UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES DE LA TERRE ET DE L'ENVIRONNEMENT

EOST/IPGS – UMR 7516

THÈSE présentée par :

Nadège LANGET

soutenue le : **9 décembre 2014**

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

Discipline/ Spécialité : Géophysique / Sismologie

Détection et caractérisation massives de phénomènes sismologiques pour la surveillance d'événements traditionnels et la recherche systématique de phénomènes rares

THÈSE dirigée par :
Mme MAGGI Alessia — Professeur, Université de Strasbourg

RAPPORTEURS :
Mme DESCHAMPS Anne — Directeur de recherches, Géoazur, Nice
Mme MANGENEY Anne — Professeur, IPGP, Paris

EXAMINATEURS :
M. GUILBERT Jocelyn — Ingénieur chercheur, CEA Bruyères-le-Châtel
Mme LAMBOTTE Sophie — Physicien Adjoint, Université de Strasbourg
M. RIVERA Luis — Professeur, Université de Strasbourg

Remerciements

> *«Quand on travaille pour plaire aux autres, on peut ne pas réussir ; mais les choses qu'on a faites pour se contenter soi-même ont toujours une chance d'intéresser quelqu'un.»*
>
> Proust

Mes premiers remerciements vont bien évidemment à ma directrice de thèse, Alessia, pour m'avoir fait confiance en me proposant ce sujet il y a trois ans. Merci surtout de m'avoir poussée à toujours ~~essayer de~~ donner le meilleur de moi-même.

Je remercie également les membres du jury qui ont accepté d'évaluer ce travail, et Thomas, d'être venu assister à ma soutenance en tant qu'invité.

Merci à toi, Christophe, de m'avoir confié les TD d'informatique des 1A pendant deux ans. J'espère qu'ils en sont tous sortis indemnes ;-)

Merci à vous, Thomas et Corentin, de m'avoir accueillie pendant un mois dans le labo de sismo de l'observatoire belge pour travailler sur les données du Kawah Ijen (qui m'auront donné bien du fil à retordre) !

Merci à mes « co-thésards » du bureau 404, Marylin, Maximilien et aussi Karim pour la bonne humeur ; et merci aussi aux doctorants des autres étages à qui je souhaite bon courage et bonne continuation pour la suite.

Merci à vous, chers parents, pour tout, depuis le début ; et à toi, ma grande sœur, d'être mon moteur de tous les instants. Merci pour votre soutien indéfectible (et merci aussi d'avoir réussi à me supporter toutes ces années !).

A tous, cette thèse est la vôtre autant que la mienne.

Table des matières

Introduction générale

Les séismes font partie des phénomènes terrestres potentiellement destructeurs qui se produisent régulièrement à l'échelle du globe. Chaque jour, des milliers de secousses de magnitudes plus ou moins fortes émettent des ondes qui se propagent dans le sous-sol et sont enregistrées continûment en surface par les sismomètres. C'est le recoupement entre les divers enregistrements d'un même phénomène qui permet d'en apprendre plus sur celui-ci.

Depuis quelques années, la multiplication du nombre de réseaux de stations sismologiques installés de manière permanente ou temporaire n'a fait qu'accroître la quantité de données disponibles pour alimenter des études toujours plus détaillées de la structure de la Terre, des ruptures complexes des failles lors des grands séismes, ou -et ce sera le sujet de cette thèse - de la distribution et caractérisation des petits séismes. Si l'on considère principalement l'étude des petits ou micro-séismes, le développement des réseaux de stations se fait essentiellement dans deux buts (Fig. 1) :

- d'abord scientifique : pour améliorer la compréhension des différents phénomènes sismologiques enregistrés. La répartition de la sismicité dans une zone d'étude ou le type de signaux enregistrés nous renseignent sur les éventuels facteurs qui peuvent favoriser le déclenchement de tel ou tel type d'événement. Elle peut aussi nous permettre de mieux appréhender les signes annonciateurs d'une éruption (dans le cas d'une sismicité volcanique) ou d'un séisme de magnitude plus forte que les autres (dans le cas des essaims sismiques)... (voir le point suivant)

- ensuite opérationnel : pour permettre la surveillance de phénomènes sismologiques naturels fréquents, comme la sismicité volcanique ou les crises sismiques (séquences précurseurs-répliques, essaims) qui affectent certaines zones du globe ; pour documenter ou suivre la sismicité induite (géothermie, mines, stockage du CO_2...) ; pour affiner l'estimation de l'aléa sismique. Le travail de dépouillement de la sismicité enregistrée effectué par les observatoires ou structures assimilées permet de fournir (*a minima*) des cartes de localisation, et la classification des types d'événements qui se sont produits, posant ainsi les bases d'une analyse scientifique plus poussée (revoir le point précédent).

Ces deux objectifs sont intrinsèquement liés entre eux et partagent le besoin d'obtenir une réponse aux questions suivantes : est-ce qu'il s'est produit un événement sismique (**détection**) ; où et quand a-t-il eu lieu (**localisation**) ; de quel type d'événement s'agit-il (**classification**).

Or, comme le nombre de séismes enregistrés est de plus en plus grand (du fait du nombre croissant de stations et du fait de l'enregistrement d'événements de plus en plus petits), il

devient difficile de tous les dépouiller manuellement, le travail étant long et fastidieux. Les outils de traitement automatiques s'imposent alors comme une aide et une complémentarité au dépouillement manuel. Ils rendent le processus à la fois plus rapide et plus systématique, ce qui est important pour l'homogénéité des résultats. En effet, bien souvent, plusieurs opérateurs dépouillent les données manuellement, mais ils n'auront jamais exactement la même manière de procéder. De même, un seul et même opérateur ne sera pas toujours constant et régulier au cours du temps. Les systèmes automatiques, eux, présentent l'avantage de procéder de la même manière en toutes circonstances (seuls les paramètres de calcul sont à définir au début), et permettent ainsi une harmonisation du dépouillement. Ils se doivent également d'être **efficaces** (on veut être sûr de détecter, localiser et classer tous les événements sismiques), **fiables** (on veut être sûr que les événements détectés et localisés sont bien des séismes et que la classification entre types de séismes est correcte) et robustes. Un algorithme automatique ne peut être à la fois le plus efficace et le plus fiable puisqu'il existe un *trade-off* entre ces deux paramètres, c'est-à-dire que le gain d'efficacité se fait aux dépens de la fiabilité et vice-versa. L'un des défis de l'automatisation est donc de rechercher le meilleur compromis permettant de la rendre la plus robuste possible.

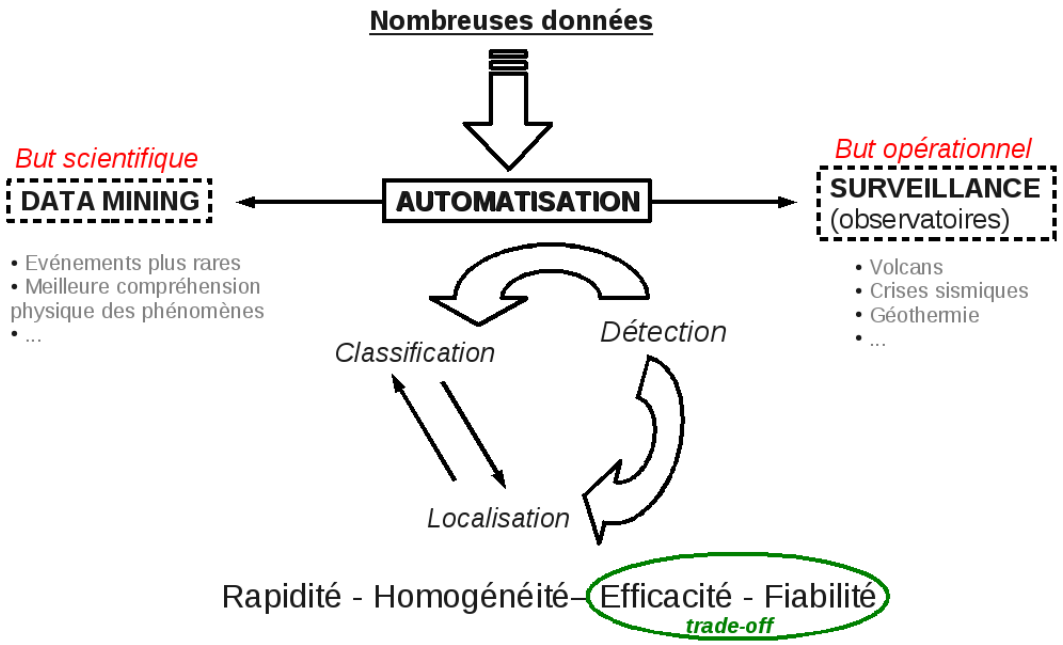


FIG. 1: Schéma de la problématique dans laquelle s'inscrit cette thèse : l'automatisation du traitement des données sismologiques.

L'automatisation en soi n'est donc pas aussi simple qu'elle n'y paraît. Lorsque les données sont traitées manuellement, l'opérateur analyse une certaine quantité de données simultanément sans s'en rendre compte. Par exemple, la détection d'un événement en confrontant les enregistrements réalisés en plusieurs endroits est relativement instantanée (dans le cas de données non bruitées). En automatique, déjà, cela suppose de réussir à détecter quelque chose (i.e. « il se passe quelque chose qui sort de l'ordinaire » - il faut donc définir « l'ordinaire ») puis à corréler ce quelque chose sur toutes les stations, sachant qu'elles n'enregistreront pas

le phénomène exactement en même temps en fonction de la distance qui les en sépare. Vient ensuite l'étape de localisation : traditionnellement, celle-ci se fait en pointant simplement les différentes phases dans le signal. Généralement, si le pointé manuel de la première arrivée (onde P) ne pose pas de problème (si, encore une fois, les données ne sont pas trop bruitées), celui des ondes S ou des autres phases peut parfois s'avérer plus délicat, ce qui rend la tâche encore plus difficile en automatique. On verra dans la suite de ce manuscrit qu'il est tout de même possible de contourner ce problème via l'utilisation d'une fonction caractéristique basée sur le moment statistique d'une distribution. Enfin, la tâche de classification se révèle ardue elle aussi car elle suppose en premier lieu une définition claire, précise et synthétique des différents types de phénomènes qui peuvent se produire, à partir d'une analyse un tant soi peu détaillée des formes d'ondes (l'analyse visuelle étant la première qui vient à l'esprit : « je vois que le signal a telle forme, je peux confirmer/infirmer son appartenance à telle ou telle classe »). On peut aisément imaginer à quel point le problème devient complexe lorsqu'il s'agit de demander à l'ordinateur de réaliser la même « gymnastique ».

L'automatisation de toutes ces étapes de traitement des données sismologiques s'apparente alors en quelque sorte à un problème de compression de l'information (quelles sont les informations essentielles prises en compte par l'opérateur lors d'une analyse manuelle ; quels mécanismes et cheminements lui permettent de déterminer le résultat ?).

De nombreux auteurs ont déjà travaillé sur la problématique d'automatisation du dépouillement des données en sismologie. Les diverses études associées seront rappelées au fil de ce manuscrit. Le travail de thèse présenté ici s'est déroulé en deux axes majeurs :

- la détection et la localisation simultanées des événements sismiques grâce à une méthode qui n'utilise pas de pointés automatiques des différentes phases du signal mais plutôt les enregistrements continus des formes d'ondes. L'algorithme, Waveloc, sera présenté dans le chapitre I.1.

- la classification des événements sismiques à partir des signaux déjà détectés et prédécoupés. Les méthodes utilisées seront détaillées dans le chapitre I.2.

Cette thèse s'est plus spécifiquement focalisée sur le traitement des données sismologiques enregistrées en domaine volcanique. L'activité volcanique s'accompagne en effet d'une activité sismique intense associée à des processus complexes qui se produisent en profondeur. L'analyse des crises sismiques, périodes où un grand nombre de séismes se produit dans un court lapse de temps, est particulièrement importante et propice à l'utilisation d'un système de traitement automatisé.

Deux exemples d'application distincts seront discutés ici : celui du volcan (effusif) du Piton de la Fournaise en partie II et du volcan (explosif) du Kawah Ijen (Indonésie) en partie III. On verra que le Piton de la Fournaise se révèle être un cas simple, « idéal » pour le traitement automatique des données, tandis que le Kawah Ijen est beaucoup plus complexe, mais très intéressant pour le développement de nouvelles stratégies d'automatisation.

Une attention particulière a aussi été apportée au choix des paramètres qui permettent de tirer le meilleur parti possible des jeux de données (toujours avec l'idée de maximiser à la fois l'efficacité et la fiabilité) et sera détaillée pour chaque exemple d'application dans les parties II et III.

Partie I

Partie théorique

Sommaire

CHAPITRE I.1

Détection et localisation automatiques des événements sismiques : Waveloc

Cette partie vise à introduire et présenter Waveloc, un algorithme de détection et de localisation automatiques des événements sismiques basé sur la migration de formes d'ondes continues. Waveloc a fait l'objet d'un article publié en 2014 dans BSSA [Langet et al., 2014] (voir l'annexe A1). Ce chapitre s'en inspire très largement.

I.1.1 Introduction

Traditionnellement, la localisation des séismes se fait grâce à l'identification des arrivées des différentes phases et par association des événements [Lee and Stewart, 1981]. L'extraction de ces arrivées réduit considérablement le volume d'informations à traiter durant le processus de localisation [Withers et al., 1998], mais nécessite le développement d'une logique complexe pour associer chacune des arrivées à un événement. Après un grand séisme ou lors d'une crise sismique ouvolcanique, les systèmes d'acquisition et de détection sont submergés par un grand nombre d'événements se produisant quasi-simultanément dans différentes parties de la zone étudiée. Or dans ces conditions, les techniques basées sur le pointé des phases et l'association des événements ne donnent pas de bons résultats [Johnson et al., 1994] : des événements sont mal associés et mal localisés, voire manqués.

Ces dernières années ont vu l'émergence de nombreuses études utilisant les formes d'ondes complètes pour localiser les séismes à l'échelle du globe, mais aussi à des échelles plus locales. Ces techniques utilisent l'information cohérente obtenue par rétro-propagation et migration en temps inverse. Shearer [1994] a initié le travail à l'échelle globale en prenant comme référence une image obtenue par sommation des signaux longue-périodes puis en recherchant dans une grille d'événements potentiels celui qui correspondait le mieux. Young et al. [1996] a réutilisé l'idée de la grille et a proposé un système de détection basé sur la corrélation des formes d'ondes. Plus récemment, une méthode de détection fondée sur la déconvolution des formes d'ondes a été développée par Ekström [2006] et a permis la détection d'événement tels que les séismes glaciaires [Ekström et al., 2003]. A une échelle plus locale, Withers et al. [1999] a été le premier à reprendre le travail de Young et al. [1996] pour développer un système de détection basé sur la corrélation des formes d'ondes locales. Toujours avec le même objectif, Kao and Shan [2004] ont proposé le *Source-Scanning Algorithm* (SSA), récemment amélioré par Liao et al. [2012], et Baker et al. [2005] a développé une méthode de localisation utilisant

la migration de Kirchhoff. La migration en temps inverse est une autre approche de rétro-propagation développée par McMechan et al. [1985], et rejointe récemment par Rietbrock and Scherbaum [1994], Gajewski and Tessmer [2005] et Larmat et al. [2006].

L'intérêt des méthodes de rétro-propagation ou de migration/sommation réside dans le fait que l'algorithme ne travaille pas plus qu'il y ait beaucoup ou peu de séismes enregistrés. Cette simplicité les rend robustes et stables, même en cas de crise sismique, ce qui est particulièrement appréciable pour les observatoires.

L'algorithme présenté ici, Waveloc, est une variante des méthodes de migration/sommation comme le SSA [Kao and Shan, 2004, Liao et al., 2012] ou de sommation des enveloppes [Gharti et al., 2010]. Il s'organise en trois étapes principales (Fig. I.1.1) :

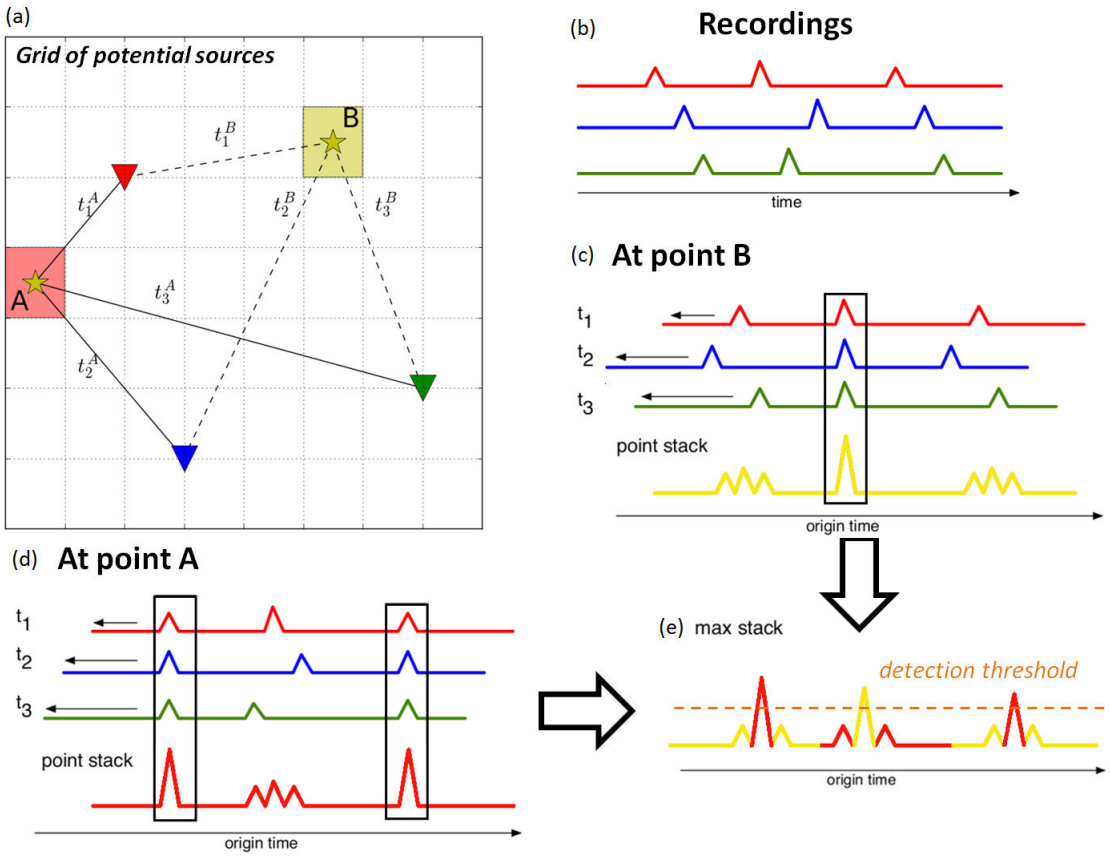


Fig. I.1.1: Illustration schématique de Waveloc. (a) Espace d'étude considéré. Trois stations (triangles inversés) enregistrent les données représentées en (b). Pour chaque point de l'espace 3D, les données sont migrées puis sommées : deux exemples sont donnés au point A (d) et au point B (c). Enfin, pour chaque échantillon de temps, le maximum atteint parmi tous les *stacks* est stocké dans une nouvelle trace unique (e), ainsi que les coordonnées des points correspondants. La détection et localisation se fait donc simultanément par simple lecture des traces.

- **Première étape :** les formes d'ondes brutes sont traitées de manière à mettre en évidence le début des signaux sismiques. On utilise pour cela une fonction caractéristique (le kurtosis).

- **Deuxième étape :** les fonctions caractéristiques sont migrées et sommées (*stackées*) selon un modèle de vitesse des ondes P connu *a priori*.

- **Troisième étape :** les maxima locaux des traces sommées en temps permettent la détection et la localisation simultanée des événements sismiques.

Les bases de Waveloc ont été développées et mises en œuvre par Maggi and Michelini [2009a,b, 2010], d'abord dans un contexte 2D avec une première application à la séquence de répliques du séisme de l'Aquila (M_w=6.3) du 6 avril 2009.

Depuis, l'algorithme a été largement amélioré et a notamment été rendu fonctionnel dans un contexte 3D. Dans cette partie, nous nous attacherons à présenter le principe de base et l'ensemble des fonctionnalités de Waveloc. La figure I.1.2 présente l'ensemble des contributions de ce travail de thèse à l'algorithme déjà existant de Waveloc.

Les exemples d'application (sismicité volcanique) qui permettent de valider la robustesse de Waveloc seront exposés dans les deux autres parties de ce travail (§II, §III). Les données réelles utilisées pour illustrer certaines parties de ce chapitre proviennent du jeu de données du Piton de la Fournaise.

