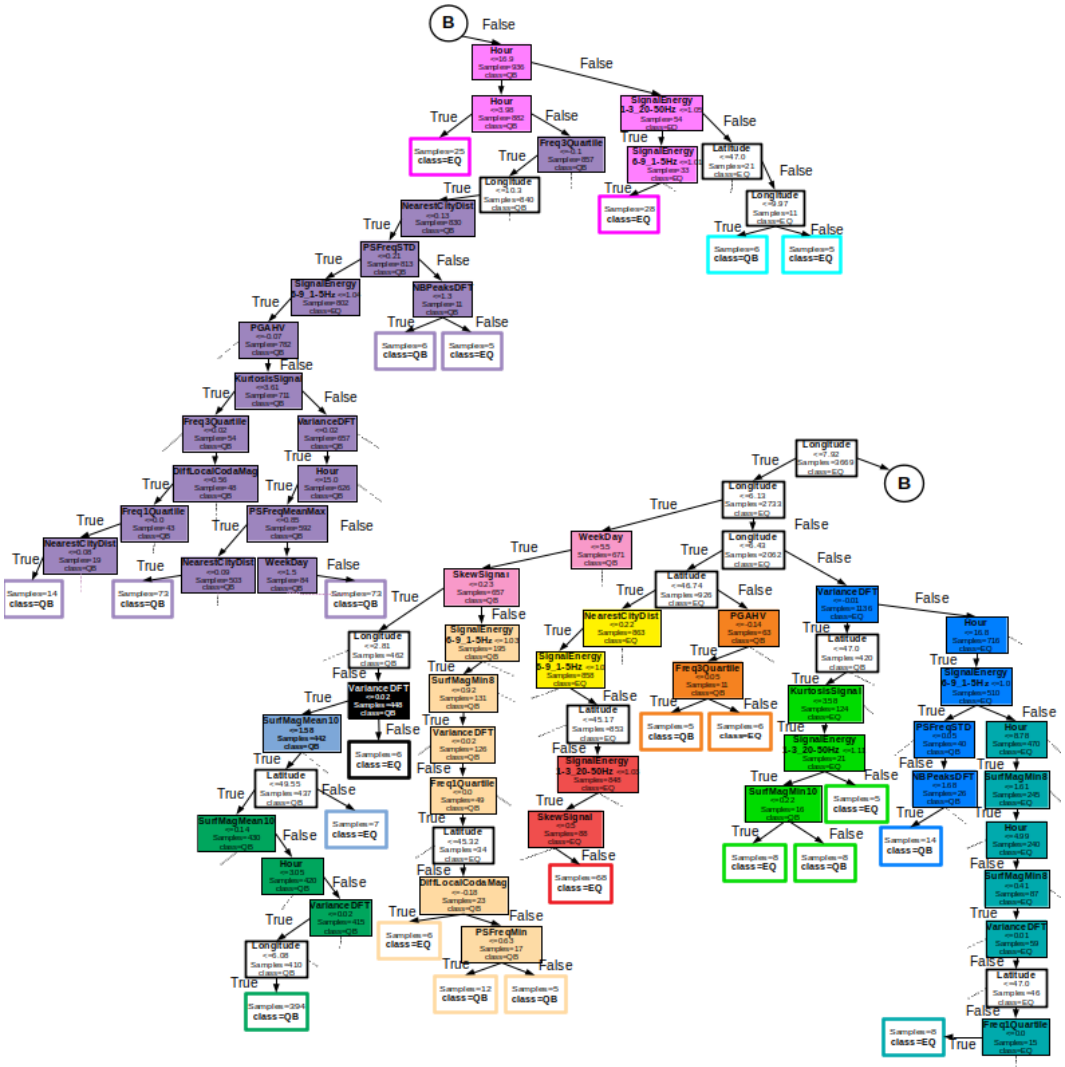
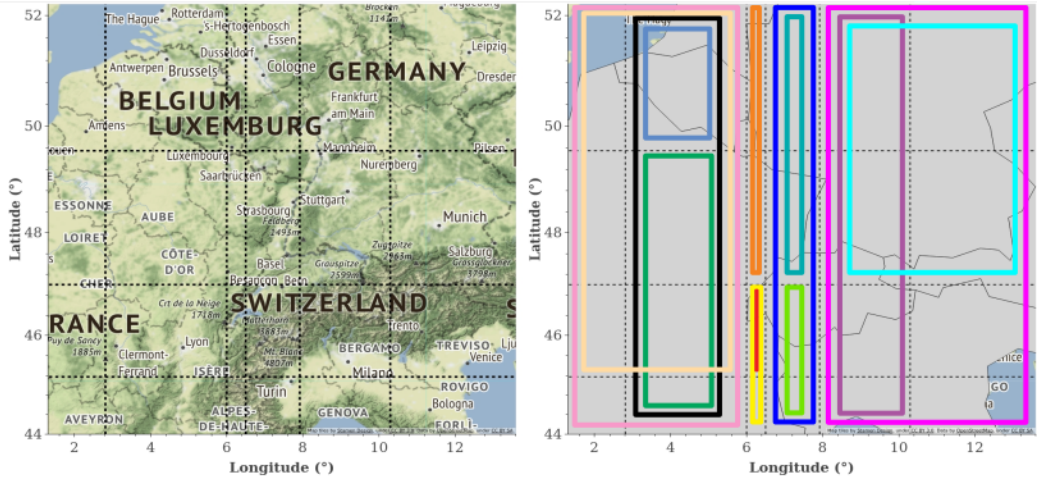


Figure S8. Simplified part of a decision tree (tree A + tree B) randomly extracted from the classifier 2. This decision tree extract was color-coded by geographical sub-region. Each geographical sub-region is delimited by the Latitude and Longitude threshold values (white rectangle outlined in black). Each geographical color group corresponds to a combination of specific signal-related features values that are used to discriminate the earthquake and quarry blast populations inside it (QB= quarry blast, EQ= earthquake.



A	WeekDay, SkewnessSignal	H	SignalEnergy1-3_20-50Hz, SkewnessSignal
B	VarianceDFT	I	Hour, VarianceDFT, NBPeaksDFT SignalEnergy6-9_1-5Hz PSFreqSTD
C	SurfMagMean10s, Hour	J	Freq1Quartile SurfMagMin8s
D	SurfMagMean10s	K	KurtosisSignal, SignalEnergy1-3_20-50Hz, SurfMagMin10s
E	Freq1Quartile SignalEnergy6-9_1-5Hz SurfMagMin8s DiffLocalCodaMag, PSFreqMeanMin	L	Hour, SignalEnergy1-3_20-50Hz SignalEnergy6-9_1-5Hz
F	PGAHV, Freq3Quartile	M	NearestCityDist, WeekDay, VarianceDFT, Freq3Quartile, Freq1Quartile, NBPeaksDFT, KurtosisSignal, PGAHV DiffLocalCodaMag, PSFreqMeanMax, PSFreqSTD
G	NearestCityDist, SignalEnergy6-9_1-5Hz	N	SignalEnergy1-3_20-50Hz

Figure S9. Map projection of each color-coded geographical sub-region represented in the Figure S10. Each sub-region is defined by an ensemble of features used to predict earthquakes and quarry blasts inside it. The dotted lines correspond to the Latitude and Longitude threshold values used in the aforementioned tree to delimit the sub-geographical regions.

5.3.3 Tableau des 361 attributs

Nbr	Short name	Description	Formula
		Event parameters	
Origin quality			
1	PhasesUsed	Number of phases used	-
2	StationsUsed	Number of stations used	-
3	StandardError	Standard error	$\sqrt{\frac{1}{N}(\sum res^2)}$ N number of residual values and res=residual value
4	MinEpicentralDist	Minimum epicentral distance	-
5	MaxEpicentralDist	Maximum epicentral distance	-
6	MeanEpicentralDist	Mean epicentral distance	-
7	MedianEpicentralDist	Median epicentral distance	-
8	EpicentralDistSTD	Epicentral Distance Standard Deviation	
9	AzimuthalGap	Azimuthal gap	-
10	ResidualMean	Residual mean	-
11	ResidualMedian	ResidualMedian	-
12	ResidualStandardDeviation	Residual standard deviation	
13	ResidualVariance	Residual variance	-
14	ResidualMin	Residual minimum	-
15	ResidualMax	Residual maximum	-
16	ClosestStationNumber	Absolute number of closest stations used to create the origin	-
17	ClosestStationProportion	Proportion of closest stations relative the total number of stations used	-

Nbr	Short name	Description	Formula
Picks			
18	SPicks	Number of S picks	-
19	SPickProportion	Proportion of S picks relative to the total picks	-
20	OnlyS	Number of isolated S picks	-
21	PSindice	Proportion of associated P-S picks	-
22	PSDiffDistCorr	Correlation Coefficient between P-S time difference and epicentral distance	-
23	PSDiffVariance	P-S time difference variance	-
24	OnsetTime-Dist-Corr	Correlation Coefficient between first onset time and epicentral distance	-
Origin uncertainty			
25	LongUncertainty	Longitude uncertainty	-
26	LatUncertainty	Latitude uncertainty	-
27	HorUncertainty	Circular confidence region given by single value of horizontal uncertainty	-
28	MinHorUncertainty	Semi-minor axis of confidence ellipse	-
29	MaxHorUncertainty	Semi-major axis of confidence ellipse	-
30	MaxAzUncertainty	Azimuth of major axis of confidence ellipse (positive to the East)	-

Nbr	Short name	Description	Formula
Origin position			
31	NearestQuarryDist	Distance to nearest quarry	-
32	NearestMineDist	Distance to nearest mining site	-
33	NearestCityDist	Distance to the nearest city	-
34	NearestGeothermalDist	Distance to the nearest geothermal power plant	-
35	NearestQuarryID	Name of nearest quarry	-
36	NearestMineID	Name of nearest mine	-
37	NearestCityID	Name of the nearest city	-
38	NearestGeothermalID	Name of nearest geothermal power plant	-
39	NearestQuarryAz	Azimuth of the nearest quarry	-
40	NearestMineAz	Azimuth of the nearest mine	-
41	NearestGeothermalAz	Azimuth of the nearest geothermal power plant	-
42	NearestCityAz	Azimuth of the nearest city	-
43	Longitude	Origin Longitude	-
44	Latitude	Origin Latitude	-
45	Depth	Origin Depth	-
46	CentroidDeviation	Origin deviation from the centroid of stations	-
Origin time			
47	Daytime	Time of the day	hour+minute/60+second/3600
48	Weekday	Day of the week	0 = Sunday, 1= Monday, 2= Tuesday, 3= Wednesday, 4= Thursday, 5= Friday, 6= Saturday

Nbr	Short name	Description	Formula
Origin magnitude			
49	MLvValue	Vertical-component local magnitude	$ML = \log(A) - \log(A0)$ A = Wood-Anderson amplitude A0= empirical calibration function
50	MLValue	Three-component local magnitude	
51	MaxAmplitudeMean	Maximum Amplitude Mean	
52	MaxAmplitudeVariance	Maximum Amplitude Variance	
53	AmplitudeDistCorr	Correlation coefficient maximum amplitude/epicentral distance	
54-63	SurfaceMagnitudeRMs8 -...- SurfaceMagnitudeRMs25	Rayleigh Surface Magnitude RMS 8-25s	$Ms = \log(a) + \frac{1}{2}\log(\sin(\Delta)) + 0.0031(\frac{20}{T})^{1.8} - 0.66\log(\frac{20}{T}) - \log(fc) - 0.43$ $8 \leqslant T \leqslant 25\,sec \quad , \quad fc \leq 0.6/T\sqrt{\Delta}$ a = amplitude of the Butterworth-filtered surface waves (zero-to-peak) Δ= epicentral distance T= period fc= filter frequency of a third-order Butterworth bandpass filter with corner frequencies 1/T-fc, 1/T+fc **(Bonner et al., 2006; Russel, 2006; Selby,2001)**
64-73	SurfaceMagnitudeMean8 -...- SurfaceMagnitudeMean25	Rayleigh Surface Magnitude mean 8-25s	
74-83	SurfaceMagnitudeMax8 -...- SurfaceMagnitudeMax25	Rayleigh Surface Magnitude maximum 8-25s	
84-93	SurfaceMagnitudeMin8 -...- SurfaceMagnitudeMin25	Rayleigh Surface Magnitude minimum 8-25s	
94-103	SurfaceMagnitudeVariance8- ...- SurfaceMagnitudeVariance25	Rayleigh Surface Magnitude variance 8-25s	
104	CodaAmplitude	Coda amplitude	$Md = -0.87 + 2.0\log(\tau) + 0.389\Delta$ τ=coda duration Δ= epicentral distance **(Koper et al., 2016; Holt et al., 2019)**
105	CodaMagnitude	Coda magnitude	
106	DiffLocalCodaMag	Difference between three-component local magnitude and coda magnitude	
107-116	SurfZHratioMean8 -...- SurfZHratioMean25	Ratio of vertical to horizontal amplitudes in Rayleigh waves mean	$\frac{Z(\omega)}{H(\omega)}$ Z(ω) = vertical amplitude H(ω) = horizontal amplitude **(Tanimoto and Rivera, 2008)**
117-126	SurfZHratioMax8 -...- SurfZHratioMax25	Ratio of vertical to horizontal amplitudes in Rayleigh waves maximum	

Nbr	Short name	Description	Formula
		Signal parameters	
Polarization analysis			
127-136	SurfaceStrike8 -...- SurfaceStrike25	Azimuth of the direction of maximum polarization of Rayleigh waves 8-25s	$\Phi = \arctan\left(\frac{\Re(yo)}{\Re(xo)}\right)$ $\Re(xo)$, $\Re(yo)$ = real parts of xo and yo coordinates of the eigenvector associated with the largest eigenvalue **(Vidale, 1986)**
137-146	SurfaceDip8 -...- SurfaceDip25	Dip of the direction of maximum polarization of Rayleigh waves 8-25s	$\delta = \arctan\left(\frac{\Re(zo)}{\sqrt{\Re(xo)^2+\Re(yo)^2}}\right)$ $\Re(xo)$, $\Re(yo)$, $\Re(zo)$ = real parts of the eigenvector (xo,yo,zo) associated with the largest eigenvalue **(Vidale, 1986)**
147-156	SurfEllipticalComponent 8 -...- SurfEllipticalComponent 25	Elliptical component of polarization of Rayleigh waves 8-25s	$PE = \frac{\sqrt{1-X^2}}{X}$ PE is the ratio of the imaginary part of the eigenvector to the real part of the eigenvector X = length of the real component of the eigenvector (x0,y0,Z0) **(Vidale, 1986)**
157-166	SurfPolarizationStrength 8 -...- SurfPolarizationStrength 25	Surface polarization strength 8-25s	$PS = 1 - \frac{\lambda 3 + \lambda 2}{\lambda 1}$ λ1 = largest eigenvalue λ2= intermediate eigenvalue λ3 = smallest eigenvalue **(Vidale, 1986)**
167-176	SurfPlanarPolarization8 -...- SurfPlanarPolarization25	Degree of planar polarization of Rayleigh waves 8-25s	$PP = 1 - \frac{\lambda 3}{\lambda 2}$ λ2= intermediate eigenvalue λ3 = smallest eigenvalue **(Vidale, 1986)**
177	SmallestEigenvalue	Smallest eigenvalue	-
178	IntermediateEigenvalue	Intermediate eigenvalue	-
179	LargestEigenvalue	Largest eigenvalue	-
180	Azimuth	Azimuth Direction of maximum polarization of signal Horizontal angular measure	$\Phi = \arctan\left(\frac{\Re(yo)}{\Re(xo)}\right)$ $\Re(xo)$, $\Re(yo)$ = real parts of xo and yo coordinates of the eigenvector associated with the largest eigenvalue **(Vidale, 1986)**

Nbr	Short name	Description	Formula
Polarization analysis			
181	Incidence	Incidence polarization of signal Vertical angular measure Direction of maximum	$\delta = \arctan\left(\frac{\Re(zo)}{\sqrt{\Re(xo)^2+\Re(yo)^2}}\right)$ $\Re(xo)$, $\Re(yo)$, $\Re(zo)$ = real parts of the eigenvector (xo,yo,zo) associated with the largest eigenvalue **(Vidale, 1986)**
182	Rectilinearity	Degree of Rectilinearity of signal	$1-\frac{\lambda 2+\lambda 3}{2\lambda 1}$ λ3= smallest eigenvalue λ2= intermediate eigenvalue λ1= largest eigenvalue **(Jurkevics, 1988)**
183	Planarity	Degree of Planarity of signal	$1-\frac{2\lambda 3}{\lambda 1+\lambda 2}$ λ3= smallest eigenvalue λ2= intermediate eigenvalue λ1= largest eigenvalue **(Jurkevics, 1988)**
P/S ratios			
184-190	PSRMS, PSMean, PSMedian PSFirstMax, PSMax PSMin PSStd	P over S maximum amplitude ratios : RMS, Mean, Median, First Maximum, Maximum, Minimum, Standard Deviation	$\frac{P_{max}}{S_{max}} = \frac{\sqrt{(P^Z_{max})^2+(P^R_{max})^2}}{\sqrt{(S^Z_{max})^2+(S^R_{max})^2+(S^T_{max})^2}}$ P_{max}^{Z} = P-wave maximum amplitude on vertical component (Z) P_{max}^{R} = P-wave maximum amplitude on radial component (R) S_{max}^{Z} = S-wave maximum amplitude on vertical component (Z) S_{max}^{Z} = S-wave maximum amplitude on radial component (R) S_{max}^{Z} = S-wave maximum amplitude on transverse component (T)
191-197	PSFreqRMS, PSFreqMean, PSFreqMedian, PSFreqFirstMax PSFreqMax PSFreqMin PSFreqStd	P over S maximum frequency ratios: RMS, Mean, Median, First Maximum, Maximum, Minimum, Standard Deviation	$\frac{P_{max}}{S_{max}} = \frac{\sqrt{(P^Z_{max})^2+(P^R_{max})^2}}{\sqrt{(S^Z_{max})^2+(S^R_{max})^2+(S^T_{max})^2}}$ P_{max}^{Z} = P-wave maximum spectral amplitude on vertical component (Z) P_{max}^{R} = P-wave maximum spectral amplitude on radial component (R) S_{max}^{Z} = S-wave maximum spectral amplitude on vertical component (Z) S_{max}^{Z} = S-wave maximum spectral amplitude on radial component (R) S_{max}^{Z} = S-wave maximum spectral amplitude on transverse component (T)

Nbr	Short name	Description	Formula
198-204	PSMeanRMS PSMeanMean PSMeanMedian PSMeanFirstMax PSMeanMax PSMeanMin PSMeanStd	RMS of P mean/max over S mean/max amplitude ratios: RMS, Mean, Median, First Maximum, Maximum, Minimum, Standard Deviation	$\frac{P_{MeanMax}}{S_{MMax}} = \frac{\sqrt{(P^Z_{MMax})^2+(P^R_{MMax})^2}}{\sqrt{(S^Z_{MMax})^2+(S^R_{MMax})^2+(S^T_{MMax})^2}}$ $P_{MMax}{}^Z$ = P-wave mean/maximum amplitude on vertical component (Z) $P_{MMax}{}^R$ = P-wave mean/maximum amplitude on radial component (R) $S_{MMax}{}^Z$ = S-wave mean/maximum amplitude on vertical component (Z) $S_{MMax}{}^Z$ = S-wave mean/maximum amplitude on radial component (R) $S_{MMax}{}^Z$ = S-wave mean/maximum amplitude on transverse component (T)
205-211	PSFreqMeanRMS PSFreqMeanMean PSFreqMeanMedian PSFreqMeanFirstMax PSFreqMeanMax PSFreqMeanMin PSFreqMeanStd	RMS of P mean/max over S mean/max frequency ratios: RMS, Mean, Median, First Maximum, Maximum, Minimum, Standard Deviation	$\frac{P_{MeanMax}}{S_{MMax}} = \frac{\sqrt{(P^Z_{MMax})^2+(P^R_{MMax})^2}}{\sqrt{(S^Z_{MMax})^2+(S^R_{MMax})^2+(S^T_{MMax})^2}}$ $P_{MMax}{}^Z$ = P-wave mean/maximum spectral amplitude on vertical component (Z) $P_{MMax}{}^R$ = P-wave mean/maximum spectral amplitude on radial component (R) $S_{MMax}{}^Z$ = S-wave mean/maximum spectral amplitude on vertical component (Z) $S_{MMax}{}^Z$ = S-wave mean/maximum spectral amplitude on radial component (R) $S_{MMax}{}^Z$ = S-wave mean/maximum spectral amplitude on transverse component (T)
Envelope			
212	MaxEnv	Maximum of the envelope	Envelope = $\sqrt{x(t)^2+H[x(t)]^2}$ x(t) = signal H[x(t)] = Hilbert-Transformed signal
213	MeanEnv	Mean of the envelope	-
214	MedianEnv	Median of the envelope	-
215	STDEnv	Standard deviation of the envelope	-
216	KurtosisEnv	Kurtosis of the envelope	Kurtosis= $\frac{1}{n}\sum_i \left(\frac{E(i)-\mu_E}{\sigma_E}\right)^4$ E(i) = envelope values σ_E= envelope standard deviation μ_E= envelope mean, n = envelope length
217	SkewnessEnv	Skewness of the envelope	Skewness = $\frac{1}{n}\sum_i \left(\frac{E(i)-\mu_E}{\sigma_E}\right)^3$ E(i) = envelope values σ_E= envelope standard deviation μ_E= envelope mean, n = envelope length

Nbr	Short name	Description	Formula
Envelope			
218	MeanMaxEnv	Ratio between the mean of the envelope and the maximum of the envelope	-
219	MedianMaxEnv	Ratio between the median of the envelope and the maximum of the envelope	-
220	EnvSum	Sum of the envelope's amplitude values	-
221	EnvSecondDerivative	Mean value of a central approximation of the second derivative of the envelope	$\frac{1}{2n}\sum_{i=1,..,n-1}(\frac{1}{2}E_{i+2}-E_{i+1}+E_i)$ E = envelope values n = envelope length
222	EnvComplexity	Complexity of the envelope (One order discrete difference mean)	$\sqrt{\sum_{n-2lag}^{i=1}(E_i-E_{i+1})^2}$ E = envelope values n = envelope length **(Batista et al., 2014)**
223	EnvMeanAbsDiff	Mean over the absolute differences between subsequent envelope values	$\frac{1}{n}\sum_{i=1,...,n-1}\lvert E_{i+1}-E_i\rvert$ E = envelope values n = envelope length
224	EnvMeanDiff	Mean over the differences between subsequent envelope values	$\frac{1}{n}\sum_{i=1,...,n-1}E_{i+1}-E_i=\frac{1}{n-1}E_n-E_1$ E = envelope values n = envelope length
225	EnvUniqueVal	Percentage of unique values, that are present in the envelope more than once	-
226	EnvAbsSumChange	Sum over the absolute value of consecutive changes in the envelope	$\sum_{i=1,...,n-1}\lvert E_{i+1}-E_i\rvert$ E = envelope values n = envelope length
227	EnvBelowMean	Number of Values in the envelope that are lower than the mean	-

Nbr	Short name	Description	Formula
Envelope			
228	EnvAboveMean	Number of Values in the envelope that are higher than the mean	-
229	EnvDuplicateMax	Number of duplicate maximum values	-
230	EnvDiffPickMax	Difference between maximum envelope value and value at P arrival	-
231	EnvAsDecTime	Ratio between ascending and descending time	$\frac{t_{max}-t_i}{t_f-t_{max}}$ t_i= time of the signal beginning t_f= time of the signal end t_{max}= time of the largest amplitude
232	EnvCorNbPeaks	Number of peaks in the autocorrelation function	-
233	EnergyCor1	Energy in the first third part of the autocorrelation function	$\int_0^{\frac{T}{3}} C(\tau)d(\tau)$ T = signal duration C = autocorrelation function
234	EnergyCor2	Energy in the remaining part of the autocorrelation function	$\int_{\frac{T}{3}}^{T} C(\tau)d(\tau)$ T = signal duration C = autocorrelation function
Waveform			
235	SkewnessSig	Skewness of the signal	Skewness= $\frac{1}{n}\sum_i \left(\frac{x(i)-\mu_x}{\sigma_x}\right)^3$ x(i) = signal values σ_x= signal standard deviation μ_x= signal mean n = signal length
236	KurtosisSig	Kurtosis of the signal	Kurtosis= $\frac{1}{n}\sum_i \left(\frac{x(i)-\mu_x}{\sigma_x}\right)^4$ x(i) = signal values σ_x= signal standard deviation μ_x= signal mean n = signal length

Nbr	Short name	Description	Formula
Waveform			
237-243	SignalEnergy_1-3Hz SignalEnergy_3-6Hz SignalEnergy_6-9Hz SignalEnergy_1-5Hz SignalEnergy_5-10HZ SignalEnergy_10-20Hz SignalEnergy_20-50Hz	Signal energy filtered in the frequency range [f1-f2]: 1-3 Hz, 3-6 Hz, 6-10 Hz, 1-5 Hz, 5-10 Hz, 10-20 Hz, 20 – 50 Hz	$\int_0^T x(\tau)d(\tau)$ x = filtered signal in the frequency range [f1-f2]
244-250	KurtoSig_1-3Hz, KurtoSig_3-6Hz, KurtoSig_6-9Hz, KurtoSig_1-5Hz, KurtoSig_5-10Hz, KurtoSig_10-20Hz, KurtoSig_20-50Hz	Signal Kurtosis 1-3 Hz, 3-6 Hz, 6-9 Hz, 1-5 Hz, 5-10 Hz, 10-20 Hz, 20-50 Hz	Kurtosis= $\frac{1}{n}\sum_i \left(\frac{x(i)-\mu_x}{\sigma_x}\right)^4$ x(i) = signal values σ_x= signal standard deviation μ_x= signal mean n = signal length
251-270	SignalEnergyRatio_1-3_3-6Hz-...-SignalEnergyRatio_10-20_20-50Hz	Signal Energy Ratio in the frequency ranges [f1-f2] and [f3-f4]: 1-3Hz/3-6Hz, …, 10-20Hz/20-50Hz	$\int_0^T x(\tau)d(\tau) \;/\; \int_0^T y(\tau)d(\tau)$ x = filtered signal in the frequency range [f1-f2] y = filtered signal in the frequency range [f3-f4]
271	SignalCCMax	Inter-station waveform similarity: maximum correlation coefficient	
272	SignalCCMean	Inter-station waveform similarity: mean correlation coefficient	-
STALTA function			
273	STALTAmax	Maximum STA/LTA ratio	$STA=\frac{1}{N_s}\sum_{j=i-N_s}^{i} CF_j$ $LTA=\frac{1}{N_l}\sum_{j=i-N_l}^{i} CF_j$ N_s= number of samples used by each STA window N_l= number of samples used by each LTA window CF_j= values of the samples
274	STALTAmin	Minimum STA/LTA ratio	-
275	STALTATriggerP	STA/LTA value at P arrival	-

Nbr	Short name	Description	Formula
STALTA function			
276	STALTATriggerS	STA/LTA value at S arrival	-
277	STALTAsum	Sum of STA/LTA values	-
278	STALTAMeanAbsDiff	Mean over the differences between subsequent STA/LTA values	$\frac{1}{n}\sum_{i=1,\dots,n-1}\lvert x_{i+1}-x_i\rvert$ x = STA/LTA values n = STA/LTA function length
279	STALTAMeanDiff	Mean over the differences between subsequent STA/LTA values	$\frac{1}{n}\sum_{i=1,\dots,n-1} x_{i+1}-x_i=\frac{1}{n-1}x_n-x_1$ x = STA/LTA values n = STA/LTA function length
280	STALTAComplexity	Complexity of the STA/LTA function	$\sqrt{\sum_{n-2lag}^{i=1}(x_i-x_{i+1})^2}$ x = STA/LTA values n = STA/LTA function length **(Batista, 2014)**
281	STALTAAbsSumChange	Sum over the absolute value of consecutive changes in the STA/LTA function	$\sum_{i=1,\dots,n-1}\lvert x_{i+1}-x_i\rvert$ x = STA/LTA values n = STA/LTA function length
282	STALTAMeanSeconDerivative	Mean value of a central approximation of the second derivative of the STA/LTA function	$\frac{1}{2n}\sum_{i=1,\dots,n-1}(\frac{1}{2}x_{i+2}-x_{i+1}+x_i)$ x = STA/LTA values n = STA/LTA function length
283	STALTAUniqueVal	Percentage of unique values, that are present in the STA/LTA function more than once	-
284	STALTABelowMean	Number of Values in the STA/LTA function that are lower than the mean	-
285	STALTAAboveMean	Number of Values in the STA/LTA function that are higher than the mean	-
286	STALTADuplicateMax	Number of duplicate maximum values	-

Nbr	Short name	Description	Formula
STALTA function			
287	STALTADiffPickMax	Difference between maximum value and value at P arrival	-
288	STALTAAbsEnergy	STA/LTA Absolute energy	$\sum_{i=1,\dots,n} x_i^2$ x = STA/ LTA function
Spectrum			
289	RMSDFT	RMS of the Discrete Fourier Transform (DFT)	Spectrum= $S(f)=x(t)+2\sum_{k=1}^{n-1} x(k)\cos(k\,\omega)$
290	InstFreq	Instantaneous frequency	-
291	MeanDFT	Mean of the Discrete Fourier Transform	-
292	MaxDFT	Maximum of the Discrete Fourier transform	-
293	MedianDFT	Median of the Discrete Fourier transform	-
294	VarianceDFT	Variance of the Discrete Fourier transform	-
295	SpecCentroid	Spectral centroid	$\frac{\sum_{1}^{N} f_i m^i}{\sum_{1}^{M} m_i}$ m_i= magnitude of bin number, f_i= central frequency at that bin, M= number of bins **(Tzanetakis et al., 2001)**
296	Freq1Quartile	Central frequency of the 1st quartile	-
297	Freq3Quartile	Central frequency of the 2nd quartile	-
298	NbPeaksDFT	Number of peaks in the DFT	-
299	MeanPeaksDFT	Mean value for the peaks	-
300-303	EnergyDFT1, EnergyDFT2, EnergyDFT3, EnergyDFT4	Spectral Energy in 0-12.5Hz, 12.5-25Hz, 25-37.5Hz, 37.5-50Hz	$\int_{f1}^{f2} \lvert S(f)^2 \rvert df$ S(f) = spectrum f1,f2= frequency range

Nbr	Short name	Description	Formula
Spectrum			
304-309	EnergyDFT1_DFT2,-...-, EnergyDFT3_DFT4	Spectral Energy ratio of [f1,f2] over [f3,f4] frequency ranges : 0-12.5Hz, 12.5-25Hz, 25-37.5Hz, 37.5-50Hz	$\int_{f1}^{f2} \lvert S(f)^2 \rvert df \;/\; \int_{f3}^{f4} \lvert S(f)^2 \rvert df$ S(f) = spectrum f1,f2= first frequency range f3,f4= second frequency range
Spectrogram			
310	MaxMeanSpec	Maximum/mean ratio of all DFTs	$mean(\sum_{t=0}^{T} \frac{max(Spec(t,f))}{mean(Spec(t,f))})$ Spec(t,f)=spectrogram
311	MaxMedianSpec	Maximum/median ratio of all DFTs	$mean(\sum_{t=0}^{T} \frac{max(Spec(t,f))}{median(Spec(t,f))})$ Spec(t,f)=spectrogram
312	KurtoMaxSpec	Kurtosis of the maximum of all DFTs	$mean(\sum_{t=0}^{T} Kurtosis(max(Spec(t,f))))$ Spec(t,f)=spectrogram
313	KurtoMedianSpec	Kurtosis of the median of all DFTs	$mean(\sum_{t=0}^{T} Kurtosis(median(Spec(t,f))))$ Spec(t,f)=spectrogram
314	NbPeaksMaxSpec	Number of peaks in the curve showing the temporal evolution of the DFTs maximum	- **(Provost et al., 2016)**
315	NbPeaksMeanSpec	Number of peaks in the curve showing the temporal evolution of the DFTs mean	- **(Provost et al., 2016)**
316	NbPeaksMedianSpec	Number of peaks in the curve showing the temporal evolution of the DFTs median	- **(Provost et al., 2016)**
317	NbPeaksCentralFreq	Number of peaks in the curve showing the temporal evolution of the DFTs central frequency	- **(Provost et al., 2016)**

Nbr	Short name	Description	Formula
Spectrogram			
318	DistMaxMeanFFTs	Mean distance between the curves of the temporal evolution of the DFTs maximum frequency and mean frequency	- **(Provost et al., 2016)**
319	DistMaxMedianFFTs	Mean distance between the curves of the temporal evolution of the DFTs maximum frequency and median frequency	- **(Provost et al., 2016)**
320	Dist1QMedianFFTs	Mean distance between the 1st quartile and the median of all DFTs as a function of time	- **(Provost et al., 2016)**
321	Dist3QMedianFFTs	Mean distance between the 3rd quartile and the median of all DFTs as a function of time	- **(Provost et al., 2016)**
322	Dist1Q3Q	Mean distance between the 3rd quartile and the 1st quartile of all DFTs as a function of time	- **(Provost et al., 2016)**
Time-frequency analysis: signal Empirical Mode Decomposition (Intrinsic Mode Functions IMFs)			
323-326	SkewnessIMFsMean SkewnessIMFsMedian SkewnessIMFsMin SkewnessIMFsMax	Skewness of all IMFS: mean, median, minimum, maximum	$x_{DFT}(t)=\sum_{i=1}^{N} IMF_i(t)+R_N(t)$ $x_{DFT}(t)$ = original signal N = number of extracted IMFs $IMF_i(t)$ = ith IMF $R_N(t)$ = final residual **(Huang et al., 1998)**
327-330	KurtosisIMFsMean KurtosisIMFsMedian KurtosisIMFsMin KurtosisIMFsMax	Kurtosis of all IMFS: mean, median, minimum, maximum	-

Nbr	Short name	Description	Formula
Time-frequency analysis: signal Empirical Mode Decomposition (Intrinsic Mode Functions IMFs)			
331-334	VarianceIMFsMean VarianceIMFsMedian VairanceIMFsMax VarianceIMFsMin	Variance of all IMFs: mean, median, minimum, maximum	-
335-338	InstFreqIMFsMean InstFreqIMFsMedian InstFreqIMFsMin InstFreqIMFsMax	Instantaneous frequency of all IMFS: mean, median, minimum, maximum	-
339-343	AmplitudesIMFsMean AmplitudesIMFsMedian AmplitudesIMFsMin AmplitudesIMFsMax AmplitudesIMFsVariance	Amplitudes of all IMFS: mean, median, minimum, maximum, variance	-
344-347	SpecCentIMFsMean SpecCentIMFsMedian SpecCentIMFsMax SpecCentIMFsMin	Spectral Centroid of all IMFs: mean, median, minimum, maximum	-
348-352	EntropyIMFsMean EntropyIMFsMedian EntropyIMFsMax EntropyIMFsMin EntropyIMFsVariance	Shannon Entropy of all IMFS: mean, median, minimum, maximum, variance	Average information contained in the probability distribution function $-\sum_{i=1}^{N} p(IMF_i)\log_2(p(IMF_i))$ $p(s_j)$ = probability of amplitude level s_j
353-357	DFAIMFsMean DFAIMFsMedian DFAIMFsMax DFAIMFsMin DFAIMFsVariance	Detrended fluctuation analysis of all IMFS: mean, median, minimum, maximum, variance	$\sqrt{\frac{1}{N}\sum_{k=1}^{N}[y(k)-y_n(k)]^2}$ with y(k) the IMF expressed as : $\sum_{i=1}^{k}[IMF(i)-\mu_{IMF}]$ μ_x=mean of IMF values, x(i) = ith IMF N = length of IMF
Ground Motions			
358	PGAHV	Peak Ground Acceleration Horizontal-to-Vertical amplitude ratio	$\frac{\sqrt{(PGA_Z)^2}}{\sqrt{(PGA_R)^2+(PGA_T)^2}}$ PGA_Z= PGA vertical component PGA_R, PGA_T= PGA horizontal components
359	PGVHV	Peak Ground Velocity Horizontal-to-Vertical amplitude ratio	$\frac{\sqrt{(PGV_Z)^2}}{\sqrt{(PGV_R)^2+(PGV_T)^2}}$ PGV_Z= PGV vertical component PGV_R, PGV_T= PGA horizontal components

Nbr	Short name	Description	Formula
360	PGA	Peak Ground Acceleration mean	-
361	PGV	Peak Ground Velocity	-

5.4 Récapitulatif

L'introduction de l'apprentissage machine supervisé pour classer les événements qui sont finalement détectés demande de délimiter les contraintes afférentes au problème de classification avec un jeu de données de petite taille mais complexe.

Outre le choix de l'algorithme d'apprentissage ainsi que la sélection de ses hyperparamètres optimaux, l'interactivité Homme-machine est une des plus grandes réponses aux contraintes du jeu de données. Du contrôle de la sélection des attributs à la validation des règles de classification, l'injection des connaissances préalables dans le système d'apprentissage offre un cadre structurel à l'espace des hypothèses possibles, augmentant les chances de capturer dans cette espace la fonction de prédiction recherchée.

La fonction de prédiction qui a été sélectionnée dans ce travail de thèse pour prédire les vrais événements et les faux événements a été sélectionnée à partir d'une combinaison finale de 13 attributs. Ces attributs retracent indirectement les critères qui vont définir ce qu'est un vrai événement dans le système de détection.

C'est un événement localisé avec précision (distribution statistique des résidus, nombre de phases utilisées, distance épicentrale minimale, écart-type à partir de la distance épicentrale moyenne) à partir d'une association cohérente de pointés (facteur de corrélation entre les premières arrivées des ondes P et la distance épicentrale) qui ont été déclenchés par variation d'amplitude (valeur maximale de la fonction STA/LTA) à partir d'un signal cohérent (estimation de l'entropie de Shannon) et non-stationnaire (différence discrète d'ordre 1 de l'enveloppe du signal) qui se détache du bruit de fond ambiant (énergie du signal dans les gammes fréquentielles 6-9 Hz et 10-20 Hz, fréquence cumulée à 25%), et dont la source apparaît moins superficielle que celle des faux événements (degré de polarisation planaire).

La fonction de prédiction qui a été sélectionnée dans ce travail de thèse pour prédire les séismes et les tirs de carrière a été sélectionnée à partir d'une combinaison finale de 22 attributs.

Les séismes et les tirs de carrière sont mieux décrits à travers les attributs qui décrivent le signal dans le domaine fréquentiel (variance des valeurs du spectre du signal, nombre de pics contenus dans le spectre, rapports spectraux entre les ondes P et S, fréquence cumulée de 25%, fréquence cumulée de 75%), puis dans le domaine temporel (coefficient d'asymétrie et d'aplatissement de la distribution des valeurs d'amplitude du signal, rapport de l'énergie du signal à différentes bandes fréquentielles). Ces arguments temporels et fréquentiels retracent la nature des ondes sismiques qui composent les signaux associés aux différents événements. Les signaux sismiques associés aux tirs de carrière présentent par exemple une intensité maximale aux faibles fréquences (1-5 Hz), principalement due aux ondes de surface.

De plus, des attributs supplémentaires apportent des informations plus ou moins indirectes sur les paramètres de la source : sa profondeur, systématiquement superficielle dans le cas des tirs de carrière (magnitudes de surface, différence magnitude de coda et magnitude locale, Z/H ratio), sa localisation épicentrale, invariablement proche d'une carrière pour les tirs (proximité de l'événement à un centre urbain, donc potentiellement d'une carrière, et son temps d'origine, inéluctablement pendant les heures ouvrées pour les tirs de carrière (heure et date de l'événement).

Ces deux classifieurs générés ont été implémentés dans un module SeisComP3 que j'ai développé. Ce module intègre les outils de l'apprentissage machine supervisé, à savoir l'algorithme d'apprentissage de Random Forest. Afin de compléter le nouveau système de détection, ce module :

- calcule les 35 attributs (13 attributs pour la discrimination des vrais et faux événements et 22 attributs pour la discrimination des séismes et des tirs de carrière, Figure 5.36d) ;
- utilise le classifieur final des vrais et faux événement généré par l'algorithme d'apprentissage de Random Forest pour prédire chaque événement détecté entrant (Figure 5.36e) ;
- supprime l'ensemble des faux événements prédits par le classifieur (label=0) de la base de données des événements (Figure 5.36e) ;
- utilise le classifieur final des séismes et des tirs de carrière généré par l'algorithme d'apprentissage de Random Forest pour prédire l'ensemble des événements qui sont prédits comme vrais événements (Figure 5.36e) ;
- labélise automatiquement les séismes (label=3) et les tirs de carrière (label=5) prédits par le classifieur précédent (Figure 5.36e).

Afin de traiter plus rapidement l'ensemble des données disponibles (notamment en cas de retraitement des données pour inclure les stations AlpArray temporaires dans la détection), le système de détection complet a été isolé dans un conteneur SINGULARITY de telle façon à pourvoir exécuter plusieurs instances de ce système de détection sur les super-ordinateurs.

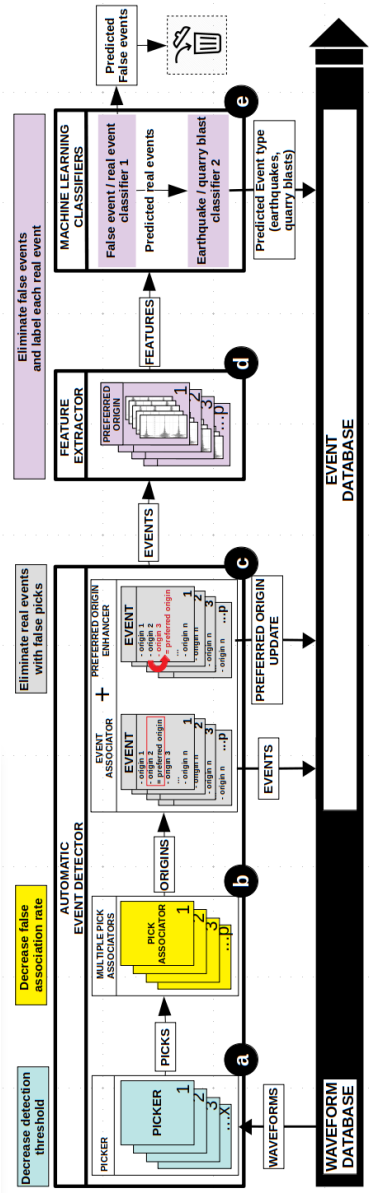


FIGURE 5.36: Procédure de détection nouvellement développée, qui vise à réduire le taux de séismes détectés avec de faux pointés et le taux de faux événements détectés, puis de discriminer les vrais événements entre eux en séismes et tirs de carrière. (d), (e) Un autre module SeisComP3 que j'ai développé est finalement implémenté pour calculer les attributs optimaux de chaque événement entrant, classer les événements en faux et vrais événements avec un premier classifieur, supprimer les faux événements prédits et classer le reste des vrais événements en tirs de carrière et séismes dont le label est automatiquement ajouté dans la base de donnée. (cf Figure 4.63 pour a,b,c).

Chapitre 6

Conclusion

> « *In summary, it has been a long journey, but that journey is not yet complete.* »—C. E. Johnson, 2020

Sommaire

6.1 La détection et la discrimination des séismes de faible magnitude, deux problèmes réciproquement liés

6.1.1 Des facteurs communs à la résolution des deux problèmes

Le dénominateur commun aux problèmes de détection et discrimination est de trouver une solution approchante des différents paramètres qui vont caractériser une source sismique inconnue, à savoir sa taille, sa nature, sa localisation et son temps d'origine. La solution approchante du problème de la détection est une localisation de cette source et une estimation de sa taille. En effet, la localisation est définie par l'hypocentre (longitude x, latitude y, profondeur z), qui correspond à la localisation physique de l'initiation du processus de rupture (**Havskov2011**), et le temps d'origine (t) qui correspond à l'heure du début de la rupture. La taille de la source est quant à elle indirectement définie par la mesure logarithmique de la magnitude (ici magnitude locale MLv). La solution approchante du problème de discrimination est une caractérisation du type de la source sismique (séisme d'origine naturel ou induit, tir de carrière, bruit d'origine anthropique).

Le point de départ de l'expression du problème de la détection est donc l'existence d'une source inconnue que l'on souhaite caractériser. Le signal, enregistré aux stations, est la seule information indirecte disponible pour résoudre le problème de détection. Ce signal est le résultat d'une combinaison des effets de la source, des effets liés au milieu de propagation des ondes sismiques émises et les effets liés au bruit enregistré aux stations. C'est donc à partir de ce signal que la localisation et la taille de la source vont être inférées.

En revanche, ce n'est pas l'existence de cette source inconnue qui va motiver l'expression du problème de la discrimination, mais sa solution hypocentrale, apportée par la résolution du problème de la détection. Par conséquent, la résolution de ces deux types de problème (détection et discrimination) se fait de manière inverse. Dans le cas de la détection, le problème s'initie à une source sismique inconnue et se résout avec la détection de l'événement qui en découle, alors que, dans le cas de la discrimination, le problème s'initie à l'événement détecté, qui est de type inconnu, pour remonter à la caractérisation de la source qui l'a créée. La discrimination est en quelque sorte la réciproque du problème de la détection (Figure 6.1).

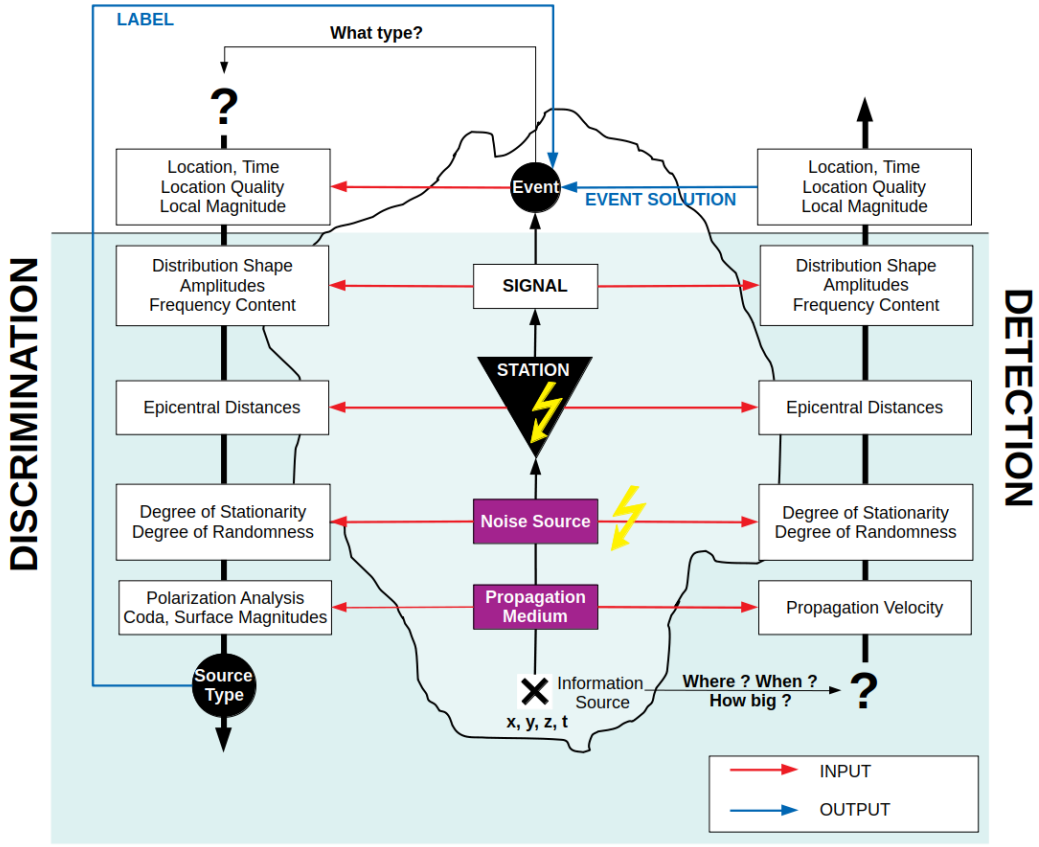


FIGURE 6.1: La détection et la discrimination, deux problèmes réciproquement liés. Chaque événement détecté dans le catalogue est simplement défini par un hypocentre, un temps d'origine, une magnitude et un label (séisme, tir de carrière). Seulement, ce dernier représente la solution finale à deux problèmes beaucoup plus complexes, celui de la détection et celui de la discrimination. Le succès de la résolution de ces deux problèmes dépend de la prise en compte de plusieurs facteurs communs : les caractéristiques globales du signal enregistré, la configuration du réseau de stations qui l'enregistre, les propriétés spécifiques du bruit enregistré aux stations ainsi que le milieu de propagation. D'où la métaphore de l'iceberg. Si le problème de détection s'initie à une source sismique inconnue et se résout avec la détection de l'événement qui en découle, le problème de la discrimination s'initie à l'événement détecté qui est de type inconnu pour remonter à la caractérisation de la source qui l'a créée.

Les problèmes de la détection et de la discrimination sont ainsi réciproquement liés par des facteurs communs qui définissent le cadre de leur résolution. Ces facteurs communs sont les propriétés des signaux sismiques détectés, la configuration du réseau de stations qui enregistrent les différents signaux, le niveau de bruit enregistré aux différentes stations, la prise en compte du milieu de propagation ainsi que les facteurs de qualité qui vont évaluer la précision de localisation de l'événement finalement détecté.

Dans le cadre de la détection, ces différents facteurs interviennent pour améliorer la qualité des pointés des temps d'arrivée des ondes P et S, la qualité du processus d'association ainsi que la sélection de l'origine préférentielle (si le catalogue produit est un catalogue multi-origine). La qualité du pointé automatique des ondes P et S est conditionnée par la prise en compte des caractéristiques du bruit enregistré aux différentes stations et des distances épicentrales. La qualité des processus d'association est déterminée par la considération de la configuration spécifique du réseau de stations (distances inter-station) et du milieu de propagation (vitesses de propagation des ondes sismiques). Enfin, la qualité de la sélection de l'origine préférentielle dépend de l'utilisation plus exhaustive de paramètres qui évaluent la précision de la localisation hypocentrale (distances épicentrales, nombre de phases, RMS des résidus, nombre de phases S, incertitudes de localisation).

La discrimination étant la réciproque du problème de la détection, pour remonter au type de la source sismique (séisme naturel, tir de carrière ou bruit d'origine anthropique) à partir de l'événement qui est détecté, c'est tout le cheminement qui a conduit à sa détection qu'il faut remonter. De ce fait, en plus des informations véhiculées par la localisation de la source apportée par la détection (longitude, latitude, heure et jour d'occurrence définis à partir du temps d'origine), ce sont également les mêmes facteurs qui vont intervenir pour optimiser le processus de discrimination des événements détectés, à savoir les distances épicentrales, les paramètres qui évaluent la précision de la localisation hypocentrale (valeurs des résidus, nombre de phases), les propriétés du bruit (caractère stationnaire, aléatoire, et impulsif), les caractéristiques du signal (forme de la distribution des amplitudes du signal, contenu fréquentiel) ainsi que le milieu de propagation (polarisation des ondes).

6.1.2 Une recherche de solutions optimales dans un espace multi-factoriel complexe

La résolution des problèmes de détection et de discrimination dans le cadre de la détection des séismes de faible magnitude est éminemment complexe. Avec la densification du réseau de stations et la diminution du seuil de détection, l'espace de recherche pour détecter les signaux associés à de potentiels événements est considérablement accru. Le taux de pointés augmente fortement car les signaux sont détectés avec de plus faibles rapports signal/bruit, les combinaisons de pointés élaborées par les processus d'association se démultiplient et les possibilités d'obtenir des solutions parasites augmentent considérablement.

La prise en compte des différents facteurs tels que la configuration du réseau de stations (distances épicentrales, distance inter-stations), les caractéristiques du signal (amplitudes, contenu fréquentiel), le niveau de bruit enregistré aux stations (variations d'amplitudes temporelles, contenu fréquentiel) ainsi que le milieu de propagation (vitesses de propagation) a donc été critique pour contraindre l'espace de solutions possibles vers des solutions de détection plus optimales.

Les problèmes de la détection et de la discrimination étant deux problèmes qui sont réciproquement liés, un espace de solutions de détection plus optimal facilite la résolution du problème de discrimination. Si par exemple les distances épicentrales ne sont pas considérées dans la procédure de détection, les risques d'émettre des pointés automatiques P ou S trop précoces ou trop tardifs par rapport aux temps d'arrivée réels des ondes sismiques P et S augmentent fortement. Par conséquent, dans ces conditions, les possibilités de générer des fausses associations sont plus grandes, d'autant plus si la distance temporelle de référence nécessaire pour clusteriser les pointés entre eux ne tient pas compte de la configuration du réseau et/ou du milieu de propagation. Or, la résolution du problème de discrimination repose sur la recherche d'un ensemble de critères qui vont solidement définir chaque événement. Si l'espace de solution des détections est dégradé (nombreuses vraies associations contaminées par du bruit, association de pointés émis trop tardivement ou précocement), c'est aussi le processus de discrimination qui se dégrade. Par exemple, la mauvaise définition des pointés P et S peut amener à calculer des rapports spectraux entre les ondes P et S erronés, diminuant la valeur discriminante de ce rapport spectral.

De plus, dans le cadre de la résolution du problème de la discrimination, les caractéristiques des signaux étant fortement influencées par les effets du bruit enregistré aux stations et du milieu de propagation, ne tenir compte que de ces dernières pour résoudre ce problème nous éloigne fortement d'une solution optimale convergente. En effet, le bruit d'origine anthropique et le signal sismique régional présentent des amplitudes, des durées et un contenu fréquentiel très souvent similaires (Hutton et al., 2010 ; Inbal et al., 2018 ; Provost et al., 2017). Il est donc difficile dans ces conditions de discriminer les faux et les vrais événements en se basant uniquement sur les caractéristiques d'un signal fortement influencé par les effets liés au bruit. La prise en compte de la solution hypocentrale apportée par la détection, qui est un paramètre de la source que l'on cherche à identifier, permet alors de compenser les effets du bruit qui atténuent fortement les effets de la source que le signal exprime.

En outre, les signaux associés aux séismes et aux tirs de carrière sont très fortement influencés par les effets liés au milieu de propagation, comme le témoignent la diversité des formes d'onde associées à ces signaux au sein d'une même classe d'événements et la similarité de ces formes d'ondes souvent remarquée entre les différentes classes d'événements. Par ailleurs, ces signaux sont fortement contaminés par le bruit enregistré aux stations, d'autant plus si les signaux détectés sont de faible amplitude. Il apparaît là encore difficile de trouver une solution convergente optimale de discrimination des séismes et des tirs de carrière en se focalisant uniquement sur les caractéristiques d'un signal qui est très fortement dominé par les effets liés au milieu de propagation mais aussi au bruit. De même, la prise en compte de la solution hypocentrale apportée par la détection (localisation épicentrale, temps d'origine) permet, en contraignant l'espace de solutions possibles pour identifier le type de la source, une compensation des effets du milieu de propagation et du bruit qui atténuent fortement les effets de la source que le signal exprime.

Trouver un espace optimal de détection et de discrimination est donc complexe car les contenus en bruit du signal sont spatio-temporellement variables, les formes d'onde associées aux signaux sont fortement soumises aux effets du milieu de propagation dans lequel les ondes se propagent. De plus, les signaux, qu'ils soient ou non associés à une même source, sont géométriquement disséminés au sein d'un réseau dense de stations qui les enregistrent. Seulement, ne pas considérer au maximum cet espace multiparamétrique, c'est probablement approximer fortement la réponse aux problèmes de détection et de discrimination, voire même la dégrader.

6.2 Les résultats de la détection et de la discrimination des séismes de faible magnitude, un reflet de la complexité d'un système multiparamétrique

6.2.1 Des résultats de détection qui reflètent les effets liés au bruit enregistré aux stations

La comparaison des détections produites par la procédure de détection développée dans ce travail de thèse, avec celles émises par le BCSF-RéNaSS pour la période juillet 2016 - décembre 2016, montre qu'un total de 2000 événements ont été détectés en plus, dont 1290 tirs de carrières et 700 séismes. Ce qui fait qu'avec les événements déjà détectés auparavant par le BCSF-RéNaSS, ce sont 2755 événements qui sont finalement détectés. Au total, 2.5 fois plus de séismes et presque 6 fois plus de tirs de carrière ont été identifiés.

Parmi les nouveaux séismes détectés, 48% d'entre eux présentent une magnitude locale MLv inférieure à 1.20 (Figure 6.2). Avec cette procédure de détection, la proportion de séismes de très faible magnitude augmente donc : deux fois plus de séismes sont désormais détectés avec des magnitudes inférieures à 1.20. Sur l'ensemble des nouveaux séismes détectés, 82% ont des magnitudes locales inférieures à 1.50.

Parmi les nouveaux tirs de carrière détectés, 55% d'entre eux ont des magnitudes locales inférieures à 1.50 et la quasi-totalité ont des magnitudes inférieures à 2.0.

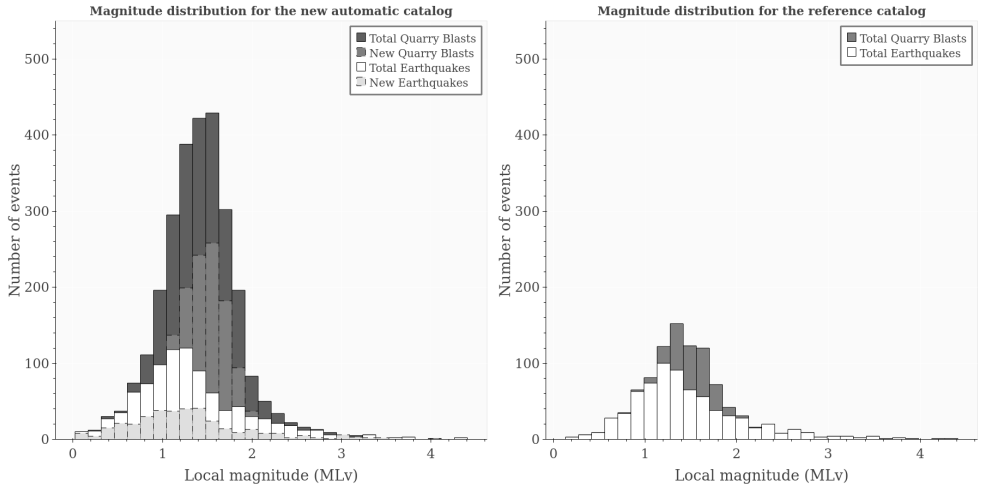


FIGURE 6.2: Distribution des magnitudes des événements détectés avec la nouvelle procédure de détection développée dans ce travail de thèse pour la période juillet 2016-décembre 2016.

L'analyse de la distribution cumulative fréquence-magnitude des séismes pour cette période de juillet à décembre 2016 montre que la magnitude de complétude, estimée grossièrement à partir de cette distribution, atteint maintenant 1.10 avec la nouvelle détection automatique, alors qu'elle était de 1.20 pour le catalogue de référence (Figure 6.3). Même si cette magnitude de complétude affiche une baisse très subtile, ce constat annonce des résultats prometteurs pour une détection future des séismes de faible magnitude plus approfondie et plus longue.

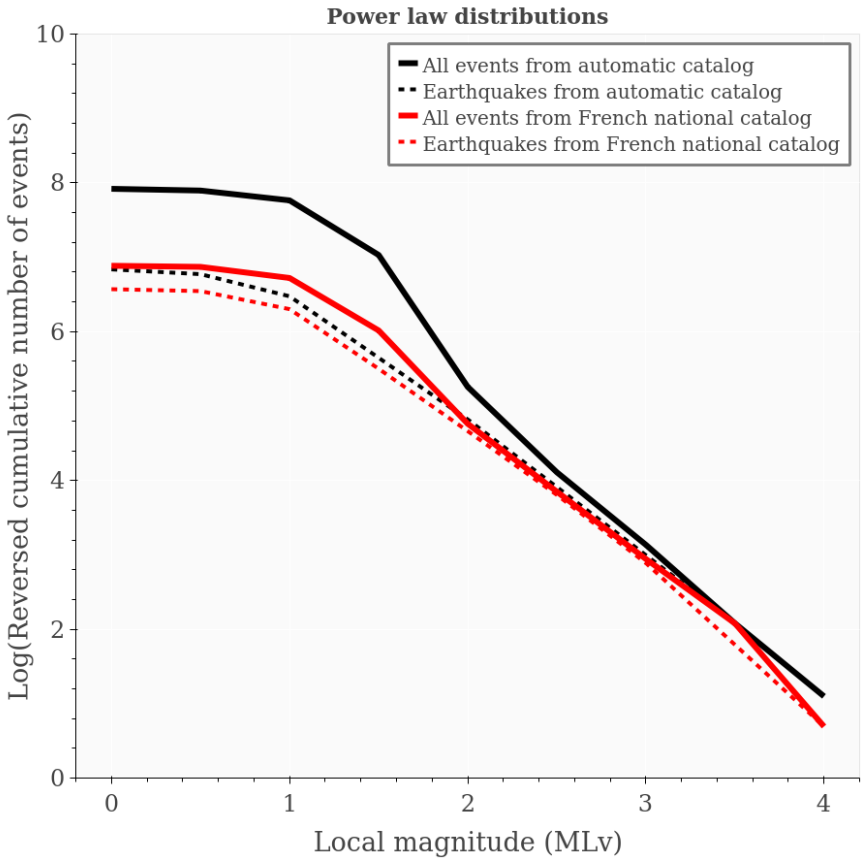


FIGURE 6.3: Distribution cumulative fréquence-magnitude des événements détectés automatiquement par la nouvelle procédure de détection pendant la période juillet 2016 -décembre 2016.Distribution cumulative fréquence-magnitude des événements détectés automatiquement par la nouvelle procédure de détection pendant la période juillet 2016 -décembre 2016.

Néanmoins, en observant le taux de détection des séismes au cours des heures de la journée, il est possible de constater que 71% des séismes contenus dans le catalogue automatique ont eu lieu avant 6 heures du matin et après 18 heures, c'est-à-dire pendant les périodes où le niveau du bruit d'origine anthropique est le plus bas (Figure 6.4). Un peu moins de 75% des tirs de carrière sont détectés entre 9 heures et 16 heures. C'est aux heures où les tirs de carrière sont majoritairement détectés que le taux de séismes capturés est le plus bas, et inversement. Il y a donc une segmentation temporelle artificielle de la détection des événements. Même si les séismes sont détectés à n'importe quel moment de la journée, il reste un déficit de détection des séismes aux périodes où sont intensément détectés les tirs de carrière.