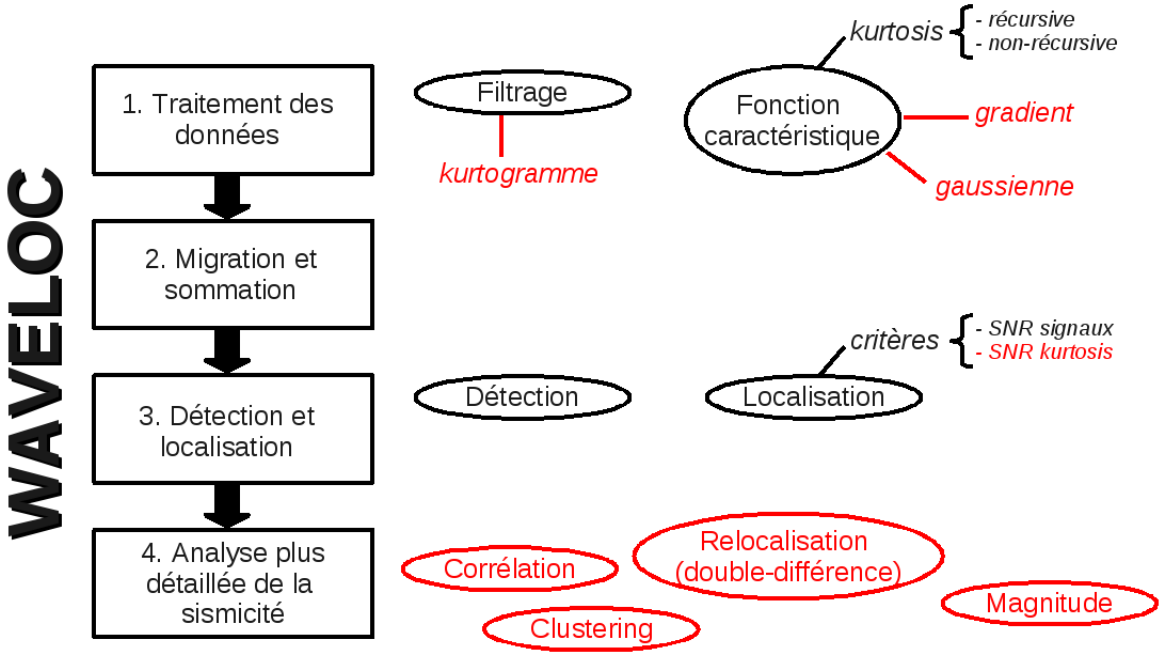


FIG. I.1.2: Répartition des contributions dans la construction de Waveloc. En rouge, les contributions apportées au cours de ce travail de thèse.

I.1.2 Mise en évidence de l'information sur les premières arrivées grâce au kurtosis

Le but de cette première étape est de simplifier l'information contenue dans les formes d'ondes en ne gardant que celle concernant les premières arrivées, d'où la nécessité d'utiliser une fonction caractéristique appropriée. Celle-ci doit être suffisamment pointue au niveau du temps d'arrivée car c'est elle qui sera utilisée par la suite dans la phase de migration.

Dans les processus de pointés automatiques, la fonction caractéristique la plus courante repose sur les variations du rapport STA/LTA (*short-term and long-term average*, [Allen, 1982]) : elle met en évidence les changements associés à l'arrivée d'un train d'ondes en prenant le rapport des moyennes d'un sismogramme mesurées dans deux fenêtres glissantes de tailles différentes.

De toutes les autres fonctions caractéristiques développées depuis, celles qui s'appuient sur des moments statistiques d'ordre supérieurs allient à la fois simplicité de calcul et forte amplification lors des changements de phases. C'est le cas notamment du kurtosis et de son gradient [Saragiotis et al., 2002, Gentili and Michelini, 2006, Küperkoch et al., 2010].

Le kurtosis est le quatrième moment statistique d'une distribution. Pour rappel, le premier moment est la moyenne ; le deuxième, la variance ; et le troisième, l'asymétrie. C'est une valeur adimensionnelle qui mesure la pointicité (valeurs positives) ou l'aplatissement (valeurs négatives) d'une distribution par rapport à une distribution normale. La définition mathématique du kurtosis K est donnée par :

$$K(x_1 \dots x_n) = \left\{ \frac{1}{n} \sum_{j=1}^{n} \left[\frac{x_j - \overline{x}}{\sigma} \right]^4 \right\}, \tag{I.1.1}$$

où $(x_1 \dots x_n)$ est une distribution de moyenne $\overline{x}$ et de variance σ^2. Le kurtosis d'une distribution gaussienne calculée avec l'équation I.1.1 vaut $+3$. On soustrait cette valeur à K de manière à ce qu'elle vaille 0.

Sur la figure I.1.3, on voit que les distributions d'amplitude d'une fenêtre prise dans le bruit (d) et dans le signal lui-même (f) sont très proches d'une distribution gaussienne. En revanche, la distribution d'amplitude associée à une fenêtre prise au niveau de la transition bruit-signal s'avère beaucoup plus pointue qu'une gaussienne : de ce fait, la valeur de kurtosis calculée sera positive et élevée. En se fondant sur ces observations, on transforme l'ensemble des formes d'ondes disponibles en traces de kurtosis $K(t)$ en calculant simplement le kurtosis dans des fenêtres glissantes et en associant sa valeur au dernier échantillon en temps de la fenêtre.

Le kurtosis mesurant une propriété statistique de la distribution d'amplitude du signal, il sera fortement affecté par tout traitement des données susceptible de modifier cette distribution. C'est le cas en particulier lors du filtrage des données. Comme le but recherché ici est d'obtenir des traces de kurtosis avec les valeurs les plus fortes possibles, il serait judicieux de filtrer les formes d'ondes dans une bande de fréquence qui maximise le kurtosis. On utilise pour cela la méthode du kurtogramme [Antoni, 2007], qui consiste à calculer les valeurs de kurtosis dans diverses bandes de fréquence et à retenir celle qui permet d'atteindre le maximum.

Un signal non-stationnaire $s(t)$ peut se décomposer de la manière suivante :

$$s(t) = \int_{-\infty}^{+\infty} e^{j2\pi ft} H(t,f) dX(f), \tag{I.1.2}$$

où $H(t,f)$ est l'enveloppe complexe de $s(t)$ et $dX(f)$ est l'incrément spectral. Le kurtosis spectral peut alors se définir par :

$$SK_s(f) = \frac{|H(t,f)|^4}{\langle |H(t,f)|^2 \rangle^2} \tag{I.1.3}$$

et donne la représentation des caractéristiques transitoires d'un signal en fonction de la fréquence.

Le kurtogramme est la représentation graphique des valeurs de SK en fonction de la fréquence f et de la largeur de bande Δf (voir Fig. I.1.4). Le couple de valeurs (f,Δf) qui correspond au maximum du kurtosis donne des indications sur les meilleurs paramètres de filtrage à utiliser lors du pré-traitement des données et avant le calcul des traces de kurtosis.

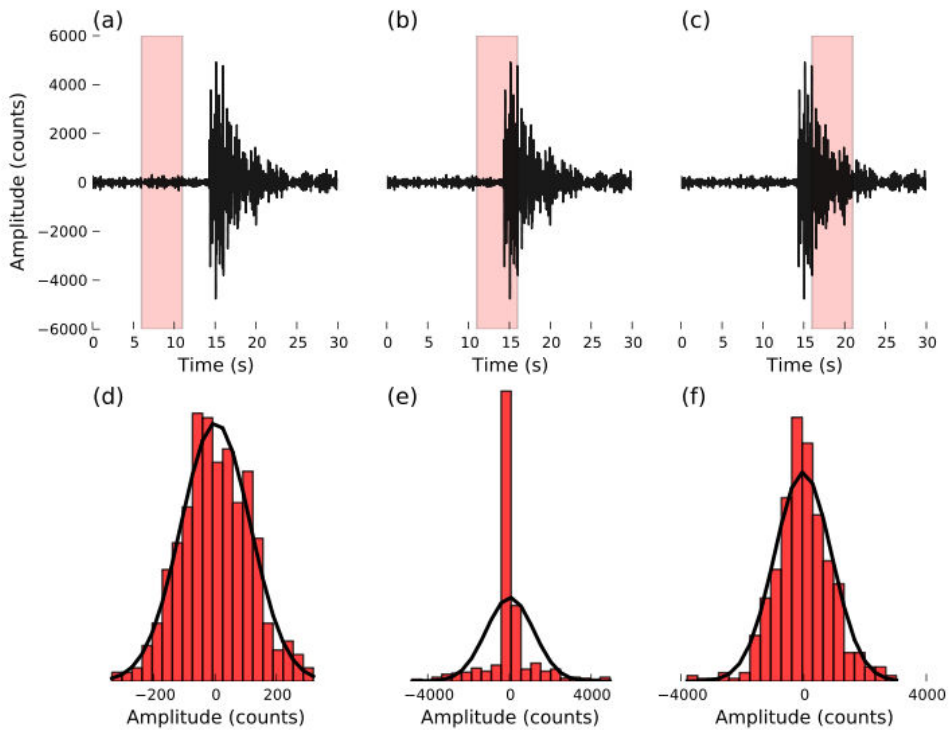


Fig. I.1.3: Distributions d'amplitude d'un signal sismique dans différentes fenêtres. (a-c) Les fenêtres de 5s considérées sont colorées en rouge : (a) bruit, (b) transition bruit-signal, (c) signal. (d-f) Distributions d'amplitude correspondants à chacune des zones colorées et superposées à leurs gaussiennes (courbes noires).

Un autre paramètre important dans le calcul du kurtosis est la taille de la fenêtre glissante choisie. La figure I.1.5 montre l'exemple d'un kurtosis calculé pour 3 fenêtres de tailles différentes. On remarque que la valeur maximale du kurtosis et son étendue augmentent lorsque la

taille de la fenêtre augmente. De plus, le maximum du kurtosis est systématiquement retardé par rapport au début du signal : cela pose problème puisqu'on souhaite utiliser cette information par la suite. La plus grande précision possible est donc requise. Pour réduire l'erreur introduite par la mesure du kurtosis, on ne considère plus seulement le kurtosis, mais la partie positive de sa dérivée première $\dot{K}_+$:

$$\dot{K}_+ = \begin{cases} \dot{K} & \text{si } \dot{K} \geq 0 \ , \\ 0 & \text{sinon.} \end{cases} \tag{I.1.4}$$

Les nouvelles traces obtenues sont présentées sur la figure I.1.5d. Elles sont beaucoup plus impulsives, donc mieux adaptées pour l'étape de migration qui suivra. On note qu'un léger décalage par rapport au début du signal subsiste.

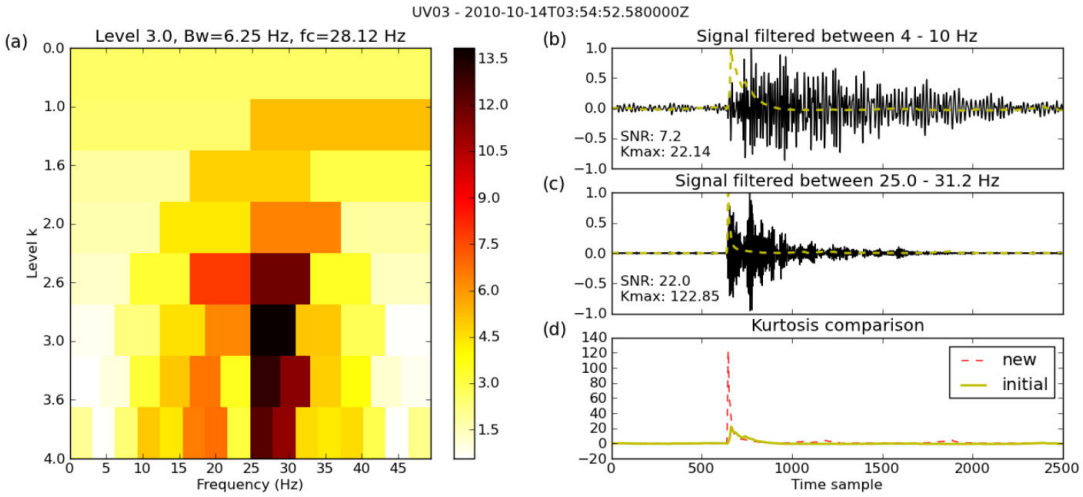


FIG. I.1.4: (a) Kurtogramme calculé à partir du signal (b). (c) Signal filtré dans la bande de fréquence dans laquelle le kurtosis est maximal (niveau 3 : fréquence centrale de 28.12 Hz et largeur de bande de 6.25 Hz). Les traces de kurtosis (lignes pointillées) sont superposées aux signaux. (d) Les deux traces de kurtosis sont représentées sur la même figure : on observe un net écart d'amplitude selon la bande de filtrage.

Le calcul des traces de kurtosis peut s'avérer très coûteux en temps. Un moyen de rendre les calculs plus rapides est d'utiliser une méthode récursive, comme celle proposée par [Chassande-Mottin, 2003]. Cependant, cette méthode peut se révéler instable lorsque que les signaux sont fortement non-stationnaires, ce qui est le cas des événements sismiques très impulsifs. On utilisera donc ici une méthode de calcul récursive moins formelle, mais restant stable.

Soit x un signal d'écart-type σ_x. A l'échantillon de temps i, la moyenne du signal et son écart-type sont définis par les relations récursives suivantes :

$$\overline{x}_i = C\overline{x}_{i-1} + (1-C)x_i, \tag{I.1.5}$$

$$\sigma_i = C\sigma_{i-1} + (1-C)(x_i - \overline{x}_i)^2, \tag{I.1.6}$$

Le kurtosis récursif se calcule alors ainsi :

$$K_i = \begin{cases} CK_{i-1} + (1-C)\frac{(x_i-\overline{x}_i)^4}{\sigma_i^2} & \text{si } \sigma_i > \sigma_x, \\ CK_{i-1} + (1-C)\frac{(x_i-\overline{x}_i)^4}{\sigma_x^2} & \text{sinon.} \end{cases} \tag{I.1.7}$$

avec la constante C telle que $C = 1 - \frac{dt}{w}$, où dt est le taux d'échantillonnage en temps et w est une grandeur liée à la taille de la fenêtre choisie pour le calcul du kurtosis. L'utilisation de σ_x dans l'équation I.1.7 stabilise le calcul quand les écart-types σ_i deviennent petits.

La comparaison des deux méthodes de calcul des kurtosis (récursive et non-récursive) a montré que, pour une taille de fenêtre équivalente, la précision demeurait identique. En revanche, les amplitudes des kurtosis récursives sont plus fortes et décroissent plus lentement : pour que les deux types de kurtosis aient des formes similaires, on a remarqué qu'il fallait que w soit environ 3 fois plus petit que la taille de la fenêtre utilisée dans le calcul "standard".

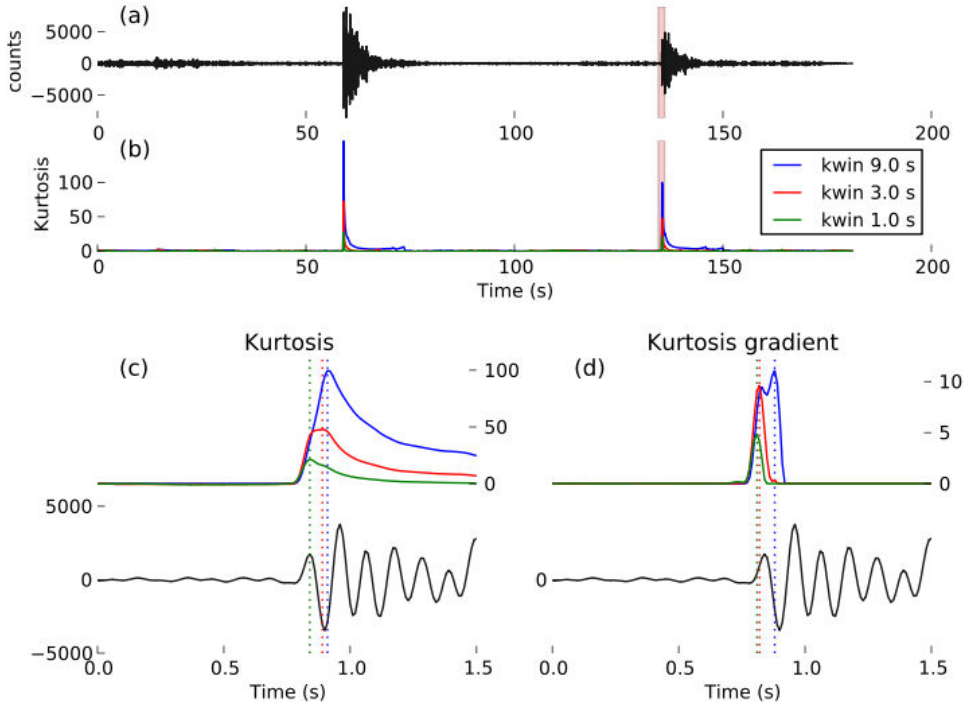


FIG. I.1.5: (a) Signal sismique brut. (b) Traces de kurtosis $K(t)$ correspondantes, calculées pour 3 tailles de fenêtres glissantes différentes (bleu : 9 s, rouge : 3 s, vert : 1 s). (c) Zoom sur la portion de signal colorée en rouge. On remarque que le maximum des kurtosis est systématiquement décalé par rapport au début du signal. (d) Même chose, avec cette fois le gradient des kurtosis $\dot{K}_+(t)$. Lorsque la fenêtre de calcul choisie n'est pas trop grande, le maximum des kurtosis se rapproche du début du signal.

I.1.3 Migration et sommation

La deuxième étape de Waveloc consiste à migrer les traces de gradient du kurtosis $\dot{K}_+$ évoquées dans la section précédente (voir FIG. I.1.1).

On rappelle que le temps d'arrivée T_k^i auquel un événement i est enregistré à la station k

est donné par la relation suivante :

$$T_k^i = \tau_i + t_k^i \tag{I.1.8}$$

où t_k^i est le temps que met l'onde émise pour parcourir la distance qui la sépare de la station k et τ_i est le temps origine de l'événement i. Les temps de trajet t_k^i sont fonction de la vitesse du milieu traversé.

L'étape de **migration** consiste à s'affranchir de l'effet de propagation entre un événement donné et les stations qui l'ont enregistré afin de faire ressortir la cohérence des informations contenues dans les premières arrivées des signaux. Cette étape nécessite l'utilisation d'un modèle de vitesse des ondes P pré-établi pour permettre le calcul des temps de trajet théoriques t_k^i.

On voit d'après l'équation I.1.8 qu'à l'issue de l'étape de migration, on aura directement accès à l'information sur les temps origine.

En pratique, on effectue la migration par simple *grid search*. L'espace d'étude est discrétisé en une grille 3D dans laquelle chaque point constitue une source sismique potentielle. Les temps de trajet théoriques de chacun de ces points à chacune des stations du réseau sont calculés grâce au modèle de vitesse (1D ou 3D) suivant la méthode de Podvin and Lecomte [1991], en utilisant la manière dont elle est implémentée dans NonLinLoc [Lomax, 2011].

En chaque point de la grille, les traces $\dot{K}_+$ sont rétro-propagées des temps de trajet correspondants (une par station), puis sommées (ou *stackées*) de manière à limiter l'information et ne garder plus qu'une seule trace en un point donné :

$$\tilde{\dot{K}}_+^{ij}(t) = \dot{K}_+^j(t + \tau^{ij}), \tag{I.1.9}$$

$$S^i(t) = \sum_j \tilde{\dot{K}}_+^{ij}(t), \tag{I.1.10}$$

où $\dot{K}_+^j(t)$ est la trace de la dérivée première positive du kurtosis à la station j, τ^{ij} est le temps de trajet du point i à la station j, $\dot{K}_+^{ij}(t)$ est la trace migrée et où $S^i(t)$ est la somme des traces de toutes les stations au point i.

On obtient finalement une grille 3D contenant en chacun de ses points i les traces sommées $S(\mathbf{x}_i; t)$, où $\mathbf{x}_i$ correspond aux coordonnées géographiques du point i et t est le temps absolu. Un maximum local important de $S(\mathbf{x}; t)$ indiquera donc qu'un événement sismique s'est produit au temps origine t et à la position $\mathbf{x}$.

Le stockage d'une telle masse d'information (une trace par point) peut vite atteindre des proportions gigantesques en fonction du nombre de points de l'espace d'étude considéré et du pas d'échantillonnage des données. Or on sait que l'information que l'on souhaite conserver ne concerne que l'information cohérente : une dernière simplification est alors nécessaire et consiste à ne garder qu'une seule trace contenant l'information pour l'ensemble du volume. Cette trace $S_{max}(t)$ s'obtient en ne gardant que les valeurs maximales de $S(\mathbf{x}; t)$ pour chaque échantillon de temps t. Parallèlement, les coordonnées correspondant à chacun des points $\mathbf{x}$ où les maximum sont atteints sont aussi stockées dans 3 traces distinctes. D'un problème où l'on avait N traces (où N est le nombre de points de la grille 3D), on passe désormais à un problème où l'on n'a plus que 4 traces à traiter, ce qui facilite considérablement le processus.