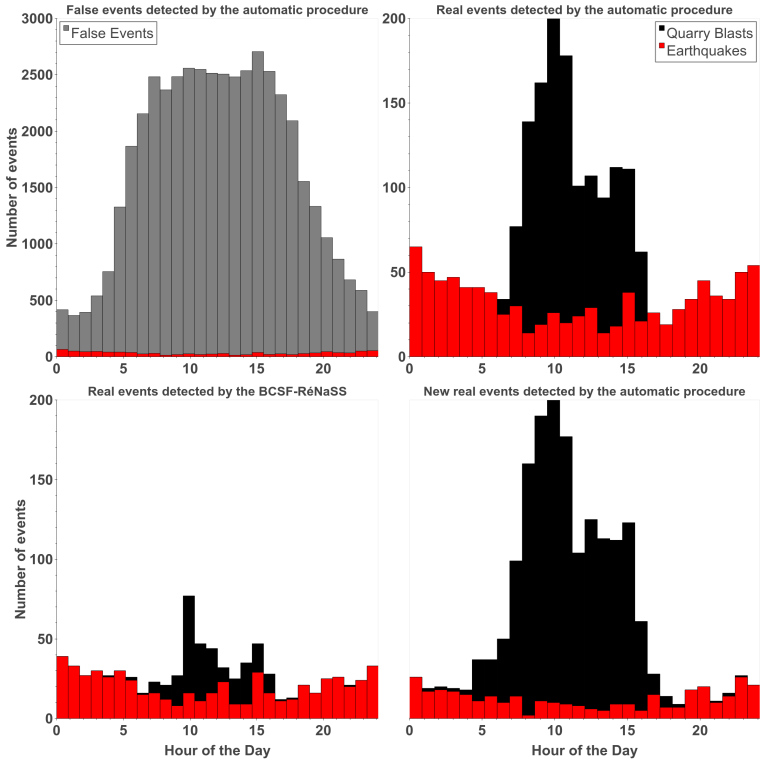


FIGURE 6.4: Comparaison des distributions des séismes et des tirs de carrière détectés automatiquement en fonction des heures de la journée avec celle du BCSF-RéNaSS pour la même période de détection (juillet 2016-décembre 2016). La distribution des faux événements est représentée également comme approximation de l'évolution du niveau de bruit anthropique au cours des heures de la journée.

De la même manière, la distribution des détections des séismes et des tirs de carrière en fonction du jour de la semaine montre une discrimination temporelle de la détection des événements, même si celle-ci est moins marquée. Sur l'ensemble des séismes détectés, environ 36% ont eu lieu un samedi ou un dimanche, c'est-à-dire pendant le week-end, période au cours de laquelle le niveau de bruit d'origine anthropique est globalement plus bas et/ou l'activité de carrière est minimale (Figure 6.5).

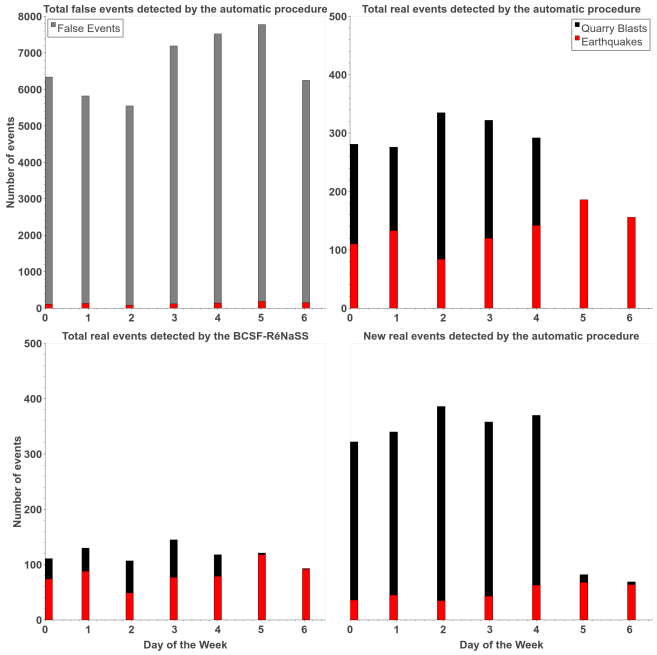


FIGURE 6.5: Distribution des séismes et des tirs de carrière détectés automatiquement en fonction du jour de la semaine. La période de détection est juillet 2016-décembre 2016. La distribution des faux événements est représentée également comme approximation de l'évolution du niveau de bruit anthropique en fonction des jours de la semaine. 0 = lundi, 1 = mardi, 2 = mercredi, 3 = jeudi, 4 = vendredi, 5 = samedi, 6 = dimanche.

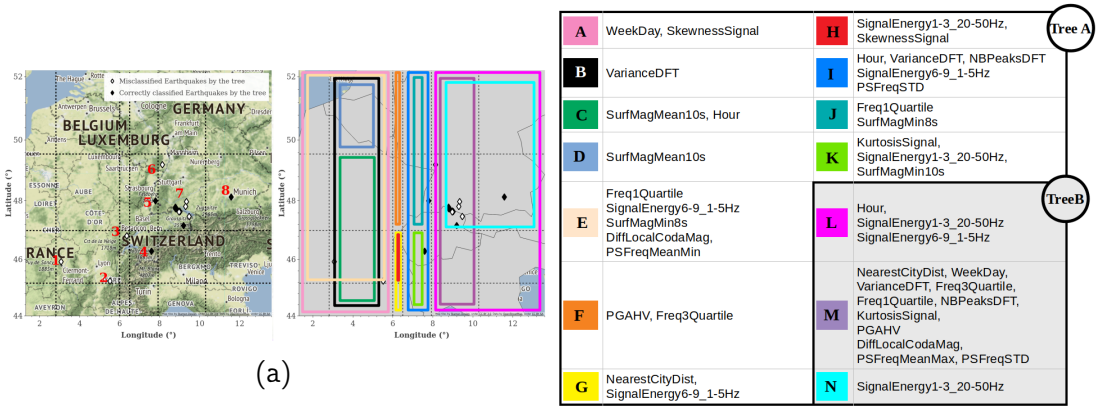
Les périodicités apparentes hebdomadaires et quotidiennes des séismes observées semblent être corrélées aux périodes de détection minimales du bruit d'origine anthropique. Comme cela a été déjà observé, ce bruit d'origine anthropique affecte durablement la détection des séismes, plus particulièrement pour les séismes de faible magnitude qui sont détectés avec de faibles rapports signal/bruit (Atef et al., 2009 ; Hao et al., 2019). Cet artefact de périodicité liée à la détectabilité des événements dans un environnement urbain brouille le comportement statistique des séismes dans la zone d'étude. Par conséquent, les résultats de la détection sont très prometteurs mais ces résultats mettent aussi en évidence qu'il reste difficile de s'affranchir complètement des effets liés au bruit enregistré aux stations.

6.2.2 Des résultats de discrimination qui reflètent les effets liés au milieu de propagation

•Une variabilité régionale de l'efficacité des discriminants

Comme exprimé dans l'article présenté au chapitre précédent, une variabilité régionale de l'efficacité des discriminants peut être mise à jour dans la classification des séismes et des tirs de carrière. Cette variabilité transparaît au niveau de l'architecture des arbres décisionnels constituant la forêt aléatoire. La localisation de l'événement (longitude et latitude) offre une contrainte qui va rythmer la constitution des différents embranchements de l'arbre. Se dessinent alors plusieurs régions géographiques qui sont chacune caractérisées par une combinaison d'attributs spécifiques et leur seuil de valeurs respectif. A partir de là, il est possible de traduire l'arrangement hiérarchique de l'arbre décisionnel en classification emboîtée, où chaque boîte correspond à un assemblage d'attributs reliés au signal, délimitée par les différentes valeurs seuils de la longitude et de la latitude.

Si je reprends l'exemple proposé dans l'article, mais en y ajoutant quelques exemples de séismes incorrectement prédits par le classifieur des séismes et des tirs de carrière pour la période septembre 2016-décembre 2016, plusieurs groupes se dessinent (Figure 6.6). D'après cet arbre décisionnel analysé, le séisme appartenant au groupe 1 se situe dans une surface géographique identifiée par les régions A, B et C. Ce dernier est alors prédit selon les critères partagés par l'ensemble des régions A, B et C, à savoir le degré d'asymétrie de la distribution des valeurs d'amplitudes du signal, le temps d'origine de l'événement détecté (jour de la semaine, heures de la journée), la variance spectrale ainsi que la magnitude de surface moyenne estimée à 10 s. Dans ce cas, ce séisme est prédit en tant que tir de carrière car il possède une magnitude de surface relativement élevée et est associé à une valeur du coefficient d'asymétrie qui s'approche de celui des tirs de carrière pour la zone d'étude.



(a) (b)

FIGURE 6.6: Projection en carte d'une partie de la classification emboîtée déduite d'un arbre décisionnel extrait aléatoirement à partir de la forêt (cf supplément de l'article pour détail de cet arbre). Chaque région géographique, exprimée à travers un code couleur, est définie par un ensemble d'attributs utilisés pour prédire les séismes et les tirs de carrière à l'intérieur de cette région. Les séismes incorrectement prédits par le classifieur pour la période septembre 2016-décembre 2016 définissent des groupes qui appartiennent à différentes régions géographiques (groupe 1 à 6). Ces séismes sont incorrectement prédits selon les critères qui sont utilisés dans chaque région géographique. Les lignes en pointillé constituent les valeurs de longitude et de latitude de référence qui ont servi à élaborer les emboîtements.

De même, le séisme du groupe 2, appartenant à la zone géographique qui regroupe les régions A, B et E, n'est pas correctement classifié à cause d'un contenu fréquentiel particulièrement bas et d'une variance spectrale moins élevée que la moyenne des séismes détectés dans la zone d'étude. Quant au séisme appartenant au groupe 3 (régions G et H), le rapport de l'accélération maximale du sol entre la composante horizontale et verticale de la station, particulièrement élevé, et le contenu relatif basse-fréquence du signal associé, plutôt haut, sont les deux critères qui vont induire une prédiction incorrecte de ce séisme. En revanche, le séisme appartenant au groupe 4 (régions I et K) est correctement prédit car les signaux qui lui sont associés présentent une énergie du signal intense à des fréquences plus élevées (6-9 Hz) typiques des séismes, tout comme celui du groupe 5 qui est également bien prédit. Le même raisonnement peut être fait pour les séismes appartenant aux trois derniers groupes (6, 7 et 8).

De cette façon, si le rapport de l'accélération maximale du sol entre la composante horizontale et verticale de la station est un attribut qui est utilisé dans l'édification du chemin décisionnel de l'arbre pour prédire un événement dans la zone géographique nommée F, cet attribut ne fait pas partir de la sélection pour l'élaboration des chemins décisionnels qui vont contribuer à prédire les événements dans la zone géographique nommée J.

Sur l'ensemble des arbres décisionnels qui composent la forêt, chaque arbre peut donc être traduit sous forme d'une classification emboîtée spécifique avec une combinaison d'attributs distinctes comme c'est le cas par exemple de deux autres classifications emboîtées élaborées à partir deux autres extrait d'arbre décisionnel différent (Figure 6.7). A partir de la comparaison des trois classifications proposées (Figures 6.6 et 6.7), quelques lignes communes peuvent être tracées.

La première est que la longitude 8°E semble à chaque fois séparer la zone d'étude en deux grands sous-groupes. La résolution des régions géographiques est plus faible à l'est de la ligne de référence 8°E. Cette observation reflète un déséquilibre de répartition des événements dans le jeu de données qui a servi à l'apprentissage. En effet, la majorité des événements dont je dispose pour entraîner l'algorithme de classification est située à l'ouest de cette ligne de référence. Cette scission se répète successivement à travers l'analyse des arbres décisionnels.

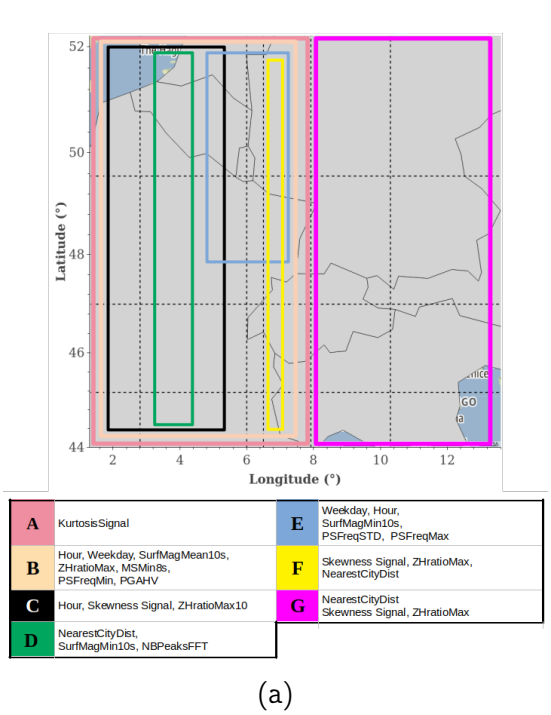
La deuxième observation est qu'il est possible d'ores et déjà de faire des recoupements entre les différentes régions géographiques révélées, à travers l'analyse grossière de ces trois extraits d'arbre décisionnel. En effet, par exemple, pour les trois classifications emboîtées, l'attribut qui décrit le coefficient d'asymétrie de la distribution des valeurs d'amplitudes du signal est intégré dans l'élaboration des chemins décisionnels qui servent à prédire les événements situés dans une zone comprise entre les longitudes 1.8°E et 5°E et les latitudes 46.8°N et 50°N.

Les attributs décrivant les magnitudes de surface à 8s (valeur minimale) et 10 s (valeurs moyenne et minimale) semblent être quant à eux impliqués dans l'édification des chemins décisionnels qui conduisent à la prédiction des événements dans la zone délimitée par les longitudes 2.8°E et 6°E et les latitudes 47°N et 49.6°N.

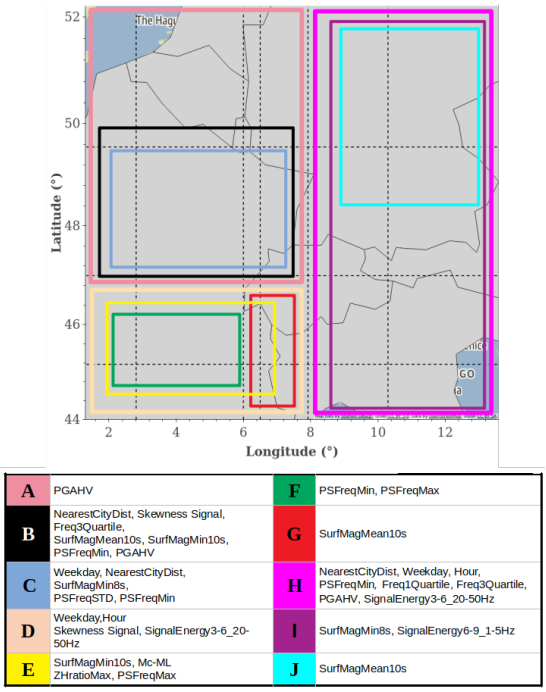
De même, le rapport de l'accélération maximale du sol entre la composante horizontale et verticale de la station est un attribut qui est utilisé dans l'édification des chemins décisionnels des trois arbres pour prédire les événements circonscrits dans la zone comprise entre les longitudes 6°E et 6.4°E et les latitude supérieures à 49.6°N.

En outre, l'attribut retraçant l'écart-type des valeurs des rapports spectraux entre les ondes P et S calculés pour l'ensemble des signaux impliqués dans la détection de chaque événement est utilisé pour prédire les événements dans la zone comprise entre les longitudes 6°E et 6.4°E et les latitudes 47°N et 49.6°N.

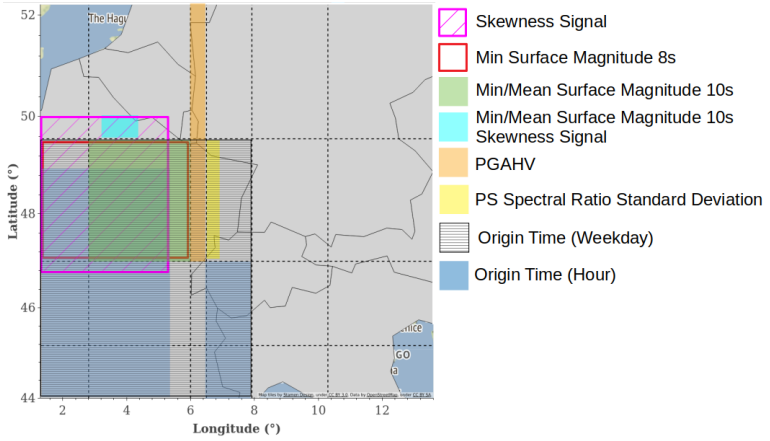
Enfin, les attributs décrivant le temps d'origine des événements (heure et date d'occurrence) sont utilisés dans l'élaboration des chemins décisionnels des trois arbres décisionnels pour des zones plus larges : entre les longitudes 2.8°E et 5.2°E et les latitudes 44°N et 49.5°N puis les longitudes 6.4°E et 8°E et les latitudes 44°N et 47°N pour les heures d'occurrence et les longitudes 1.8°E et 8°E et les latitudes 44°N et 49°N pour le jour d'occurrence.



(a)



(b)



(c)

FIGURE 6.7: Exemple de cartographie de la régionalisation de l'effet des attributs sur la prédiction des séismes et des tirs de carrière. (a),(b) Projection en carte de deux classifications emboîtées déduites chacune d'un extrait d'arbre décisionnel tiré aléatoirement parmi l'ensemble des 500 arbres décisionnels qui composent la forêt aléatoire. Chaque région géographique, exprimée à travers un code couleur, est définie par un ensemble d'attributs utilisés pour prédire les séismes et les tirs de carrière à l'intérieur de cette région. (c) Projection en carte du résultat de la combinaison des trois classifications emboîtées présentées dans les Figures 6.6 et 6.7a, b. Chaque zone géographique commune, partageant le même échantillonnage d'attributs pour la prédiction, est représentée par un code couleur spécifique. Les lignes en pointillés constituent les valeurs de longitude et de latitude de référence qui ont servi à élaborer les emboîtements sur l'ensemble des figures présentées.

Par conséquent, si l'ensemble des informations véhiculées par chaque zone géographique issue de la fusion des trois extraits d'arbre décisionnels sont combinées ensemble, il est possible de révéler une zone commune, délimitée par les longitudes 2.8°E et 5.2°E et les latitudes 47°N et 49°N. Dans cette zone mise en relief, les séismes et les tirs de carrière sont prédits grâce à un pool d'attributs constitués par l'heure et la date d'occurrence des événements, le coefficient d'asymétrie des distributions des valeurs des amplitudes des signaux associés à chaque événement ainsi que les magnitudes de surface à 8 s et 20 s (valeurs minimales et/ou moyennes).

Les premiers résultats des classifications emboîtées offrent pour l'instant une image très incomplète de l'étendue de la variabilité de l'efficacité des discriminants sur les différentes régions géographiques. Seulement, l'ébauche de cartographie très simplifiée de la régionalisation des effets des attributs effectuée dans ce travail de thèse met en évidence le potentiel réel des résultats de l'apprentissage machine pour révéler une cartographie complète de cette régionalisation.

Ainsi, une piste intéressante à approfondir est d'élaborer une classification emboîtée exhaustive sur l'ensemble de la forêt aléatoire de façon à pouvoir finement cartographier les combinaisons de discriminants partagés par une même zone géographique. Une information riche est contenue dans cette forêt, qui ne demande qu'à être exploitée. Seulement, pour l'exploiter efficacement, une procédure automatique d'analyse des différents arbres doit être mise en place. Certains auteurs ont par ailleurs déjà élaboré des outils d'analyse automatique des arbres décisionnels. Ce qui constitue une première approche (Lapuschkin et al., 2019 ; Samek, 2020.

•Une variabilité locale de l'efficacité des discriminants

La variabilité des discriminants peut être aussi observée plus localement. Si je prends l'exemple de la séquence d'événements qui a eu lieu au nord du lac Konstanz en Allemagne au cours de la période septembre 2016-décembre 2016, 61 séismes ont d'abord été identifiés. Ces séismes sont caractérisés par une similarité de formes d'onde manifeste au premier abord (cf Figure 5.30 pour la visualisation de la similarité des formes d'onde enregistrée à la station SLE).

Seulement, malgré la forte similarité de ces formes d'ondes et de localisations épicentrales, le classifieur des séismes et des tirs de carrière ne prédit pas ces événements avec la même probabilité (Figure 6.8). Un total de 16 séismes vont même être prédits comme des tirs de carrière pour les raisons déjà évoquées dans le chapitre précédent.

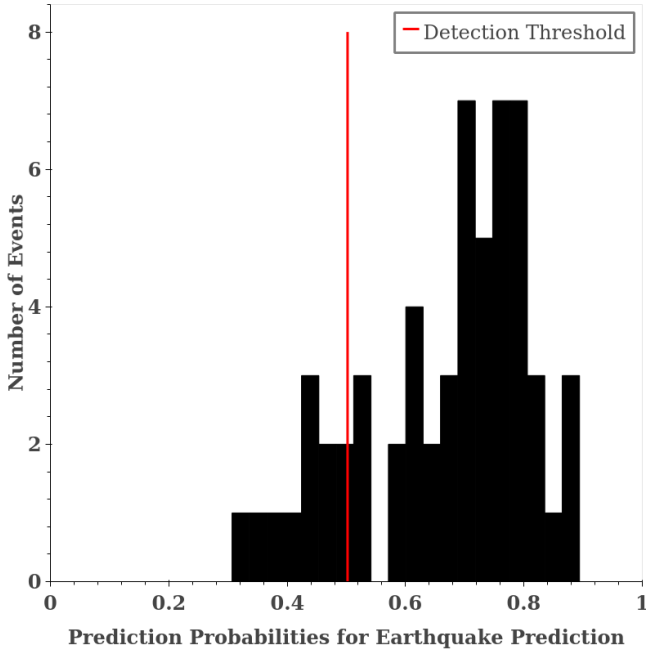
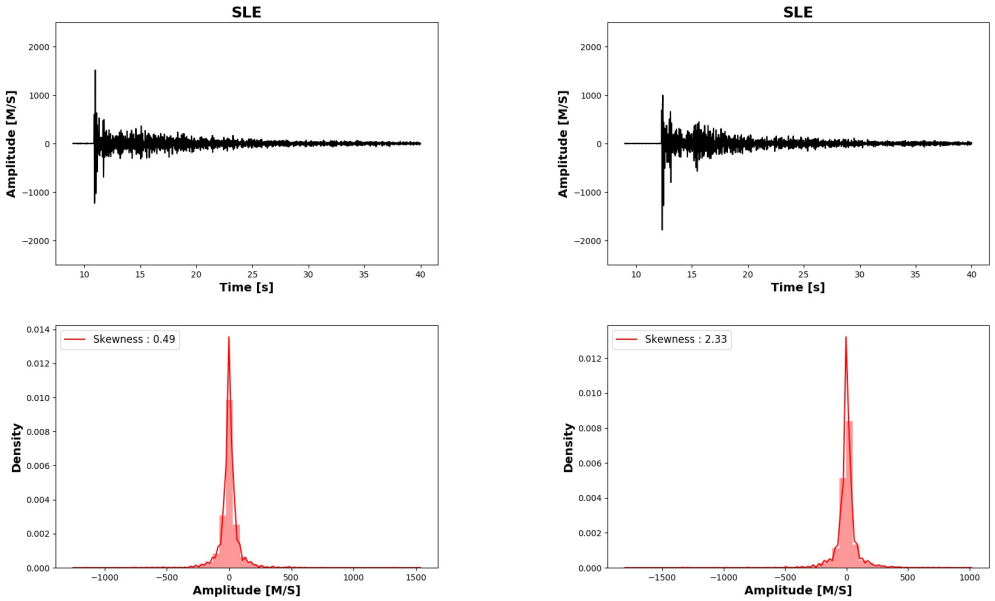


FIGURE 6.8: Distribution de la famille de séismes localisés au nord du lac Konstanz en Allemagne en fonction des probabilités de prédiction du classifieur des séismes et des tirs de carrière. Une probabilité de 1 signifie que la totalité des arbres de la forêt aléatoire a prédit l'événement comme étant un séisme. Une probabilité de 0 signifie qu'aucun des arbres de la forêt aléatoire n'a prédit l'événement comme étant un séisme (donc les 500 arbres ont prédit dans ce cas l'événement comme étant un tir de carrière). Un événement est prédit comme séisme à partir d'une probabilité de 0.502.

D'une manière globale, l'ensemble des 61 séismes repérés au nord du lac Konstanz sont associés à des signaux dont la valeur moyenne du coefficient d'asymétrie (0.31 ± 0.20) tend à se rapprocher de celle des tirs de carrière de la zone d'étude (0.20 ± 0.49), plutôt que de celle des séismes (0.55 ± 0.81). Les séismes les mieux prédits (probabilité de prédiction > 0.72) par le classifieur des séismes et des tirs de carrière sont d'ailleurs reliés aux valeurs les plus élevées du coefficient d'asymétrie de cette distribution (Figure 6.9).

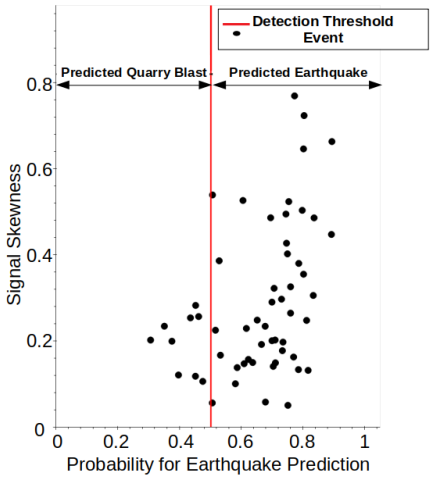


(a) Signal correspondant à un séisme ayant eu lieu le 21 novembre 2016 à 04h30 (MLv 1.50) et prédit avec une probabilité d'être un séisme faible (0.37).

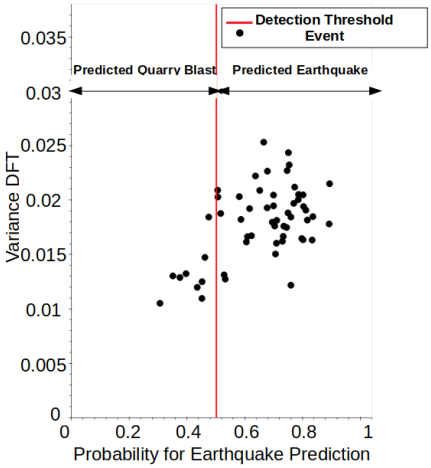
(b) Signal correspondant à un séisme ayant eu lieu le 13 novembre 2016 à 15h12 (MLv 1.65) et prédit avec une probabilité d'être un séisme élevé (0.80).

FIGURE 6.9: Coefficients d'asymétrie et formes d'onde associés à deux signaux enregistrés sur la composante verticale de la station SLE et correspondant chacun à un séisme appartenant à l'ensemble des 61 séismes identifiés au Nord du lac Konstanz en Allemagne (distance épicentrale 20 km). Les valeurs les plus élevées du coefficient d'asymétrie sont à relier avec des événements prédits avec une forte probabilité d'être assimilés à des séismes.

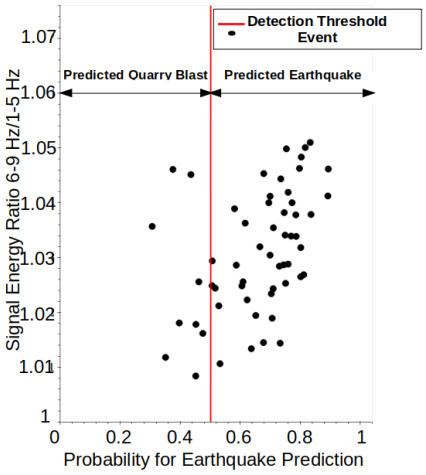
De plus, comme il a été écrit précédemment, les séismes qui sont prédits incorrectement par le classifieur sont reliés à des signaux qui ont une variance spectrale généralement inférieure aux autres événements correctement prédits (Figure 6.10b) et une énergie relative plus intense aux fréquences comprises entre 1 et 5 Hz (Figure 6.10c).



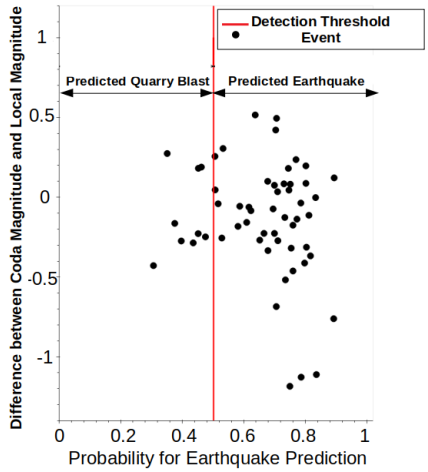
(a) Distribution des valeurs absolues du coefficient d'asymétrie.



(b) Distribution des valeurs de la variance spectrale (le spectre du signal est normalisé par sa valeur maximale).



(c) Distribution des valeurs des rapport d'énergie du signal entre les bandes fréquentielles 6-9 Hz et 1-5 Hz.



(d) Distribution des valeurs issues de la différence entre la magnitude de coda et la magnitude locale.

FIGURE 6.10: Distribution des valeurs de 4 attributs utilisés pour prédire les séismes et tirs de carrière (coefficient d'asymétrie, variance spectrale, rapport d'énergie du signal entre les bandes fréquentielles 6-9 Hz et 1-5 Hz, différence entre la magnitude de coda et la magnitude locale) en fonction des probabilités de prédiction émises par le classifieur. Les valeurs sont extraites de la totalité des 61 séismes détectés au Nord du lac Konstanz en Allemagne entre septembre 2016 et décembre 2016. Une probabilité de 1 signifie que la totalité des arbres de la forêt aléatoire a prédit l'événement comme étant un séisme. Une probabilité de 0 signifie qu'aucun des arbres de la forêt aléatoire n'a prédit l'événement comme étant un séisme (donc cela signifie que les 500 arbres ont prédit l'événement en tant que tir de carrière).

En outre, en observant la répartition des valeurs de la différence entre la magnitude de coda et la magnitude locale en fonction des probabilités de prédiction du classifieur, une tendance s'affirme. En effet, la différence entre la magnitude de coda et la magnitude locale diminue à mesure que les probabilités de prédiction augmentent (Figure 6.10d).

Or, il a été constaté que la différence moyenne entre la magnitude de coda et la magnitude locale est une fonction sensible de la profondeur de la source (Koper et al., 2016). Plusieurs pistes ont d'ailleurs été proposées pour tenter d'expliquer pourquoi les événements les plus superficiels possèdent des codas de plus longue durée, comme la présence d'un guide d'ondes à faible vitesse proche de la surface ou des chutes de contrainte plus faibles (Holt et al., 2019).

Par conséquent, si cet attribut (différence entre magnitude de coda et magnitude locale) témoigne indirectement de la profondeur des événements, cela signifie alors que le classifieur prédit plus difficilement correctement les événements superficiels de ces essaims de séismes : une plus forte valeur de cet attribut est corrélée avec une valeur de probabilité faible. Ce classifieur prédira plus facilement les séismes superficiels comme étant des tirs de carrière car leurs signaux présentent des propriétés similaires à ceux des tirs de carrière pour les attributs considérés, d'autant plus que la particularité de ces 61 séismes est de partager des valeurs de coefficient d'asymétrie de la distribution des valeurs d'amplitude des signaux associés similaires à celles des tirs de carrière.

De ce fait, les erreurs de classification pour ces essaims de séismes ont manifesté une valeur discriminante différentielle des attributs en fonction de l'événement considéré au sein même de chaque essaim. Cette variabilité plus locale de l'effet des discriminants limite donc localement la performance de la prédiction. Celle-ci révèle également indirectement, à travers les probabilités de prédiction, un paramètre de la source, à savoir ici sa profondeur. Enfin, ce résultat signale que les critères utilisés pour classer les événements reflètent indirectement les effets du milieu de propagation que les signaux manifestent.

6.3 Une procédure de détection des séismes de faible magnitude encore à optimiser

6.3.1 Approfondir l'interactivité Homme-machine au sein des observatoires sismologiques

La procédure de détection qui est développée dans ce travail de thèse, élaborée sous SeisComP3, a l'avantage d'être transposée facilement en opérationnel. De plus, elle fournit des résultats de classification prometteurs, qui ne se déparent pas des résultats de la classification manuelle, ou bien d'une classification automatique élaborée à partir d'un jeu pointé manuellement.

En effet, si je compare les résultats de la classification automatique des séismes et des tirs de carrière élaborée à partir du jeu de données produit automatiquement (période septembre 2016 -décembre 2016) aux résultats de la classification automatique élaborée à partir de ce même jeu de données, mais repris manuellement, il est possible de constater que le classifieur prédit correctement un taux équivalent de séismes, quel que soit le jeu de données (Table 6.1).

De plus, pour le jeu automatique non repris manuellement, la performance prédictive du classifieur vis-à-vis des tirs de carrière se dégrade légèrement (de l'ordre de 2%), même si les résultats restent très honorables (94.76% de tirs de carrière bien classés).

TABLE 6.1: Comparaison des performances prédictives du classifieur de séismes et de tirs de carrière vis-à-vis du jeu d'événements détectés automatiquement entre septembre 2016 et décembre 2016 et le même jeu d'événements repris manuellement

	Manually Reviewed Automatic Data	**Automatic data**
Specificity (%)	96.82 ± 0.24	94.76 ± 0.34
Sensitivity (%)	96.55 ± 0.22	96.04 ± 0.30
Precision (%)	96.00 ± 0.29	94.68 ± 0.33
F-Measure	0.963 ± 0.002	0.953 ± 0.002

[a] Spécificité : le taux de tirs de carrière correctement prédits (soit le rapport des vrais négatifs sur la somme des vrais négatifs et des faux positifs). Sensitivité : le taux de séismes correctement prédits (soit le rapport des vrais positifs sur la somme des vrais positifs et des faux négatifs). Précision : la proportion de séismes correctement prédits relativement à toutes les détections positives (le rapport des vrais positifs sur la somme des vrais positifs et faux positifs). La mesure F : un résumé statistique qui combine la précision et la sensitivité ($2 \times \text{prcision} \times \text{sensitivit}/(\text{prcision} + \text{sensitivit})$).

Ce dernier résultat souligne indirectement la difficulté accrue d'obtenir des pointés automatiques de très bonne qualité lorsqu'il s'agit de détecter des signaux sismiques associés aux tirs de carrière. En effet, ces signaux sont détectés dans des périodes où le bruit d'origine anthropique est le plus élevé et avec de faibles rapports signal/bruit. Toutefois, la forte proportion de tirs de carrière bien prédits avec le jeu automatique met tout de même en évidence l'apport significatif de la procédure de détection pour qualitativement détecter l'ensemble des vrais événements de la zone d'étude.

Comme décrit dans l'article présenté précédemment, l'intégration de cette procédure de détection au sein des observatoires sismologiques a des avantages certains.

Elle permet d'abord la détection des vrais événements en éliminant plus de 99% des faux événements. En effet, sans intégration de l'apprentissage machine dans le flux de détection, la procédure de détection génère près de 50 000 événements sur 4 mois. De ce fait, l'introduction du module de discrimination SeisComP3, que j'ai développé, dans le système de détection final élimine la fatigue physiologique liée aux faux événements, tout en maintenant un taux de séismes détectés très satisfaisant (environ 93 %).

De plus, les probabilités de prédiction apportées par le classifieur des séismes et des tirs de carrière offrent une base intéressante pour revoir manuellement les résultats finaux de la détection automatique. Seulement, il faudrait comprendre plus précisément la nature de l'interaction Homme-machine dans le cadre de cette revue manuelle des événements. En effet, revoir manuellement les événements en se basant sur la valeur de la probabilité de prédiction est une approche simple. Comme écrit dans l'article précédent, la revue manuelle de l'ensemble des événements discriminés avec une probabilité comprise entre 0.4 et 0.7 n'est pas une lourde tâche et conduit même à une amélioration très forte du taux d'événements correctement classés.

Cependant, avec cette approche, il y a un risque de conformation forte aux résultats prédits par la fonction de prédiction, en particulier pour les probabilités en dehors de la gamme 0.40-0.70. Pour éviter ces effets potentiellement négatifs, il apparaît indispensable d'étudier de manière approfondie comment l'humain se comporte dans un processus décisionnel qui intègre la machine. Au demeurant, si une trop grande conformation vis-à-vis de l'apprentissage machine peut dégrader les solutions finales, le rejet systématique des résultats de l'apprentissage machine sous prétexte que l'algorithme se "trompe" souvent ou qu'il apprend "mal" est également contre-productif. En définitive, si l'algorithme apprend "mal", c'est que l'espace d'hypothèses initiales pour rechercher la fonction de prédiction optimale n'est pas suffisamment contraint.

L'analyse comportementaliste de l'Homme vis-à-vis de la machine, comme par exemple comparer une population d'analystes qui classe les événements sans l'apprentissage machine et une autre population qui classe les événements avec, puis comprendre les choix élaborés par les deux populations, reste indispensable pour optimiser l'interactivité Homme-machine. Il s'agit de tirer profit de l'apport manifeste de l'apprentissage machine dans la discrimination, tout en assurant une veille permanente de l'Homme sur les résultats prodigués par cet apprentissage.

6.3.2 Tendre vers l'erreur de généralisation la plus petite possible

Si dans ce travail j'ai recherché à élaborer des classifieurs (un classifieur pour les vrais et faux événements et un classifieur pour les séismes et les tirs de carrière) qui minimisent au maximum l'erreur de généralisation, d'autres angles sont à considérer pour augmenter la performance prédictive de ces derniers, et asseoir leur validité.

Ces classifieurs ont été testés en dehors de la zone d'étude, sur un jeu d'événements détectés dans la zone des Pyrénées françaises. C'est d'ailleurs avec ce jeu d'événements qu'il a été confirmé que les différents paramètres qui vont décrire l'enveloppe du signal (statistique et forme de la distribution des valeurs de l'enveloppe, complexité, etc) dégradent la prédiction des séismes. En effet, étant donné la forte variabilité des formes d'onde au sein même d'une classe d'événements et entre les classes d'événements, une introduction détaillée des paramètres qui vont définir le signal dans le domaine temporel apporte beaucoup de confusions.

Ainsi, utiliser ces classifieurs pré-entraînés sur d'autres jeux d'événements détectés dans d'autres environnements peut apporter une validité aux résultats proposés dans ce travail de thèse. De plus, détecter les événements dans la zone d'étude en retirant les stations temporaires AlpArray, pourra aussi être un autre garant de la robustesse des deux classifieurs. Tester le pouvoir prédictif de ces classifieurs avec un jeu d'entraînement plus grand est aussi intéressante pour évaluer leur stabilité.

De plus, comme écrit précédemment, la variabilité régionale des discriminants des séismes et des tirs de carrière mérite d'être approfondie à travers une étude complète de l'ensemble des arbres décisionnels, de façon à savoir s'il est possible d'élaborer une cartographie globale de l'efficacité de ces derniers. Or, si ces discriminants s'avèrent être d'efficacité variable en fonction des régions géographiques, il serait intéressant de comprendre plus exactement ce qu'ils révèlent : les effets de la source ? les effets du milieu de propagation ? les effets du bruit enregistré à certaines stations spécifiques ?

Enfin, la discrimination des vrais événements reste à être affinée notamment en se penchant sur d'autres classes d'événements comme la sismicité induite par la géothermie profonde. Si cette classe d'événements est minoritaire par rapport aux séismes et aux tirs de carrière, elle présente des enjeux non négligeables (économiques, scientifiques, sociologiques) pour la compréhension des risques sismiques associés à cette activité géothermique. Seulement, la résolution de ce problème de classification des séismes induits par la géothermie profonde est une tâche complexe à accomplir.

C'est un problème qui est dès le départ complexe puisque le jeu de données disponible dans la zone d'étude est de taille petite et pollué par d'autres événements qui sont étiquetés comme induits mais qui sont en fait purement liés à une activité minière (effondrement de toit de mines par exemple). C'est en élaborant un premier apprentissage à partir de ce jeu de données que je me suis aperçue de l'inclusion de ces événements. La sélection automatique des attributs qui est produite considère la proximité de l'événement à la mine la plus proche avec une importance relative non négligeable de l'ordre de 3% (Figure 6.11).

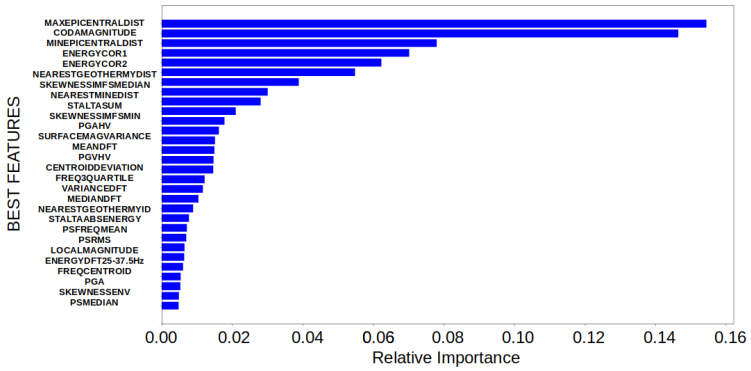


FIGURE 6.11: Première sélection d'attributs produite par élimination récursive pour l'élaboration d'un classifieur qui puisse également identifier les séismes induits par la géothermie profonde parmi l'ensemble des autres vrais événements détectés dans la zone d'étude. Cette sélection a été élaborée à partir d'un jeu d'événements détectés au cours de l'année 2016-2017. La valeur discriminante de cette sélection n'a pas été testée, ni validée.

Par ailleurs, si la discrimination de la sismicité induite par la géothermie peut être établie à travers des corrélations spatiales et temporelles avec les injections qui l'ont produite (VERDON et al., 2019), ces critères peuvent être non suffisants lorsque des séismes d'origine naturelle sont régulièrement détectés dans la zone des puits d'injection.

Ainsi, ces séismes induits par la géothermie profonde sont reconnus pour être reliés à des caractéristiques de la source, des relations magnitude-fréquence et des mouvements du sol similaires aux séismes dits naturels (Schoenball et al., 2015 ; Atkinson, 2020 ; Atkinson et al., 2020. Ils partagent également avec ces derniers des mécanismes de rupture assez semblables, comme peuvent le témoigner l'analyse de leurs mécanismes focaux double-couple qui semblent compatibles avec les champs de contrainte régionaux (Clarke et al., 2019 ; T. S. Eyre et al., 2019 ; Lei, Z. Wang et al., 2019 ; Z. Zhang et al., 2019. Ce qui rend encore plus difficile la résolution du problème de discrimination de cette sismicité induite, au-delà de la qualité du jeu d'entraînement disponible.

6.4 Bilan

Le problème de la détection des séismes de faible magnitude dans une région continentale stable, telle que la zone d'étude de ce travail de thèse, est intimement associée à la notion de détectabilité des signaux sismiques dans un système multiparamétrique. Ces signaux sont le résultat de la combinaison des effets de la source, souvent atténués, du milieu de propagation et du bruit enregistré aux stations.

Lorsque les seuils de détection sont diminués pour détecter les signaux avec de faible rapports signal/bruit, les effets du bruit sur ces signaux à détecter s'amplifient. Si la diminution des seuils des détection est combinée avec un réseau de stations plus dense, comme c'est le cas de la zone d'étude, la complexité des chemins de propagation des ondes sismiques est plus facilement capturée dans toutes les directions de l'espace, et, à l'échelle du réseau, les effets du milieu de propagation sur les signaux à détecter s'intensifient. Par conséquent, si ce sont des signaux de faible amplitude qui sont détectés, les effets de la source deviennent très vite atténués.

Le problème de la détection des séismes de faible magnitude émerge alors, et il s'agit de comprendre comment décoder le signal pour en extraire les informations atténuées de la source sismique, c'est-à-dire comment diminuer les effets liés au bruit et au milieu de propagation.

La procédure de détection que j'ai développée vise à diminuer ces deux effets. Les effets liés au bruit se manifestent d'emblée par le pointé automatique des ondes sismiques P et S. En effet, les algorithmes de pointé automatique, implémenté dans le système de détection de SeisComP3 que je cherche à optimiser, se basent sur des variations d'amplitudes, voire de fréquence et de phase pour détecter les temps d'arrivée des ondes sismiques. Ces algorithmes peuvent donc reconnaître indistinctement des signaux sismiques cohérents associés à un événement comme des signaux associés purement à du bruit impulsif d'origine anthropique, d'autant plus si tous ces signaux sont de même amplitude, de même durée et de même contenu fréquentiel.

Afin de limiter ces effets liés au bruit, le filtrage fréquentiel du signal, le début de la fenêtre temporelle utilisée pour détecter les temps d'arrivée des ondes P et S et la valeur du rapport signal/bruit minimal autorisée pour pointer les phases S ont été adaptés aux conditions de bruit enregistrées spécifiquement aux stations. De plus, les phases S étant pointées une fois que les pointés des phases P ont été émis, la taille et le pas de la fenêtre temporelle utilisés pour détecter l'arrivée des ondes S sont fortement conditionnés par la distance épicentrale et doivent donc être ajustés en fonction de cette distance. Dans ce cas-ci, ce sont les effets du milieu de propagation qui sont pris en compte.

Le processus d'association étant fondé sur le regroupement de temps d'arrivée compatibles dans une fenêtre temporelle donnée, celui-ci est donc fortement soumis aux effets du milieu de propagation. Si les vitesses des ondes dans le milieu ne sont pas correctement définies et si la configuration du réseau de stations est négligée, la probabilité de créer des combinaisons de pointés avec des pointés parasites s'élève. Pour diminuer les effets liés au milieu de propagation, les distances inter-station et plusieurs vitesses de propagation des ondes sismiques ont été explorées.

Le système de détection de SeisComP3 produisant un catalogue multi-origine, une sélection préférentielle d'une origine est réalisée pour chaque événement. Afin de réaliser une sélection optimale, il s'agit de choisir l'origine qui minimise à la fois les effets du bruit et les effets du milieu de propagation. Pour cela, estimer des paramètres supplémentaires (seuil de RMS, seuil du nombre de phases, distances épicentrales, valeurs des résidus, incertitudes de localisation latitudinales et longitudinales, nombre de phases S) qui vont définir la précision de la localisation de l'origine est une étape nécessaire à la détection finale optimale des séismes de faible magnitude.

A l'issue des différentes étapes, la réduction des effets liés au bruit enregistré aux stations et au milieu de propagation améliore la détection des vrais événements mais n'empêche pas la détection des faux événements. Or, dans le cadre de ce travail de thèse, c'est facilement 50 000 faux événement qui sont détectés en 4 mois. De ce fait, même si la détection des vrais événements a été améliorée, les effets de cette amélioration sont fortement limités par la détection outrancière des faux événements.

Un deuxième problème se soulève, celui de la discrimination des événements. Discriminer un événement revient à disséquer complètement ce dernier de manière à extraire l'information sur la nature de la source qui l'a engendré. Or, un événement c'est à la fois une solution hypocentrale et épicentrale, un temps d'origine, une magnitude, une combinaison de pointés, une combinaison de signaux, mais c'est aussi des incertitudes puisque ce dernier est détecté à partir d'une source inconnue. Un événement c'est donc un très grand espace de solutions possibles difficiles à décrypter avec le seul cerveau humain. L'utilisation de l'apprentissage machine a alors permis de gérer plus efficacement cet espace des possibles.

Si les effets liés au bruit et au milieu de propagation ont fortement conditionné le succès de la détection, ces derniers effets vont aussi apporter une contrainte forte à la résolution du problème de discrimination. Or, l'apprentissage machine permet assez bien de gérer ces effets en segmentant les différentes informations dans un espace d'attributs indépendants, où chaque attribut véhicule une partie de la réponse au problème posé de la discrimination. Seulement, la difficulté ici est de trouver dans cet espace la combinaison optimale d'attributs qui va pouvoir retracer le plus fidèlement possible l'information véhiculée par la source, sans risque de sur-apprentissage ou de sous-apprentissage de l'algorithme utilisé.

De ce fait, afin d'optimiser la résolution du problème de discrimination des événements par apprentissage machine, l'interactivité Homme-machine a été privilégiée dans la construction de cette apprentissage, en plus du choix raisonnée de l'algorithme d'apprentissage et de la configuration de son espace d'hyperparamètres. Cette interactivité vise à détecter les corrélations parasites élaborées par le système d'apprentissage et estimer la validité des règles de classification émises.

La solution optimale obtenue pour discriminer les faux événements des vrais événements ne s'est pas basée majoritairement sur les caractéristiques du signal, les signaux associés à ces différents événements étant très souvent d'amplitude similaire, de durée équivalente et de contenu fréquentiel semblable.

Un faux événement est donc mieux classé à partir d'une combinaison d'attributs qui le définit comme étant un événement généré à partir d'une association incohérente de pointés (facteur de corrélation entre les premières arrivées des ondes P et la distance épicentrale), qui ont été déclenchés suite à une forte variation d'amplitude (valeur maximale de la fonction STA/LTA), à partir de signaux aléatoires (estimation de l'entropie de Shannon) et relativement stationnaires (différence discrète d'ordre 1 de l'enveloppe du signal) et dont la source est superficielle (fort degré de polarisation planaire) et mal localisée (distribution statistique des résidus, nombre de phases utilisées, distance épicentrale minimale, écart-type à partir de la distance épicentrale moyenne).

En revanche, les caractéristiques du signal dans le domaine fréquentiel (variance, nombre de pics, fréquence cumulée à 25%, fréquence cumulée à 75%, rapports spectraux entre les ondes P et S) puis le domaine temporel (coefficient d'asymétrie et d'aplatissement de la distribution des valeurs d'amplitude du signal, rapport de l'énergie du signal à différentes bandes fréquentielles) ont été largement utilisés dans la classification des séismes et des tirs de carrière. En effet, ces différentes caractéristiques expriment la nature des différentes phases sismiques qui composent les signaux associés aux séismes et aux tirs de carrière. De plus, des informations complémentaires très indirectes sur la profondeur de ces événements (différence entre magnitude de coda et magnitude locale, magnitudes de surface, rapports Z/H) ainsi que le lieu et le temps d'occurrence

des événements (proximité de l'événement à un centre urbain, donc potentiellement d'une carrière, heure et date de l'événement) viennent compléter le diagnostic.

La procédure de détection des séismes de faible magnitude qui est développée au cours de ce travail de thèse apportent des résultats prometteurs. Cette procédure détecte 2.5 fois plus de séismes dont 48% ont des magnitudes locales MLv inférieures à 1.20. Cette détection supplémentaire de séismes amène à diminuer la magnitude de complétude qui atteint une valeur de 1.10 au lieu de 1.20. A ce niveau de magnitude c'est une différence subtile mais qui catalyse un début d'infléchissement qui n'était pas encore observé malgré l'apport de nouvelles stations. L'intégration des stations AlpArray semblent apporter une plus-value qui reste à être confirmée plus largement.

De plus, la procédure de classification des événements est également prometteuse : elle élimine plus de 99% des faux événements et manque très peu de séismes (moins de 7%) parmi les vrais événements, elle discrimine aussi correctement environ 95% des tirs de carrière et 96% des séismes.

Si ces résultats sont bel et bien prometteurs, la procédure de détection développée gagnerait en robustesse si elle était testée sur d'autres jeux de données ou dans d'autres conditions de monitoring (comme par exemple sans les stations AlpArray), et si la caractérisation de l'interaction Homme-machine était plus approfondie pour augmenter les bénéfices de l'intéractivité dans l'affinage finale de la discrimination.

Ces résultats expriment aussi qu'il reste difficile de se détacher complètement des effets liés au bruit et au milieu de propagation. Si la procédure détecte plus de séismes de faible magnitude, le profil de détection des séismes maintient une périodicité apparente qui est fortement liée au niveau de bruit d'origine anthropique enregistrée au cours de la journée : le taux de séismes reste le plus élevé aux périodes de journées où le niveau de bruit est minimal.

De plus, l'utilisation de l'apprentissage machine semble mettre à jour une variabilité spatiale dans l'efficacité des attributs du signal utilisés pour discriminer les séismes et les tirs de carrière. Cela signifie que, face à un milieu de propagation hétérogène et complexe, la réponse du système semble être une régionalisation de l'effet des discriminants. Cependant, pour confirmer cela, une cartographie fine de cette variabilité spatiale est indispensable pour mieux comprendre ce que ces discriminants expriment régionalement ou plus localement, à savoir s'ils expriment une signature sismique spécifique, gouvernée par des hétérogénéités géologiques et/ou des effets localisés du bruit enregistré aux stations et/ou des effets du milieu de propagation et/ou des effets dûs à la profondeur de la source.

En effet, pouvoir distinguer explicitement dans la signature du signal sismique ce qui révèle spécifiquement de la source ou des autres effets mentionnés, serait une avancée majeure pour mieux contraindre les profondeurs hypocentrales, en donnant des informations indirectes sur la profondeur de la source, et pour clairement identifier si cette zone héberge des caractéristiques sismiques bien définies. L'identification soit de sources types associées à une zone précise, soit de formes d'ondes récurrentes ouvre alors des fenêtres d'études permettant de mieux caractériser le fonctionnement sismotectonique de la zone d'étude.

Bibliographie

Abdrakhmatov, K. E., R. T. Walker, G. E. Campbell, A. S. Carr, A. Elliott, C. Hillemann et al. (2016). « Multisegment rupture in the 11 July 1889 Chilik earthquake (Mw 8.0-8.3), Kazakh Tien Shan, interpreted from remote sensing, field survey, and paleoseismic trenching ». In : *Journal of Geophysical Research : Solid Earth* 121 (6), p. 4615-4640. doi : 10.1002/2015JB012763.

Agarwal, S. (2018). *Understanding Generalization Error : Bounds and Decompositions*.

Agnon, A. (2014). « Pre-instrumental earthquakes along the Dead Sea rift ». In : *Modern Approaches in Solid Earth Sciences* 6. doi : 10.1007/978-94-017-8872-4_8.

Akaike, H. (1971). « Autoregressive model fitting for control ». In : *Annals of the Institute of Statistical Mathematics* 23 (1), p. 163-180. doi : 10.1007/BF02479221.

Allen, R. (1978). « Automatic earthquake recognition and timing from single traces ». In : *Bulletin of the Seismological Society of America* 68, p. 1521-1532.

— (1982). « Automatic phase pickers : Their present use and future prospects ». In : *Bulletin of the Seismological Society of America* 72, p. 225-242.

Allen, T. I. (mar. 2020). « Seismic hazard estimation in stable continental regions : Does psha meet the needs for modern engineering design in Australia ? » In : *Bulletin of the New Zealand Society for Earthquake Engineering* 53 (1), p. 22-36. doi : 10.5459/BNZSEE.53.1.22-36.

Almeida, M. M., B. S. Schlüter, J. A. F. van Loenhout, S. S. Thapa, K. C. Kumar, R. Singh et al. (déc. 2020). « Changes in patient admissions after the 2015 Earthquake : a tertiary hospital-based study in Kathmandu, Nepal ». In : *Scientific Reports* 10 (1), p. 400043. doi : 10.1038/s41598-020-61901-7.

Amini, M.-R. (2015). « Apprentissage machine : De la theorie a la pratique. Concepts fondamentaux en Machine Learning. » In :

Amorese, D., J. R. Grasso et P. A. Rydelek (jan. 2010). « On varying b-values with depth : Results from computer-intensive tests for Southern California ». In : *Geophysical Journal International* 180 (1), p. 347-360. doi : 10.1111/j.1365-246X.2009.04414.x.

Arlot, S. et R. Genuer (juil. 2014). « Analysis of purely random forests bias ». In :

Atef, A. H., K. H. Liu et S. S. Gao (août 2009). « Apparent weekly and daily earthquake periodicities in the western United States ». In : *Bulletin of the Seismological Society of America* 99 (4), p. 2273-2279. doi : 10.1785/0120080217.

Atkinson, G. M. (2020). « Special section : Observations, Mechanisms, and Hazards of Induced Seismicity The Intensity of Ground Motions from Induced Earthquakes with Implications for Damage Potential ». In : *Bulletin of the Seismological Society of America* 110 (5), p. 2366-2379. doi : 10.1785/0120190166.

Atkinson, G. M., D. W. Eaton et N. Igonin (mai 2020). « Developments in understanding seismicity triggered by hydraulic fracturing ». In : *Nature Reviews Earth Environment* 1 (5), p. 264-277. doi : 10.1038/s43017-020-0049-7.

Audin, L., J.-P. Avouac, M. Flouzat et J.-L. Plantet (mar. 2002). « Fluid-driven seismicity in a stable tectonic context : The Remiremont fault zone, Vosges, France ». In : *Geophysical Research Letters* 29 (6), p. 13-1-13-4. doi : 10.1029/2001GL012988.

Avouac, J.-P. (juil. 2011). « Earthquakes : The lessons of Tohoku-Oki ». In : *Nature* 475 (7356), p. 300-301. doi : 10.1038/nature10265.

Bachmann, C. E., S. Wiemer, B. P. Goertz-Allmann et J. Woessner (2012). « Influence of pore-pressure on the event-size distribution of induced earthquakes ». In : *Geophysical Research Letters* 39 (9). doi : 10.1029/2012GL051480.

Belkin, M., D. Hsu, S. Ma et S. Mandal (août 2019). « Reconciling modern machine-learning practice and the classical bias–variance trade-off ». In : *Proceedings of the National Academy of Sciences of the United States of America* 116 (32), p. 15849-15854. doi : 10.1073/pnas.1903070116.

Bellman, R. (1961). « A Mathematical Formulation of Variational Processes of Adaptive Type ». In :

Bergen, K. J. et G. C. Beroza (mar. 2019). « Earthquake Fingerprints : Extracting Waveform Features for Similarity-Based Earthquake Detection ». In : *Pure and Applied Geophysics* 176 (3), p. 1037-1059. doi : 10.1007/s00024-018-1995-6.

Bergstra, J. et Y. Bengio (2012). « Random Search for Hyper-Parameter Optimization Yoshua Bengio ». In : *Journal of Machine Learning Research* 13, p. 281-305.

Bezada, M. J. et J. Smale (nov. 2019). « Lateral Variations in Lithospheric Mantle Structure Control the Location of Intracontinental Seismicity in Australia ». In : *Geophysical Research Letters* 46 (22), p. 12862-12869. doi : 10.1029/2019GL084848.

Bilham, R. (2010). « Lessons from the Haiti earthquake ». In : *Nature* 463 (7283), p. 878-879. doi : 10.1038/463878a.

Boncio, P., F. Liberi, M. Caldarella et F. C. Nurminen (2018). « Width of surface rupture zone for thrust earthquakes : implications for earthquake fault zoning ». In : *Hazards Earth Syst. Sci* 18, p. 241-256. doi : 10.5194/nhess-18-241-2018.

Bondar, I., S. C. Myers, E. R. Engdahl et E. A. Bergman (mar. 2004). « Epicentre accuracy based on seismic network criteria ». In : *Geophysical Journal International* 156 (3), p. 483-496. doi : 10.1111/j.1365-246X.2004.02070.x.

Bonner, J. L., D. C. Pearson et W. S. Blomberg (2003). « Azimuthal variation of short-period Rayleigh waves from cast blasts in Northern Arizona ». In : *Bulletin of the Seismological Society of America* 93 (2), p. 724-736. doi : 10.1785/0120020115.

Bouchon, M., V. Durand, D. Marsan, H. Karabulut et S. J. (2013). « The long precursory phase of most large interplate earthquakes ». In : *Nature Geoscience* 6 (4), p. 299-302. doi : 10.1038/ngeo1770.

Bouchon, M., H. Karabulut, M. Aktar, S. Ozalaybey, J. Schmittbuhl et M.-P. Bouin (2011). « Extended Nucleation of the 1999 M w 7.6 Izmit Earthquake ». In :

Bourouis, S. et P. Bernard (mai 2007). « Evidence for coupled seismic and aseismic fault slip during water injection in the geothermal site of Soultz (France), and implications for seismogenic transients ». In : *Geophysical Journal International* 169 (2), p. 723-732. doi : 10.1111/j.1365-246X.2006.03325.x.

Bowman, J., G. Gibson et T. Jones (1990). « Aftershocks of the 1988 January 22 Tennant Creek, Australia intraplate earthquakes : evidence for a complex thrust-fault geometry ». In : *Geophys. J. Int* 100, p. 87-97.

Bratt, S. et C. Bache (1988). « Location Estimation Using Regional Array Data ». In : p. 780-798.

Breiman, L. (oct. 2001). « Random forests ». In : *Machine Learning* 45 (1), p. 5-32. doi : 10.1023/A:1010933404324.

Brodsky, E. E. (oct. 2019a). « Determining whether the worst earthquake has passed ». In : *Nature* 574 (7777), p. 185-186. doi : 10.1038/d41586-019-02972-z.

Brodsky, E. E. et T. Lay (2014). « Recognizing foreshocks from the 1 April 2014 Chile earthquake ». In : *Science* 344 (6185), p. 700-702. doi : 10.1126/science.1255202.

Brodsky, E. E. (2019b). « The importance of studying small earthquakes ». In : *Science* 364 (6442), p. 736-737. doi : 10.1126/science.aax2490.

Bulut, F., M. Bohnhoff, M. Aktar et G. Dresen (2007). « Characterization of aftershock-fault plane orientations of the 1999 İzmit (Turkey) earthquake using high-resolution aftershock locations ». In : *Geophysical Research Letters* 34 (20), p. L20306. doi : 10.1029/2007GL031154.

Bungum, H., E. S. Husebye et F. Ringdal (1971). « The NORSAR Array and Preliminary Results of Data Analysis ». In : *Geophysical Journal of the Royal Astronomical Society* 25 (3), p. 115-126. doi : 10.1111/j.1365-246X.1971.tb02334.x.

Calais, E., A. M. Freed, R. Van Arsdale et S. Stein (juil. 2010). « Triggering of New Madrid seismicity by late-Pleistocene erosion ». In : *Nature* 466 (7306), p. 608-611. doi : 10.1038/nature09258.