La figure I.1.1 illustre schématiquement ce qui vient d'être expliqué : prenons l'exemple d'un réseau composé de 3 stations (triangles inversés) ayant enregistré 3 événements sismiques.

Chaque point de l'espace 3D considéré est perçu comme un hypocentre potentiel : les formes d'ondes enregistrées à chaque station sont donc corrigées (= migrées) du temps de trajet station-source. Deux exemples sont donnés en deux points différents (points A et B). Une fois l'étape de migration effectuée, l'ensemble des traces est sommé : si un événement s'est effectivement produit au point considéré, l'information se sommera de manière cohérente et sera amplifiée ; si aucun événement ne s'est produit, les traces ne se sommeront pas de manière constructive. Ainsi, on voit qu'au point A, deux des événements enregistrés semblent se sommer de manière cohérente (le premier et le dernier), alors qu'au point B, c'est l'événement intermédiaire qui ressort. Ceci signifie en d'autres termes que les événements 1 et 3 se sont produits en A à deux dates différentes, et que l'événement 2 s'est produit en B. Si on prenait n'importe quel autre point de l'espace, la sommation des traces ne serait constructive pour aucun des 3 événements. Enfin, la dernière étape consiste à ne conserver que les valeurs maximales de l'ensemble des traces sommées (e).

On a choisi ici d'illustrer ce propos avec un exemple où 2 événements se produisent au même endroit, mais en deux dates distinctes, afin de montrer qu'une seule migration est nécessaire en chaque point pour trouver l'ensemble des événements s'étant produit en ce point.

La figure I.1.6 présente quant à elle un exemple synthétique montrant comment la cohérence de l'information est exploitée. On a utilisé pour ce test la géométrie du réseau de stations du Piton de la Fournaise, ainsi que le modèle de vitesse associé (plus de détails seront donnés dans la partie étant consacrée spécifiquement à ce volcan - §II). Les formes d'ondes synthétiques ont été construites en utilisant de simples pulses triangulaires pour simuler les traces $\dot{K}_+(t)$. La largeur des pulses a été choisie de façon à s'accorder avec celles des vraies formes d'ondes (0.1 s).

Les trois coupes (a),(d) et (f) correspondent aux coupes dans l'espace contenant l'ensemble des traces sommées $S(\mathbf{x};t)$ prises selon les coordonnées de l'hypocentre synthétique. Elles montrent que les traces migrées forment des anneaux dont l'intersection correspond à l'hypocentre. En fonction de la géométrie du réseau, ces anneaux se recoupent plus ou moins : la zone correspondant à l'hypocentre sera donc plus ou moins étendue. Sur la sous-figure (f), par exemple, on observe des extensions dans les directions y et z. Avec des données réelles, on s'attend évidemment à ce que la focalisation ne soit pas parfaite, c'est-à-dire à avoir des taches plutôt que des points. Ceci fournira des indications sur la précision que l'on peut espérer obtenir en utilisant cette méthode.

Les 4 traces qui sont conservées à l'issue de l'étape de migration-sommation sont représentées dans les sous-figures (b),(c),(e),(g). Elles montrent que le pic d'amplitude maximale de la trace $S_{max}(t)$ (b) se produit au bon temps origine et que les intersections avec les traces $\mathbf{x}(t)$ associées donnent également les bonnes coordonnées. C'est d'ailleurs cette correspondance entre les 4 traces qui sera utilisée dans la prochaine étape de détection et localisation.

Notons que les valeurs de $\mathbf{x}(t)$ prises en dehors du temps où un événement se produit n'ont pas de signification particulière puisque que les coordonnées du maximum local changent à chaque échantillon de temps et peuvent correspondre à n'importe quel point de la grille. La plus forte dispersion observée en z est liée à la fois à l'incertitude sur la mesure de la profondeur due à la géométrie du réseau et au *trade-off* qui existe entre la profondeur et le temps origine.

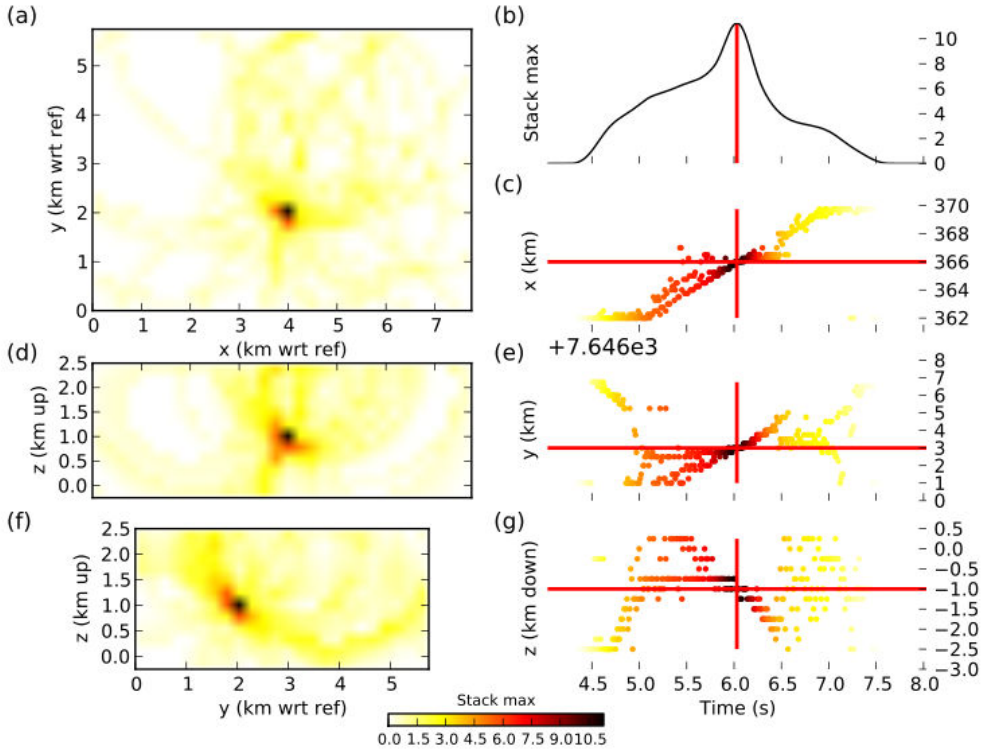


FIG. I.1.6: Test synthétique réalisé en utilisant la géométrie du réseau de stations du Piton de la Fournaise. Les coupes (a) (plan x-y),(d) (plan x-z),(f) (plan y-z) sont faites dans l'espace contenant l'ensemble des traces sommées $S(\mathbf{x};t)$ au temps origine de l'événement synthétique et montrent la cohérence de l'information au niveau de l'hypocentre. (b) Trace $S_{max}(t)$ contenant les maximum de l'ensemble des traces sommées prise dans une fenêtre de 4 s centrée sur le temps origine du synthétique (trait vertical). Les sous-figures (c),(e),(g) représentent les 3 composantes de $\mathbf{x}(t)$ dans la même fenêtre de temps que (b) (un point par échantillon de temps). Les coordonnées de l'hypocentre synthétique sont symbolisées par les lignes horizontales et correspondent à la valeur prise par les coordonnées respectives au temps où le *stack* atteint son maximum (traits verticaux). Le code couleur entre les sous-figures (a),(d),(f) et (c),(e),(g) est le même. Notez que la grandeur z indique l'altitude en (d) et (f) pour faciliter la lecture, mais indique la profondeur en (g).

I.1.4 Détection et localisation

On désigne par **détection** la détermination de la présence d'un événement sismique et par **localisation** la détermination de ses paramètres hypocentraux (coordonnées spatiales et temps origine) et des incertitudes associées.

Comme le montre la figure I.1.6, le maximum local de $S_{max}(t)$ correspond au temps origine du signal. De ce fait, l'étape de détection ne consiste qu'à choisir un seuil au-delà duquel on considèrera qu'un événement s'est effectivement produit. La valeur du seuil dépend de l'impulsivité des événements, du nombre de stations disponibles et doit être ajusté en comparant visuellement $S_{max}(t)$ aux formes d'ondes d'une partie représentative des données.

Pour l'étape de localisation, on utilise l'information stockée dans $S_{max}(t)$ et dans $\mathbf{x}(t)$. Lorsque un événement est détecté sur $S_{max}(t)$, le temps auquel le maximum est atteint correspond au temps origine t_0 de l'événement. On définit également t_1 et t_2 pris à 95% de la valeur du maximum local et qui constituent une mesure de l'incertitude. La moyenne et l'écart-type des coordonnées $\mathbf{x}(t)$ dans l'intervalle $[t_1,t_2]$ donnent respectivement les coordonnées de l'hypocentre et leurs incertitudes.

L'intérêt principal de développer une méthode comme Waveloc est de s'affranchir de l'étape d'association des phases commune aux nombreuses méthodes de localisation de pointé automatique et qui peut se révéler critique lorsqu'un grand nombre d'événements se produit dans un lapse de temps court (essaims sismiques par exemple). On a donc effectué quelques tests synthétiques pour évaluer la capacité de notre algorithme à séparer des événements proches dans le temps (voir FIG. I.1.7). La résolution en temps dépend d'un certain nombre de paramètres, comme la manière dont le kurtosis est calculé (méthode retenue, taille de la fenêtre). Elle dépend aussi du seuil de détection choisi (qui, rappelons-le, dépend lui-même de l'impulsivité des signaux et du nombre de stations présentes dans le réseau) et de la profondeur du creux entre les 2 pics successifs de la trace $S_{max}(t)$: en effet, dans le cas de la figure I.1.7, si on avait choisi un seuil très peu élevé (autour de la valeur 4 par exemple), l'algorithme de détection n'aurait détecté qu'un seul événement au lieu des 2. La résolution en temps que l'on peut espérer atteindre est donc très dépendante d'un certain nombre de paramètres environnants. Dans l'exemple d'illustration présenté ici, on s'est efforcé de choisir des paramètres relativement "standards".

Finalement, le test synthétique montre qu'il est possible de distinguer les deux événements proches dans le temps, mais que leur détection sur une seule trace peut s'avérer moins facile en fonction des paramètres choisis. Il faudrait envisager le développement d'un algorithme plus complexe prenant en compte l'ensemble des traces sommées pour espérer améliorer ce point et détecter à coup sûr les deux événements.

D'autres tests synthétiques ont été effectués et ont permis de montrer que Waveloc est capable de localiser des événements même si le rapport signal-sur-bruit (SNR) est peu élevé. La figure I.1.8 est à comparer avec la figure I.1.6 pour laquelle aucun bruit n'a été rajouté aux données.

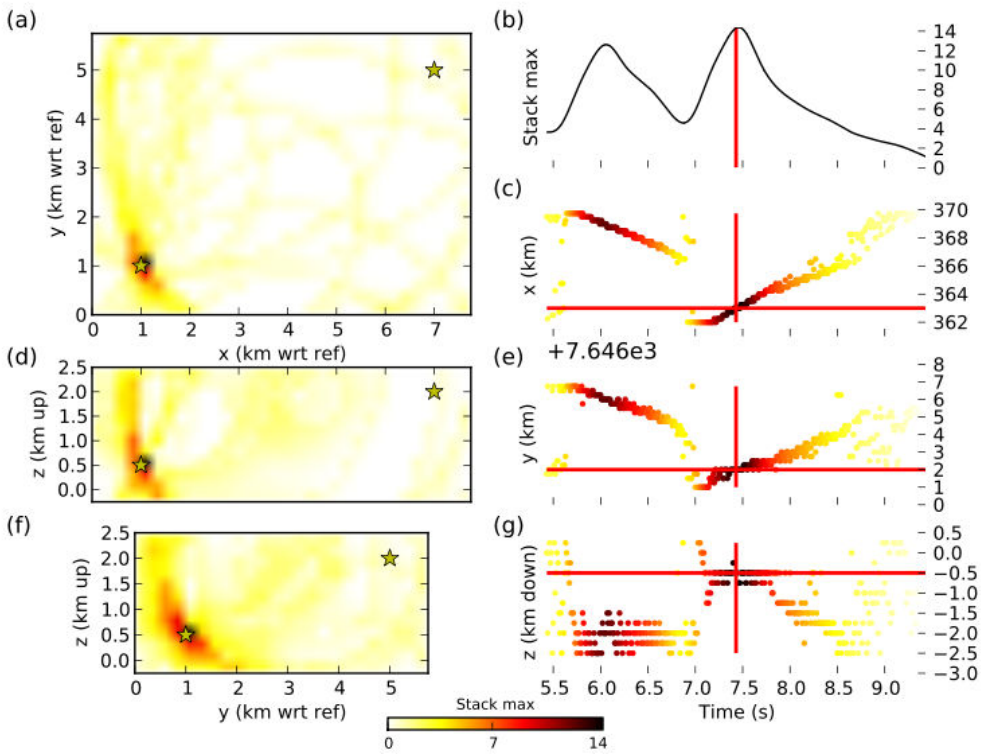


FIG. I.1.7: Test synthétique visant à montrer la capacité de Waveloc à séparer des événements voisins dans le temps (l'un se produisant à 6.0 s, l'autre à 7.4 s), mais distincts dans l'espace (étoiles). La figure a été réalisée pour l'événement se produisant à 7.4 s. Les deux événements sont clairement identifiables : en (b), les deux pics sont bien distincts. En (c),(e) et (g), les coordonnées de chacun d'eux ne se mélangent pas avec celles de l'autre. Pour le détail concernant la description des sous-figures, se référer à la figure I.1.6.

On avait vu sur la figure I.1.5d que les traces de gradient du kurtosis $\dot{K}_+(t)$ étaient asymétriques, avec une durée de phase décroissante plus longue que celle de la montée. Ceci peut créer des problèmes lors de l'étape de sommation et engendrer une forte dispersion des coordonnées dans l'intervalle de temps autour du temps origine. On a alors essayé de remplacer les traces $\dot{K}_+(t)$ par une série de gaussiennes de demi-largeurs comparables à celle des kurtosis et centrées sur chacun des maxima locaux. Utiliser ces gaussiennes revient en quelque sorte à lisser les traces de kurtosis. La plupart du temps, elles facilitent la sommation de l'information cohérente car elles sont plus larges que les kurtosis initiales. On pourrait choisir de calculer les kurtosis sur une fenêtre plus longue, mais, comme on l'a déjà évoqué, cela accentue à la fois le biais sur le début du signal et l'asymétrie, d'où une baisse importante de la précision sur la détermination des paramètres hypocentraux. L'utilisation des gaussiennes est donc un excellent moyen d'augmenter le nombre de détections sans perdre pour autant la précision sur les localisations (voire en l'améliorant).

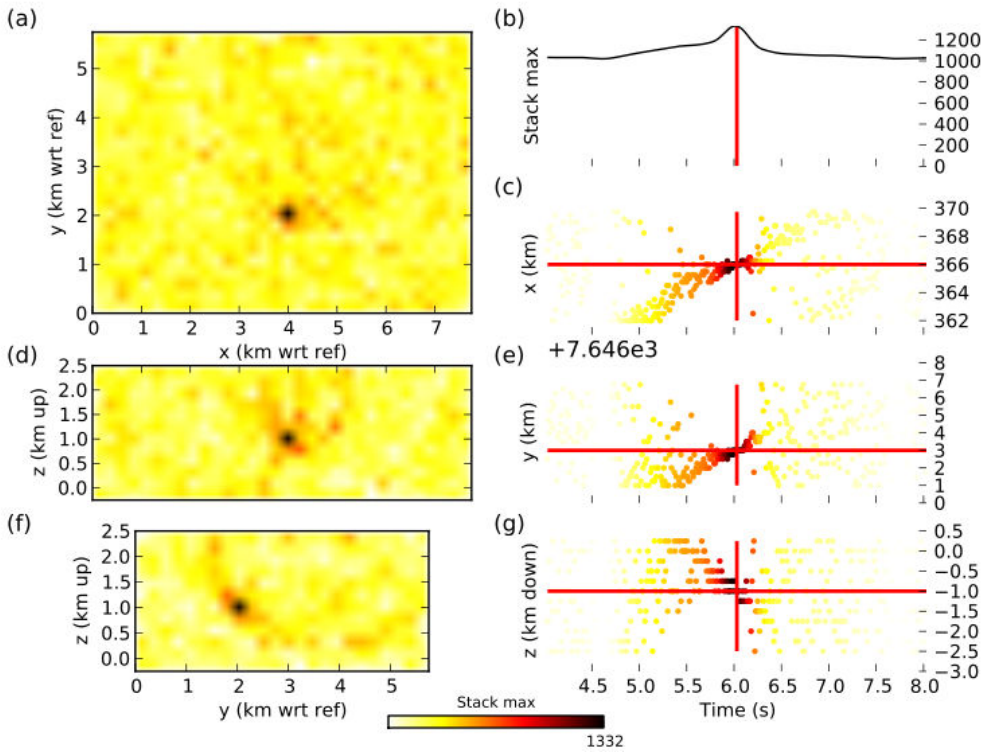


FIG. I.1.8: Test synthétique visant à montrer la capacité de Waveloc à localiser un événement avec un SNR de 1.5. Pour le détail concernant la description des sous-figures, se référer à la figure I.1.6.

I.1.5 Calcul des magnitudes locales

La dernière étape d'une étude de sismicité un peu détaillée consiste à calculer les magnitudes des événements. On a ajouté dans Waveloc un module de calcul des magnitudes locales d'après la définition faite par Bakun and Joyner [1984] :

$$M_l = \log\ A + \log \frac{\Delta}{100} + 0.00301(\Delta - 100) + 3 \tag{I.1.11}$$

où A est l'amplitude du pic qui serait mesurée sur un sismomètre Wood-Anderson et Δ est la distance hypocentrale.

Connaître les magnitudes permet de calculer la b-value qui est un paramètre statistique important correspondant à la pente de la distribution fréquence-magnitude donnée par la loi de Gutenberg-Richter :

$$\log_{10} N = a - bM \tag{I.1.12}$$

où N est le nombre cumulé d'événements de magnitude supérieure ou égale à M.

La b-value constitue en quelque sorte une mesure du nombre de petits séismes relativement au nombre de grands séismes. On suppose communément que la b-value a une valeur

constante dans une zone d'étude donnée et qu'elle est proche de 1, mais de nombreuses études ont montré qu'il existe en réalité une variation spatio-temporelle de ce paramètre [Kulhanek, 2005, Wiemer, 2001].

Il existe plusieurs méthodes (plus ou moins complexes) de calcul de la b-value. Le point délicat est surtout la détermination de la magnitude de complétude, c'est-à-dire la magnitude au-delà de laquelle on considère que le réseau de stations a été capable d'enregistrer tous les événements [Rierola, 2000, Woessner and Wiemer, 2005]. Dans Waveloc, on a simplement défini cette magnitude comme étant celle correspondant à la fin du plateau observée sur la distribution de Gutenberg-Richter.

I.1.6 Corrélation et relocalisation par double-différence

A l'issue des étapes décrites dans les sections précédentes, Waveloc est en mesure de fournir un premier catalogue de sismicité. Cependant, il est toujours possible d'améliorer la précision en réalisant une analyse un peu plus poussée de la sismicité, passant entre autres par relocalisation des événements.

Deux nouvelles fonctionnalités ont été ajoutées à Waveloc dans ce but :
- la recherche des événements de formes d'ondes similaires par inter-corrélation [Poupinet et al., 1984].
- la relocalisation par double-différence, dans une version "basique" par rapport à celle développée par Waldhauser and Ellsworth [2000].

Corrélation

Rappelons que deux événements sont supposés similaires lorsqu'ils se produisent au même endroit avec des mécanismes au foyer similaires. Ceci se traduit par des formes d'ondes similaires. La recherche de ces multiplets se fait par simple inter-corrélation des formes d'ondes. On rappelle que l'expression mathématique de l'inter-corrélation entre deux traces $x_1(t)$ et $x_2(t)$ est la suivante :

$$C(\Delta t) = \int_{-\infty}^{+\infty} x_1(t) x_2(t - \Delta t) dt \tag{I.1.13}$$

$C(\Delta t)$ mesure la similitude entre la trace $x_1(t)$ et la trace $x_2(t)$ décalée de Δt.

En pratique, on procède comme suit :

- pour chaque événement du catalogue de sismicité de Waveloc, on mesure les coefficients de corrélation et les décalage en temps Δt avec tous les autres événements du catalogue en domaine temporel. La taille de la fenêtre dans laquelle s'effectue l'inter-corrélation est soigneusement choisie (temps pris avant et après le temps origine de l'événement) : de manière générale, on remarque que prendre une trop grande fenêtre n'améliore pas la corrélation et rend les calculs plus longs.

- lorsque le coefficient de corrélation excède une valeur seuil donnée (suffisamment élevée), on effectue aussi la corrélation en domaine de Fourier. Ceci permet en effet d'avoir une meilleure précision sur le calcul du décalage en temps Δt qui est alors mesuré à partir de la pente de la phase déroulée du spectre [Poupinet et al., 1984, Schaff et al., 2004].