Camelbeeck, T., K. Vanneste, P. Alexandre, K. Verbeeck, T. Petermans, P. Rosset et al. (2007). « Relevance of active faulting and seismicity studies to assessments of long-term earthquake activity and maximum magnitude in intraplate northwest Europe, between the Lower Rhine Embayment and the North Sea ». In : *Special Paper of the Geological Society of America* 425, p. 193-224. doi : 10.1130/2007.2425(14).

Cappa, F., L. De Barros, N. Wynants-Morel, Y. Guglielmi, J. Birkholzer et C. Nussbaum (2019). « From Aseismic Slip to Seismicity During Fluid Injection Controlled by Interactions Between Stress Perturbation, Permeability Increases and Fault Structure ». In :

Chang, C. H., Y. M. Wu, L. Zhao et F. T. Wu (août 2007). « Aftershocks of the 1999 Chi-Chi, Taiwan, earthquake : The first hour ». In : *Bulletin of the Seismological Society of America* 97 (4), p. 1245-1258. doi : 10.1785/0120060184.

Chang, K. W., H. Yoon, Y. H. Kim et M. Y. Lee (déc. 2020). « Operational and geological controls of coupled poroelastic stressing and pore-pressure accumulation along faults : Induced earthquakes in Pohang, South Korea ». In : *Scientific Reports* 10 (1), p. 1-12. doi : 10.1038/s41598-020-58881-z.

Chen, P.-F., P.-L. Su, Y.-L. Chen, Y.-D. Peng et L.-F. Chen (2019). « Hualien earthquake sequence through catalog compilation ». In : *Terr. Atmos. Ocean. Sci* 30, p. 399-409. doi : 10.3319/TAO.2018.11.15.02.

Chen, X. et P. M. Shearer (jan. 2016). « Analysis of Foreshock Sequences in California and Implications for Earthquake Triggering ». In : *Pure and Applied Geophysics* 173 (1), p. 133-152. doi : 10.1007/s00024-015-1103-0.

Chen, Y., M. Liu et G. Luo (juin 2020). « Complex temporal patterns of large earthquakes : Devil's staircases ». In : *Bulletin of the Seismological Society of America* 110 (3), p. 1064-1076. doi : 10.1785/0120190148.

Chin, C. S. a. (2011). « The origin of the Haitian cholera outbreak strain ». In : *National English Journal of Medecine*, p. 33-42.

Christensen, B. C. et J. F. Blanco Chia (2017). « Raspberry Shake- A World-Wide Citizen Seismograph Network ». In : *AGUFM* 2017.

Chui, G. (oct. 2009). « Shaking up earthquake theory : Geological faults are not behaving as scientists once expected. Glennda Chui reports on efforts to forge a new understanding of quake behaviour ». In : *Nature* 461 (7266), p. 870-873.

Clark, D. J., S. Brennand, G. Brenn, M. C. Garthwaite, J. Dimech, T. I. Allen et al. (avr. 2020). « Surface deformation relating to the 2018 Lake Muir earthquake sequence, southwest Western Australia : new insight into stable continental region earthquakes ». In : *Solid Earth* 11 (2), p. 691-717. doi : 10.5194/se-11-691-2020.

Clark, D. et A. Mcpherson (2011). « Australia's seismogenic neotectonic record : A case for heterogeneous intraplate deformation Southern Thomson Project View project National seismic hazard assessment of Australia View project ». In :

Clark, D., A. McPherson, T. Allen et M. De Kool (fév. 2014). « Coseismic surface deformation caused by the 23 March 2012 Mw 5.4 Ernabella (Pukatja) earthquake, central Australia : Implications for fault scaling relations in cratonic settings ». In : *Bulletin of the Seismological Society of America* 104 (1), p. 24-39. doi : 10.1785/0120120361.

Clark, D., A. McPherson et R. Van Dissen (2012). « Long-term behaviour of Australian stable continental region (SCR) faults ». In : *Tectonophysics* 566-567, p. 1-30. doi : 10.1016/j.tecto.2012.07.004.

Clarke, H., J. P. Verdon, T. Kettlety, A. F. Baird et J. M. Kendall (sept. 2019). In : *Seismological Research Letters* 90 (5), p. 1902-1915. doi : 10.1785/0220190110.

Clayton, R. W., T. Heaton, M. Kohler, C. M., G. R. et J. Bunn (sept. 2015). « Community seismic network : A dense array to sense earthquake strong motion ». In : *Seismological Research Letters* 86 (5), p. 1354-1363. doi : 10.1785/0220150094.

Cliet, C. et M. Dubesset (fév. 1987). « THREE-COMPONENT RECORDINGS : INTEREST FOR LAND SEISMIC SOURCE STUDY. » In : *Geophysics* 52 (8), p. 1048-1059. doi : 10.1190/1.1442370.

Cochran, E. Z., A. Kohler, D. D. Given, S. Guiwits, J. Andrews, M.-M. Meier et al. (jan. 2018). « Earthquake early warning shakealert system : Testing and certification platform ». In : *Seismological Research Letters* 89 (1), p. 108-117. doi : 10.1785/0220170138.

Crotwell, H. P., T. J. Owens et J. Ritsema (mar. 1999). « The TauP Toolkit : Flexible Seismic Travel-time and Ray-path Utilities ». In : *Seismological Research Letters* 70 (2), p. 154-160. doi : 10.1785/gssrl.70.2.154.

De Barros, L., C. J. Bean, M. Zecevic, F. Brenguier et A. Peltier (sept. 2013). « Eruptive fracture location forecasts from high-frequency events on Piton de la Fournaise Volcano ». In : *Geophysical Research Letters* 40 (17), p. 4599-4603. doi : 10.1002/grl.50890.

De Barros, L., F. Cappa, A. Deschamps et P. Dublanchet (2020). « Imbricated Aseismic Slip and Fluid Diffusion Drive a Seismic Swarm in the Corinth Gulf, Greece ». In : *Geophysical Research Letters* 47 (9). doi : 10.1029/2020GL087142.

De Barros, L., Y. Guglielmi, D. Rivet, F. Cappa et L. Duboeuf (déc. 2018). « Seismicity and fault aseismic deformation caused by fluid injection in decametric in-situ experiments ». In : *Comptes Rendus - Geoscience* 350 (8), p. 464-475. doi : 10.1016/j.crte.2018.08.002.

Delahaye, E. J., J. Townend, M. E. Reyners et G. Rogers (2009). « Microseismicity but no tremor accompanying slow slip in the Hikurangi subduction zone, New Zealand ». In : *Earth and Planetary Science Letters* 277 (1-2), p. 21-28. doi : 10.1016/j.epsl.2008.09.038.

Díaz, J., M. Ruiz, P. S. Sánchez-Pastor et P. Romero (2017). « Urban Seismology : On the origin of earth vibrations within a city ». In : *Scientific Reports* 7.15296. doi : 10.1038/s41598-017-15499-y.

Dickey, J., B. Borghetti, W. Junek et R. Martin (jan. 2019). « Beyond correlation : A path-invariant measure for seismogram similarity ». In : *Seismological Research Letters* 91 (1), p. 356-369. doi : 10.1785/0220190090.

Ding, K., J. T. Freymueller, P. He, Q. Wang et C. Xu (2019). « Glacial Isostatic Adjustment, Intraplate Strain, and Relative Sea Level Changes in the Eastern United States ». In : *Journal of Geophysical Research : Solid Earth* 124 (6), p. 6056-6071. doi : 10.1029/2018JB017060.

Dolan, J. F., D. D. Bowman et C. G. Sammis (2007). « Long-range and long-term fault interactions in Southern California ». In : *Geology* 35 (9), p. 855-858. doi : 10.1130/G23789A.1.

Doshi-Velez, F. et B. Kim (fév. 2017). « Towards A Rigorous Science of Interpretable Machine Learning ». In :

Dowla, F., S. Taylor et S. Anderson (1990). « Seismic discrimination with artificial neural networks : Preliminary results with regional spectral data ». In :

Draelos, T. J., M. G. Peterson, H. A. Knox, B. J. Lawry, K. E. Phillips-Alonge, A. E. Ziegler et al. (2018). « Dynamic tuning of seismic signal detector trigger levels for local networks ». In : *Bulletin of the Seismological Society of America* 108.3, p. 1346-1354. doi : 10.1785/0120170200.

Drouin, A., G. Letarte, F. Raymond, M. Marchand, J. Corbeil et F. Laviolette (2019). « Interpretable genotype-to-phenotype classifiers with performance guarantees ». In : *Scientific Reports* 9 (1), p. 1-13. doi : 10.1038/s41598-019-40561-2.

Ellsworth, W. L., M. V. Matthews, R. M. Nadeau, S. P. Nishenko, P. A. Reasenberg et R. W. Simpson (1999). « A Physically Based Earthquake Recurrence Model for Estimation of Long-Term Earthquake Probabilities ». In :

Ende, M. P. A. et J.-P. Ampuero (fév. 2020). « On the Statistical Significance of Foreshock Sequences in Southern California ». In : *Geophysical Research Letters* 47 (3). doi : 10.1029/2019GL086224.

Enescu, B., J. Mori et M. Miyazawa (avr. 2007). « Quantifying early aftershock activity of the 2004 mid-Niigata Prefecture earthquake (Mw 6.6) ». In : *Journal of Geophysical Research : Solid Earth* 112 (B4). doi : 10.1029/2006JB004629.

Ester, M., H.-P. Kriegel, J. Sander et X. Xu (1996). « A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise ». In :

Eyre, R., F. De Luca et F. Simini (déc. 2020). « Social media usage reveals recovery of small businesses after natural hazard events ». In : *Nature Communications* 11 (1), p. 1-10. doi : 10.1038/s41467-020-15405-7.

Eyre, T. S., D. W. Eaton, M. Zecevic, D. D'Amico et D. Kolos (juil. 2019). In : *Geophysical Journal International* 218 (1), p. 534-546. doi : 10.1093/gji/ggz168.

Fah, D., M. Gisler, B. Jaggi, P. Kastli, T. Lutz, V. Masciadri et al. (2009). « The 1356 Basel earthquake : an interdisciplinary revision ». In :

Geophys. J. Int 178, p. 351-374. DOI : 10.1111/j.1365-246X.2009.04130.x.

FAH, D., J. R. MOORE, J. BURJANEK, I. IOSIFESCU, L. DALGUER, F. DUPRAY et al. (2012). « Coupled seismogenic geohazards in Alpine regions ». In : *Bollettino di Geofisica Teorica ed Applicata* 53 (4), p. 485-508. DOI : 10.4430/bgta0048.

FIEDLER, B., S. HAINZL, G. ZOLLER et M. HOLSCHNEIDER (oct. 2018). « Detection of gutenberg-richter b-value changes in earthquake time series ». In : *Bulletin of the Seismological Society of America* 108 (5), p. 2778-2787. DOI : 10.1785/0120180091.

FLETCHER, J. et O. J. TERAN (2017). « The role of a keystone fault in triggering the complex El Mayor-Cucapah earthquake rupture ». In : DOI : 10.1038/ngeo2660.

FRECHET, J. (1978). « Sismicité du Sud-Est de la France et une nouvelle méthode de zonage sismique ». In :

FROHLICH, C. et S. D. DAVIS (jan. 1993). « Teleseismic b values ; or, much ado about 1.0 ». In : *Journal of Geophysical Research* 98 (B1), p. 631-644. DOI : 10.1029/92JB01891.

FU, J., X. WANG, Z. LI, H. MENG, J. WANG, W. WANG et al. (2019). « Automatic phase-picking method for detecting earthquakes based on the signal-to-noise-ratio concept ». In : *Seismological Research Letters* 91 (1), p. 334-342. DOI : 10.1785/0220190043.

FUCHS, F., F. M. SCHNEIDER, P. KOLINSKY, S. SERAFIN et G. BOKELMANN (déc. 2019). « Rich observations of local and regional infrasound phases made by the AlpArray seismic network after refinery explosion ». In : *Scientific Reports* 9 (1), p. 1-14. DOI : 10.1038/s41598-019-49494-2.

GAGNEPAIN-BEYNEIX, J., H. HAESSLER et T. MODIANO (mai 1982). « The pyrenean earthquake of february 29, 1980 : An example of complex faulting ». In : *Tectonophysics* 85 (3-4), p. 273-290. DOI : 10.1016/0040-1951(82)90106-8.

GALLEN, S. F. et J. R. THIGPEN (2018). « Lithologic Controls on Focused Erosion and Intraplate Earthquakes in the Eastern Tennessee Seismic Zone ». In : *Geophysical Research Letters* 45 (18), p. 9569-9578. DOI : 10.1029/2018GL079157.

GARDONIO, B., R. JOLIVET, E. CALAIS et H. LECLERE (2018). « The April 2017 Mw 6.5 Botswana Earthquake : An Intraplate Event Triggered by Deep Fluids ». In : *Geophysical Research Letters* 45 (17), p. 8886-8896. DOI : 10.1029/2018GL078297.

GENTILI, S. et A. MICHELINI (2006). « Automatic picking of P and S phases using a neural tree ». In : *Journal of Seismology* 10 (1), p. 39-63. DOI : 10.1007/s10950-006-2296-6.

GILPIN, L. H., D. BAU, B. Z. YUAN, A. BAJWA, M. SPECTER et L. KAGAL (mai 2018). « Explaining Explanations : An Overview of Interpretability of Machine Learning ». In : *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018*, p. 80-89.

GITTERMAN, Y., V. PINSKY et A. SHAPIRA (1998). « Spectral classification methods in monitoring small local events by the Israel seismic network ». In : *Journal of Seismology* 2 (3), p. 237-256. DOI : 10.1023/A:1009738721893.

GODANO, C., E. LIPPIELLO et L. DE ARCANGELIS (2014). « Variability of the b value in the Gutenberg-Richter distribution ». In : *Geophysical Journal International* 199 (3). DOI : 10.1093/gji/ggu359.

GOEBEL, T. H. W., D. SCHORLEMMER, T. W. BECKER, G. DRESEN et C. G. SAMMIS (2013). « Acoustic emissions document stress changes over many seismic cycles in stick-slip experiments ». In : *Geophysical Research Letters* 40 (10), p. 2049-2054. DOI : 10.1002/grl.50507.

GOMBERG, J. S., K. M. SHEDLOCK et S. W. ROECKER (1990). « THE EFFECT OF S-WAVE ARRIVAL TIMES ON THE ACCURACY OF HYPOCENTER ESTIMATION ». In : *Bulletin of the Seismological Society of America* 80 (6), p. 1605-1628.

GREENHALGH, S., D. SOLLBERGER, C. SCHMELZBACH et M. RUTTY (jan. 2018). « Single-station polarization analysis applied to seismic wavefields : A tutorial ». In : *Advances in Geophysics* 59, p. 123-170. DOI : 10.1016/bs.agph.2018.09.002.

GRIFFI, J. D., T. I. ALLE et M. C. GERSTENBERGE (mar. 2020). « Seismic hazard assessment in australia : Can structured expert elicitation achieve consensus in the "land of the Fair Go" ? » In : *Seismological Research Letters* 91 (2), p. 859-873. DOI : 10.1785/0220190186.

GRIGOLI, F., S. CESCA, M. VASSALLO et T. DAHM (juil. 2013). « Automated seismic event location by travel-time stacking : An application to mining induced seismicity ». In : *Seismological Research Letters* 84 (4), p. 666-677. DOI : 10.1785/0220120191.

GRIGOLI, F., L. SCARABELLO, M. BOSE, B. WEBER, S. WIEMER et J. F. CLINTON (2018). « Pick- and waveform-based techniques for real-time detection of induced seismicity ». In : *Geophysical Journal International* 213 (2), p. 868-884. DOI : 10.1093/gji/ggy019.

GROOS, J. C. et J. R. R. RITTER (2009). « Time domain classification and quantification of seismic noise in an urban environment ». In : *Geophysical Journal International* 179 (2), p. 1213-1231. DOI : 10.1111/j.1365-246X.2009.04343.x.

GRUNBERG, M., C. GRELIER, R. PESTOURIE et F. ENGELS (2018). « French National Network of Seismic Survey (ReNaSS) contributions to SeisComP3 : application to parameter fine tuning, real time multiple origins fusion and data parallelism to process huge dataset. » In : *American Geophysical Union*.

GRUTZNER, C., E. CARSON, R. T. WALKER, E. J. RHODES, A. MUKAMBAYEV, D. MACKENZIE et al. (2017). « Assessing the activity of faults in continental interiors : Palaeoseismic insights from SE Kazakhstan ». In : *Earth and Planetary Science Letters* 459, p. 93-104. DOI : 10.1016/j.epsl.2016.11.025.

Guglielmi, Y., F. Cappa, J.-P. Avouac, P. Henry et D. Elsworth (juin 2015). « Seismicity triggered by fluid injection-induced aseismic slip ». In : *Science* 348 (6240), p. 1224-1226. doi : 10.1126/science.aab0476.

Gulia, L., T. Tormann, S. Wiemer, M. Herrmann et S. Seif (fév. 2016). « Short-term probabilistic earthquake risk assessment considering time-dependent b values ». In : *Geophysical Research Letters* 43 (3), p. 1100-1108. doi : 10.1002/2015GL066686.

Gulia, L. et S. Wiemer (oct. 2019). « Real-time discrimination of earthquake foreshocks and aftershocks ». In : *Nature* 574 (7777), p. 193-199. doi : 10.1038/s41586-019-1606-4.

Gurrola, B. H., J. B. Minster, H. Given, F. Vernon et R. Aster (1990). « ANALYSIS OF HIGH-FREQUENCY SEISMIC NOISE IN THE WESTERN UNITED STATES AND EASTERN KAZAKHSTAN ». In : *Bulletin of the Seismological Society of America* 80 (4), p. 951-970.

Gutenberg, B. et C. F. Richter (1944). « Frequency of earthquakes in California ». In : *Bulletin of the Seismological Society of America* 34, p. 185-188.

Haessler, H. et H. Hoang-Trong (1985). « La crise sismique de Remiremont (Vosges) de decembre 1984 : implications tectoniques regionales ». In :

Haessler, H., P. Hoang-Trong, R. Schick, G. Schneider et K. Strobach (sept. 1980). « The September 3, 1978, Swabian Jura earthquake ». In : *test Tectonophysics* 68 (1-2), p. 1-14. doi : 10.1016/0040-1951(80)90005-0.

Hainzl, S. (2004). « Seismicity patterns of earthquake swarms due to fluid intrusion and stress triggering ». In : *Geophysical Journal International* 159 (3), p. 1090-1096. doi : 10.1111/j.1365-246X.2004.02463.x.

Hainzl, S., T. Fischer et T. Dahm (2012). « Seismicity-based estimation of the driving fluid pressure in the case of swarm activity in Western Bohemia ». In : *Geophysical Journal International* 191 (1), p. 271-281. doi : 10.1111/j.1365-246X.2012.05610.x.

Hanks, T. C. (juin 1992). « Small earthquakes, tectonic forces ». In : *Science* 256 (5062), p. 1430-1432. doi : 10.1126/science.256.5062.1430.

Hao, J., J. Zhang et Z. Yao (oct. 2019). « Evidence for diurnal periodicity of earthquakes from midnight to daybreak ». In : *National Science Review* 6 (5), p. 1016-1023. doi : 10.1093/nsr/nwy117.

Harris, D. B. (2006). « Subspace Detectors : Theory ». In : doi : 10.2172/900081.

Hatch, R. L., R. E. Abercrombie, C. J. Ruhl et K. D. Smith (2020). « Evidence of Aseismic and Fluid-Driven Processes in a Small Complex Seismic Swarm Near Virginia City, Nevada ». In : *Geophysical Research Letters* 47 (4). doi : 10.1029/2019GL085477.

Helmstetter, A. (2005). « Importance of small earthquakes for stress transfers and earthquake triggering ». In : *Journal of Geophysical Research* 110 (B5). doi : 10.1029/2004JB003286.

Henrion, E., F. Masson, C. Doubre, P. Ulrich et M. Meghraoui (oct. 2020). « Present-day deformation in the Upper Rhine Graben from GNSS

data ». In : *Geophysical Journal International* 223 (1), p. 599-611. DOI : 10.1093/gji/ggaa320.

HETÉNYI, G., I. MOLINARI, J. CLINTON, G. BOKELMANN, I. BONDÁR, W. C. CRAWFORD et al. (2018). « The AlpArray Seismic Network : A Large-Scale European Experiment to Image the Alpine Orogen ». In : *Surveys in Geophysics* 39.5, p. 1009-1033. DOI : 10.1007/s10712-018-9472-4.

HIROSE, H., T. MATSUZAWA, T. KIMURA et H. KIMURA (2014). « The Boso slow slip events in 2007 and 2011 as a driving process for the accompanying earthquake swarm ». In : *Geophysical Research Letters* 41 (8), p. 2778-2785. DOI : 10.1002/2014GL059791.

HOLT, M. M., K. D. KOPER, W. YECK, D. S., Z. LI, J. MARK HALE et al. (oct. 2019). « On the Portability of ML–Mc as a Depth Discriminant for Small Seismic Events Recorded at Local Distances ». In : *Bulletin of the Seismological Society of America* 109 (5), p. 1661-1673. DOI : 10.1785/0120190096.

HSU, Y. J., M. SIMONS, J.-P. AVOUAC, J. GALETEKA, K. SIEH, M. CHLIEH et al. (juin 2006). « Frictional afterslip following the 2005 Nias-Simeulue earthquake, Sumatra ». In : *Science* 312 (5782), p. 1921-1926. DOI : 10.1126/science.1126960.

HUSEN, S. et J. L. HARDEBECK (2010). « Theme IV - Understanding Seismicity Catalogs and their Problems Earthquake Location Accuracy ». In : DOI : 10.5078/corssa-55815573.

HUSEN, S., E. KISSLING, N. DEICHMANN, S. WIEMER, D. GIARDINI et M. BAER (2003). « Probabilistic earthquake location in complex three-dimensional velocity models : Application to Switzerland ». In : *Journal of Geophysical Research : Solid Earth* 108 (B2). DOI : 10.1029/2002jb001778.

HUTTON, K., J. WOESSNER et E. HAUKSSON (2010). « Earthquake monitoring in southern California for seventy-seven years (1932-2008) ». In : *Bulletin of the Seismological Society of America* 100.2, p. 423-446. DOI : 10.1785/0120090130.

IMPROTA, L., D. LATORRE, L. MARGHERITI, A. NARDI, A. MARCHETTI, A. M. LOMBARDI et al. (déc. 2019). « Multi-segment rupture of the 2016 Amatrice-Visso-Norcia seismic sequence (central Italy) constrained by the first high-quality catalog of Early Aftershocks ». In : *Scientific Reports* 9 (1), p. 1-13. DOI : 10.1038/s41598-019-43393-2.

INBAL, A., T. CRISTEA-PLATON, J. P. AMPUERO, G. HILLERS, D. AGNEW et S. E. HOUGH (2018). « Sources of Long-Range Anthropogenic Noise in Southern California and Implications for Tectonic Tremor Detection ». In : *Bulletin of the Seismological Society of America* 108.6, p. 3511-3527. DOI : 10.1785/0120180130.

ISHIMOTO, M. et K. IIDA (1939). « Observations of Earthquakes Registered with the Micro Seismograph Constructed Recently ». In : *Bulletin of the Earthquake Research Institute* 17, p. 443-478.

ISRAELSSON, H. (1990). « CORRELATION OF WAVEFORMS FROM CLOSELY SPACED REGIONAL EVENTS ». In : *Bulletin of the Seismological Society of America* 80 (6), p. 2177-2193.

Jamtveit, B., Y. Ben-Zion, F. Renard et H. Austrheim (2018). « Earthquake-induced transformation of the lower crust ». In : *Nature* 556 (7702), p. 487-491. doi : 10.1038/s41586-018-0045-y.

Johnson, C. E. (mar. 2020). « Beyond earthworm : Keeping the promise ». In : *Seismological Research Letters* 91 (2 A), p. 581-584. doi : 10.1785/0220190198.

Johnson, C. E., A. G. Lindh et B. Hirshorn (1997). « Robust Regional Phase Association ». In : *Open-File Report*. doi : 10.3133/OFR94621.

Johnson, C., A. Bittenbinder, B. Bogaert, L. Dietz et W. Kohler (1995). « Earthworm : a flexible approach to seismic network processing. » In : *IRIS newsletter* 14 (2), p. 1-4.

Johnson, J. M. et T. M. Khoshgoftaar (déc. 2019). « Survey on deep learning with class imbalance ». In : *Journal of Big Data* 6 (1), p. 27. doi : 10.1186/s40537-019-0192-5.

Jousset, P., T. Reinsch, T. Ryberg, H. Blanck, A. Clarke, R. Aghayev et al. (2018). « Dynamic strain determination using fibre-optic cables allows imaging of seismological and structural features ». In : *Nature Communications* 9 (1), p. 1-11. doi : 10.1038/s41467-018-04860-y.

Kagan, Y. Y. (1999). « Is Earthquake Seismology a Hard, Quantitative Science ? » In : *Seismicity Patterns, their Statistical Significance and Physical Meaning*, p. 233-258. doi : 10.1007/978-3-0348-8677-2_3.

— (nov. 2002). « Modern California earthquake catalogs and their comparison ». In : *Seismological Research Letters* 73 (6), p. 921-929. doi : 10.1785/gssrl.73.6.921.

— (août 2004). « Short-term properties of earthquake catalogs and models of earthquake source ». In : *Bulletin of the Seismological Society of America* 94 (4), p. 1207-1228. doi : 10.1785/012003098.

— (2010). « Statistical distributions of earthquake numbers : Consequence of branching process ». In : *Geophysical Journal International* 180 (3), p. 1313-1328. doi : 10.1111/j.1365-246X.2009.04487.x.

Kagan, Y. Y., D. D. Jackson et R. Geller (nov. 2012). « Characteristic earthquake model, 1884-2011, R.I.P. » In : *Seismological Research Letters* 83 (6), p. 951-953. doi : 10.1785/0220120107.

Kao, H. et S. J. Shan (2004). « The Source-Scanning Algorithm : Mapping the distribution of seismic sources in time and space ». In : *Geophysical Journal International* 157 (2), p. 589-594. doi : 10.1111/j.1365-246X.2004.02276.x.

Kato, A. et S. Nakagawa (2014). « Multiple slow-slip events during a foreshock sequence of the 2014 Iquique, Chile Mw 8.1 earthquake ». In : *Geophysical Research Letters* 41 (15), p. 5420-5427. doi : 10.1002/2014GL061138.

Kato, H., K. Obara, T. Igarashi, H. Tsuruoka, S. Nakagawa et N. Hirata (2012). « Propagation of slow slip leading up to the 2011 Mw 9.0 Tohoku-Oki earthquake ». In : *Science* 335 (6069), p. 705-708. doi : 10.1126/science.1215141.

KERANEN, K. M. et M. WEINGARTEN (2018). « Induced Seismicity ». In : *Annual Review of Earth and Planetary Sciences* 46 (1), p. 149-174. DOI : 10.1146/annurev-earth-082517-010054.

KING, C., A. QUIGLEY et D. CLARK (2019). « Surface-Rupturing Historical Earthquakes in Australia and Their Environmental Effects : New Insights from Re-Analyses of Observational Data ». In : *Geosciences* 9 (10), p. 408. DOI : 10.3390/geosciences9100408.

KISSLING, E., U. KRADOLFER et H. MAURER (1995). « Velest User's Guide ». In : *Int. Report, Inst. Geophys.*, p. 1-26.

KLOSE, C. D. (sept. 2010). « Human-triggered earthquakes and their impacts on human security ». In : *Nature Precedings*. DOI : 10.1038/npre.2010.4745.3.

KOHOUTOVA, L., J. HEO, S. CHA, S. LEE, T. MOON, T. D. WAGER et al. (avr. 2020). « Toward a unified framework for interpreting machine-learning models in neuroimaging ». In : *Nature Protocols* 15 (4), p. 1399-1435. DOI : 10.1038/s41596-019-0289-5.

KONG, Q., D. T. TRUGMAN, Z. E. ROSS, M. J. BIANCO, B. J. MEADE et P. GERSTOFT (2019). « Machine Learning in Seismology : Turning Data into Insights ». In : *Seismological Research Letters* 90.1, p. 3-14. DOI : 10.1785/0220180259.

KOPER, K. D., J. C. PECHMANN, R. BURLACU, K. L. PANKOW, J. STEIN, J. M. HALE et al. (oct. 2016). « Magnitude-based discrimination of man-made seismic events from naturally occurring earthquakes in Utah, USA ». In : *Geophysical Research Letters* 43 (20), p. 10, 638-10, 645. DOI : 10.1002/2016GL070742.

KOPPEN, M. (2000). « The curse of dimensionality ». In :

KRADOLFER, U. et M. BAER (1987). « Automatic phase pickers : their present use and future prospects ». In : *Bull. Seism. Soc. Am.* 72, p. 225-242.

KRAFT, T., P.-M. MAI, S. WIEMER, N. DEICHMANN, J. RIPPERGER, P. KASTLI et al. (août 2009). « Enhanced Geothermal Systems : Mitigating Risk in Urban Areas ». In : *Eos, Transactions American Geophysical Union* 90 (32), p. 273. DOI : 10.1029/2009EO320001.

KUMAR, G. P., P. MAHESH, M. NAGAR, E. MAHENDER, V. KUMAR, K. MOHAN et al. (2017). « Role of deep crustal fluids in the genesis of intraplate earthquakes in the Kachchh region, northwestern India ». In : *Geophysical Research Letters* 44 (9), p. 4054-4063. DOI : 10.1002/2017GL072936.

KUPERKOCH, L., T. MEIER, A. BRUSTLE, J. LEE et W. FRIEDERICH (2012). « Automated determination of S-phase arrival times using autoregressive prediction : Application to local and regional distances ». In : *Geophysical Journal International* 188 (2), p. 687-702. DOI : 10.1111/j.1365-246X.2011.05292.x.

KWIATEK, G. et Y. BEN-ZION (2016). « Theoretical limits on detection and analysis of small earthquakes ». In : *Journal of Geophysical Research : Solid Earth* 121 (8), p. 5898-5916. DOI : 10.1002/2016JB012908.

LAHR, J. C. (1989). « HYPOELLIPSE/version 2.0 ; a computer program for determining local earthquake hydrocentral parameters, magnitude, and first

motion pattern ». In : *Open-File Report*, p. 89-116. DOI : 10.3133/OFR89116.

LAPUSCHKIN, S., S. WALDCHEN, A. BINDER, G. MONTAVON, W. SAMEK et K. R. MÜLLER (déc. 2019). « Unmasking Clever Hans predictors and assessing what machines really learn ». In : *Nature Communications* 10 (1), p. 1-8. DOI : 10.1038/s41467-019-08987-4.

LAY, T. (mar. 2012). « Seismology : Why giant earthquakes keep catching us out ». In : *Nature* 483 (7388), p. 149-150. DOI : 10.1038/483149a.

LAY, T. et H. KANAMORI (2011). « Insights from the great 2011 Japan earthquake Additional resources for Physics Today ». In : *Citation : Phys. Today* 64 (12), p. 33. DOI : 10.1063/PT.3.1361.

LECLERE, H. et E. CALAIS (2019). « A Parametric Analysis of Fault Reactivation in the New Madrid Seismic Zone : The Role of Pore Fluid Overpressure ». In : *Journal of Geophysical Research : Solid Earth* 124 (10), p. 10630-10648. DOI : 10.1029/2018JB017181.

LEE, E. J., D. MU, W. WANG et P. CHEN (août 2020). « Weighted template-matching algorithm (Wtma) for improved foreshock detection of the 2019 ridgecrest earthquake sequence ». In : *Bulletin of the Seismological Society of America* 110 (4), p. 1832-1844. DOI : 10.1785/0120200020.

LEE, W. H. K. et J. C. LAHR (1972). « HYP071 : A COMPUTER PROGRAM FOR DETERMINING HYPOCENTER, MAGNITUDE, AND FIRST MOTION PATTERN OF LOCAL EARTHQUAKES* ». In :

LEI, X., D. HUANG, J. SU, G. JIANG, X. WANG, H. WANG et al. (déc. 2017). « Fault reactivation and earthquakes with magnitudes of up to Mw4.7 induced by shale-gas hydraulic fracturing in Sichuan Basin, China ». In : *Scientific Reports* 7 (1), p. 1-12. DOI : 10.1038/s41598-017-08557-y.

LEI, X., Z. WANG et J. SU (mai 2019). « The December 2018 ML 5.7 and January 2019 mL 5.3 earthquakes in South Sichuan basin induced by shale gas hydraulic fracturing ». In : *Seismological Research Letters* 90 (3), p. 1099-1110. DOI : 10.1785/0220190029.

LEVANDOWKI, W., R. B. HERRMANN, R. BRIGGS, B. O. et G. R. (2018). « An updated stress map of the continental United States reveals heterogeneous intraplate stress ». In : *Nature Geoscience* 11, p. 433-437. DOI : 10.1038/s41561-018-0120-x.

LEVANDOWSKI, W., M. ZELLMAN et R. BRIGGS (avr. 2017). « Gravitational body forces focus North American intraplate earthquakes ». In : *Nature Communications* 8 (1), p. 14314. DOI : 10.1038/ncomms14314.

LEWIS, M. A. et P. GERSTOFT (2012). « Shear wave anisotropy from cross-correlation of seismic noise in the Parkfield pilot hole ». In : *Geophysical Journal International* 188 (2), p. 626-630. DOI : 10.1111/j.1365-246X.2011.05285.x.

LI, Z., M. A. MEIER, E. HAUKSSON, Z. ZHAN et J. ANDREWS (2018). « Machine Learning Seismic Wave Discrimination : Application to Earthquake Early Warning ». In : *Geophysical Research Letters* 45.10, p. 4773-4779. DOI : 10.1029/2018GL077870.

Lindsey N., J., E. R. Martin, D. S. Dreger, B. Freifeld, S. Cole, S. R. James et al. (déc. 2017). « Fiber-Optic Network Observations of Earthquake Wavefields ». In : *Geophysical Research Letters* 44 (23), p. 11, 792-11, 799. doi : 10.1002/2017GL075722.

Linville, L., K. Pankow et T. Draelos (avr. 2019). « Deep Learning Models Augment Analyst Decisions for Event Discrimination ». In : *Geophysical Research Letters* 46 (7), p. 3643-3651. doi : 10.1029/2018GL081119.

Lippiello, E., F. Giacco, W. Marzocchi, C. Godano et L. De Arcangelis (oct. 2015). « Mechanical origin of aftershocks ». In : *Scientific Reports* 5 (1), p. 1-7. doi : 10.1038/srep15560.

Liu, M. et S. Stein (2016). « Mid-continental earthquakes : Spatiotemporal occurrences, causes, and hazards ». In : *Earth-Science Reviews* 162, p. 364-386. doi : 10.1016/j.earscirev.2016.09.016.

Llenos, A. L., J. J. McGuire et Y. Ogata (2009). « Modeling seismic swarms triggered by aseismic transients ». In : *Earth and Planetary Science Letters* 281 (1-2), p. 59-69. doi : 10.1016/j.epsl.2009.02.011.

Loer, K., N. Riahi et E. H. Saenger (2018). « Three-component ambient noise beamforming in the Parkfield area ». In : *Geophysical Journal International* 213 (3), p. 1478-1491. doi : 10.1093/gji/ggy058.

Lomax, A. (avr. 2008). « Location of the focus and tectonics of the focal region of the California earthquake of 18 April 1906 ». In : *Bulletin of the Seismological Society of America* 98 (2), p. 846-860. doi : 10.1785/0120060405.

Lomax, A. et A. Curtis (2001). « Fast probabilistic earthquake location in 3D models using Oct-Tree importance sampling Early-est : tsunami early warning, rapid location and magnitudes for large earthquakes View project Immersive wave experimentation View project ». In :

Lomax, A., J. Virieux, P. Volant et C. Berge-Thierry (2000). *Probabilistic Earthquake Location in 3D and Layered Models*. doi : 10.1007/978-94-015-9536-0_5.

Louppe, G., L. Wehenkel, A. Sutera et P. Geurts (2013). « Understanding variable importances in forests of randomized trees ». In : *Proceedings of the 26th International Conference on Neural Information Processing Systems* 1.

Lundberg, S. et S.-I. Lee (mai 2017). « A Unified Approach to Interpreting Model Predictions ». In : *Advances in Neural Information Processing Systems* 2017-December, p. 4766-4775.

Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair et al. (jan. 2020). « From local explanations to global understanding with explainable AI for trees ». In : *Nature Machine Intelligence* 2 (1), p. 56-67. doi : 10.1038/s42256-019-0138-9.

Luo, G. et M. Liu (juin 2012). « Multi-timescale mechanical coupling between the San Jacinto fault and the San Andreas fault, southern California ». In : *Lithosphere* 4 (3), p. 221-229. doi : 10.1130/L180.1.

M, M. (2007). « A test for checking earthquake aperiodicity estimates from small samples ». In : *Natural Hazards and Earth System Science*, p. 399-404.

MAEDA, N. (1985). « A method for reading and checking phase times in auto-processing system of seismic wave data ». In : *Zisin* 38, p. 365-379.

MAGGI, A., V. FERRAZZINI, C. HIBERT, F. BEAUDUCEL, P. BOISSIER et A. AMEMOUTOU (mai 2017). « Implementation of a multistation approach for automated event classification at Piton de la Fournaise volcano ». In : *Seismological Research Letters* 88 (3), p. 878-891. DOI : 10.1785/0220160189.

MALFANTE, M., M. D. MURA, J. P. METAXIAN, J. I. MARS, O. MACEDO et A. INZA (2018). « Machine Learning for Volcano-Seismic Signals : Challenges and Perspectives ». In : *IEEE Signal Processing Magazine* 35 (2), p. 20-30. DOI : 10.1109/MSP.2017.2779166.

MALIN, P. E., M. BOHNHOFF, F. BLÜMLE, G. DRESEN, P. MARTINEZ-GARZON, M. NURLU et al. (déc. 2018). « Microearthquakes preceding a M4.2 Earthquake Offshore Istanbul ». In : *Scientific Reports* 8 (1), p. 16176. DOI : 10.1038/s41598-018-34563-9.

MANDELBROT, B. B. (1982). « The Fractal Geometry of Nature ». In : *San Francisco : Freeman*.

MARKHVIDA, M., B. WALSH, S. HALLEGATTE et J. BAKER (1999). « Quantification of disaster impacts through household well-being losses ». In : *Nature Sustainability*, p. 537-574. DOI : 10.1038/s41893-020-0508-7.

MARSAN, D., A. HELMSTETTER, M. BOUCHON et P. DUBLANCHET (2014). « Foreshock activity related to enhanced aftershock production ». In : *Geophysical Research Letters* 41 (19), p. 6652-6658.

MARTINEZ-GARZON, P., Y. BEN-ZION, I. ZALIAPIN et M. BOHNHOFF (oct. 2019). « Seismic clustering in the Sea of Marmara : Implications for monitoring earthquake processes ». In : *Tectonophysics* 768, p. 228176. DOI : 10.1016/j.tecto.2019.228176.

MATOS, C., S. CUSTODIO, J. BATLO, J. ZAHRADNIK, P. ARROUCAU, G. SILVEIRA et al. (avr. 2018). « An Active Seismic Zone in Intraplate West Iberia Inferred From High-Resolution Geophysical Data ». In : *Journal of Geophysical Research : Solid Earth* 123 (4), p. 2885-2907. DOI : 10.1002/2017JB015114.

MATTHEWS, M. V., W. L. ELLSWORTH et P. A. REASENBERG (août 2002). « A Brownian model for recurrent earthquakes ». In : *Bulletin of the Seismological Society of America* 92 (6), p. 2233-2250. DOI : 10.1785/0120010267.

MCBREARTY, I. W., A. A. DELOREY et P. A. JOHNSON (2019). « Pairwise association of seismic arrivals with convolutional neural networks ». In : *Seismological Research Letters* 90 (2A), p. 503-509. DOI : 10.1785/0220180326.

MCNAMARA, D. E. et R. P. BULAND (2004). « Ambient Noise Levels in the Continental United States ». In : *Bulletin of the Seismological Society of America* 96.4. DOI : 10.1785/012003001.

MCNUTT, S. R. (2005). « VOLCANIC SEISMOLOGY ». In : *Annual Review of Earth and Planetary Sciences* 33 (1), p. 461-491. DOI : 10.1146/annurev.earth.33.092203.122459.

MEGHRAOUI, M., B. DELOUIS, M. FERRY, D. GIARDINI, P. HUGGENBERGER, I. SPOTTKE et al. (sept. 2001). « Active normal faulting in the upper rhine

graben and paleoseismic identification of the 1356 basel earthquake ». In : *Science* 293 (5537), p. 2070-2073. DOI : 10.1126/science.1010618.

MEIER, M., Z. E. ROSS, A. RAMACHANDRAN, A. BALAKRISHNA, S. NAIR, P. KUNDZICZ et al. (jan. 2019). « Reliable Real-Time Seismic Signal/Noise Discrimination With Machine Learning ». In : *Journal of Geophysical Research : Solid Earth* 124 (1), p. 788-800. DOI : 10.1029/2018JB016661.

MENG, H. et Y. BEN-ZION (2018). « Detection of small earthquakes with dense array data : example from the San Jacinto fault zone, southern California ». In : *Geophysical Journal International* 212, p. 442-457. DOI : 10.1093/gji/ggx404.

MIGNAN, A. (fév. 2014). « The debate on the prognostic value of earthquake foreshocks : A meta-analysis ». In : *Scientific Reports* 4 (1), p. 1-5. DOI : 10.1038/srep04099.

MIGNAN, A. et J. WOESSNER (2012). « Estimating the magnitude of completeness for earthquake catalogs, Community Online Resource for Statistical Seismicity Analysis, » in : *Community Online Resource for Statistical Seismicity Analysis* (April), p. 1-45. DOI : 10.5078/corssa-00180805.

MORI, J. et R. E. ABERCROMBIE (1997). « Depth dependence of earthquake frequency-magnitude distributions in California : Implications for rupture initiation ». In : *Journal of Geophysical Research : Solid Earth* 102 (B7), p. 15081-15090. DOI : 10.1029/97jb01356.

MOUSSET, E., Y. CANSI, R. CRUSEM et Y. SOUCHET (mar. 1996). « A connectionist approach for automatic labeling of regional seismic phases using a single vertical component seismogram ». In : *Geophysical Research Letters* 23 (6), p. 681-684. DOI : 10.1029/95GL03811.

MURPHY, B. S., M. LIU et G. D. EGBERT (2019). « Insights Into Intraplate Stresses and Geomorphology in the Southeastern United States ». In : *Geophysical Research Letters* 46 (15), p. 8711-8720. DOI : 10.1029/2019GL083755.

NADEAU, R. M. et L. R. JOHNSON (1998). « Seismological Studies at Parkfield VI : Moment Release Rates and Estimates of Source Parameters for Small Repeating Earthquakes ». In : *Bulletin of the Seismological Society of America* 88 (3), p. 790-814.

NEELY, J. S., S. STEIN, M. MERINO et J. ADAMS (nov. 2018). « Have we seen the largest earthquakes in eastern North America ? » In : *Physics of the Earth and Planetary Interiors* 284, p. 17-27. DOI : 10.1016/j.pepi.2018.09.005.

NISHIKAWA, T. et S. IDE (2017). « Detection of earthquake swarms at subduction zones globally : Insights into tectonic controls on swarm activity ». In : *Journal of Geophysical Research : Solid Earth* 122 (7), p. 5325-5343. DOI : 10.1002/2017JB014188.

NORMILE, D. (2012). « Report : Fukushima A Manmade Disaster ». In : *Science*.

ST-ONGE, A. (2011). « Akaike information criterion applied to detecting first arrival times on microseismic data ». In : *SEG Technical Program Expanded Abstracts* 30 (1), p. 1658-1662. DOI : 10.1190/1.3627522.

Pacheco, J. F. et L. R. Sykes (1992). « Seismic moment catalog of large shallow earthquakes, 1900 to 1989 ». In : *Bulletin of the Seismological Society of America* 82 (3), p. 1306-1349.

Parsons, L., E. Haque et H. Liu (2004). « Subspace Clustering for High Dimensional Data : A Review ». In :

Peng, Z., J. E. Vidale et H. Houston (sept. 2006). « Anomalous early aftershock decay rate of the 2004 Mw6.0 Parkfield, California, earthquake ». In : *Geophysical Research Letters* 33 (17). doi : 10.1029/2006GL026744.

Peng, Z. et P. Zhao (2009). « Migration of early aftershocks following the 2004 Parkfield earthquake ». In : *Nature Geoscience* 2. doi : 10.1038/NGEO697.

Perol, T., M. Gharbi et M. Denolle (fév. 2018). « Convolutional neural network for earthquake detection and location ». In : *Science Advances* 4 (2), e1700578. doi : 10.1126/sciadv.1700578.

Pesicek, J. D., D. Child, B. Artman et K. Cieslik (2014). « Picking versus stacking in a modern microearthquake location : Comparison of results from a surface passive seismic monitoring array in oklahoma ». In : *Geophysics* 79 (6). doi : 10.1190/GEO2013-0404.1.

Pestourie, R., M. Grunberg et V. Maurer (2017). « Contribution of the SeisComP 3 scanloc module to the monitoring of microseismicity induced by deep geothermal projects ». In : *5th European Geothermal Workshop, Karlsruhe*.

Peterson, J. (1993). « Observations and modelling of seismic background noise. » In : *Seismological Research Letters* 79 (2), p. 1-94. doi : 10.3133/ofr93322.

Pfungst, O. (1911). « Clever Hans (the horse of Mr. von Osten) a contribution to experimental animal and human psychology ». In : *New York : Henry Holt.*

Pino, N. A., V. Convertito et R. Madariaga (2019). « Clock advance and magnitude limitation through fault interaction : the case of the 2016 central Italy earthquake sequence ». In : *Scientific Reports* 9 (1). doi : 10.1038/s41598-019-41453-1.

Poli, P., J. Boaga, I. Molinari, V. Cascone et L. Boschi (déc. 2020). « The 2020 coronavirus lockdown and seismic monitoring of anthropic activities in Northern Italy ». In : *Scientific Reports* 10 (1), p. 1-8. doi : 10.1038/s41598-020-66368-0.

Provost, F., C. Hibert et J.-P. Malet (2017). « Automatic classification of endogenous landslide seismicity using the Random Forest supervised classifier ». In : *Geophysical Research Letters* 44.1, p. 113-120. doi : 10.1002/2016GL070709.

Rabin, M., C. Sue, A. Walpersdorf, P. Sakic, J. Albaric et B. Fores (2018). « Present-Day Deformations of the Jura Arc Inferred by GPS Surveying and Earthquake Focal Mechanisms ». In : *Tectonics* 37 (10), p. 3782-3804.

Raschka, S. (nov. 2018). « Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning ». In :

Ratzov, G., A. Cattaneo, N. Babonneau, J. Deverchere, K. Yelles, R. Bracene et al. (fév. 2015). « Holocene turbidites record earthquake supercycles at a slow-rate plate boundary ». In : *Geology* 43 (4), p. 331-334. doi : 10.1130/G36170.1.

Reid, H. F. (1910). « The Mechanics of the Earthquake, The California Earthquake of April 18, 1906 ». In : *Report of the State Investigation Commission* 2, p. 16-28.

Reina Ortiz, M., N. K. Le, V. Sharma, I. Hoare, E. Quizhpe, E. Teran et al. (déc. 2017). « Post-earthquake Zika virus surge : Disaster and public health threat amid climatic conduciveness ». In : *Scientific Reports* 7 (1), p. 1-10. doi : 10.1038/s41598-017-15706-w.

Reverso, T., D. Marsan, A. Helmstetter et B. Enescu (juin 2016). « Background seismicity in Boso Peninsula, Japan : Long-term acceleration, and relationship with slow slip events ». In : *Geophysical Research Letters* 43 (11), p. 5671-5679. doi : 10.1002/2016GL068524.

Rhoades, D. A. (2010). « Lessons and questions from thirty years of testing the Precursory swarm hypothesis ». In : *Pure and Applied Geophysics* 167 (6), p. 629-644. doi : 10.1007/s00024-010-0071-7.

Riahi, N. et P. Gerstoft (2015). « The seismic traffic footprint : Tracking trains, aircraft, and cars seismically ». In : *Geophysical Research Letters* 42 (8), p. 2674-2681. doi : 10.1002/2015GL063558.

Roland, E. et J. J. McGuire (sept. 2009). « Earthquake swarms on transform faults ». In : *Geophysical Journal International* 178 (3), p. 1677-1690. doi : 10.1111/j.1365-246X.2009.04214.x.

Romanowicz, B., D. Dreger, M. Pasyanos et R. Uhrhammer (août 1993). « Monitoring of strain release in central and northern California using broadband data ». In : *Geophysical Research Letters* 20 (15), p. 1643-1646. doi : 10.1029/93GL01540.

Ross, Z. E., D. T. Trugman, E. Hauksson et P. M. Shearer (2019). « Searching for hidden earthquakes in Southern California ». In : *Science* 364 (6442), p. 767-771. doi : 10.1126/science.aaw6888.

Ross, Z. E., E. S. Cochran, D. T. Trugman et J. D. Smith (juin 2020). « 3D fault architecture controls the dynamism of earthquake swarms ». In : *Science* 368 (6497), p. 1357-1361. doi : 10.1126/science.abb0779.

Ross, Z. E., M.-A. Meier et E. Hauksson (2018). « P-Wave Arrival Picking and First-Motion Polarity Determination With Deep Learning ». In : *Journal of Geophysical Research : Solid Earth* 123 (6), p. 5120-5129. doi : 10.1029/2017JB015251.

Ross, Z. E., C. Rollins, E. S. Cochran, E. Hauksson, J.-P. Avouac et Y. Ben-Zion (août 2017). « Aftershocks driven by afterslip and fluid pressure sweeping through a fault-fracture mesh ». In : *Geophysical Research Letters* 44 (16), p. 8260-8267. doi : 10.1002/2017GL074634.

Rothe, J.-P. et E. Peterschmitt (1950). « etude seismique des explosions d'Haslach ». In : *Ann. Inst. Phys. Globe Strasbourg 3e part Geophysique,* 5, p. 13-28.

ROUET-LEDUC, B., C. HULBERT et P. A. JOHNSON (jan. 2019). « Continuous chatter of the Cascadia subduction zone revealed by machine learning ». In : *Nature Geoscience* 12 (1), p. 75-79. DOI : 10.1038/s41561-018-0274-6.

RUEDEN, L. von, S. MAYER, K. BECKH, B. GEORGIEV, S. GIESSELBACH, R. HEESE et al. (2019). « Informed Machine Learning – A Taxonomy and Survey of Integrating Knowledge into Learning Systems ». In :

SAENGER, E. K., S. M. SCHMALHOLZ, M.-A. LAMBERT, T. T. NGUYEN, A. TORRES, S. METZGER et al. (2009). « A passive seismic survey over a gas field : Analysis of low-frequency anomalies ». In : DOI : 10.1190/1.3078402.

SALDITCH, L., S. STEIN, J. NEELY, . D. SPENCER, E. M. BROOKS, A. AGNON et al. (jan. 2020). « Earthquake supercycles and Long-Term Fault Memory ». In : *Tectonophysics* 774, p. 228289. DOI : 10.1016/j.tecto.2019.228289.

SAMEK, W. (jan. 2020). « Learning with explainable trees ». In : *Nature Machine Intelligence* 2 (1), p. 16-17. DOI : 10.1038/s42256-019-0142-0.

SASAKI, Y., J. AIDA, T. TSUJI, S. KOYAMA, T. TSUBOYA, T. SAITO et al. (déc. 2019). « Pre-disaster social support is protective for onset of post-disaster depression : Prospective study from the Great East Japan Earthquake Tsunami ». In : *Scientific Reports* 9 (1), p. 1-10. DOI : 10.1038/s41598-019-55953-7.

SATO, Y. et T. ARIMOTO (2016). « Five years after Fukushima : scientific advice in Japan ». In : *Palgrave Communications* 2 (1), p. 16-25.

SCHAUMBERG, A. J., W. C. JUAREZ-NICANOR, S. J. CHOUDHURY, L. G. PASTRIÁN, B. S. PRITT, M. PRIETO-POZUELO et al. (2020). « Interpretable multimodal deep learning for real-time pan-tissue pan-disease pathology search on social media ». In : *Modern Pathology*. DOI : 10.1038/s41379-020-0540-1.

SCHOENBALL, M., N. C. DAVATZES et J. . . GLEN (août 2015). « Differentiating induced and natural seismicity using space-time-magnitude statistics applied to the Coso Geothermal field ». In : *Geophysical Research Letters* 42 (15), p. 6221-6228. DOI : 10.1002/2015GL064772.

SCHOLZ, C. H. (fév. 1968). « Microfracturing and the inelastic deformation of rock in compression ». In : *Journal of Geophysical Research* 73 (4), p. 1417-1432. DOI : 10.1029/jb073i004p01417.

SCHORLEMMER, D., S. WIEMER et M. WYSS (sept. 2005). « Variations in earthquake-size distribution across different stress regimes ». In : *Nature* 437 (7058), p. 539-542. DOI : 10.1038/nature04094.

SCHRAMOWSKI, P., W. STAMMER, S. TESO, A. BRUGGER, F. HERBERT, X. SHAO et al. (août 2020). « Making deep neural networks right for the right scientific reasons by interacting with their explanations ». In : *Nature Machine Intelligence* 2 (8), p. 476-486. DOI : 10.1038/s42256-020-0212-3.

SEYDOUX, L., N. M. SHAPIRO, J. DE ROSNY, F. BRENGUIER et M. LANDES (2016). « Detecting seismic activity with a covariance matrix analysis of data recorded on seismic arrays ». In : *Geophysical Journal International* 204 (3), p. 1430-1442. DOI : 10.1093/gji/ggv531.

Shalev-Shwartz, S. et S. Ben-David (2014). « Understanding Machine Learning : From Theory to Algorithms ». In : doi : https://doi.org/10.1017/CBO9781107298019.

Shearer, P. M. (juin 2012). « Self-similar earthquake triggering, Båth's law, and foreshock/aftershock magnitudes : Simulations, theory, and results for southern California ». In : *Journal of Geophysical Research : Solid Earth* 117 (B6), n/a-n/a. doi : 10.1029/2011JB008957.

Shebalin, P. et C. Narteau (déc. 2017). « Depth dependent stress revealed by aftershocks ». In : *Nature Communications* 8 (1), p. 1-8. doi : 10.1038/s41467-017-01446-y.

Sheen, D.-H., J. S. Shin, T.-S. Kang et C. E. Baag (sept. 2009). « Low frequency cultural noise ». In : *Geophysical Research Letters* 36 (17), p. L17314. doi : 10.1029/2009GL039625.

Shelly, D. R., J. L. Hardebeck, W. L. Ellsworth et D. P. Hill (2016). « A new strategy for earthquake focal mechanisms using waveform-correlation-derived relative polarities and cluster analysis : Application to the 2014 Long Valley Caldera earthquake swarm ». In : *Journal of Geophysical Research : Solid Earth* 121 (12), p. 8622-8641. doi : 10.1002/2016JB013437.

Shi, Y. et B. A. Bolt (1982). « THE STANDARD ERROR OF THE MAGNITUDE-FREQUENCY b VALUE ». In : *Bulletin of the Seismological Society of America* 72 (5), p. 1677-1687.

Shipton, Z. K., M. Meghraoui et L. Monro (2017). « Seismic slip on the west flank of the upper rhine graben (france-germany) : Evidence from tectonic morphology and cataclastic deformation bands ». In : *Geological Society Special Publication* 432 (1), p. 147-161. doi : 10.1144/SP432.12.

Sieh, K., M. Stuiver et D. Brillinger (1989). « A More Precise Chronology of Earthquakes Produced by the San Andreas Fault in Southern California ». In : *JOURNAL OF GEOPHYSICAL RESEARCH* 94 (B1).

Sleeman, R. et T. Van Eck (juin 1999). « Robust automatic P-phase picking : An on-line implementation in the analysis of broadband seismogram recordings ». In : t. 113, p. 265-275. doi : 10.1016/S0031-9201(99)00007-2.

Soto-Cordero, L., A. Meltzer et J. C. Stachnik (jan. 2018). « Crustal structure, intraplate seismicity, and seismic hazard in the mid-Atlantic United States ». In : *Seismological Research Letters* 89 (1), p. 241-252. doi : 10.1785/0220170084.

Steim, J. M. (2015). *Theory and Observations - Instrumentation for Global and Regional Seismology*. doi : 10.1016/B978-0-444-53802-4.00023-3.

Stein, S., R. J. Geller et M. Liu (2012). « Why earthquake hazard maps often fail and what to do about it ». In : doi : 10.1016/j.tecto.2012.06.047.

Stein, S., M. Liu, T. Camelbeeck, M. Merino, A. Landgraf, E. Hintersberger et al. (2017). « Challenges in assessing seismic hazard in intraplate Europe ». In : *Geological Society Special Publication* 432 (1), p. 13-28. doi : 10.1144/SP432.7.

Stump, B., C. Hayward, C. Hetzer et R. M. Zhou (2001). « UTILIZATION OF SEISMIC AND INFRASOUND SIGNALS FOR CHARACTERIZING MINING EXPLOSIONS ». In :

Sun, C., A. Shrivastava, S. Singh et A. Gupta (juil. 2017). « Revisiting Unreasonable Effectiveness of Data in Deep Learning Era ». In : *Proceedings of the IEEE International Conference on Computer Vision*, p. 843-852.

Tarantola, A. et B. Valette (1982). « Generalized Nonlinear Inverse Problems Solved Using the Least Squares Criterion ». In : *Reviews of Geophysics and Space Physics* 20 (2), p. 219-232.

Terrinha, P., L. Matias, J. Vicente, J. Duarte, J. Luis, L. Pinheiro et al. (2009). « Morphotectonics and strain partitioning at the Iberia-Africa plate boundary from multibeam and seismic reflection data ». In : *Marine Geology* 267 (3-4), p. 156-174. doi : 10.1016/j.margeo.2009.09.012.

Thouvenot, F., J. Frechet, L. Jenatton et J.-F. Gamond (2003). « The Belledonne Border Fault : identification of an active seismic strike-slip fault in the western Alps ». In : *Geophysical Journal International* 155, p. 174-192.

Thurber, C. H. (1985). « Nonlinear earthquake location : Theory and examples ». In : *Bulletin of the Seismological Society of America* 75 (3), p. 779-790.

Tiira, T. (1999). « Detecting teleseismic events using artificial neural networks ». In : *Computers and Geosciences* 25 (8), p. 929-938. doi : 10.1016/S0098-3004(99)00056-4.

Toda, S. et R. S. Stein (juin 2018). « Why aftershock duration matters for probabilistic seismic hazard assessment ». In : *Bulletin of the Seismological Society of America* 108 (3), p. 1414-1426. doi : 10.1785/0120170270.

Tokuda, T. et H. Shimada (déc. 2019). « Classes of low-frequency earthquakes based on inter-time distribution reveal a precursor event for the 2011 Great Tohoku Earthquake ». In : *Scientific Reports* 9 (1), p. 1-14. doi : 10.1038/s41598-019-45765-0.

Tong, P., D. Yang, Q. Liu, X. Yang et J. Harris (2016). « Acoustic wave-equation-based earthquake location ». In : *Geophysical Journal International* 205 (1), p. 464-478. doi : 10.1093/gji/ggw026.

Tormann, T., S. Wiemer et A. Mignan (2014). « Systematic survey of high-resolution b value imaging along Californian faults : Inference on asperities ». In : *Journal of Geophysical Research : Solid Earth* 119 (3), p. 2029-2054. doi : 10.1002/2013JB010867.

Trnkoczy, A. (1999). « Topic Understanding and parameter setting of STA/LTA trigger algorithm Author Amadej Trnkoczy (formerly Kinemetrics SA) ». In : p. 1-20. doi : 10.2312/GFZ.NMSOP-2_IS_8.1.

Trugman, D. T. et Z. E. Ross (2019). « Pervasive Foreshock Activity Across Southern California ». In : *Geophysical Research Letters* 46 (15), p. 8772-8781. doi : 10.1029/2019GL083725.

Trunk, G. V. (1979). « A Problem of Dimensionality : A Simple Example ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* (3), p. 306-307. doi : 10.1109/TPAMI.1979.4766926.

Turcotte, D. L. (juin 1997). « Fractals and Chaos in geology and geophysics ». In : *Cambridge University Press*, p. 214-215. doi : http://dx.doi.org/10.1017/CBO9781139174695.

Tuttle, M. P., E. S. Schweig, J. D. Sims, R. H. Lafferty, L. W. Wolf et M. L. Haynes (août 2002). « The earthquake potential of the New Madrid seismic zone ». In : *Bulletin of the Seismological Society of America* 92 (6), p. 2080-2089. doi : 10.1785/0120010227.

Utheim, T., J. Havskov, M. Ozyazicioglu, J. Rodriguez et E. Talavera (mai 2014). « Rtquake, a real-time earthquake detection system integrated with seisan ». In : *Seismological Research Letters* 85 (3), p. 735-742. doi : 10.1785/0220130175.

Utsu, T., Y. Ogata, S. Ritsuko et M. Matsuura (1995). « The Centenary of the Omori Formula for a Decay Law of Aftershock Activity. » In : *Journal of Physics of the Earth* 43 (1), p. 1-33. doi : 10.4294/jpe1952.43.1.

Vallage, A. et L. Bollinger (mai 2020). « Testing Fault Models in Intraplate Settings : A Potential for Challenging the Seismic Hazard Assessment Inputs and Hypothesis ? » In : *Pure and Applied Geophysics* 177 (5), p. 1879-1889. doi : 10.1007/s00024-019-02129-z.

Vallee, M., J.-M. Nocquet, J. Battaglia, Y. Font, M. Segovia, M. Regnier et al. (juin 2013). « Intense interface seismicity triggered by a shallow slow slip event in the Central Ecuador subduction zone ». In : *Journal of Geophysical Research : Solid Earth* 118 (6), p. 2965-2981. doi : 10.1002/jgrb.50216.

Vanneste, K., T. Camelbeeck et K. Verbeeck (avr. 2013). « Model of composite seismic sources for the Lower Rhine Graben, northwest Europe ». In : *Bulletin of the Seismological Society of America* 103 (2 A), p. 984-1007. doi : 10.1785/0120120037.