Clustering

Une fois que tous les coefficients de corrélation et les délais en temps Δt de toutes les paires d'événements possibles ont été calculés, on prépare les données pour l'étape suivante de relocalisation en les regroupant dans des groupes d'événements de formes d'ondes similaires (*clustering*). Pour que deux événements se retrouvent dans le même groupe, deux critères doivent être remplis : un coefficient de corrélation suffisamment grand attestant de leur similarité (valeur seuil à déterminer), et un nombre minimum de stations où cette valeur seuil est effectivement mesurée. Le *clustering* s'effectue ici avec un algorithme de parcours en profondeur (*depth-first search algorithm* (DFS) en anglais), par opposition à un algorithme de parcours en largeur (*breadth-first search algorithm* (BFS) en anglais) [Cormen et al., 2001]. Il consiste à explorer chaque branche jusqu'à ce qu'elle se termine (voir Fig. I.1.9). L'avantage d'un tel algorithme tient surtout dans sa rapidité d'exécution.

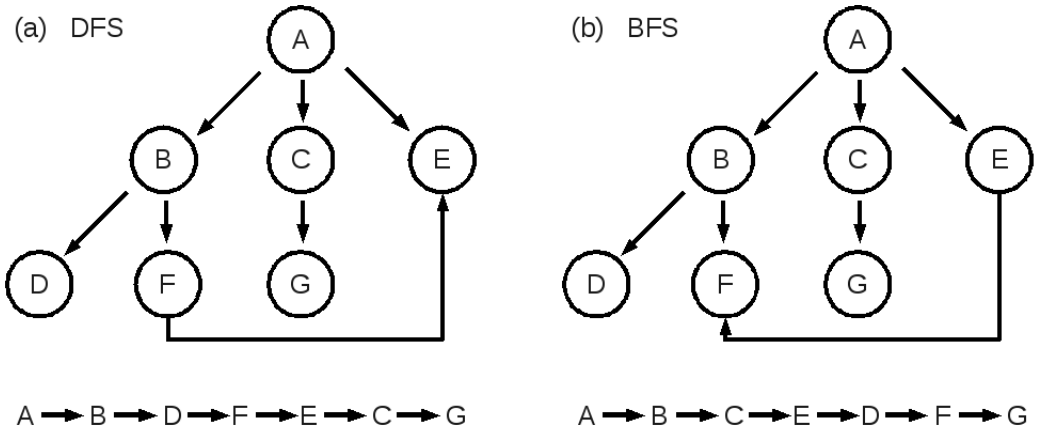


Fig. I.1.9: (a) Algorithme de DFS. (b) Algorithme de BFS. On considère 7 événements : l'événement A est fortement corrélé à B, C et E. B est corrélé avec A, D et F ; C avec A et G ; D avec B ; E avec A et F ; F avec B et E ; et G avec C. En partant du sommet A, l'algorithme DFS va explorer d'abord trouver B, puis D. En D, la branche se termine, il revient au niveau précédent et examine le deuxième événement corrélé à B (F)... et ainsi de suite. L'algorithme BFS, quant à lui, explore tous les "niveaux" d'un coup : il trouve d'abord tous les événements corrélés à A, puis à B ...

Relocalisation par double-différence

Les groupes d'événements similaires étant maintenant constitués, il est possible de passer à l'étape de relocalisation par double-différence. Cette méthode a été développée par Waldhauser and Ellsworth [2000] et consiste à relocaliser les événements appartenant à un même groupe relativement entre-eux. Elle tire partie du fait que si la distance qui sépare deux événements sismiques est très petite devant la distance hypocentre-station, alors les parcours des

rais seront similaires sur presque tout le parcours. Par conséquent, la différence de temps de trajet mesurée entre deux événements enregistrés à une même station sera simplement liée à la distance séparant les 2 hypocentres.

La relation entre le temps de trajet et la localisation d'un événement n'est pas linéaire (éq. I.1.8). Le problème de localisation d'un séisme peut être linéarisé en mettant en relation le résidu des temps de trajet r avec les perturbations Δm des 4 paramètres hypocentraux :

$$\frac{\partial t_k^i}{\partial m}\Delta m^i = r_k^i \tag{I.1.14}$$

où $r_k^i = (t^{obs} - t^{th})_k^i$ et $\Delta m^i = (\Delta x^i, \Delta y^i, \Delta z^i, \Delta t^i)$, avec i l'indice de l'événement et k l'indice de la station.

L'équation de double-différence s'obtient simplement en différenciant l'équation précédente. Pour un milieu de vitesse non homogène, on aura donc :

$$\frac{\partial t_k^i}{\partial m}\Delta m^i - \frac{\partial t_k^j}{\partial m}\Delta m^j = dr_k^{ij} \tag{I.1.15}$$

où $dr_k^{ij} = (t_k^i - t_k^j)^{obs} - (t_k^i - t_k^j)^{th}$ est la différence de temps de trajet entre les temps de trajet relatifs observés et théoriques pour toutes les paires d'événements (i, j).

Finalement, il reste à résoudre un système linéaire de la forme :

$$\mathbf{WGm} = \mathbf{Wd} \tag{I.1.16}$$

où $\mathbf{G}$ est une matrice de taille (M,$4N$) contenant les dérivées partielles, $\mathbf{m}$ est le vecteur modèle de taille ($4N$,1) contenant les perturbations de paramètres hypocentraux, $\mathbf{d}$ est le vecteur données de taille (M,1) contenant les double-différences et $\mathbf{W}$ est une matrice de taille (M,M) contenant des coefficients de pondération. N désigne le nombre d'événements à relocaliser. M est le nombre de double-différences : il correspond au nombre de paires possibles dans une liste de N événements multiplié par le nombre de stations du réseau S, soit :

$$M = \frac{1}{2}\frac{N!}{(N-2)!}S$$

Ici, les temps de trajet relatifs $(t_k^i - t_k^j)^{obs}$ correspondent aux décalages en temps Δt mesurés pour chaque paire d'événements (i, j) par la corrélation et la matrice de pondération $\mathbf{W}$ contient simplement les coefficients de corrélation. Les temps de trajet théoriques $(t_k^i - t_k^j)^{th}$ sont calculés à partir du modèle de vitesse des ondes P. Le système d'équations est finalement résolu en utilisant la méthode des moindres carrés.

I.1.7 Conclusion partielle

Tous les outils présentés permettent la détection et localisation simultanées des événements sismiques, fournissant ainsi un catalogue de sismicité. L'analyse un peu plus détaillée des données permet finalement l'établissement de cartes de micro-sismicité dont la répartition peut renseigner, voire "imager" les failles et les ruptures en profondeur.

A priori, une telle méthode doit être capable de détecter et localiser n'importe quel signal sismique enregistré sur plusieurs stations à la même date, indépendamment de son type. Le chapitre suivant est une introduction aux méthodes qui peuvent permettre de classer les divers événements sismiques.

CHAPITRE I.2

Classification automatique

I.2.1 Introduction

I.2.1.1 Généralités

Qu'est-ce que la classification et pourquoi ?

La classification consiste à classer un ensemble d'éléments selon des caractéristiques qu'ils partagent entre-eux. Les problèmes de classification sont légion, y compris dans la vie de tous les jours. Par exemple, l'identification des *spams* à partir de quelques mots-clé est un problème de classification. En sciences, les domaines d'application sont variés, allant des logiciels de reconnaissance de la parole ou de l'écriture à l'aide aux diagnostics médicaux.

L'objectif, en développant des techniques automatiques de classification, est de "mimer" la manière dont le cerveau humain traite les informations pour effectuer une classification, et utiliser la puissance de calcul des ordinateurs pour traiter un grand nombre d'informations. En effet, si l'on prend l'exemple de la reconnaissance de l'écriture, notre cerveau sait déchiffrer facilement une lettre donnée. Et pourtant, une même lettre écrite deux fois par la même personne ne sera jamais strictement identique. Le problème de reconnaissance (et, par extension, de classification) est donc beaucoup plus compliqué qu'il n'y paraît : il s'agit de fournir au classifieur suffisamment d'informations discriminantes, mais non identiques, afin de lui permettre d'effectuer une classification correcte. Les sources d'information doivent être les plus diverses possibles pour permettre le traitement d'un grand nombre de situations.

FIG. I.2.1: La même lettre "a" écrite trois fois par la même personne. Pour nous, aucun problème de lecture pour savoir que c'est bien un "a". Mais comment faire en sorte que l'ordinateur reconnaisse dans les trois cas la bonne lettre ? Quelles informations sont nécessaires ?

Comment classer des éléments automatiquement ? : quelques définitions et principes

L'exemple le plus simple d'apprentissage automatique est sans nul doute celui de la régression linéaire : à partir d'un jeu de données constitué de couples de valeurs (x, y), on cherche la loi (le modèle) qui explique au mieux y connaissant x. Une fois celle-ci connue, il est possible de prédire la valeur de y d'une variable x nouvellement introduite dans le jeu de données.

Un problème de classification repose sur le même principe, sauf que les techniques ne sont pas les mêmes puisqu'il s'agit non plus de prédire la valeur de y d'un élément de valeur x, mais la classe à laquelle il appartient. Dans le cas d'un problème de classification binaire, y ne peut donc prendre que deux valeurs : 0 ou 1.

Les méthodes d'apprentissage automatique sont diverses. Parmi les principales, et parmi celles qui nous intéressent, on peut citer :

- l'apprentissage **supervisé** qui suppose que l'utilisateur a déjà défini les classes manuellement sur une certaine proportion du jeu de données qui constitue le *training set* (ou jeu d'apprentissage). Il faut donc entraîner le système (Fig. I.2.2) sur ce *training set* pour qu'il apprenne et détermine la loi qui permet de classer les données au mieux. Le problème peut, une fois cette étape terminée, se résumer à un calcul de probabilités : « si je sais qu'un élément x vaut tant, quelle est sa probabilité d'appartenance à telle ou telle classe ? » ;

- l'apprentissage **non supervisé** (comme le *clustering* par exemple), qui consiste à trouver une structure dans un jeu de données dont on ne connaît rien *a priori*. L'algorithme se fonde uniquement sur les informations contenues dans le jeu de données et doit les agencer de manière cohérente par lui-même.

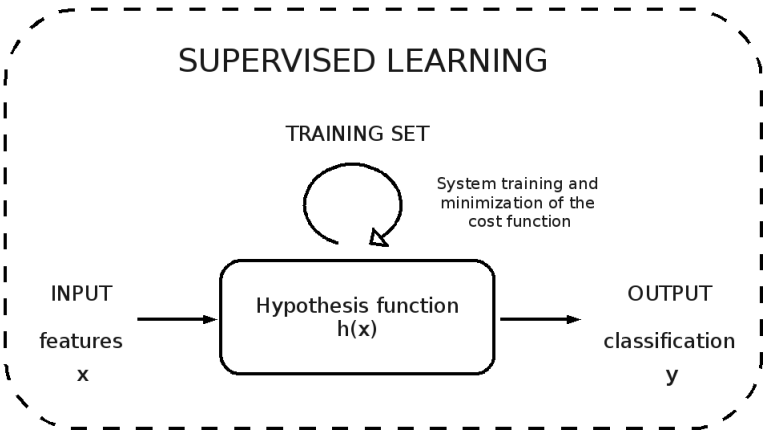


Fig. I.2.2: Principe de l'apprentissage supervisé : schéma d'une unité logistique.

Quel que soit le type de méthode utilisé, il est nécessaire de définir un certain nombre de **caractéristiques** ou **attributs** caractérisant chacun des éléments de l'ensemble que l'on veut classer. Ceux-ci servent de paramètres d'entrée (*input*) au système automatisé. Soit n le

nombre de caractéristiques d'un élément i du jeu de données, alors celui-ci peut s'écrire sous la forme d'un vecteur $\mathbf{x}^{(i)} = [x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}]$. Soit m la taille du jeu de données. On définit la matrice $\mathbf{X}$ contenant les m vecteurs caractéristiques du jeu de données telle que :

$$\mathbf{X} = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)}] = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \ldots & x_1^{(m)} \\ x_2^{(1)} & x_2^{(2)} & \ldots & x_2^{(m)} \\ \vdots & \vdots & \ldots & \vdots \\ x_n^{(1)} & x_n^{(2)} & \ldots & x_n^{(m)} \end{bmatrix} \tag{I.2.1}$$

C'est cette matrice des caractéristiques $\mathbf{X}$ qui sera donnée en entrée au classifieur.

Dans le cas de l'apprentissage supervisé, le système "apprend" (ou s'entraîne) à partir d'une classification déjà connue. On définit alors également le vecteur $\mathbf{y} = [y^{(1)}, y^{(2)}, ..., y^{(m)}]$ de longueur m contenant la classe attribuée à chacun des éléments i du jeu de données. Le but du problème de discrimination sera de trouver les paramètres θ_n d'une fonction hypothèse permettant d'expliquer $\mathbf{y}$ connaissant $\mathbf{X}$ tout en minimisant l'erreur. En d'autres termes, on cherche la fonction $h_\theta(\mathbf{x}^{(i)})$ qui donne la probabilité que l'élément i appartienne à la classe $y = k$ connaissant $\mathbf{x}^{(i)}$ et avec les n paramètres θ_n :

$$h_\theta^{(k)}(\mathbf{x}^{(i)}) = p(y^{(i)} = k|\mathbf{x}^{(i)}; \theta_n). \tag{I.2.2}$$

Un exemple simple

La figure I.2.3 donne l'exemple d'un jeu de données constitué de 2 classes A et B. Le but d'un algorithme de classification supervisée va être de trouver l'expression mathématique du séparateur entre A et B, connaissant la répartition des valeurs $x_j^{(i)}$ prises par chacun des éléments i du jeu de données pour un attribut j donné. Cette répartition est représentée sous forme d'histogrammes. Ceux-ci sont ensuite modélisés par les fonctions densité de probabilité (courbes). Celles-ci sont estimées à partir de la modélisation d'une gaussienne dans le cas des exemples synthétiques. En pratique, sur les données réelles, on utilise la méthode d'estimation par noyau développée par Parzen [1962] avec un noyau gaussien. C'est une méthode non-paramétrique qui calcule la densité f en chaque échantillon de la variable de la manière suivante :

$$f(x) = \frac{1}{mh} \sum_{i=1}^{m} K\left(\frac{x - x_i}{h}\right) \tag{I.2.3}$$

où m est le nombre d'éléments, h est le paramètre de lissage et K est le noyau gaussien de la forme

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$$

L'objectif du problème de classification va être de déterminer les paramètres du séparateur qui délimite les deux classes (ligne verte tiretée).

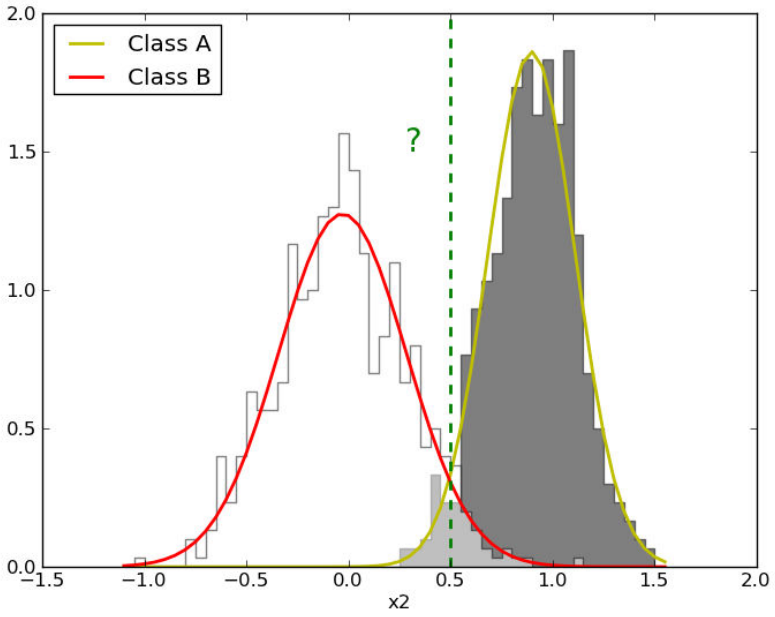


FIG. I.2.3: Exemple d'un jeu de données synthétiques dans lequel il y a deux classes. Les histogrammes donnent la fonction de répartition de l'attribut et les courbes colorées correspondent aux densités de probabilité associées (simples gaussiennes). La ligne verte pointillée symbolise le séparateur que l'on recherche.

Quelques méthodes de classification automatique

Parmi les méthodes de classification supervisée les plus connues, on peut citer la régression logistique (LR - *Logistic Regression* en anglais), la SVM (*Support Vector Machine* en anglais), ou les réseaux de neurones artificiels (ANN - *Artificial Neural Network* en anglais). Les ANN ne sont qu'une généralisation « grande échelle » de la régression logistique. Leur nom provient du fait que le réseau d'unités logistiques (une unité logistique = un neurone, voir FIG. I.2.2) est construit selon un système de couches pour lesquelles la sortie de la couche précédente correspond à l'entrée de la suivante, permettant ainsi un grand nombre de combinaisons (= de ramifications).

Dans ce travail de thèse, on a utilisé deux méthodes d'apprentissage supervisé (la régression logistique et la SVM) et une méthode d'apprentissage non supervisé (les K-moyennes). Ces méthodes seront décrites en détails dans la section sur les méthodes (§I.2.2).

La section suivante est un bref rappel des études qui ont déjà été menées dans le domaine de la classification automatique appliquée à la sismologie.

I.2.1.2 Pourquoi faire de la classification automatique en sismologie ?

Les sismomètres enregistrent toutes les vibrations qui se propagent dans le sous-sol, y compris celles causées par les activités humaines. Selon la problématique étudiée en sismologie, l'information que l'on souhaite garder ne concerne pas tous les types de signaux. Lors de l'élaboration des catalogues de sismicité, par exemple, les signaux anthropiques ne sont pas répertoriés. Dans les études de sismicité volcanique ou de microsismicité, tous les signaux enregistrés ne sont pas exclusivement locaux (on enregistrera inévitablement les grands séismes tectoniques lointains et les tectoniques locaux) : il s'agit alors de séparer les événements externes de ceux internes au système étudié. En effet, on constate que la pluralité observée parmi les phénomènes locaux est généralement révélatrice des processus qui les génèrent, d'où l'intérêt d'une classification.

Des outils de classification automatique existent déjà en sismologie et sont développés depuis quelques décennies. Ainsi, les premiers réseaux de neurones artificiels (ANN) appliqués à la sismologie l'ont été pour discrimer les séismes naturels des explosions nucléaires souterraines [Dowla et al., 1990, Pulli, 1990] à partir des informations contenues dans l'arrivée des différentes phases $P_g, L_g \ldots$. Depuis, l'idée a fait son chemin et a été reprise dans une problématique similaire par Fedorenko et al. [1998], Musil and Plešinger [1996], Ursino et al. [2011], mais avec des attributs sismiques différents (calcul de l'enveloppe des formes d'onde, caractéristiques spectrales, amplitudes, ...). Parallèlement, l'application des méthodes de classification automatique ne s'est pas limitée à la simple discrimination des signaux naturels et des signaux anthropiques, et s'est étendue plus particulièrement à la sismicité en domaine volcanique. L'intérêt de ce cas spécifique réside dans le fait que la diversité des signaux sismiques enregistrés peut traduire différents états du volcan étudié. Leur analyse et leur identification peut conduire à une meilleure connaissance et à une meilleure compréhension des phénomènes les générant. Ceci est d'autant plus important que la corrélation entre activité sismique et volcanique peut permettre, à terme, la mise en évidence de signes précurseurs des éruptions. Falsaperla et al. [1996] a ainsi été l'un des premiers à utiliser un ANN pour classer les quatre différents types d'explosions observés sur le Stromboli avec un taux de réussite atteignant les 90%. Les ANN ont ensuite été appliqués à d'autres volcans, avec des thématiques parfois légèrement différentes en fonction du volcan et de l'objectif de l'étude. Par exemple, les études de Falsaperla et al. [1996] et Langer and Falsaperla [2003] sur le volcan du Stromboli s'attèlent à ne classer que les événements de type explosion. Etant donné que d'autres types d'événements sont aussi enregistrés (les tremors volcaniques notamment), dans un problème de classification plus "grande échelle", les explosions pourraient constituer une classe à elles seules. Ceci montre que la définition précise d'un problème de discrimination conditionne ce problème en lui donnant l'orientation souhaitée (dans le cas du Stromboli, une corrélation entre les différents types d'explosion et les cratères les produisant avait été observée). Ainsi, les nombreux exemples d'application qui ont suivis sur la Soufrière de Montserrat [Langer et al., 2003, 2006], sur le Vésuve [Scarpetta et al., 2005], sur le Krakatau [Ibs-von Seht, 2008] ou le Villarrica (Chili) [Curilem et al., 2009] ont permis de démontrer la force des ANN pour la classification des événements sismiques, mais aussi, et surtout, le développement et la recherche de nouveaux attributs sismiques.