Vapnik, V. (1999). « Principles of Risk Minimization for Learning Theory ». In :

Verdon, J. P., B. J. Baptie et J. J. Bommer (juil. 2019). « An improved framework for discriminating seismicity induced by industrial activities from natural earthquakes ». In : *Seismological Research Letters* 90 (4), p. 1592-1611. doi : 10.1785/0220190030.

Vidale, J. E. et P. M. Shearer (2006). « A survey of 71 earthquake bursts across southern California : Exploring the role of pore fluid pressure fluctuations and aseismic slip as drivers ». In : *Journal of Geophysical Research : Solid Earth* 111 (5). doi : 10.1029/2005JB004034.

Voyles, J. R., M. M. Holt, J. M. Hale, K. D. Koper, R. Burlacu et D. J. Chambers (2019). « A new catalog of explosion source parameters in the Utah region with application to ML–MC-based depth discrimination at local distances ». In : *Seismological Research Letters* 91 (1), p. 222-236. doi : 10.1785/0220190185.

Wagner, G. S. et T. J. Owens (1996). « Signal Detection Using Multi-Channel Seismic Data ». In : *Bulletin of the Seismological Society of America* 86 (1A), p. 221-231.

Wainberg, M., M. D., A. Delong et B. J. Frey (2018). « Deep learning in biomedicine ». In : *Nature Biotechnology* 36 (9), p. 829-838. doi : 10.1038/nbt.4233.

Walker, R. T., M. M. Khatib, A. Bahroudi, A. Rodes, C. Schnabel, M. Fattahi et al. (2015). « Co-seismic, geomorphic, and geologic fold growth associated with the 1978 Tabas-e-Golshan earthquake fault in eastern Iran ». In : *Geomorphology* 237, p. 98-118. doi : 10.1016/j.geomorph.2013.02.016.

Wallace, R. E. (1987). « Grouping and migration of surface faulting and variations in slip rates on faults in the Great Basin province ». In : *Bulletin of the Seismological Society of America* 77 (3), p. 868-876.

Wang, J., I. G. Main et R. M. W. Musson (2017). « Earthquake clustering in modern seismicity and its relationship with strong historical earthquakes around Beijing, China ». In : *Geophysical Journal International* 211 (2), p. 1005-1018. doi : 10.1093/gji/ggx326.

Wang, J. et T.-L. Teng (1995). « Artificial Neural Network-Based Seismic Detector ». In : *Bulletin of the Seismological Society of America* 85 (1), p. 308-319.

Wei, S., J.-P. Avouac, K. W. Hudnut, A. Donnellan, J. W. Parker, R. W. Graves et al. (juil. 2015). « The 2012 Brawley swarm triggered by injection-induced aseismic slip ». In : *Earth and Planetary Science Letters* 422, p. 115-125. doi : 10.1016/j.epsl.2015.03.054.

Wiemer, S. et M. Wyss (2002). « Mapping spatial variability of the frequency-magnitude distribution of earthquakes ». In : *Advances in Geophysics* 45, p. 259-302.

— (1997). « Mapping the frequency-magnitude distribution in asperities : An improved technique to calculate recurrence times ? » In : *Journal of Geophysical Research : Solid Earth* 102 (B7), p. 15115-15128. doi : 10.1029/97jb00726.

Williams, E. F., M. Fernandez-Ruiz, M. G., V. R., Z. Zhan, C. Gonzalez-Herraez et al. (déc. 2019). « Distributed sensing of microseisms and teleseisms with submarine dark fibers ». In : *Nature Communications* 10 (1), p. 1-11. doi : 10.1038/s41467-019-13262-7.

Xiao, H., Z. C. Eilon, C. Ji et T. Tanimoto (2020). « COVID-19 societal response captured by seismic noise in China and Italy ». In :

Xu, W., G. Feng, L. Meng, A. Zhang, J.-P. Ampuero, R. Bürgmann et al. (2018). « Transpressional Rupture Cascade of the 2016 Mw 7.8 Kaikoura Earthquake, New Zealand ». In : *Journal of Geophysical Research : Solid Earth* 123 (3). doi : 10.1002/2017JB015168.

Yang, H., L. Zhu et R. Chu (déc. 2009). « Fault-plane determination of the 18 april 2008 mount Carmel, Illinois, earthquake by detecting and relocating aftershocks ». In : *Bulletin of the Seismological Society of America* 99 (6), p. 3413-3420. doi : 10.1785/0120090038.

Yao, D., Y. Huang, Z. Peng et R. R. Castro (2020). « Detailed Investigation of the Foreshock Sequence of the 2010 Mw 7.2 El Mayor-Cucapah Earthquake ». In : *Journal of Geophysical Research : Solid Earth* 125 (6). doi : 10.1029/2019JB019076.

Yeck, W. L., J. M. Patton, C. E. Johnson, D. Kragness, H. M. Benz, P. S. Earle et al. (août 2019). « GLASS3 : A standalone multiscale seis-

mic detection associator ». In : *Bulletin of the Seismological Society of America* 109 (4), p. 1469-1478. DOI : 10.1785/0120180308.

YIN, X. Z., J. H. CHEN, Z. PENG, X. MENG, Q. Y. LIU, B. GUO et al. (2018). « Evolution and Distribution of the Early Aftershocks Following the 2008 Mw 7.9 Wenchuan Earthquake in Sichuan, China ». In : *Journal of Geophysical Research : Solid Earth* 123 (9), p. 7775-7790. DOI : 10.1029/2018JB015575.

YOON, C. E., K. J. BERGEN, K. RONG, H. ELEZABI, W. L. ELLSWORTH, G. C. BEROZA et al. (2019). « Unsupervised Large-Scale Search for Similar Earthquake Signals ». In : *Bulletin of the Seismological Society of America* 109.4, p. 1451-1468. DOI : 10.1785/0120190006.

YOON, C. E., O. O'REILLY, K. J. BERGEN et G. C. BEROZA (2015). « Earthquake detection through computationally efficient similarity search ». In : *Science Advances* 1 (11). DOI : 10.1126/sciadv.1501057.

YOON, C. E., Y. HUANG, W. L. ELLSWORTH et G. C. BEROZA (2017). « Seismicity During the Initial Stages of the Guy-Greenbrier, Arkansas, Earthquake Sequence ». In : *Journal of Geophysical Research : Solid Earth* 122 (11), p. 9253-9274. DOI : 10.1002/2017JB014946.

YOUNG, C. J., E. P. CHAEL, M. M. WITHERS et R. C. ASTER (1996). « A Comparison of the High-Frequency (> 1 Hz) Surface and Subsurface Noise Environment at Three Sites in the United States ». In : *Bulletin of the Seismological Society of America* 86 (5), p. 1516-1528.

YU, T. et H. ZHU (mar. 2020). « Hyper-Parameter Optimization : A Review of Algorithms and Applications ». In :

ZALIAPIN, I., A. GABRIELOV, V. KEILIS-BOROK et H. WONG (2008). « Clustering analysis of seismicity and aftershock identification ». In : *Physical Review Letters* 101 (1), p. 018501. DOI : 10.1103/PhysRevLett.101.018501.

ZHANG, M. et L. WEN (2015). « Earthquake characteristics before eruptions of Japan's Ontake volcano in 2007 and 2014 ». In : *Geophysical Research Letters* 42 (17), p. 6982-6988. DOI : 10.1002/2015GL065165.

ZHANG, P. Z. (mai 2013). « Beware of slowly slipping faults ». In : *Nature Geoscience* 6 (5), p. 323-324. DOI : 10.1038/ngeo1811.

ZHANG, X., J. ZHANG, Y. C., S. LIU, Z. CHEN et W. LI (déc. 2020). « Locating induced earthquakes with a network of seismic stations in Oklahoma via a deep learning method ». In : *Scientific Reports* 10 (1), p. 1-12. DOI : 10.1038/s41598-020-58908-5.

ZHANG, Z., Y. LIN, Z. ZHOU et T. CHEN (2019). « Adaptive Filtering for Event Recognition from Noisy Signal : An Application to Earthquake Detection ». In : p. 3327-3331. DOI : 10.1109/ICASSP.2019.8683688.

ZHENG, Y. et R. R. STEWART (1992). « Polarization filter : Design and testing ». In :

ZHOU, T., H. FU, G. CHEN, J. SHEN et L. SHAO (2020). « Hi-Net : Hybrid-Fusion Network for Multi-Modal MR Image Synthesis ». In : *IEEE Transactions on Medical Imaging* 39 (9), p. 2772-2781. DOI : 10.1109/TMI.2020.2975344.

Zhou, Y., H. Yue, S. Zhou et Q. Kong (2019). « Hybrid event detection and phase-picking algorithm using convolutional and recurrent neural networks ». In : *Seismological Research Letters* 90 (3), p. 1079-1087. doi : 10.1785/0220180319.

Zhu, W. et G. C. Beroza (jan. 2019). « PhaseNet : A deep-neural-network-based seismic arrival-time picking method ». In : *Geophysical Journal International* 216 (1), p. 261-273. doi : 10.1093/gji/ggy423.

Annexes

Annexe A

Distribution de la sismicité historique et expérimentale de la zone du Graben du Rhin Supérieur

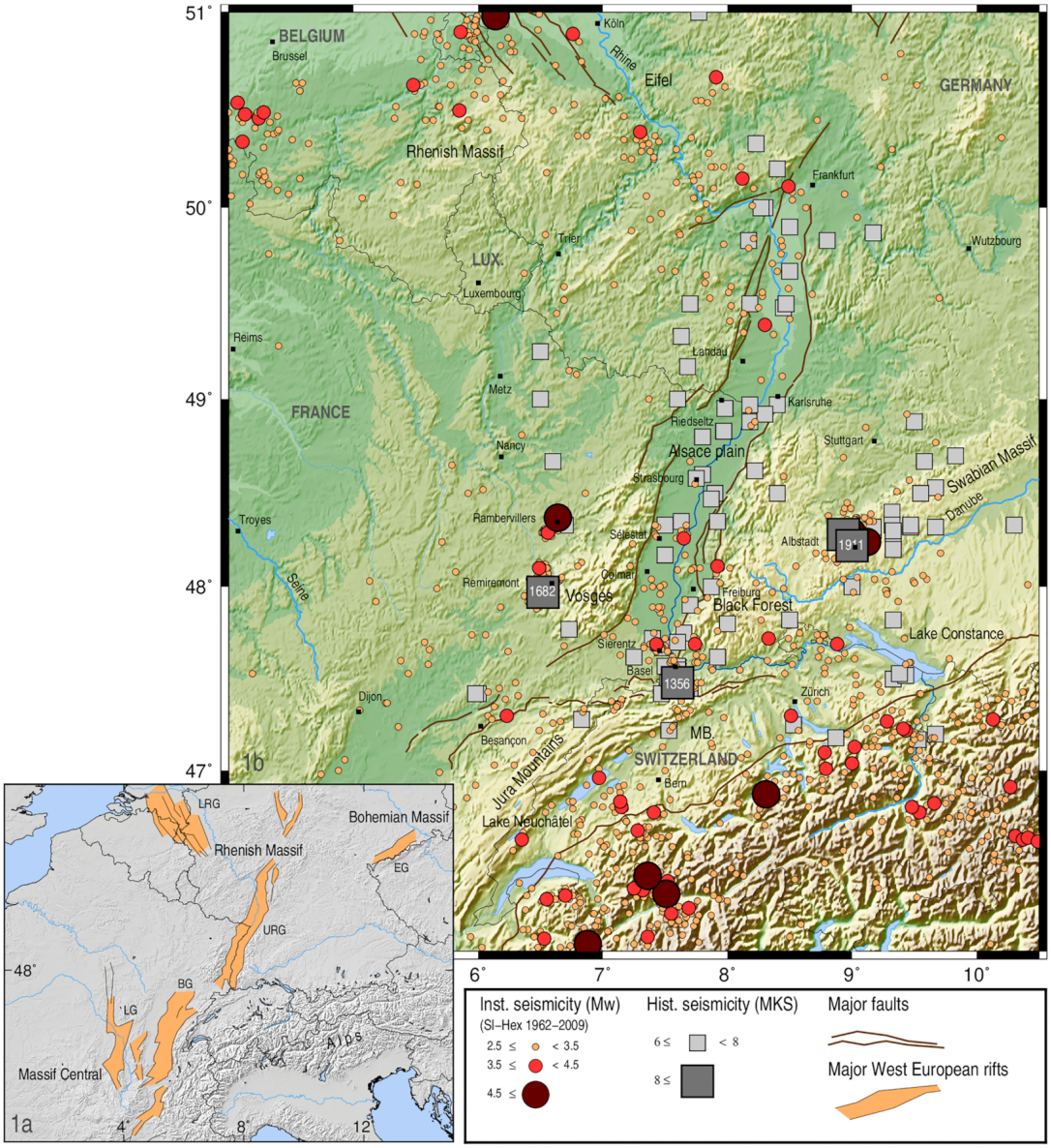


FIGURE A.1: (a) Segments majeurs du rift Cénozoïque ouest-européen, représentés en orange ECRIS. (b) Sismicité historique et instrumentale de la zone du Graben Supérieur (catalogue SI-Hex, Cara et al. 2015). Les failles majeures sont représentées par des lignes marrons pour les deux figures. D'après Henrion et al., 2020.

Annexe B

Distribution de la sismicité extraite du catalogue RéNaSS et du réseau de détection utilisé pour la période 2012-2019

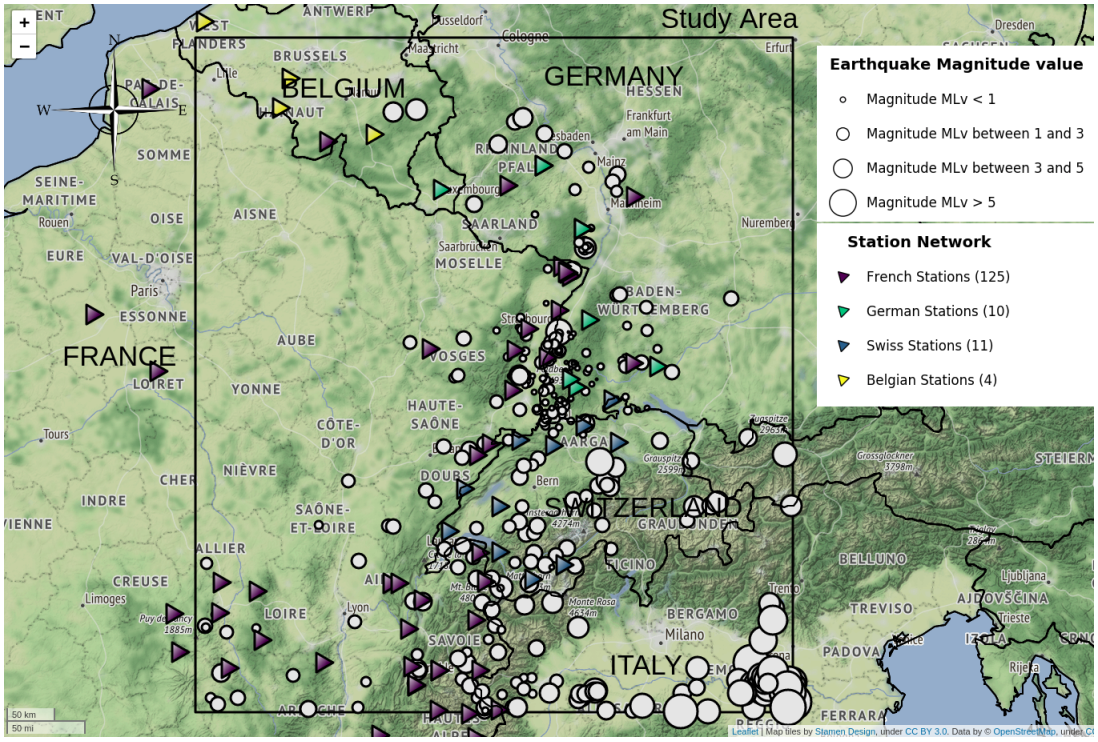


FIGURE B.1: Distribution de la sismicité et du réseau de détection utilisé par le RéNaSS pour l'année **2012**. Localisations des stations et des séismes ainsi que magnitudes des séismes extraites de la base de données RéNaSS selon un protocole FDSN à l'adresse http://renass-sc1.u-strasbg.fr:8080

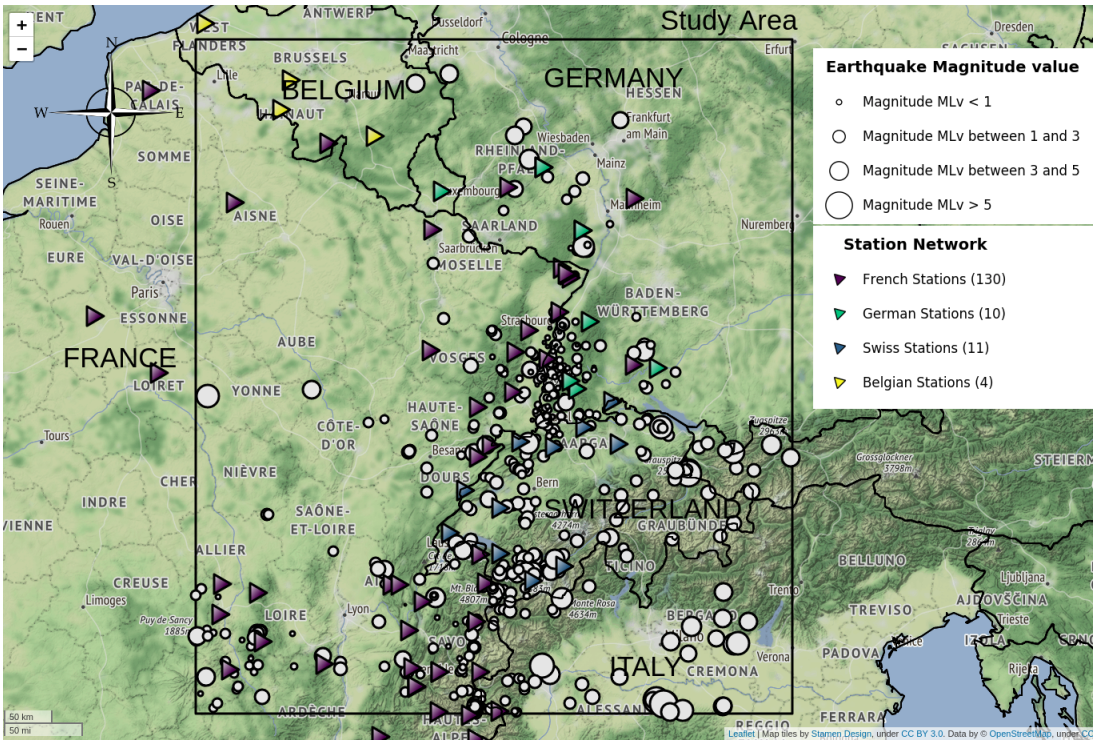


FIGURE B.2: Distribution de la sismicité et du réseau de détection utilisé par le RéNaSS pour l'année **2013**.

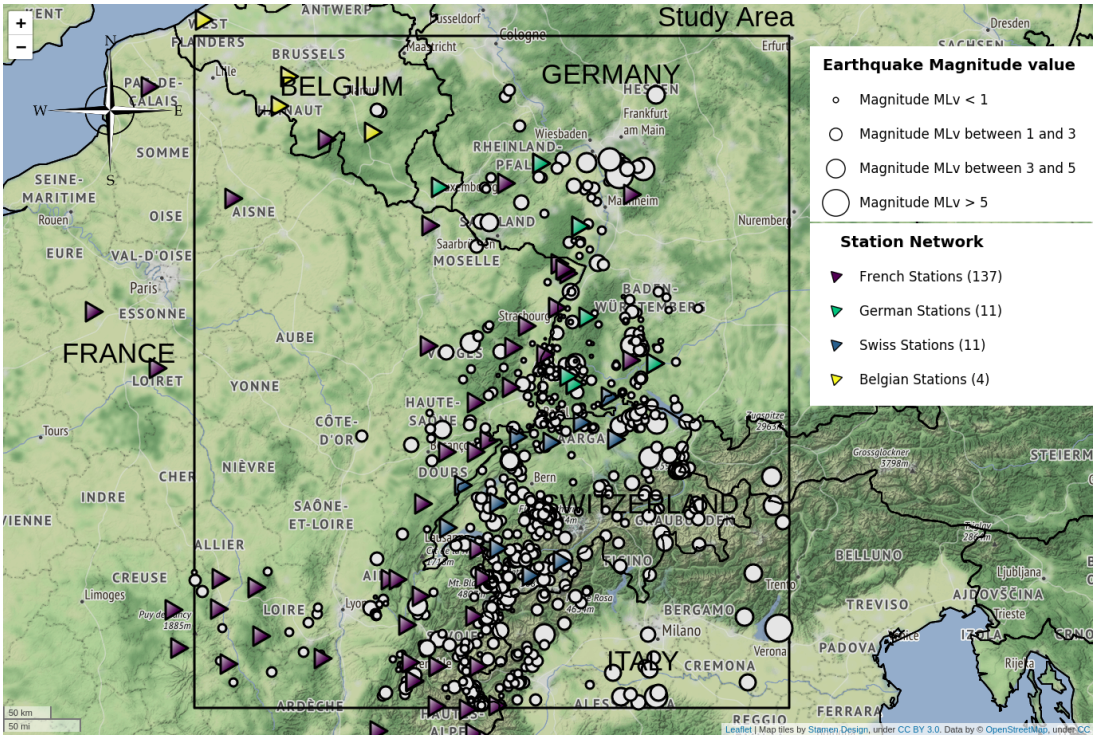


FIGURE B.3: Distribution de la sismicité et du réseau de détection utilisé par le RéNaSS pour l'année **2014**.

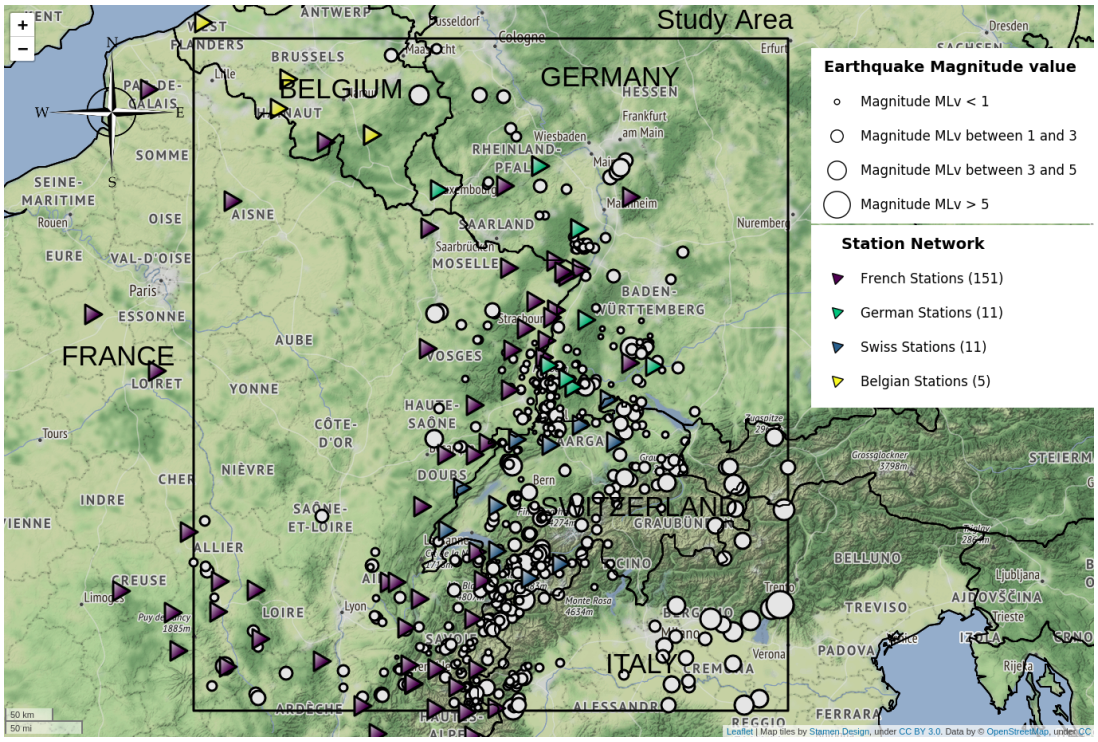


FIGURE B.4: Distribution de la sismicité et du réseau de détection utilisé par le RéNaSS pour l'année **2015**.

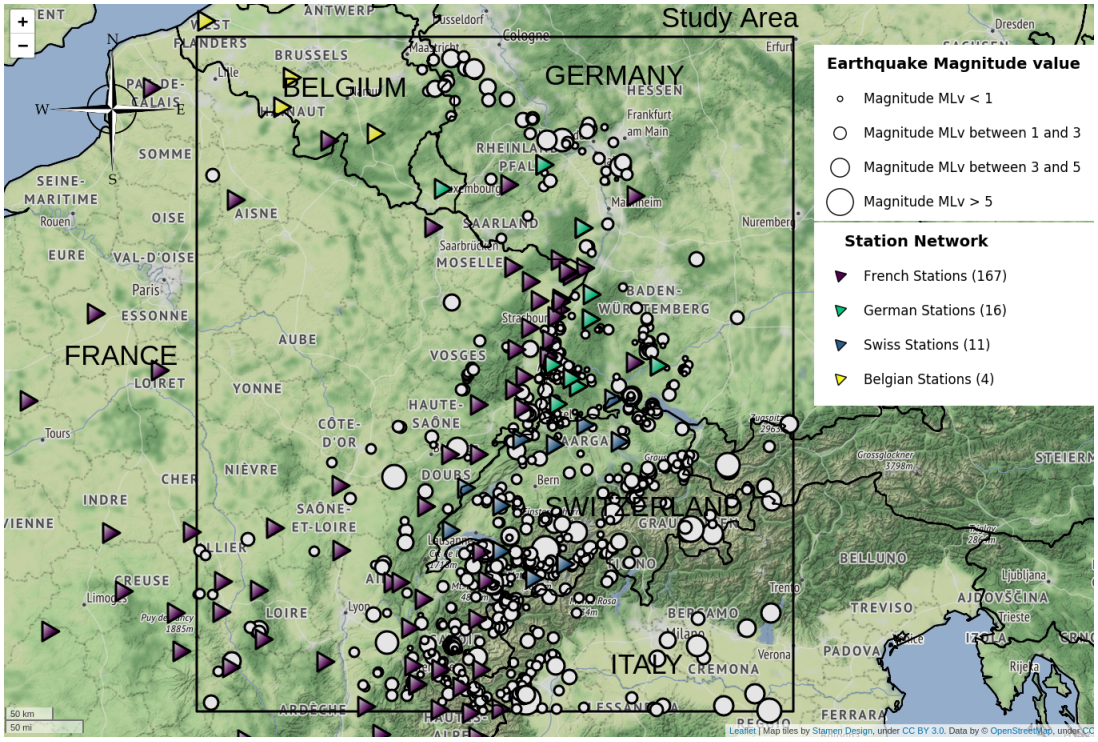


FIGURE B.5: Distribution de la sismicité et du réseau de détection utilisé par le RéNaSS pour l'année **2016**.

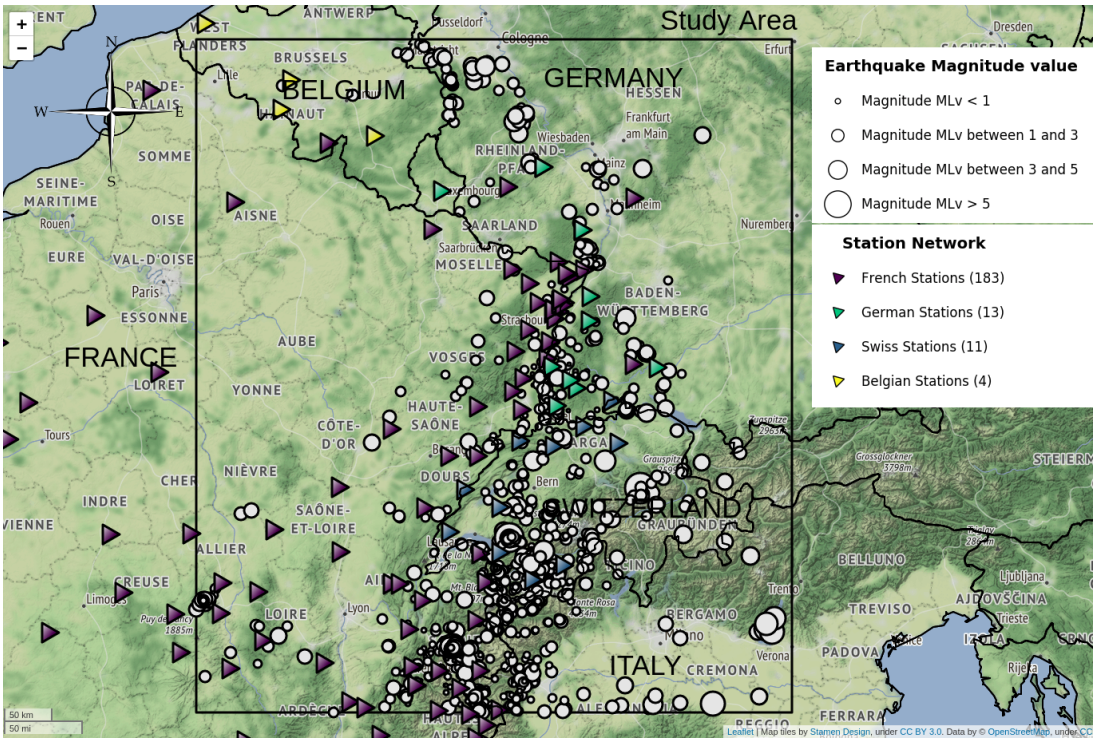


FIGURE B.6: Distribution de la sismicité et du réseau de détection utilisé par le RéNaSS pour l'année **2017**.

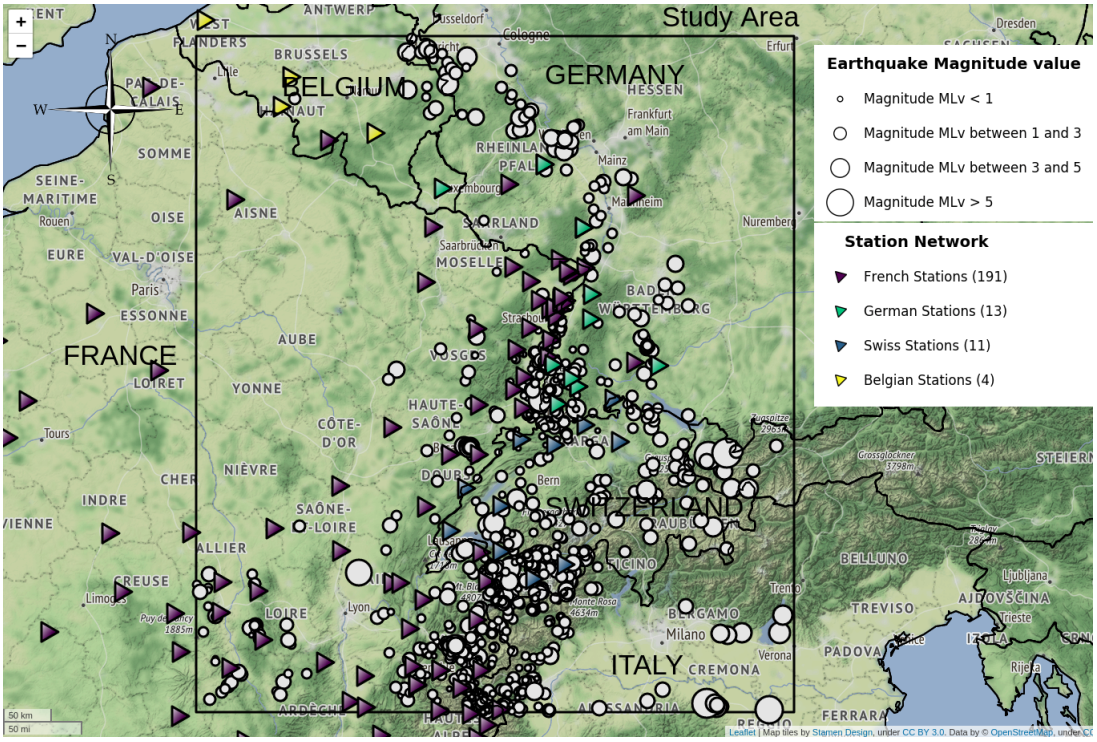


FIGURE B.7: Distribution de la sismicité et du réseau de détection utilisé par le RéNaSS pour l'année **2018**.