Si au travers des nombreuses applications effectuées avec les ANN, ceux-ci ont pu prouver leur efficacité, ils ne restent pas moins faciles à mettre en œuvre. Leur désavantage principal

demeure le choix de la structure du réseau (nombre de couches et nombre d'unités par couches). Curilem et al. [2009] s'est affranchi du problème en couplant l'ANN avec un algorithme génétique, permettant ainsi de sélectionner les meilleures combinaisons d'attributs et de définir la structure de l'ANN et l'algorithme d'apprentissage les plus adaptés à leur problème.

D'autres auteurs ont développé d'autres méthodes. Parmi celles-ci, on peut citer les cartes auto-adaptatives (SOM - *Self-Organizing Map* en anglais) [Esposito et al., 2008] qui fonctionnent en apprentissage non-supervisé ; mais surtout les modèles de Markov cachés (HMM - *Hidden Markov Model* en anglais). L'exemple d'application le plus courant des HMM est la reconnaissance automatique de la parole. Ohrnberger [2001] a été le premier à l'utiliser en sismologie pour la classification des événements sismiques enregistrés sur le Merapi (Indonésie), en s'appuyant sur les similitudes qui existent entre une onde acoustique et une onde sismique : dans les deux cas, le problème consiste à détecter et classer les signaux transitoires enregistrés. L'intérêt principal de cette méthode en sismologie est la possibilité de détecter et classer simultanément les événements sismiques en analysant les sismogrammes continus (alors que les ANN supposent de travailler avec des signaux préalablement détectés et découpés). Un deuxième avantage est la capacité des HMM à reconnaître les événements même bruités ou "cachés" dans des événements longue-période. Le principe d'un algorithme de HMM est le suivant : il suppose d'abord que les séquences de vecteurs d'attributs qui caractérisent un événement sont générés par un modèle de Markov. Ce dernier est défini par plusieurs états, un changement d'état se produisant à chaque échantillon de temps. Le passage d'un état à un autre est liée à sa probabilité de transition. Lors de la phase d'entraînement du système, ce sont donc ces probabilités de transition qui doivent être déterminées, ainsi que les probabilités de sortie, c'est-à-dire celles qui permet de choisir la classe la plus probable. Cette méthode a fait ses preuves en sismologie avec divers exemples d'application en domaine volcanique (Merapi [Ohrnberger, 2001, Beyreuther et al., 2012], île de la Déception (Antarctique) [Benítez et al., 2007], Stromboli et Etna [Ibáñez et al., 2009], Soufrière de Montserrat [Hammer et al., 2012]), mais aussi pour le réseau sismique bavarois [Beyreuther and Wassermann, 2008].

Enfin, des techniques de logique floue (*fuzzy logic* en anglais) ont aussi été mises en œuvre récemment pour la discrimination des éboulements et des événements volcano-tectoniques sur le Piton de la Fournaise à la Réunion avec un taux de réussite supérieur à 90% [Hibert et al., 2014].

Un seul cas d'utilisation de la SVM en sismologie a été répertorié pour la discrimination des différents types de tremors volcaniques enregistrés sur le volcan italien de l'Etna [Masotti et al., 2006].

Ce bref rappel des autres études de classification qui ont été effectuées en sismologie permet de mettre en évidence leur diversité, tant dans les méthodes choisies que dans leurs objectifs (classification à plus ou moins grande échelle, dans un but purement scientifique de meilleure compréhension des phénomènes ou non).

Les deux méthodes d'apprentissage supervisé présentées dans cette thèse ont été choisies pour les raisons suivantes :

- la **régression logistique**, parce qu'elle est simple à mettre en œuvre et permet de bien appréhender les concepts inhérents aux problèmes de discrimination automatique. Il faut rappeler que les ANN correspondent simplement à l'agrégation et la structuration de plusieurs unités de régression logistique. Ils sont évidemment plus efficaces que la régression logistique seule, de part leur grande adaptabilité aux divers problèmes de classification,

y compris (et surtout) non-linéaires. Mais cette capacité d'adaptation constitue aussi leur point faible, car il s'agit de trouver la meilleure structure possible parmi toutes celles potentielles. Un des désavantages notable de la LR est qu'elle nécessite d'avoir un nombre de données et de caractéristiques suffisant pour pouvoir prétendre à un niveau de stabilité satisfaisant.

– la **SVM**, parce qu'elle permet justement de résoudre des problèmes de classification non-linéaires bien mieux que ne le ferait la régression logistique, mais en évitant la complexité des ANN.

Ces deux méthodes ont aussi été préférées aux HMM car ceux-ci sont plus particulièrement adaptés à l'analyse des signaux continus puisqu'ils supposent que l'information contenue dans un état présent du signal renseigne sur son état futur indépendamment de son état passé. Peu d'importance est ainsi accordée au sens physique réel des attributs.

I.2.2 Description des méthodes d'apprentissage supervisé

Dans cette section, on présente plus en détail la théorie qui se cache derrière chacune des méthodes utilisées. Elle s'appuie en partie sur le cours en ligne de *Machine Learning* donné par l'université de Stanford [Ng, 2012].

I.2.2.1 Régression logistique

Principe

La régression logistique est une méthode de d'apprentissage supervisé qui utilise une fonction hypothèse de la forme suivante :

$$h_\theta(\mathbf{x}) = g(\theta_0 x_0 + \theta_1 x_1 + \ldots + \theta_n x_n) = g(\theta^T \mathbf{x}) \tag{I.2.4}$$

où g est la fonction logistique ou fonction sigmoïde (Fig. I.2.4) :

$$g(z) = \frac{1}{1 + exp(-z)}. \tag{I.2.5}$$

Cette fonction est bornée par les valeurs 0 et 1, ce qui facilite l'interprétation de sa valeur comme la probabilité que l'élément pris en considération appartienne ou pas à la classe cible. Dans la suite de cette explication théorique, nous nous posons dans le cas d'une classification binaire (c'est à dire à deux classes seulement). La généralisation à plusieurs classes sera abordée par la suite.

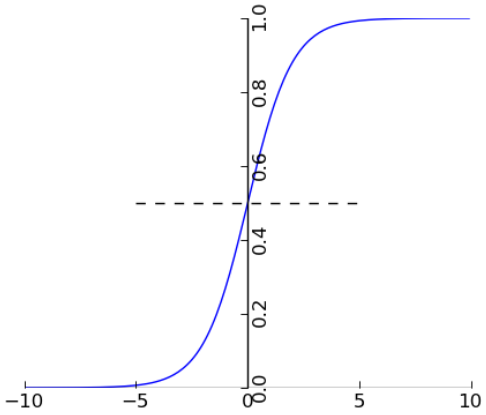


FIG. I.2.4: Fonction sigmoïde.

Le but du problème de classification se résume à déterminer les paramètres θ_n qui vont minimiser une fonction coût du type moindres carrés :

$$\text{Cout}(h_\theta(\mathbf{X}), \mathbf{y}) = \frac{1}{2}(h_\theta(\mathbf{X}) - \mathbf{y})^2, \tag{I.2.6}$$

où les valeurs du vecteur $\mathbf{y}$ sont 1 si cet élément appartient à la classe cible, et 0 autrement.

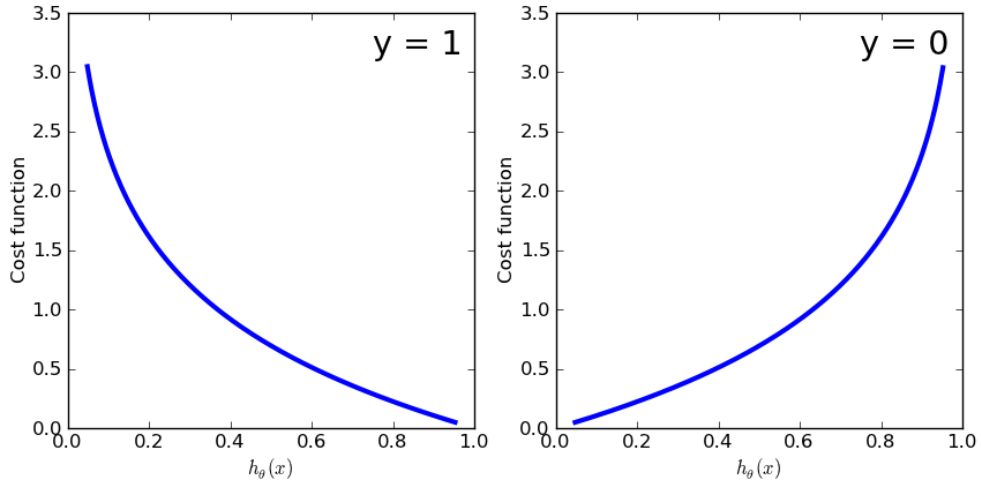


FIG. I.2.5: Fonction coût pour $\mathbf{y}$=1 et $\mathbf{y}$=0. $h_\theta(\mathbf{X})$ est la fonction hypothèse et donne la probabilité que $\mathbf{x}$ appartienne à la classe cible qui vaut 1. (Gauche) Pour $\mathbf{y}$=1, le coût est nul lorsque $h_\theta(\mathbf{X})$=1. (Droite) Pour $\mathbf{y}$=0, le coût est très élevé lorsque $h_\theta(\mathbf{X})$=1.

La fonction logistique g n'étant pas linéaire, son élévation au carré dans la fonction coût crée une fonction non convexe : il existe de multiples miminum locaux. On n'est donc pas assuré de converger vers le minimum absolu pendant la phase de minimisation et on préfère

minimiser à la place la fonction suivante (FIG. I.2.5) :

$$\log(\text{Cout}(h_\theta(\mathbf{X}), \mathbf{y})) = \begin{cases} -\log(h_\theta(\mathbf{X})) & \text{si } \mathbf{y} = 1, \\ -\log(1 - h_\theta(\mathbf{X})) & \text{si } \mathbf{y} = 0. \end{cases} \tag{I.2.7}$$

$$= -\mathbf{y}\log(h_\theta(\mathbf{X})) - (1 - \mathbf{y})\log(1 - h_\theta(\mathbf{X})). \tag{I.2.8}$$

En généralisant aux m échantillons du *training set*, on veut donc minimiser la fonction suivante :

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log(h_\theta(\mathbf{x}^{(i)})) + (1 - y^{(i)})\log(1 - h_\theta(\mathbf{x}^{(i)}))\right] \tag{I.2.9}$$

où $i = 1 \dots m$ est l'indice d'un élément du jeu de données.

La minimisation de J peut se faire grâce à différentes méthodes d'optimisation, telles que la méthode de descente du gradient, la méthode du gradient conjugué etc., connaissant les dérivées partielles par rapport à chaque attribut j (allant de 1 à n) :

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m}\left[(h_\theta(\mathbf{x}^{(i)}) - y^{(i)})x_j^{(i)}\right] \tag{I.2.10}$$

Dans un souci de facilité d'implémentation, nous avons utilisé la minimisation par descente du gradient à pas fixe. Cette minimisation est peu performante en termes de rapidité d'exécution. Néanmoins, le nombre de données étant encore limité dans les deux exemples d'application présentés dans les prochaines parties, il n'a pas été jugé nécessaire et utile d'optimiser cette rapidité d'exécution (mais il faudrait y songer pour de futurs développements).

Une fois les paramètres de la fonction hypothèse déterminés, on peut définir la **surface de séparation** (*decision boundary* en anglais) qui permettra de séparer les éléments de chacune des classes. Pour illustrer, nous allons prendre l'exemple d'un problème de discrimination de deux classes pour des éléments caractérisés par deux attributs x_1 et x_2. La fonction hypothèse est du type :

$$h_\theta(\mathbf{x}^{(i)}) = g(\theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)})$$

Connaissant les paramètres θ, on peut décider que les éléments du jeu de données appartiendront à la classe $\mathbf{y} = 1$ si $h_\theta(\mathbf{X}) \geq 0.5$, ce qui revient à considérer que :

$$\begin{cases} y^{(i)} = 1 \text{ si } h_\theta(\mathbf{x}^{(i)}) \geq 0.5, & \text{soit } \theta_0 + \theta_1 x_1 + \theta_2 x_2 \geq 0 \\ y^{(i)} = 0 \text{ si } h_\theta(\mathbf{x}^{(i)}) < 0.5, & \text{soit } \theta_0 + \theta_1 x_1 + \theta_2 x_2 < 0 \end{cases} \tag{I.2.11}$$

Le séparateur linéaire sera donc défini par la fonction suivante : $db(x_1, x_2) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$, ce qui revient à tracer dans le plan (x_1, x_2) la droite d'équation :

$$x_2 = -\frac{1}{\theta_2}(\theta_1 x_1 + \theta_0)$$

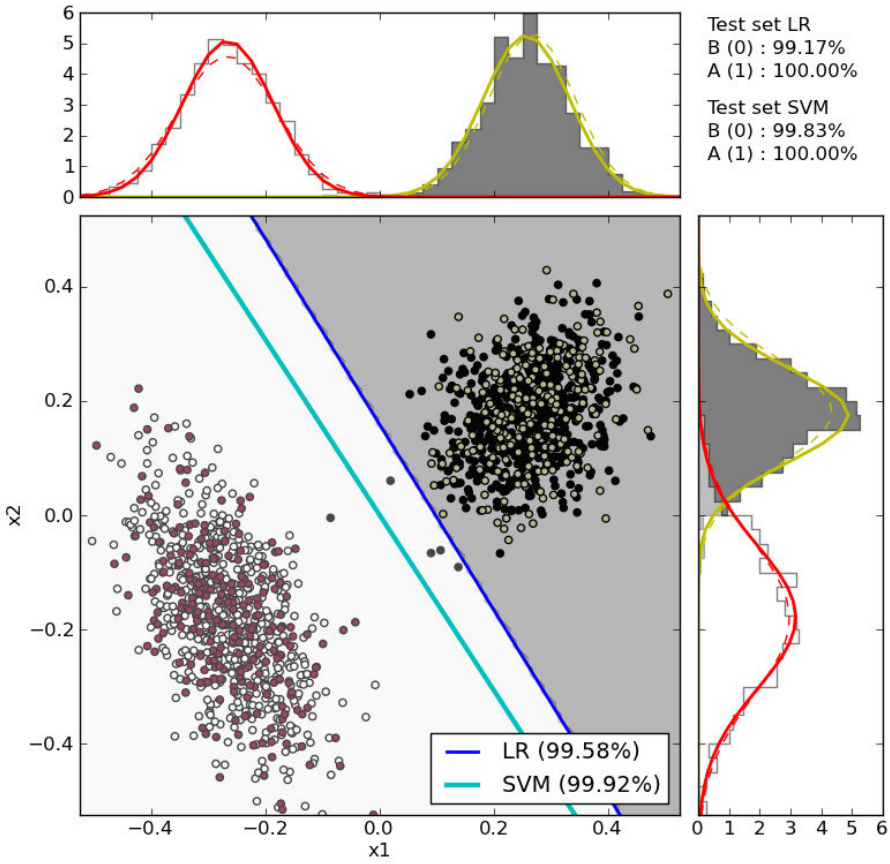


FIG. I.2.6: Séparateurs pour un exemple simple de 2 classes et 2 attributs (x_1 et x_2) : en bleu foncé, celui de la régression logistique ; en bleu clair, celui de la SVM. Les éléments de la classe 0 sont en blanc ; ceux de la classe 1 en noir. Les éléments constitutifs du *training set* sont aussi représentés, respectivement en rouge et jaune. La couleur de fond cartographie la séparation d'après les résultats de la régression logistique (tous les éléments situés dans la partie plus foncés seront classés automatiquement dans la classe 1). Les densités de probabilité (gaussiennes) associées à chacun des attributs sont également représentées de part et d'autre du graphe.

Un exemple d'illustration est donné dans la figure I.2.6. Il a été réalisé sur des données synthétiques générées de la manière suivante (voir A4 en annexes) :

- les deux classes sont chacune issue du tirage aléatoire d'une distribution gaussienne multivariable (2 attributs).
- les deux classes sont de même taille et comportent 600 éléments chacune (soit un *test set* de 1200 éléments).
- le *training set* fait 40% de la taille du *test set*, soit 480 éléments. Le tirage de ces éléments s'est fait de manière strictement identique à celle du *test set*. Les proportions entre les deux classes sont toujours respectées.

Il est également possible de déterminer des séparateurs non linéaires en considérant non plus les valeurs d'attributs simples, mais leurs puissances, ou bien en multipliant plusieurs

valeurs d'attributs entre eux. Par exemple, on pourrait très bien considérer des fonctions hypothèse du type :

$$h_\theta(\mathbf{x}) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2)$$

ou encore :

$$h_\theta(\mathbf{x}) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2)$$

L'avantage d'un séparateur non linéaire par rapport à un séparateur linéaire est qu'il devrait s'adapter encore mieux aux données. Le désavantage est le risque de sur-apprentissage (ce problème est discuté plus en détail dans la section I.2.2.1).

Précision et rappel

La fonction hypothèse déterminée à l'issue du processus d'apprentissage prend des valeurs comprises entre 0 et 1 (c'est une probabilité) :

$$0 \leq h_\theta(x) \leq 1$$

Cependant, le seuil de décision à partir duquel un élément donné va être classé dans une classe plutôt que l'autre reste à définir. Il peut prendre n'importe quelle valeur *a priori*. Pour aider à choisir le seuil le mieux adapté au problème, on utilise deux indicateurs : la précision et le rappel.

La **précision** donne la fraction d'une classe prédite par rapport à l'ensemble de la classe réelle (se référer à la figure I.2.7a pour la définition des vrais positifs, faux positifs...) :

$$\text{Précision} = \frac{\text{Vrais positifs}}{\text{Positifs prédits}} = \frac{\text{Vrais positifs}}{\text{Vrais positifs} + \text{Faux positifs}} \tag{I.2.12}$$

Le **rappel** (ou *recall* en anglais) donne la fraction d'une classe correctement prédite (donc le taux de bonne classification) :

$$\text{Rappel} = \frac{\text{Vrais positifs}}{\text{Positifs réels}} = \frac{\text{Vrais positifs}}{\text{Vrais positifs} + \text{Faux négatifs}} \tag{I.2.13}$$

$$\tag{I.2.14}$$

Il existe un *trade-off* entre la précision et le rappel (Fig. I.2.7b,I.2.9). L'idéal serait de choisir le seuil qui permet le meilleur équilibre entre les deux paramètres ou de l'adapter en fonction de ce que l'on souhaite :

- si on veut être sûr de bien prédire les éléments de la classe 1, alors il faut augmenter le seuil, ce qui se traduit par une précision plus élevée et un rappel plus faible (on veut s'assurer que tous les éléments classés en 1 sont bien de classe 1) ;

- si, en revanche, on veut éviter de manquer de classer des éléments dans la classe 1, il faut diminuer le seuil, ce qui se traduit par un rappel plus élevé et une précision moindre (on veut classer tous les éléments de la classe 1 en classe 1).

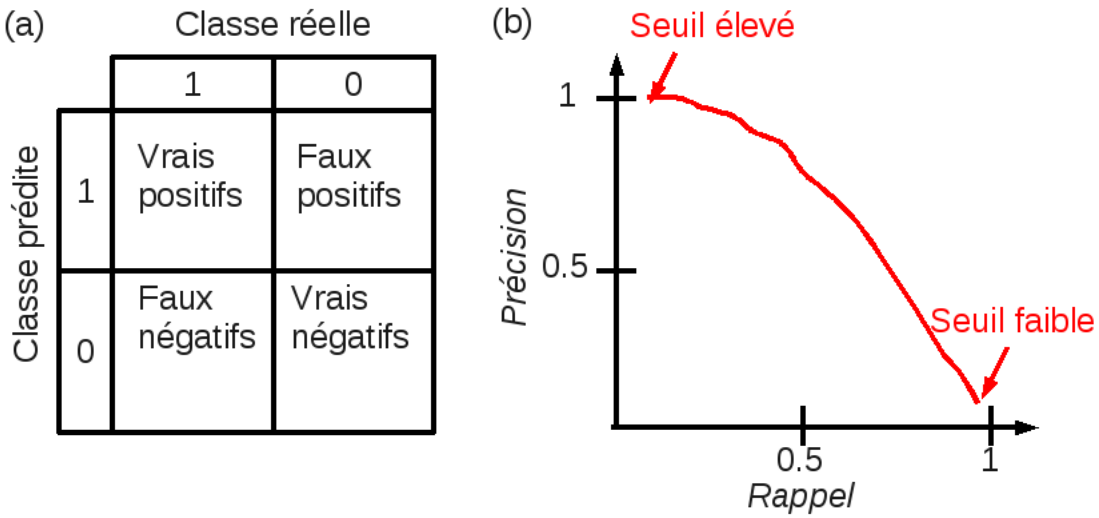


FIG. I.2.7: (a) Définition des termes vrais positifs, faux positifs, faux négatifs et vrais négatifs. (b) Illustration du *trade-off* existant entre la précision et le rappel.

Pour réussir à déterminer le seuil le plus adapté à notre cas automatiquement, on calcule alors un troisième indicateur, le **score $\mathbf{F_1}$** :

$$F_1 = 2\frac{PR}{P+R} \tag{I.2.15}$$

où P est la précision et R est le rappel. En pratique, il suffit donc de tester différents seuils lors de la phase d'entraînement du système et de choisir celui qui maximise le score F_1.

Dans l'exemple de la figure I.2.6, on avait choisi le séparateur qui classait les éléments de la classe 1 avec une probabilité supérieure à 0.5. On vient de voir que ce choix est arbitraire et qu'il peut être fait plus judicieusement. La figure I.2.8 illustre ce propos en présentant les différentes surfaces de décision calculées pour la discrimination des deux classes d'un jeu de données synthétiques. Dans ce cas, les équations des séparateurs sont donnés par l'équation suivante :

$$x_2 = -\frac{\theta_1}{\theta_2}x_1 - \frac{1}{\theta_2}\left(\log\frac{1-t}{t} - \theta_0\right)$$

où t est le seuil de probabilité choisi.

Au seuil 0.9, par exemple, on voit qu'on classera bien tous les éléments de la classe 1 (points noirs). Avec un seuil de 0.1, en revanche, quasiment tous les éléments de la classe 0 (points blancs) seront bien classés.

Les courbes de précision et de rappel calculées pour ce même jeu de données synthétiques sont représentées sur la figure I.2.9.

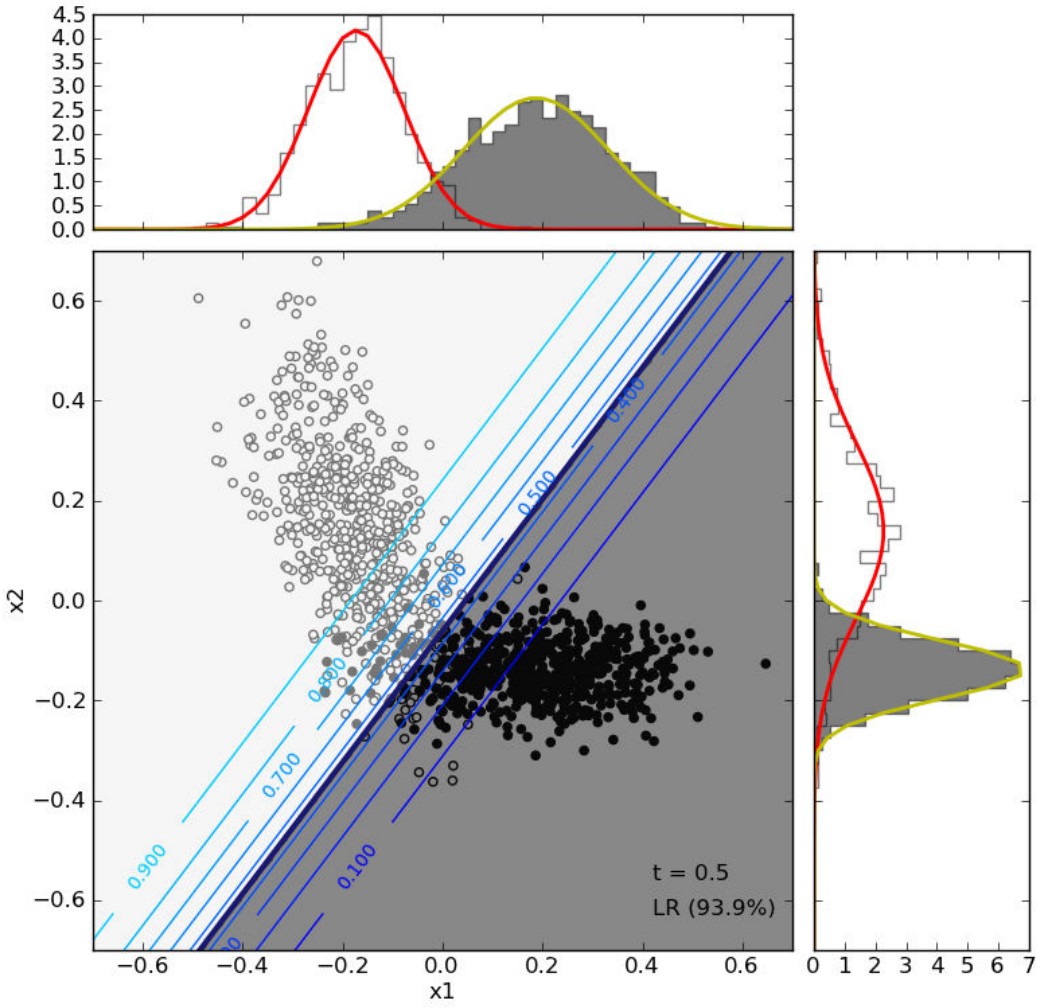


FIG. I.2.8: Séparateurs pour un exemple simple de 2 classes et 2 attributs ($x1$ et $x2$) en fonction du seuil de probabilité choisi. Le seuil donnant la meilleure classification finale vaut 0.5. Les éléments de la classe 0 sont en blanc ; ceux de la classe 1 en noir. La couleur de fond cartographie l'appartenance à la classe 1 (foncé) et 0 (clair) d'après les résultats de la régression logistique pour un seuil égal à 0.5. Les densités de probabilité associées à chacun des attributs sont également représentées de part et d'autre du graphe.

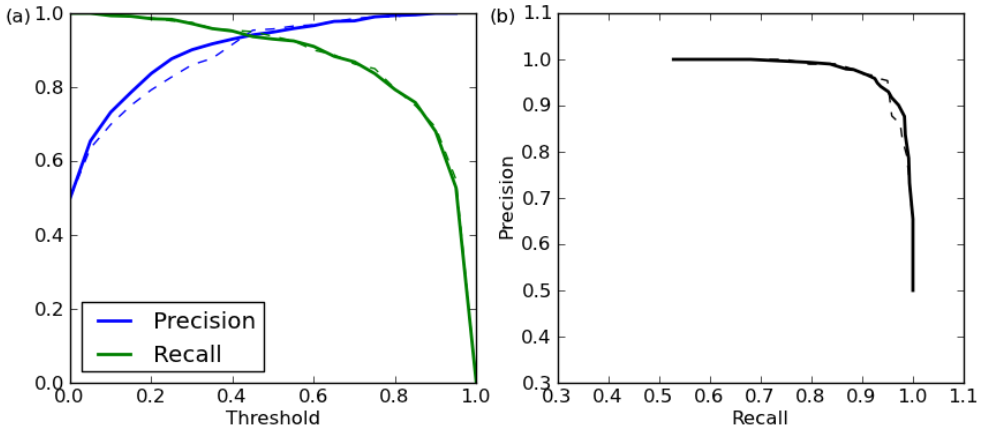


FIG. I.2.9: (a) Courbes de précision (bleu) et rappel (vert) obtenues pour le *test set* (traits pleins) et le *training set* (traits tiretés) en fonction du seuil de probabilité. (b) Courbe de précision en fonction du rappel. Les courbes ont été calculées sur des données synthétiques (voir FIG. I.2.8).

Régularisation et courbes d'apprentissage

L'un des problèmes qui peut facilement se rencontrer en apprentissage automatique supervisé est celui du sur-apprentissage ou du sous-apprentissage. En cas de **sur-apprentissage**, l'erreur sur le *training set* est minime mais la variance est élevée (mauvaise généralisation au *test set*). En cas de **sous-apprentissage**, c'est la situation inverse qui se produit, avec une variance faible, mais une erreur élevée (FIG. I.2.10). Il faut donc contrôler du mieux possible le *trade-off* existant entre le biais (ou erreur) et la variance.

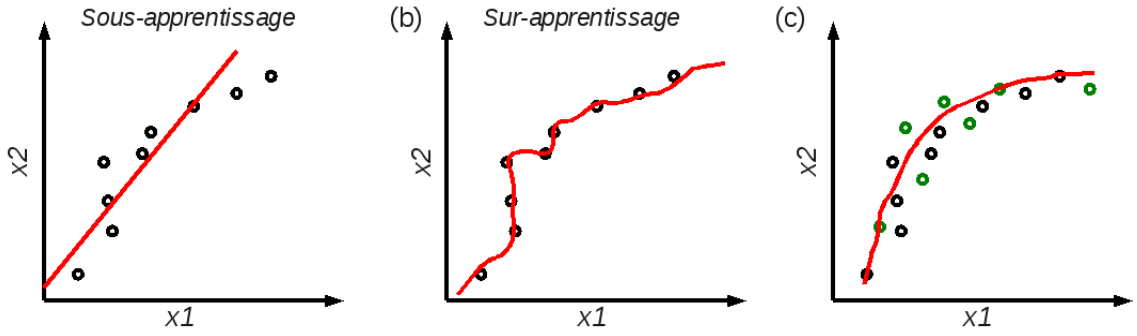


FIG. I.2.10: Illustrations du sous-apprentissage (a) et du sur-apprentissage (b) en comparaison de la situation souhaitée (c). En (c), les ronds verts représentent les éléments qui ont été ajoutés au *training set* par rapport aux situations (a) et (b).

Afin de pallier à ces problèmes, il est nécessaire de régulariser le problème, en introduisant un coefficient de régularisation λ qui a pour but de contrôler le *trade-off* entre la minimisation de l'erreur et l'augmentation des valeurs des paramètres θ (on veut garder les θ les plus petits

possibles). On ajoute donc un terme à la fonction qu'on cherche à minimiser $J(\theta)$ (I.2.9) :

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log(h_\theta(\mathbf{x}^{(i)})) + (1-y^{(i)})\log(1-h_\theta(\mathbf{x}^{(i)}))\right] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2 \qquad \text{(I.2.16)}$$

Une autre manière de contrôler le *trade-off* entre erreur et variance est de jouer sur le degré polynomial d du séparateur recherché (rappelons que celui-ci n'est pas nécessairement linéaire) ou sur le nombre de paramètres utilisés.

Ces deux paramètres cruciaux, λ et d, sont déterminés à partir du *training set* lors de la phase d'entraînement du système. Celui-ci est alors séparé en 3 sous-ensembles (voir Fig. I.2.12) :

- le *training set*, constitué de 60% du *training set* initial. C'est sur ce jeu que se fait réellement le processus d'apprentissage et la minimisation de la fonction coût J_{train}. Pour éviter les confusions, on nommera ce sous-ensemble *sub-training set*.

- le *cross-validation set*, constitué de 20% du *training set* initial. C'est à partir de ce jeu que λ et d sont retenus : pour chaque d et chaque λ , les paramètres θ sont calculés à partir du *sub-training set*. Ceux-ci sont utilisés pour calculer la fonction coût $J_{CV}(\theta)$ avec les données du jeu de cross-validation. Une fois que les calculs ont été effectués pour plusieurs valeurs de d ou de λ, les valeurs pour lesquelles la fonction coût du jeu de cross-validation est minimisée sont retenues.

- le *test set*, constitué de 20% du *training set* initial.Pour éviter les confusions, on le nommera *sub-test set*. C'est sur ce jeu que l'erreur de généralisation est estimée. Elle se définit comme suit (avec t le seuil de décision) :

$$\text{Erreur}(h_\theta(\mathbf{x}), y) = \begin{cases} 1 & \text{si } h_\theta(\mathbf{x}) \geq t \text{ et } y = 0 \\ & \text{ou si } h_\theta(\mathbf{x}) < t \text{ et } y = 1, \\ 0 & \text{sinon.} \end{cases} \qquad \text{(I.2.17)}$$

Pour évaluer la validité de la fonction hypothèse trouvée à l'issue de l'étape d'apprentissage, il est intéressant d'afficher des courbes d'apprentissage (Fig. I.2.11) qui peuvent nous aider à diagnostiquer puis solutionner les problèmes rencontrés dans l'algorithme. En effet, le but est de réussir à trouver la solution la plus satisfaisante, avec une erreur et une variance minimisées : on cherche un juste équilibre entre une situation de sur-apprentissage et une situation de sous-apprentissage. La figure I.2.11 présente les 3 situations (a) optimale, b) sous-apprentissage, c) sur-apprentissage). Elle montre qu'il est plus facile de coller aux données quand on en a peu, et que cela devient plus difficile lorsque le nombre d'échantillons augmente (J_{train} augmente). Mais plus il y a de données, plus on peut bien généraliser, d'où la décroissance de J_{CV}.

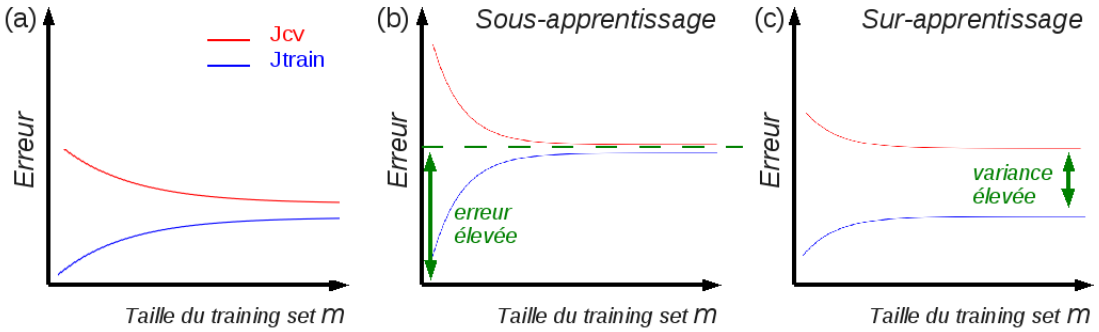


FIG. I.2.11: Courbes d'apprentissage : évolution des fonctions coût du *training set* et du set de cross-validation en fonction du nombre d'échantillons dans le *training set*. (a) Situation équilibrée. (b) Situation où l'erreur est élevée (sous-apprentissage). (c) Situation où la variance est élevée (sur-apprentissage).

Comme évoqué précédemment, on peut solutionner en partie les problèmes de sous- et sur-apprentissage en jouant sur un certain nombre de paramètres :

- le nombre d'éléments m du *training set* : augmenter m diminue le risque de sur-apprentissage (voir FIG. I.2.10c).
- le nombre d'attributs n : en cas de sur-apprentissage, il faut l'abaisser afin d'éviter un ajustement excessif aux données ; au contraire, en cas de sous-apprentissage, il faut en ajouter afin de réduire les erreurs de classification.
- le degré polynomial d de la fonction hypothèse : augmenter le degré revient à ajouter des attributs (voir point précédent).
- le coefficient de régularisation λ : il a pour but de réduire la magnitude des paramètres θ régissant la fonction hypothèse. Si la valeur de λ est trop grande, on sera en situation de sous-apprentissage ; dans le cas contraire (λ trop petit, ce qui équivaut à avoir un nombre trop important d'attributs), on est en sur-apprentissage.

Un récapitulatif des paragraphes précédents est proposé dans le tableau I.2.1.

Situation	**Diagnostic**				**Solutions proposées**			
	Erreur	Variance	J_{train}	J_{CV}	m	n	d	λ
Sur-apprentissage	faible	élevée	faible	élevé	↗	↘	↘	↗
Sous-apprentissage	élevée	faible	élevé	élevé	↘	↗	↗	↘

TAB. I.2.1: Récapitulatif des deux situations possibles dans un problème d'apprentissage automatisé et solutions possibles pour y remédier.

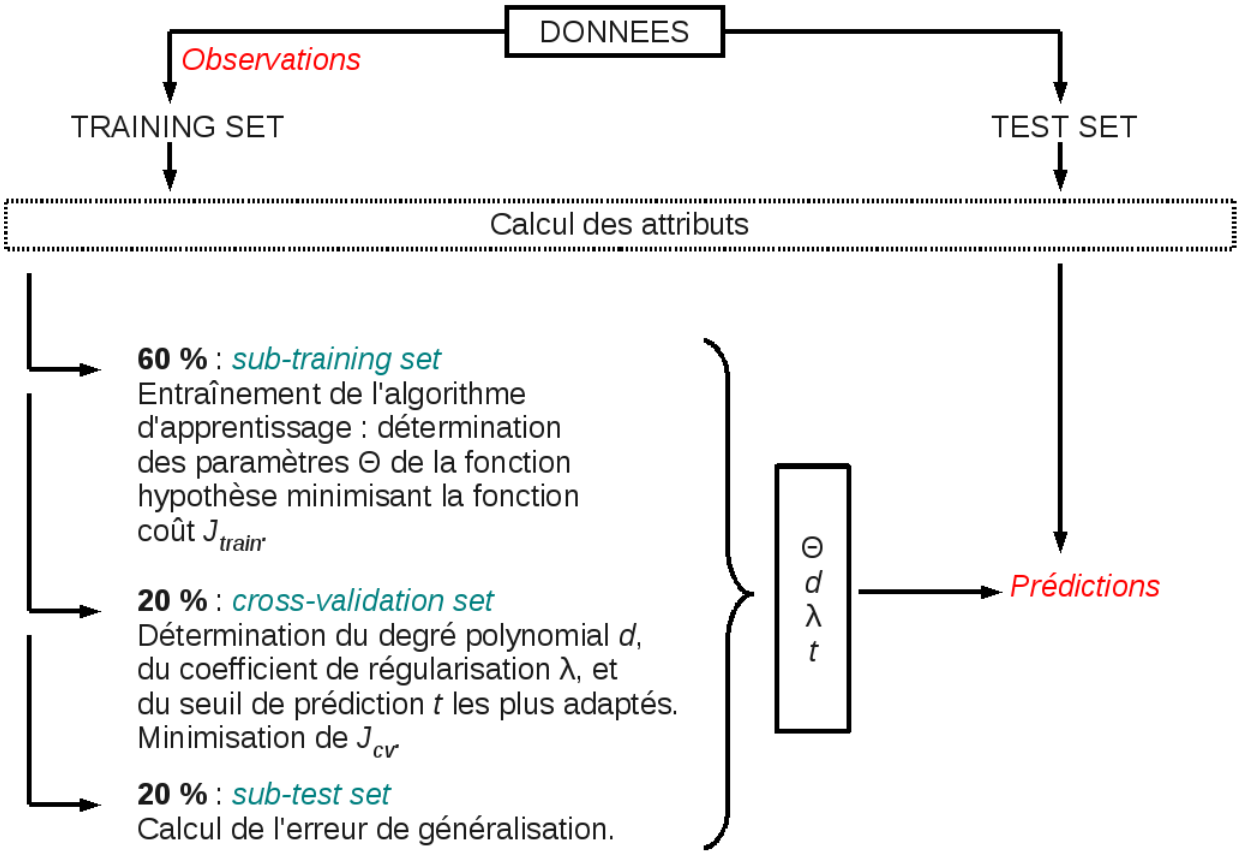


FIG. I.2.12: Schéma récapitulatif du processus de régression logistique.

I.2.2.2 SVM : Support Vector Machine

La SVM (*Support Vector Machine* en anglais, *Machine à Vecteurs de Support* ou *Séparateur à Vaste Marge* en français) est une deuxième méthode d'apprentissage supervisé. Contrairement à la régression logistique dans laquelle on cherche le séparateur qui minimise la fonction coût globale $J(\theta)$, la SVM recherche le séparateur optimal, c'est-à-dire celui qui maximise la marge entre les échantillons et le séparateur. Il est donc unique, là où il peut exister une multitude de séparateurs potentiels pour la régression logistique.

L'intérêt et l'avantage de la SVM par rapport à la régression logistique (et, plus généralement, par rapport aux ANN) réside dans sa capacité à résoudre plus facilement des problèmes de classification non-linéaires, donc plus complexes. Pour essayer d'expliquer au mieux la théorie à l'origine de la SVM, on traitera d'abord le cas simple linéaire, en procédant par analogie avec la régression logistique. Ensuite, on s'attaquera au cas non-linéaire avec l'utilisation d'une fonction noyau gaussienne.

Cas d'un problème linéaire

On rappelle que dans la régression logistique, on cherchait à minimiser la fonction suivante (I.2.16) :

$$J(\theta) = \frac{1}{m}\underbrace{\sum_{i=1}^{m}\left[-y^{(i)}\log(h_\theta(\mathbf{x}^{(i)})) - (1-y^{(i)})\log(1-h_\theta(\mathbf{x}^{(i)}))\right]}_{=A} + \frac{\lambda}{m}\underbrace{\frac{1}{2}\sum_{j=1}^{n}\theta_j^2}_{=B} \tag{I.2.18}$$

Cette fonction est constituée de deux termes A et B. Dans la régression logistique, on recherche simplement les paramètres θ tels que $\min_\theta A + \lambda B$: la meilleure classification possible des éléments du *training set* étant obtenue en minimisant le terme A, il faut donc aussi maintenir le terme λB (donc les coefficients θ) le plus petit possible. Dans la SVM, le problème se pose juste un peu différemment puisque que l'on cherche θ tels que $\min_\theta CA + B$. Cette redéfinition du problème de minimisation implique un contrôle du *trade-off* différent entre l'adéquation aux données et les paramètres θ décrivant la surface de décision.