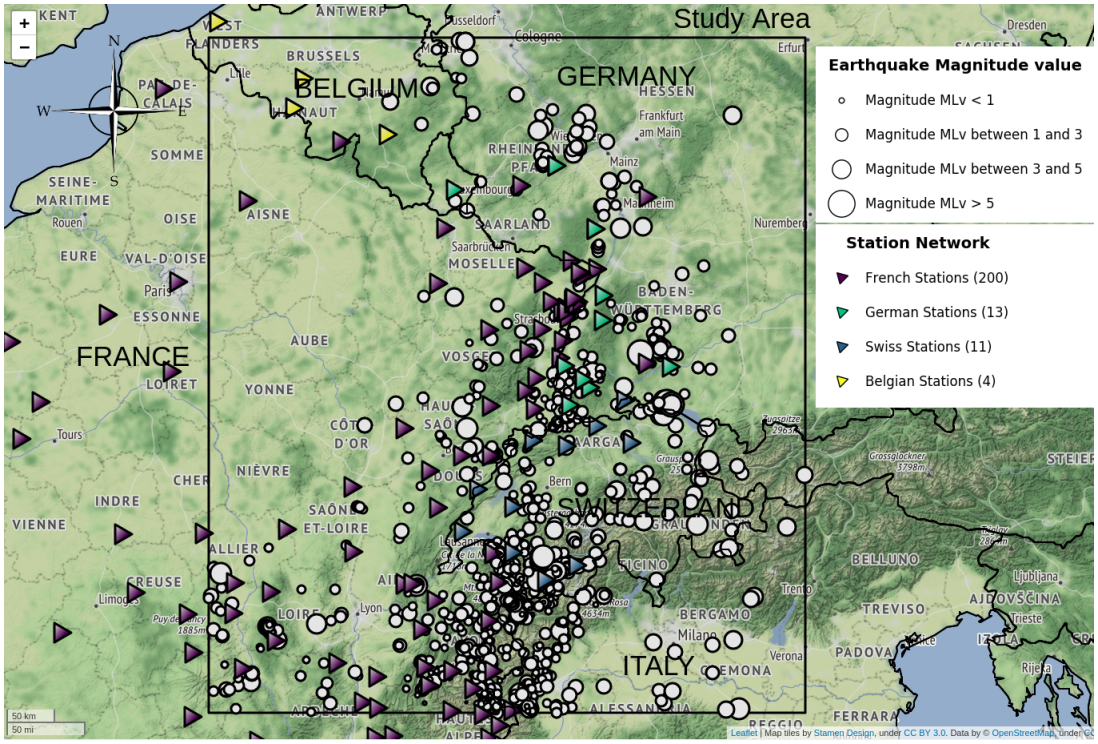


FIGURE B.8: Distribution de la sismicité et du réseau de détection utilisé par le RéNaSS pour l'année **2019**.

Annexe C

Distribution du nombre de pointés manuels effectués pour l'année 2016 en fonction des stations AlpArray

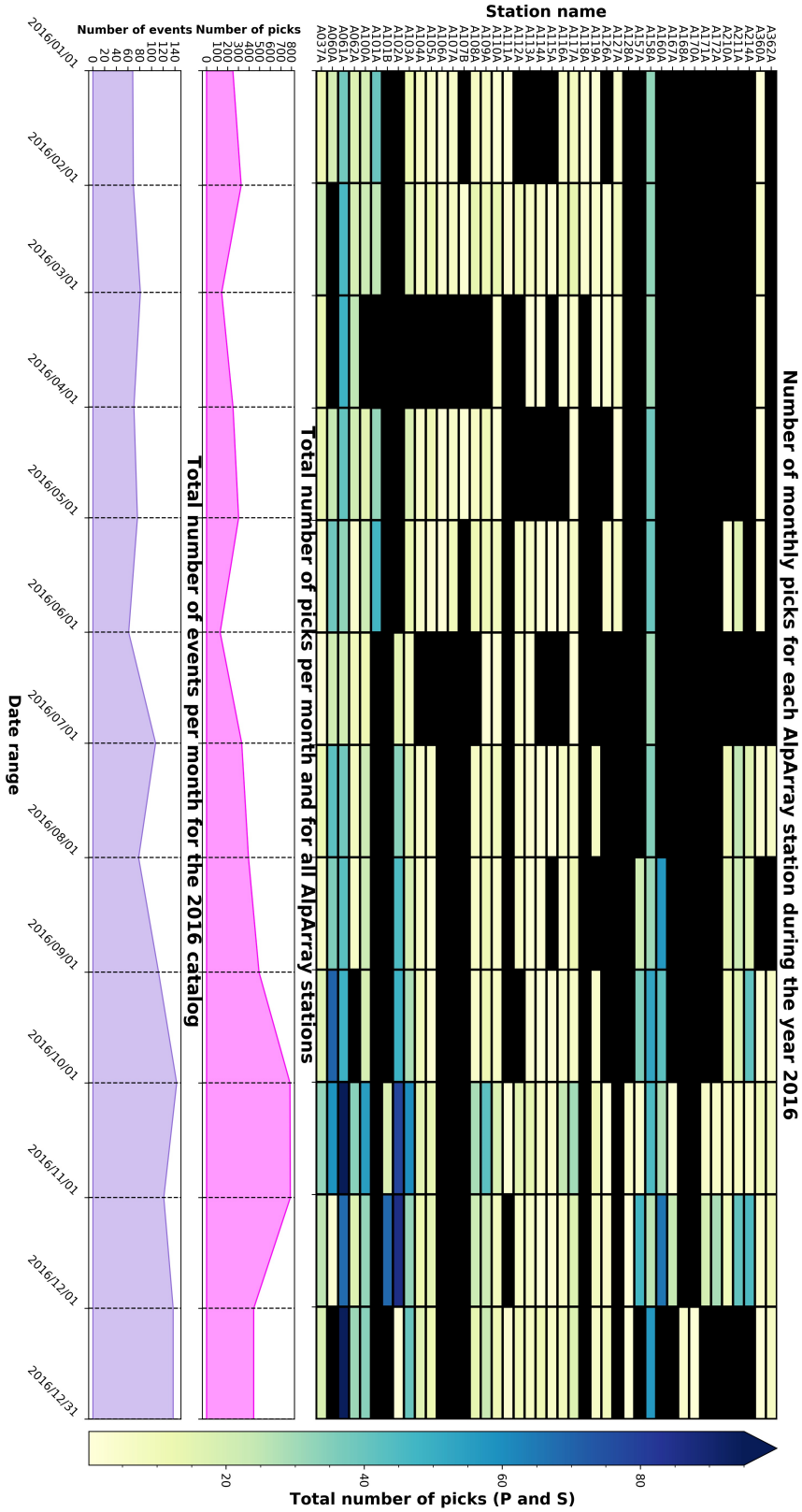


FIGURE C.1: Distribution du nombre de pointés manuels effectués pour l'année 2016 en fonction des stations AlpArray.

Annexe D

Modèles de vitesse utilisés pour les solutions épicentrales et hypocentrales proposées dans le chapitre 4.

TABLE D.1: Modèle de vitesse multicouche utilisé pour les solutions épicentrales et hypocentrales proposées dans les Figures 4.29 et 4.33 pour le tir de la carrière de Dotternhausen identifié le 15 juillet 2016 à 10h25 (MLv 1.7).

Depth	P- and	S-wave velocity	Density
0.0	3.83	2.26	2.7
1.0	3.88	2.29	2.7
1.0	3.88	2.29	2.7
2.0	4.41	2.61	2.7
2.0	4.41	2.61	2.7
5.0	4.71	2.69	2.7
5.0	4.71	2.69	2.7
8.0	5.45	3.11	2.7
8.0	5.45	3.11	2.7
11.0	5.45	3.12	2.7
11.0	5.45	3.12	2.7
14.0	5.63	3.21	2.7
14.0	5.63	3.21	2.7
17.0	5.99	3.42	2.7
17.0	5.99	3.42	2.7
20.0	6.78	3.82	2.7
20.0	6.78	3.82	2.7
22.0	6.85	3.86	2.7
22.0	6.85	3.86	2.7
24.0	6.92	3.87	2.7
24.0	6.92	3.87	2.7
26.0	7.26	4.06	2.7
26.0	7.26	4.06	2.7
28.0	7.54	4.21	2.7
28.0	7.54	4.21	2.7
30.2	8.01	4.40	2.7
mantle			
30.2	8.15	4.40	3.3
60.0	8.15	4.55	3.3
410	8.90	4.7	3.5
410	9.10	4.9	3.7
670	10.2	5.5	4.0
670	10.7	5.9	4.4
2891	13.7	7.2	5.6
outer-core			
2891	8.0	0.0	9.9
5149.5	10.3	0.0	12.
inner-core			
5149.5	11	3.5	12.7
6371	11.3	3.7	13

TABLE D.2: Modèle de vitesse à 3 couches utilisé pour les solutions épicentrales et hypocentrales proposées dans la Figure 4.30 pour le tir de la carrière de Dotternhausen identifié le 15 juillet 2016 à 10h25 (MLv 1.7).

Depth	P- and	S-wave velocity	Density
0.0	4.24	2.38	2.7
2.4	4.24	2.38	2.7
2.4	5.72	3.04	2.7
20.1	5.72	3.94	2.7
20.1	7.39	3.85	2.7
30.2	7.39	3.85	2.7
mantle			
30.2	8.15	4.40	3.3
60.0	8.15	4.55	3.3
410	8.90	4.7	3.5
410	9.10	4.9	3.7
670	10.2	5.5	4.0
670	10.7	5.9	4.4
2891	13.7	7.2	5.6
outer-core			
2891	8.0	0.0	9.9
5149.5	10.3	0.0	12.2
inner-core			
5149.5	11	3.5	12.7
6371	11.3	3.7	13

TABLE D.3: Modèle de vitesse à multicouche utilisé pour les solutions épicentrales et hypocentrales proposées dans les Figures 4.31 et 4.35 pour le séisme qui a eu lieu le 16 juillet 2016 à 02h36 dans les Pré-alpes Suisses (MLv 2.7).

Depth (km)	P-wave and	S-wave velocity (km/s)	Density
0.0	4.65	2.68	2.7
1.0	4.65	2.68	2.7
1.0	4.84	2.75	2.7
2.0	4.84	2.75	2.7
2.0	4.89	2.71	2.7
5.0	4.89	2.71	2.7
5.0	5.28	2.90	2.7
8.0	5.28	2.90	2.7
8.0	5.31	2.91	2.7
11.0	5.31	2.91	2.7
11.0	5.57	3.02	2.7
14.0	5.57	3.02	2.7
14.0	5.60	3.03	2.7
17.0	5.60	3.03	2.7
17.0	5.84	3.09	2.7
20.0	5.84	3.09	2.7
20.0	6.09	3.20	2.7
22.0	6.09	3.20	2.7
22.0	6.18	3.25	2.7
24.0	6.18	3.25	2.7
24.0	7.10	3.69	2.7
26.0	7.65	3.92	2.7
28.0	7.65	3.92	2.7
28.0	8.09	4.12	2.7
30.2	8.15	4.12	2.7
mantle			
30.2	8.15	4.40	3.3
60.0	8.15	4.55	3.3
410	8.90	4.7	3.5
410	9.10	4.9	3.7
670	10.2	5.5	4.0
670	10.7	5.9	4.4
2891	13.7	7.2	5.6
outer-core			
2891	8.0	0.0	9.9
5149.5	10.3	0.0	12.2
inner-core			
5149.5	11	3.5	12.7
6371	11.3	3.7	13

TABLE D.4: Modèle de vitesse à 3 couches utilisé pour les solutions épicentrales et hypocentrales proposées dans les Figures 4.32 et 4.35 pour le séisme qui a eu lieu le 16 juillet 2016 à 02h36 dans les Pré-alpes Suisses (MLv 2.7).

Depth (km)	P- and	S-wave velocity (km/s)	Density
0.0	4.24	2.52	2.7
2.4	4.24	2.52	2.7
2.4	5.72	3.18	2.7
20.1	5.72	3.18	2.7
20.1	7.39	3.69	2.7
30.2	7.39	3.69	2.7
mantle			
30.2	8.15	4.40	3.3
60.0	8.15	4.55	3.3
410	8.90	4.7	3.5
410	9.10	4.9	3.7
670	10.2	5.5	4.0
670	0.7	5.9	4.4
2891	13.7	7.2	5.6
outer-core			
2891	8.0	0.0	9.9
5149.5	10.3	0.0	12.2
inner-core			
5149.5	11	3.5	12.7
6371	11.3	3.7	13

TABLE D.5: Modèle de vitesse à 3 couches utilisé pour les solutions épicentrales et hypocentrales proposées dans les Figure 4.32 et 4.36 pour le tir de la caiirère de Dotternhausen identifié le 15 juillet 2016 à 10h25 (MLv 1.7).

Depth (km)	P-wave and	S-wave velocity (km/s)	Density
0.0	5.46	3.23	2.7
2.4	5.46	3.23	2.7
2.4	6.13	3.5	2.7
20.1	6.13	3.5	2.7
20.1	6.91	3.86	2.7
30.2	6.91	3.86	2.7
mantle			
30.2	8.15	4.40	3.3
60.0	8.15	4.55	3.3
410	8.90	4.7	3.5
410	9.10	4.9	3.7
670	10.2	5.5	4.0
670	10.7	5.9	4.4
2891	13.7	7.2	5.6
outer-core			
2891	8.0	0.0	9.9
5149.5	10.3	0.0	12.2
inner-core			
5149.5	11	3.5	12.7
6371	11.3	3.7	13

TABLE D.6: Modèle de vitesse à 3 couches utilisé pour les solutions épicentrales et hypocentrales proposées dans la Figure 4.34 pour le tir de la caiirère de Dotternhausen identifié le 15 juillet 2016 à 10h25 (MLv 1.7).

Depth (km)	P-wave and	S-wave velocity (km/s)	Density
0.0	4.28	2.53	2.7
2.4	4.28	2.53	2.7
2.4	5.78	3.3	2.7
20.1	5.78	3.3	2.7
20.1	6.79	3.79	2.7
30.2	6.79	3.79	2.7
mantle			
30.2	8.15	4.40	3.3
60.0	8.15	4.55	3.3
410	8.90	4.7	3.5
410	9.10	4.9	3.7
670	10.2	5.5	4.0
670	10.7	5.9	4.4
2891	13.7	7.2	5.6
outer-core			
2891	8.0	0.0	9.9
5149.5	10.3	0.0	12.2
inner-core			
5149.5	11	3.5	12.7
6371	11.3	3.7	13

TABLE D.7: Modèle de vitesse à 3 couches utilisé pour les solutions épicentrales et hypocentrales proposées dans la Figure 4.37 pour le tir de la carrière de Dotternhausen identifié le 15 juillet 2016 à 10h25 (MLv 1.7).

Depth (km)	P-wave and	S-wave velocity (km/s)	Density
0.0	4.65	2.80	2.7
1.0	4.65	2.80	2.7
1.0	4.84	2.89	2.7
2.0	4.84	2.89	2.7
2.0	4.89	2.91	2.7
5.0	4.89	2.91	2.7
5.0	5.28	3.03	2.7
8.0	5.28	3.03	2.7
8.0	5.31	3.05	2.7
11.0	5.31	3.05	2.7
11.0	5.57	3.10	2.7
14.0	5.57	3.10	2.7
14.0	5.60	3.14	2.7
17.0	5.60	3.14	2.7
17.0	5.84	3.17	2.7
20.0	5.84	3.17	2.7
20.0	6.09	3.27	2.7
22.0	6.09	3.27	2.7
22.0	6.18	3.28	2.7
24.0	6.18	3.28	2.7
24.0	7.10	3.64	2.7
26.0	7.65	3.90	2.7
28.0	7.65	3.90	2.7
28.0	8.09	4.10	2.7
30.2	8.15	4.14	2.7
mantle			
30.2	8.15	4.40	3.3
60.0	8.15	4.55	3.3
410	8.90	4.7	3.5
410	9.10	4.9	3.7
670	10.2	5.5	4.0
670	10.7	5.9	4.4
2891	13.7	7.2	5.6
outer-core			
2891	8.0	0.0	9.9
5149.5	10.3	0.0	12.2
inner-core			
5149.5	11	3.5	12.7
6371	11.3	3.7	13

TABLE D.8: Modèle de vitesse à 3 couches utilisé pour les solutions épicentrales et hypocentrales proposées dans les Figures 4.38 et 4.39 pour le tir de la carrière de Dotternhausen identifié le 15 juillet 2016 à 10h25 (MLv 1.7).

Depth (km)	P-wave and	S-wave velocity (km/s)	Density
0.0	4.73	2.81	2.7
2.4	4.73	1.81	2.7
2.4	6.07	3.46	2.7
20.1	6.07	3.46	2.7
20.1	7.19	4.01	2.7
30.2	7.19	4.01	2.7
mantle			
30.2	8.15	4.40	3.3
60.0	8.15	4.55	3.3
410	8.90	4.7	3.5
410	9.10	4.9	3.7
670	10.2	5.5	4.0
670	10.7	5.9	4.4
2891	13.7	7.2	5.6
outer-core			
2891	8.0	0.0	9.9
5149.5	10.3	0.0	12.2
inner-core			
5149.5	11	3.5	12.7
6371	11.3	3.7	13

Annexe E

Modèles de vitesse testées pour optimiser les processus d'association (chapitre 4).

TABLE E.1: Modèle de vitesse à 3 couches n°11.

Depth (km)	P-wave and	S-wave velocity (km/s)	Density
0.0	4.70	2.78	2.7
2.4	4.70	2.78	2.7
2.4	5.75	3.29	2.7
20.1	5.75	3.29	2.7
20.1	7.30	4.08	2.7
30.2	7.30	4.08	2.7
mantle			
30.2	8.15	4.40	3.3
60.0	8.15	4.55	3.3
410	8.90	4.7	3.5
410	9.10	4.9	3.7
670	10.2	5.5	4.0
670	10.7	5.9	4.4
2891	13.7	7.2	5.6
outer-core			
2891	8.0	0.0	9.9
5149.5	10.3	0.0	12.2
inner-core			
5149.5	11	3.5	12.7
6371	11.3	3.7	13

TABLE E.2: Modèle de vitesse à 3 couches n°25

Depth (km)	P-wave and	S-wave velocity (km/s)	Density
0.0	3.54	2.09	2.7
2.4	3.54	2.09	2.7
2.4	5.84	3.34	2.7
20.1	5.84	3.34	2.7
20.1	7.30	4.08	2.7
30.2	7.30	4.08	2.7
mantle			
30.2	8.15	4.40	3.3
60.0	8.15	4.55	3.3
410	8.90	4.7	3.5
410	9.10	4.9	3.7
670	10.2	5.5	4.0
670	10.7	5.9	4.4
2891	13.7	7.2	5.6
outer-core			
2891	8.0	0.0	9.9
5149.5	10.3	0.0	12.2
inner-core			
5149.5	11	3.5	12.7
6371	11.3	3.7	13

TABLE E.3: Modèle de vitesse à 3 couches n°31

Depth (km)	P-wave and	S-wave velocity (km/s)	Density
0.0	5.21	3.08	2.7
2.4	5.21	3.08	2.7
2.4	5.72	3.27	2.7
20.1	5.72	3.27	2.7
20.1	7.37	4.12	2.7
30.2	7.37	4.12	2.7
mantle			
30.2	8.15	4.40	3.3
60.0	8.15	4.55	3.3
410	8.90	4.7	3.5
410	9.10	4.9	3.7
670	10.2	5.5	4.0
670	10.7	5.9	4.4
2891	13.7	7.2	5.6
outer-core			
2891	8.0	0.0	9.9
5149.5	10.3	0.0	12.2
inner-core			
5149.5	11	3.5	12.7
6371	11.3	3.7	13

TABLE E.4: Modèle de vitesse à 3 couches n°38.

Depth (km)	P-wave and	S-wave velocity (km/s)	Density
0.0	4.07	2.41	2.7
2.4	4.07	2.41	2.7
2.4	5.73	3.27	2.7
20.1	5.73	3.27	2.7
20.1	7.45	4.16	2.7
30.2	7.45	4.16	2.7
mantle			
30.2	8.15	4.40	3.3
60.0	8.15	4.55	3.3
410	8.90	4.7	3.5
410	9.10	4.9	3.7
670	10.2	5.5	4.0
670	10.7	5.9	4.4
2891	13.7	7.2	5.6
outer-core			
2891	8.0	0.0	9.9
5149.5	10.3	0.0	12.2
inner-core			
5149.5	11	3.5	12.7
6371	11.3	3.7	13

TABLE E.5: Modèle de vitesse multicouche n°10.

Depth (km)	P-wave and	S-wave velocity (km/s)	Density
0.0	4.65	2.71	2.7
1.0	4.65	2.71	2.7
1.0	4.84	2.80	2.7
2.0	4.84	2.80	2.7
2.0	4.89	2.83	2.7
5.0	4.89	2.83	2.7
5.0	5.28	2.91	2.7
8.0	5.28	2.91	2.7
8.0	5.31	3.00	2.7
11.0	5.31	3.00	2.7
11.0	5.57	3.04	2.7
14.0	5.57	3.04	2.7
14.0	5.60	3.04	2.7
17.0	5.60	3.04	2.7
17.0	5.84	3.14	2.7
20.0	5.84	3.14	2.7
20.0	6.09	3.23	2.7
22.0	6.09	3.23	2.7
22.0	6.18	3.25	2.7
24.0	6.18	3.25	2.7
24.0	7.10	3.69	2.7
26.0	7.65	3.94	2.7
28.0	7.65	3.94	2.7
28.0	8.09	4.10	2.7
30.2	8.15	4.10	2.7
mantle			
30.2	8.15	4.40	3.3
60.0	8.15	4.55	3.3
410	8.90	4.7	3.5
410	9.10	4.9	3.7
670	10.2	5.5	4.0
670	10.7	5.9	4.4
2891	13.7	7.2	5.6
outer-core			
2891	8.0	0.0	9.9
5149.5	10.3	0.0	12.2
inner-core			
5149.5	11	3.5	12.7
6371	11.3	3.7	13

TABLE E.6: Modèle de vitesse multicouche n°24.

Depth (km)	P-wave and	S-wave velocity (km/s)	Density
0.0	4.65	2.75	2.7
1.0	4.65	2.75	2.7
1.0	4.84	2.86	2.7
2.0	4.84	2.86	2.7
2.0	4.89	2.89	2.7
5.0	4.89	2.89	2.7
5.0	5.28	3.01	2.7
8.0	5.28	3.01	2.7
8.0	5.31	3.03	2.7
11.0	5.31	3.03	2.7
11.0	5.57	3.18	2.7
14.0	5.57	3.18	2.7
14.0	5.60	3.20	2.7
17.0	5.60	3.20	2.7
17.0	5.84	3.33	2.7
20.0	5.84	3.33	2.7
20.0	6.09	3.45	2.7
22.0	6.09	3.45	2.7
22.0	6.18	3.48	2.7
24.0	6.18	3.48	2.7
24.0	7.10	3.96	2.7
26.0	7.65	3.96	2.7
28.0	7.65	4.27	2.7
28.0	8.09	4.31	2.7
30.2	8.15	4.31	2.7
mantle			
30.2	8.15	4.40	3.3
60.0	8.15	4.55	3.3
410	8.90	4.7	3.5
410	9.10	4.9	3.7
670	10.2	5.5	4.0
670	10.7	5.9	4.4
2891	13.7	7.2	5.6
outer-core			
2891	8.0	0.0	9.9
5149.5	10.3	0.0	12.2
inner-core			
5149.5	11	3.5	12.7
6371	11.3	3.7	13

TABLE E.7: Modèle de vitesse multicouche n°25.

Depth (km)	P-wave and	S-wave velocity (km/s)	Density
0.0	4.20	2.48	2.7
1.0	4.41	2.61	2.7
1.0	4.41	2.61	2.7
2.0	4.47	2.64	2.7
2.0	4.47	2.64	2.7
5.0	4.57	2.64	2.7
5.0	4.57	2.64	2.7
8.0	4.76	2.72	2.7
8.0	4.76	2.72	2.7
11.0	5.61	3.21	2.7
11.0	5.61	3.21	2.7
14.0	5.66	3.23	2.7
14.0	5.66	3.23	2.7
17.0	5.80	3.31	2.7
17.0	5.80	3.31	2.7
20.0	6.55	3.69	2.7
20.0	6.55	3.69	2.7
22.0	6.60	3.74	2.7
22.0	6.60	3.74	2.7
24.0	6.98	3.90	2.7
24.0	6.98	3.90	2.7
26.0	7.27	4.06	2.7
28.0	7.27	4.34	2.7
28.0	7.77	4.34	2.7
30.2	7.99	4.40	2.7
mantle			
30.2	8.15	4.40	3.3
60.0	8.15	4.55	3.3
410	8.90	4.7	3.5
410	9.10	4.9	3.7
670	10.2	5.5	4.0
670	10.7	5.9	4.4
2891	13.7	7.2	5.6
outer-core			
2891	8.0	0.0	9.9
5149.5	10.3	0.0	12.2
inner-core			
5149.5	11	3.5	12.7
6371	11.3	3.7	13

TABLE E.8: Modèle de vitesse multicouche n°27.

Depth (km)	P-wave and	S-wave velocity (km/s)	Density
0.0	3.59	2.12	2.7
1.0	4.95	2.93	2.7
1.0	4.95	2.93	2.7
2.0	5.25	3.10	2.7
2.0	5.25	3.10	2.7
5.0	5.39	3.08	2.7
5.0	5.39	3.08	2.7
8.0	5.45	3.11	2.7
8.0	5.45	3.11	2.7
11.0	5.62	3.21	2.7
11.0	5.62	3.21	2.7
14.0	5.77	3.29	2.7
14.0	5.77	3.29	2.7
17.0	6.17	3.53	2.7
17.0	6.17	3.53	2.7
20.0	6.22	3.55	2.7
20.0	6.22	3.55	2.7
22.0	6.37	3.56	2.7
22.0	6.37	3.56	2.7
24.0	6.45	3.60	2.7
24.0	6.45	3.60	2.7
26.0	7.39	4.13	2.7
28.0	7.41	4.14	2.7
28.0	7.41	4.14	2.7
30.2	7.60	4.25	2.7
mantle			
30.2	8.15	4.40	3.3
60.0	8.15	4.55	3.3
410	8.90	4.7	3.5
410	9.10	4.9	3.7
670	10.2	5.5	4.0
670	10.7	5.9	4.4
2891	13.7	7.2	5.6
outer-core			
2891	8.0	0.0	9.9
5149.5	10.3	0.0	12.2
inner-core			
5149.5	11	3.5	12.7
6371	11.3	3.7	13

Annexe F

Modèle de vitesse testé pour la détection automatique des événements dans la zone d'étude exposée dans le chapitre 4.

TABLE F.1: Modèle de vitesse tiré de l'inversion des paramètres hypocentraux et de vitesse sous VELEST à partir du modèle Haslach.

Depth (km)	P-wave and	S-wave velocity (km/s)
0.0	4.65	2.76
1.0	4.65	2.76
1.0	4.84	2.88
2.0	4.84	2.88
2.0	4.89	2.91
5.0	4.89	2.91
5.0	5.28	3.14
8.0	5.28	3.14
8.0	5.31	3.16
11.0	5.31	3.16
11.0	5.57	3.27
14.0	5.57	3.27
14.0	5.60	3.29
17.0	5.60	3.29
17.0	5.84	3.41
20.0	5.84	3.41
20.0	6.09	3.58
22.0	6.09	3.58
22.0	6.18	3.63
24.0	6.18	3.63
24.0	7.10	4.07
26.0	7.65	4.20
28.0	7.65	4.20
28.0	8.09	4.31
30.2	8.15	4.40

Alexandra RENOUARD

Détection automatique et classification basée sur l'apprentissage machine des séismes de faible magnitude dans une région continentale stable

Résumé

La compréhension des mécanismes qui gouvernent l'occurrence et la distribution de la sismicité faible à modérée des régions continentales stables est entravée par les capacités limitées des algorithmes traditionnels à détecter les petits séismes dans des environnements anthropisés, malgré le déploiement intensif des réseaux de stations. Cette thèse développe une procédure de détection automatique des séismes de faible magnitude à travers SeisComP3 et le Calcul de Haute Performance. Cette nouvelle procédure réduit la contamination des séismes détectés par du bruit sismique en tenant compte des niveaux de bruit enregistré aux stations, de la géométrie du réseau de stations et du milieu de propagation des ondes sismiques. En incorporant un algorithme d'apprentissage machine supervisé, elle discrimine efficacement les séismes détectés, des tirs de carrière et des faux événements associés à du bruit. Les résultats sont prometteurs : 50% de séismes de magnitude inférieure à 1.2 sont détectés en plus. Ce travail vise à une plus large exploration de l'apprentissage machine dans les observatoires sismologiques.

Mots clés : détection, discrimination, apprentissage machine supervisé, intelligence artificielle, tirs de carrière, bruit sismique, séismes de faible magnitude, calcul de haute performance

Résumé en anglais

Understanding the mechanisms responsible for the occurrence of low-to-moderate seismicity in stable continental regions is hampered by the limited capabilities of the algorithms used to detect small-magnitude earthquakes in anthropogenic environments, and despite extensive station deployment. This thesis work develops an automatic detection procedure via SeisComP3 and High Performance Computing. This new procedure takes into account the station noise level, the station network geometry and the seismic wave propagation medium to reduce the detection rate of earthquakes contaminated by seismic noise. By incorporating a supervised machine learning algorithm, it also robustly discriminates all detected earthquakes from quarry blasts and noise-related events. The detection results are promising: compared to the reference French National Catalog for the same time period, twice as many earthquakes with magnitudes less than 1.2 are detected. This work also promotes a broader implication of hybrid intelligence monitoring within seismological observatories.

Keywords : detection, discrimination, supervised machine learning, artificial intelligence, quarry blasts, seismic noise, small-magnitude earthquakes, High Performance Computing