Dans la régression logistique, la classe d'appartenance d'un élément i est déterminée de la manière suivante (avec un seuil de décision valant 0.5) :

$$\begin{cases} y^{(i)} = 1 \text{ si } \theta^T\mathbf{x}^{(i)} \geq 0 \\ y^{(i)} = 0 \text{ si } \theta^T\mathbf{x}^{(i)} < 0 \end{cases} \tag{I.2.19}$$

Traitons le cas dans lequel le terme A est proche de 0, c'est-à-dire l'adéquation aux données est quasiment parfaite. La prédiction de la valeur de $h_\theta(\mathbf{x})$ doit donc être directe : il faut « exagérer » la limite de décision (voir Fig. I.2.13). Ainsi, on aura :

$$\begin{cases} y^{(i)} = 1 \text{ si } \theta^T\mathbf{x}^{(i)} \geq 1 \\ y^{(i)} = 0 \text{ si } \theta^T\mathbf{x}^{(i)} \leq -1 \end{cases} \tag{I.2.20}$$

Pour mieux comprendre en quoi ceci permet de maximiser la marge entre les échantillons, prenons un exemple simple où $n = 2$ attributs et $\theta_0 = 0$. On a donc $\theta = [\theta_1, \theta_2]$. Pour simplifier le problème encore plus, on considèrera également que le terme A est nul, ce qui implique que C est très grand (en pratique, c'est mieux si C n'est pas trop grand). La SVM cherche alors à minimiser le terme

$$B = \frac{1}{2}(\theta_1^2 + \theta_2^2) = \frac{1}{2}||\theta||^2$$

De plus, pour un élément i du jeu de données, on sait aussi que $\theta^T\mathbf{x}^{(i)} = p^{(i)}||\theta||$ où $p^{(i)}$ est la longueur de la projection du vecteur $\mathbf{x}^{(i)}$ sur le vecteur θ.

L'équation I.2.20 peut donc se réécrire ainsi :

$$\begin{cases} y^{(i)} = 1 \text{ si } p^{(i)}||\theta|| \geq 1 \\ y^{(i)} = 0 \text{ si } p^{(i)}||\theta|| \leq -1 \end{cases} \tag{I.2.21}$$

La figure I.2.14 présente deux cas de figures. En (a), la surface de décision (ligne rouge) est celle qui aurait pu être calculée par la régression logistique par exemple. En (b), la surface de décision est celle calculée par la SVM et elle maximise la marge entre les échantillons du jeu de données.

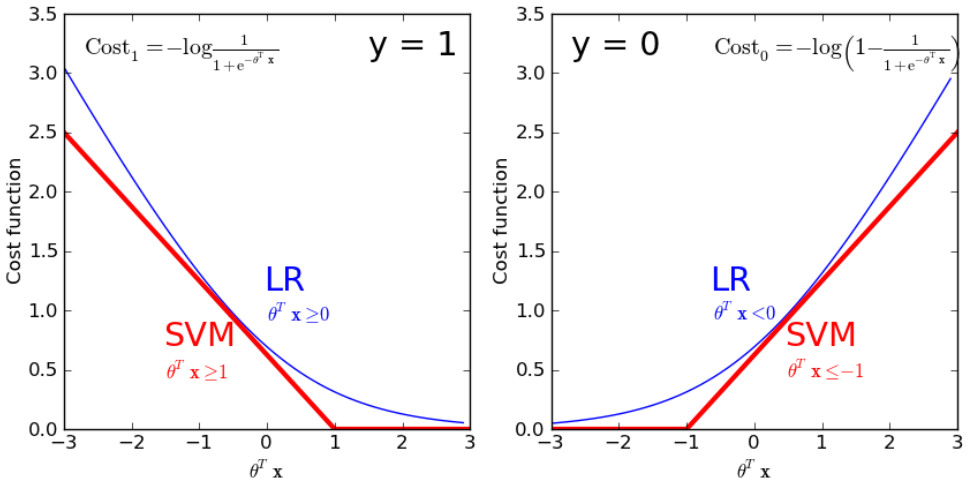


FIG. I.2.13: Fonctions coût pour y=1 (gauche) et y=0 (droite) pour la régression logistique (bleu) et la SVM (rouge).

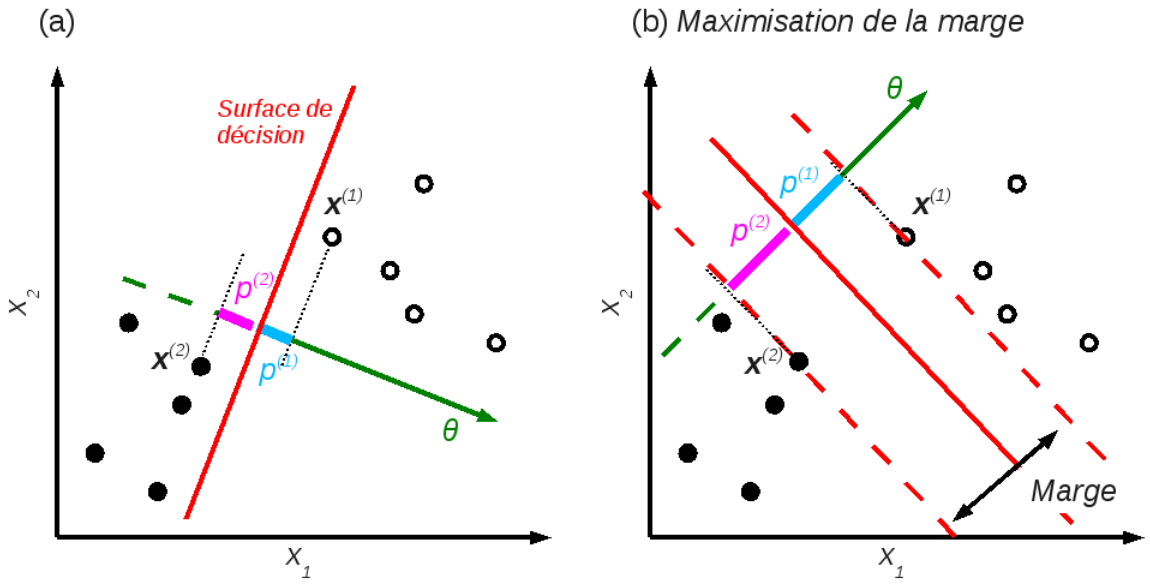


FIG. I.2.14: Schéma explicatif de la maximisation de la marge. (a) Cas avec un séparateur (en rouge) qui pourrait être déterminé par la régression logistique (parmi toutes les possibilités existantes). (b) Cas avec le séparateur maximisant la marge déterminé par la SVM. Les lignes tiretées délimitent la marge. Les longueurs $p^{(1)}$ (bleu) et $p^{(2)}$ (rose) correspondent à la projection des vecteurs $\mathbf{x}^{(1)}$ et $\mathbf{x}^{(2)}$ respectivement sur le vecteur θ (en vert).

Dans le cas (a), les longueurs de projection $p^{(1)}$ et $p^{(2)}$ sont petites. Les conditions $p^{(1)}||\theta|| \geq 1$ (avec $p^{(1)} > 0$) et $p^{(2)}||\theta|| \leq -1$ (avec $p^{(2)} < 0$) ne sont remplies que si $||\theta||$ est très grand. Or ceci ne concorde pas avec le fait que l'on veuille minimiser le terme B défini précédemment.

Dans le cas (b), en revanche, les longueurs de projection $p^{(i)}$ sont plus grandes et autorisent

des θ plus petits pour respecter les conditions de décision. Le terme B sera donc bien mieux minimisé.

Cet exemple montre bien comment la SVM agit : une de ses hypothèses implique la maximisation des longueurs de projection sur le séparateur des éléments les plus proches du séparateur. Autrement dit, le séparateur retenu est celui qui se trouve le plus loin possible de tous les éléments du *training set*. Dans l'exemple de la figure I.2.14, les vecteurs $\mathbf{x}^{(1)}$ et $\mathbf{x}^{(2)}$ constituent alors les deux **vecteurs de support**, d'où le nom de la méthode : plus généralement, ceux-ci correspondent aux vecteurs d'attributs des éléments les plus proches de la surface de décision. La marge est, quant à elle, égale à $p^{(1)}$ et $p^{(2)}$.

La comparaison des surfaces de décision trouvées par la régression logistique et par la SVM pour un jeu de données synthétiques est visible sur la figure I.2.6.

Cas d'un problème non linéaire

L'intérêt principal de la SVM réside dans sa capacité à traiter des problèmes de classification non-linéaires. On part en effet du postulat que plus l'espace des caractéristiques dans lequel on travaille est grand, plus les chances de trouver une surface de décision entre les éléments constituants différentes classes sont élevées. Une première étape est donc d'augmenter la taille de l'espace des caractéristiques qui est fourni en entrée au système automatique en calculant de nouveaux attributs. Ceci se fait via l'utilisation d'une fonction noyau (*kernel* en anglais) qui permet notamment d'effectuer des mesures de similarité entre éléments. Le redimensionnement de l'espace des caractéristiques va en quelque sorte permettre à la SVM de "linéariser" le problème afin de trouver le séparateur maximisant la marge.

Ainsi, le principe consiste à calculer, connaissant $\mathbf{X}$, de nouvelles caractéristiques qui vont dépendre de leur proximité à des repères notés $\mathbf{l}^{(i)}$ (l pour *landmark*). Ces repères $\mathbf{l}^{(i)}$ sont choisis exactement aux mêmes positions que les échantillons du *training set* $\mathbf{x}^{(i)}$, ce qui revient finalement à mesurer la distance de chacun des éléments par rapport au *training set*. Ainsi, connaissant les paires $(\mathbf{x}^{(i)}, y^{(i)})$ avec i allant de 1 à m, on choisit $\mathbf{l}^{(1)} = \mathbf{x}^{(1)}$, $\mathbf{l}^{(2)} = \mathbf{x}^{(2)}$, $\dots \mathbf{l}^{(m)} = \mathbf{x}^{(m)}$. Puis on utilise la fonction noyau qui calcule la proximité (ou similarité) des échantillons du *training set* à un repère $\mathbf{l}^{(j)}$ donné (avec j compris entre 1 et m). On a choisi ici de prendre une fonction noyau gaussienne, d'où la mesure de similarité suivante, pour un élément i du *training set* (i allant de 1 à m) :

$$f_j^{(i)} = \text{similarity}(\mathbf{x}^{(i)}, \mathbf{l}^{(j)}) = e^{\left(-\frac{|\mathbf{x}^{(i)} - \mathbf{l}^{(j)}|^2}{2\sigma^2}\right)} = e^{\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k^{(i)} - l_k^{(j)})^2\right)} \tag{I.2.22}$$

Comme les indices i et j sont compris dans l'intervalle $[1, m]$, on obtient une matrice $\mathbf{F}$ carrée de dimension (m,m) contenant les nouveaux vecteurs d'attributs $\mathbf{f}_j$ de la forme :

$$\mathbf{F} = \begin{bmatrix} f_1^{(1)} & f_1^{(2)} & \dots & f_1^{(m)} \\ f_2^{(1)} & f_2^{(2)} & \dots & f_2^{(m)} \\ \vdots & \vdots & \dots & \vdots \\ f_m^{(1)} & f_m^{(2)} & \dots & f_m^{(m)} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \\ \mathbf{f}_m \end{bmatrix} \tag{I.2.23}$$

On comprend donc que la fonction noyau permet de redéfinir l'espace des caractéristiques de dimension (n,m) en un espace plus grand de dimension (m,m) (voir Fig. I.2.15) et où les

nouveaux attributs donnent une mesure de similarité entre tous les éléments constitutifs du *training set*. Ainsi, pour un élément $\mathbf{x}^{(j)}$ proche d'un repère $\mathbf{l}^{(i)}$, alors $\mathbf{f}_i \simeq 1$. En revanche, pour $\mathbf{x}^{(j)}$ éloigné de $\mathbf{l}^{(i)}$, alors $\mathbf{f}_i \simeq 0$.

On voit aussi que cette mesure de similarité dépend de la largeur de la gaussienne choisie : plus celle-ci est étroite, plus les mesures seront proches de 0 rapidement, tandis que plus elle est large, plus la décroissance dans les mesures sera lente.

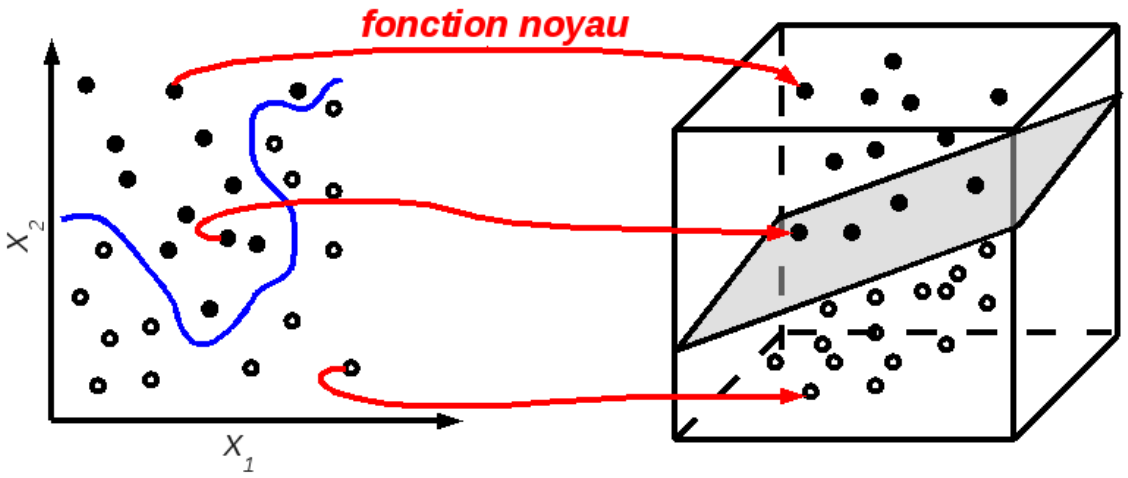


FIG. I.2.15: Transformation de l'espace des caractéristiques en un espace de plus grande dimension dans lequel il sera plus facile de séparer les données linéairement.

Enfin, on donne en entrée à la SVM la nouvelle matrice d'attributs $\mathbf{F}$. Le problème d'optimisation reste le même. On cherche toujours à minimiser la fonction suivante (en remplaçant $\mathbf{x}^{(i)}$ par $\mathbf{f}^{(i)}$) :

$$J_{SVM}(\theta) = \frac{C}{m}\sum_{i=1}^{m}\left[-y^{(i)}\log(h_\theta(\mathbf{f}^{(i)})) - (1-y^{(i)})\log(1-h_\theta(\mathbf{f}^{(i)}))\right] + \frac{1}{2m}\sum_{j=1}^{m}\theta_j^2 \quad \text{(I.2.24)}$$

Il est important de rappeler que, comme pour la régression logistique, des problèmes de sous- ou sur-apprentissage peuvent survenir (voir §I.2.2.1) car il n'est pas forcément possible de trouver un séparateur linéaire séparant parfaitement les données, même en redimensionnant l'espace d'entrée. Le séparateur trouvé sera alors celui qui minimise le nombre d'éléments mal classés. Ainsi, la constante C, qui a un comportement inversement proportionnel au coefficient de régularisation λ, sert à contrôler le *trade-off* entre l'adéquation aux données et la complexité du problème. De grandes valeurs de C peuvent mener à un sur-apprentissage ; alors que de petites valeurs seront révélatrices d'un sous-apprentissage.

Le choix de σ est aussi un facteur déterminant. En effet, il agit sur la largeur de la gaussienne. Lorsque σ^2 est grand, alors $\mathbf{f}_i$ sera modifié plus doucement quand on s'éloigne du repère $\mathbf{l}^{(i)}$: on risque d'avoir une faible variance, mais une erreur importante (sous-apprentissage). Au contraire, si σ^2 est petit, les caractéristiques $\mathbf{f}_i$ vont varier plus lentement : on risque d'avoir une erreur minime, mais une variance élevée (sur-apprentissage).

I.2.2.3 Cas multiclasse (> 2)

Lorsque le nombre de classes sort du cadre binaire, on utilise une approche du type "un contre tous". Cette approche est utilisable avec tout classificateur binaire. L'approche « un contre tous » est une approche en deux étapes. La première étape consiste à calculer une fonction hypothèse $h_\theta^{(k)}(\mathbf{X})$ pour chacune des classes k afin de connaître la probabilité que $y = k$. Autrement dit, on décompose le problème en une multitude de classifications binaires. Lorsque toutes les fonctions hypothèses ont été déterminées, il ne reste plus qu'à calculer les probabilités d'appartenance d'un élément donné pour chacune des classes. La classe finale sera donc celle ayant la plus forte probabilité (voir la figure I.2.16).

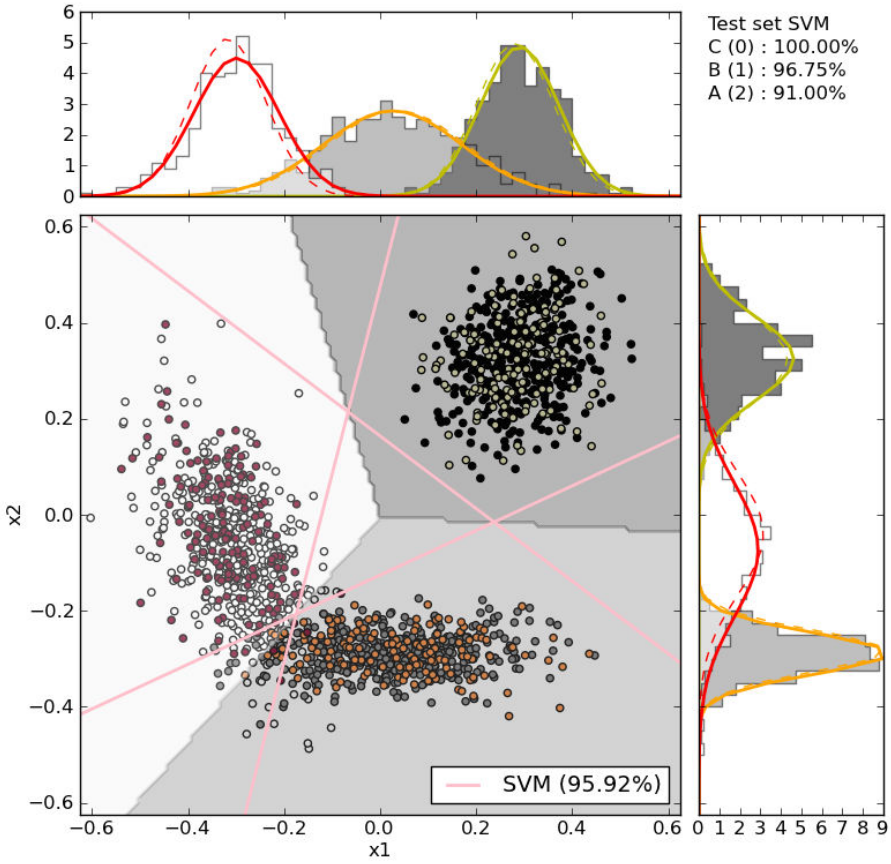


FIG. I.2.16: Exemple de classification pour un jeu de données synthétique avec trois classes se séparant bien et deux attributs x_1 et x_2. Les droites roses correspondent aux séparateurs déterminés par la SVM. Il y a trois classes, donc trois séparateurs : classe 0 contre le reste ; classe 1 contre le reste et classe 2 contre le reste. La carte d'appartenance à chaque classe, déterminée après calcul des probabilités, est donnée en fond en échelle de gris. Les histogrammes et densités de probabilité de chacun des 2 attributs et pour les 3 classes sont aussi représentés de part et d'autre de la figure.

Notons également que dans le cas multiclasse, le choix du seuil évoqué en §I.2.2.1 revêt moins d'importance, puisqu'au final la classe attribuée à un élément correspond à la classe

ayant la plus forte probabilité.

I.2.2.4 Présentation des résultats

Une fois que la loi de discrimination a été déterminée par l'une ou l'autre des deux méthodes, il ne reste plus qu'à prédire les valeurs du vecteur $\mathbf{y}$. Les résultats sont ensuite présentés sous forme de matrice de confusion (Fig. I.2.17). Une **matrice de confusion** décompose les résultats en observations et en prédictions : pour chacune des classes, on peut connaître la répartition en classes manuelles des classes prédites, et, réciproquement, la répartition en classes automatiques des classes initiales. Une lecture par ligne donnera donc, pour une classe « vraie » donnée, la répartition des éléments de cette classe dans toutes les classes existantes à l'issue de la classification. Une lecture par colonne donnera donc, pour une classe prédite donnée, la répartition par classe d'origine des éléments classés automatiquement. La matrice de confusion idéale est une matrice diagonale. Elle signifierait que tous les éléments ont été correctement classés.

Dans les exemples qui suivront dans les deux prochains chapitres de cette thèse, on a divisé chaque ligne i de la matrice de confusion par le nombre d'éléments de la classe i afin de lire plus facilement les résultats sous forme de pourcentages en donnant directement les valeurs du rappel (§I.2.2.1) sur la diagonale, c'est-à-dire les taux de classification correcte pour chaque classe.

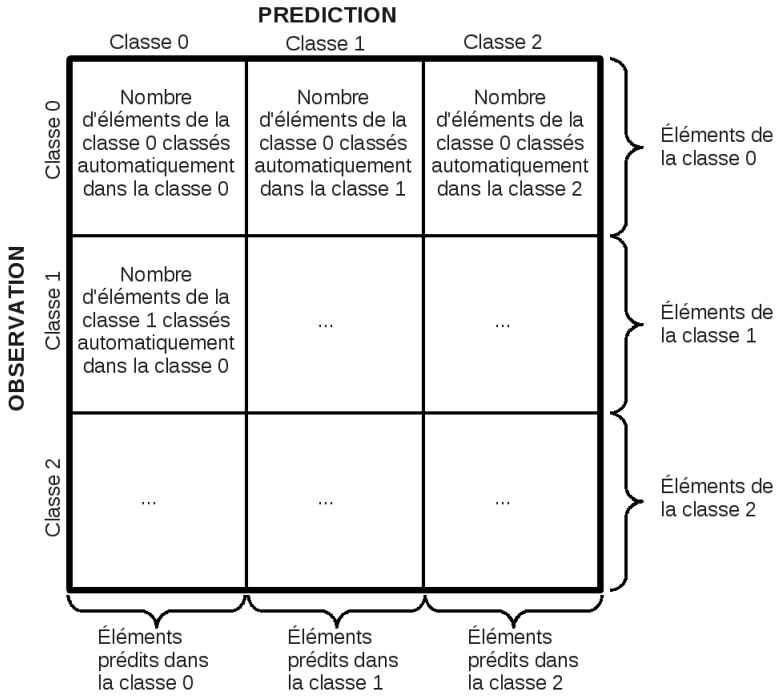


Fig. I.2.17: Matrice de confusion.

I.2.3 Description d'une méthode d'apprentissage non supervisé : les K-moyennes

La méthode des K-moyennes (*K-means* en anglais) est la méthode la plus courante d'apprentissage non supervisé : l'algorithme doit trouver une structure dans un jeu de données dont la classification est inconnue *a priori*. K désigne ainsi le nombre de clusters dans lesquels on veut regrouper les données.

La méthode des K-moyennes est itérative. Elle nécessite au départ le tirage aléatoire de K points du jeu de données (Fig. I.2.18a). Ceux-ci sont considérés comme les K centroïdes des clusters que l'on souhaite former et sont notés $\mu_{\mathbf{k}}$ où $k = 1 \dots K$. Chaque $\mu_{\mathbf{k}}$ est un vecteur de longueur n (où n correspond au nombre d'attributs). L'algorithme se décompose ensuite en deux étapes principales :

- une étape au cours de laquelle chaque élément du jeu de données est associé à un cluster. Les attributions se font par simple minimisation d'une fonction coût qui calcule la distance de chaque point à chacun des K centroïdes (I.2.25). Un point donné appartiendra donc au cluster qui a le centroïde le plus proche (Fig. I.2.18b) ;

- une étape au cours de laquelle les centroïdes sont recalculés à partir des assignations effectuées dans l'étape précédente, par simple calcul de la moyenne des éléments appartenant à chacun des K groupes trouvés (Fig. I.2.18c). Puis l'algorithme passe de nouveau à l'étape 1.

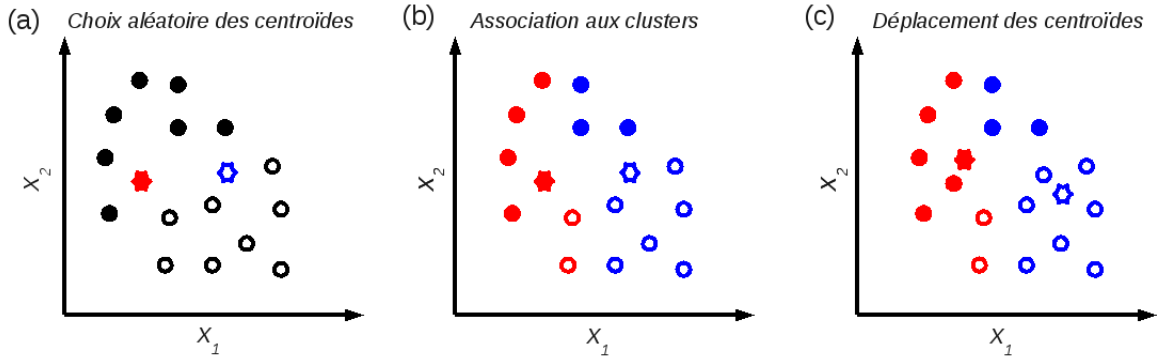


Fig. I.2.18: Illustration des étapes de l'algorithme des K-moyennes. Ici, K vaut 2. Pour aider à mieux comprendre le processus, on a représenté par des ronds pleins et des ronds évidés les deux classes, mais il ne faut pas oublier que dans un cas réel, la classification *a priori* n'est pas connue... (a) Initialisation des centroïdes par tirage aléatoire de 2 échantillons du jeu de données (étoiles rouge et bleue). (b) Chaque échantillon est associé à l'un des clusters en fonction de sa distance avec les centroïdes. (c) Les centroïdes sont recalculés d'après les assignations précédentes, puis on repasse à l'étape (b).

Soient $c^{(i)}$ le numéro du cluster associé à un élément $\mathbf{x}^{(i)}$ ($c^{(i)}$ peut donc prendre des valeurs de 1 à K) ; $\mu_{\mathbf{k}}$, le centroïde du cluster k ; et $\mu_{c^{(i)}}$ le centroïde du cluster auquel l'élément $\mathbf{x}^{(i)}$ est associé, alors la fonction coût que l'on cherche à minimiser dans l'algorithme des K-moyennes

est la suivante :

$$J(c^{(1)},\dots,c^{(m)},\mu_{\mathbf{1}},\dots,\mu_{\mathbf{k}}) = \frac{1}{m}\sum_{i=1}^{m}||\mathbf{x}^{(i)} - \mu_{c^{(i)}}||^2 \tag{I.2.25}$$

En résumé, dans l'étape ①, la fonction coût est minimisée selon $c^{(i)}$, en gardant $(\mu_{\mathbf{1}},\dots,\mu_{\mathbf{k}})$ fixe, alors que dans l'étape ②, elle est minimisée selon $\mu_{\mathbf{k}}$, avec $(c^{(1)},\dots,c^{(m)})$ fixe.

On peut trouver deux désavantages à cette méthode : d'abord l'initialisation aléatoire des centroïdes (si les deux points choisis sont suffisamment éloignés dès le départ, il n'y aura pas de problème ; mais si ceux-ci sont proches, il est possible de tomber dans des minima locaux), mais aussi le choix du nombre de clusters K, qui n'est pas évident si le jeu de données que l'on a est totalement inconnu. On peut pallier à ces deux inconvénients en effectuant un certain nombre de tirages des centroïdes initiaux et en testant également différentes valeurs de K. Dans les deux cas, on gardera la configuration finale qui minimise le coût.

I.2.4 Extraction des attributs sismiques

I.2.4.1 Introduction

Comme on l'avait expliqué dans l'introduction de cette partie (I.2.1.1), il nous faut définir un certain nombre de **caractéristiques** qui décrivent les données afin de les donner en entrée au système de classification, quel qu'il soit.

Au cours des dernières années, l'essor des méthodes de classification appliquées plus particulièrement à la sismicité volcanique a engendré le calcul d'attributs diversifiés selon les auteurs. En 1990, les problèmes de discrimination des explosions nucléaires [Dowla et al., 1990, Pulli, 1990] nécessitaient l'identification de certaines phases du signal. La problématique de la sismicité volcanique étant légèrement différente (et c'est celle qui nous intéresse ici), d'autres auteurs ont tenté de calculer de nouveaux attributs. Dans un premier temps, Falsaperla et al. [1996] avait décidé de travailler directement sur les signaux bruts en ne gardant qu'un nombre d'échantillons limité pour chaque événement, indépendamment de la longueur du signal. Il avait trouvé que les résultats de la classification n'étaient pas concluants et les avait imputés à une trop forte complexité. Dans la même optique, l'utilisation de la fonction d'autocorrélation donnait des résultats corrects, tandis que celle de l'enveloppe du signal ne s'avérait pas concluante. L'utilisation des attributs spectraux (calcul du spectrogramme) permettait finalement d'atteindre de très bons résultats, prouvant qu'ils contiennent de l'information utile. Depuis cette première étude, d'autres manières d'utiliser et de traiter l'information (spectrale principalement) contenue dans les sismogrammes ont été développées.

Dans cette section, on s'attachera à présenter chacun des attributs sismiques extrait. La plupart d'entre-eux sont issus de la littérature et ont déjà été utilisés dans d'autres problèmes de classification. Les exemples d'illustration sont tirés du jeu de données du volcan du Piton de la Fournaise (voir la partie II y étant consacrée pour plus de détails). Celui-ci est constitué uniquement de deux classes (d'où des figures plus claires) : les événements de type volcano-tectonique (VT) et les éboulements (EB). Sur les figures représentant les densités de probabilité (PDF - *Probability Density Function* en anglais), les courbes bleues sont celles des VT et les rouges, celles des EB. On rappelle que celles-ci ont été calculées en utilisant une méthode

d'estimation avec un noyau gaussien. Les courbes tiretées correspondent aux PDFs calculées à partir des données du *training set* ; les pleines à celles du *test set*. On précise que les données du *training set* utilisé ne correspondent pas à un tirage aléatoire de l'ensemble des données.

Un événement de chaque type a aussi été choisi pour donner des exemples d'illustration de certains attributs. Ce sont toujours les mêmes tout au long de cette section.

Dans certains cas, la discrimination entre les différentes données n'est pas évidente (les PDFs se recouvrent entre elles) et on peut être amené à les normaliser par le logarithme (voir figure I.2.22). On le précisera lorsque ce sera le cas.

I.2.4.2 Attributs calculés en domaine temporel

Les sismogrammes contiennent l'information en temps et en amplitude (forme d'onde) d'un événement sismique. Les premières informations qu'un opérateur humain remarque en regardant les enregistrements concernent alors la durée du signal et sa forme (impulsivité, valeurs et distribution des amplitudes...- voir la figure I.2.19). L'objectif ici est donc de réunir un certain nombre de caractéristiques qui rendent compte de ces observations "directes".

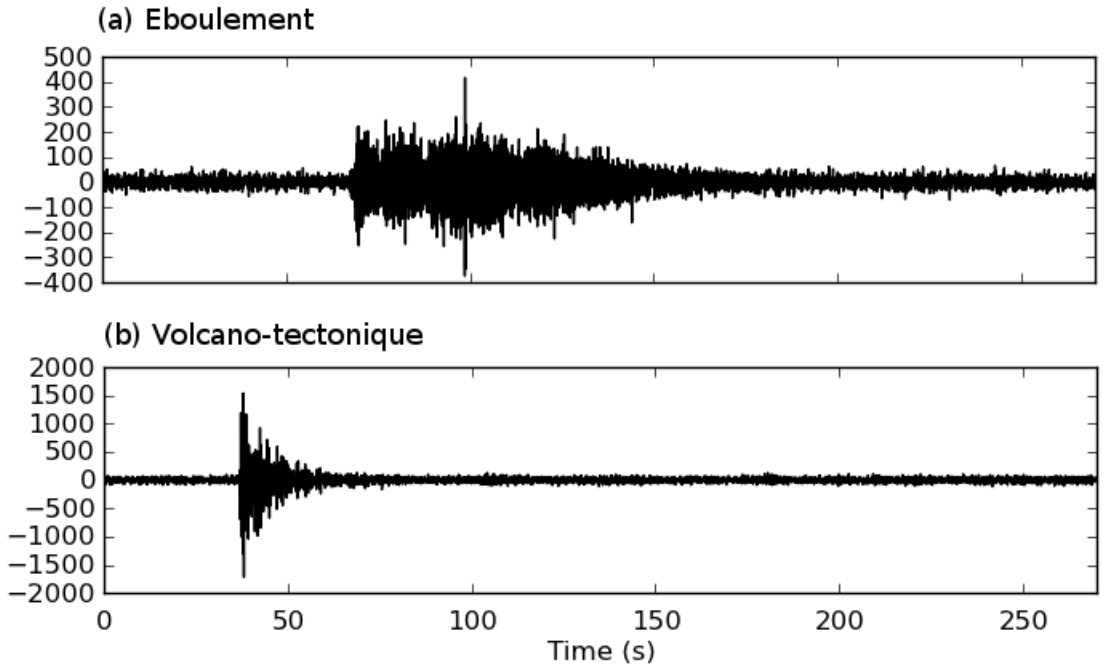


FIG. I.2.19: Formes d'ondes d'un éboulement (a) et d'un événement volcano-tectonique (b) enregistrées sur le Piton de la Fournaise. On peut déjà noter des différences dans l'allure générale des 2 sismogrammes (durée du signal, impulsivité, amplitudes...)

Durée

Le premier attribut qu'il paraît naturel de calculer est la durée du signal. Celle-ci peut constituer un indicateur important de la nature d'un événement donné. Dans le cas de la discrimination des EB et des VT sur le Piton de la Fournaise, une durée relativement courte (quelques secondes) est caractéristique d'un VT, tandis qu'une durée de plusieurs dizaines de secondes est caractéristique d'un EB (voir Fig. I.2.21a).

Si la mesure manuelle de la durée d'un signal avec une bonne précision n'est pas des plus évidentes (en fonction du bruit présent dans les données notamment), elle l'est encore moins en automatique. La détermination de la durée du signal passe par deux étapes : d'abord trouver le début du signal, puis la fin. Les durées des enregistrements pré-découpés dont on dispose doivent être suffisamment grandes pour permettre une bonne mesure.

Pour déterminer le début du signal, on a utilisé deux techniques en fonction du jeu de données avec lequel on travaillait. Dans le cas du Kawah Ijen, on disposait des catalogues de sismicité, donc des temps origine des événements, et on a découpé les signaux dans des fenêtres de taille fixe autour du temps origine. Puis on a utilisé l'outil du gradient de kurtosis discuté dans la première partie de ce chapitre (§I.1.2) pour mettre en évidence la première arrivée du signal. Pour s'assurer de la meilleure précision possible, on considère que le temps du maximum atteint dans une une fenêtre réduite autour du temps origine (qui est connu) correspond au début du signal.

Dans le cas du Piton de la Fournaise, comme on ne disposait que des signaux pré-découpés, sans autre information préalable, on a préféré utiliser une autre approche pour déterminer le début du signal. On a ainsi utilisé l'information contenue dans le spectrogramme des signaux et, plus particulièrement, la courbe obtenue par la sommation de toutes les fréquences en un échantillon de temps donné (les détails sur les calculs des spectrogrammes seront abordés plus loin au cours de ce chapitre). La fonction caractéristique résultant de la sommation donne une idée de la distribution de l'énergie du signal au cours du temps : elle contient donc bien l'information sur le début et la fin du signal. La méthode de pointé du début du signal qui suit est reprise de Baillard et al. [2014], Hibert et al. [2014] et consiste à (voir Fig. I.2.20) :

1. calculer la somme cumulée (courbe jaune) de la fonction caractéristique (*frequency stack* - courbe noire) ;
2. calculer la droite de régression (en bleu) entre les deux valeurs extrémales de la somme cumulée ;
3. soustraire cette droite à la somme cumulée (courbe rouge) ;
4. puis rechercher le minimum absolu.

Enfin, pour déterminer la fin du signal, quel que soit le jeu de données, on utilise cette même fonction caractéristique issue du spectrogramme. On définit simplement un niveau d'amplitude en-dessous duquel on estime qu'on est retourné dans le bruit. Le niveau est défini automatiquement en calculant la moyenne de la fonction caractéristique prise dans les 0.2 premières secondes du signal disponible.

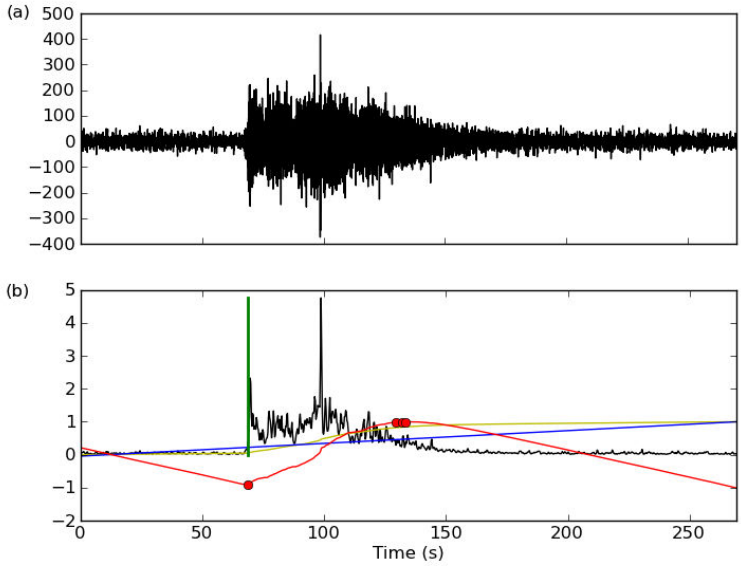


FIG. I.2.20: Détermination du début du signal d'un éboulement à partir de la courbe résultant de la sommation des fréquences de son spectrogramme. (a) Signal brut. (b) Fonction caractéristique (en noir) et sa somme cumulée (en jaune). La droite de régression (en bleu) associée est soustraite à la somme cumulée : on obtient ainsi la courbe rouge dont on recherche les minima locaux (points rouges). Le minimu absolu correspond au début du signal (trait vertical vert).

Conformément à ce qui est observé sur la figure I.2.19, les PDFs calculées pour l'attribut de durée montrent que les EB durent plus longtemps que les VT (voir FIG. I.2.21a). Une autre remarque que l'on peut faire concerne la différence de l'intervalle des valeurs prises par les EB dans le *training set* (courbe rouge tiretée) et le *test set* (courbe rouge continue)... Cela risque de poser problème lors de la généralisation, puisque le système aura été entraîné sur des données non représentatives.

Indicateurs de la forme d'onde

Connaissant la durée du signal, il est possible de capturer d'autres attributs qui vont rendre compte de la forme du signal. Hibert et al. [2014] utilise le rapport des phases de croissance et de décroissance de l'enveloppe (noté AsDec) pour rendre compte de la forme plus ou moins émergente du signal :

$$\text{AsDec} = \frac{t_{max} - t_i}{t_f - t_{max}} \tag{I.2.26}$$

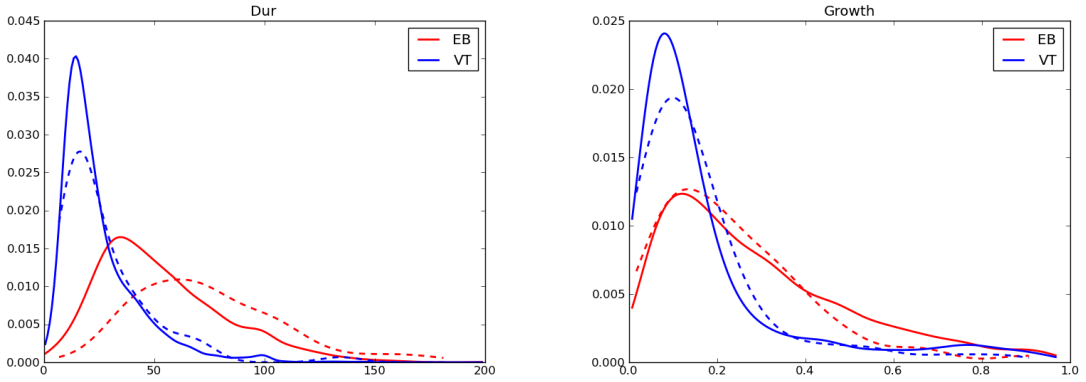
où t_{max} est le temps auquel l'enveloppe lissée atteint son maximum.

Dans le même ordre d'idée, on calcule également le rapport de la phase de croissance de l'enveloppe sur la durée totale du signal :

$$\text{Growth} = \frac{t_{max} - t_i}{t_f - t_i}. \tag{I.2.27}$$

On s'attend à ce que les valeurs de cet attribut soient plus petites pour les VT que pour les EB, puisque les VT sont très impulsifs et atteignent rapidement de fortes amplitudes. Cependant, les PDFs (voir Fig. I.2.21b) ne montrent pas une grande différence entre EB et VT.

Il faut également souligner que, comme indiqué par leur définition, Growth et AsDec sont corrélés entre-eux.



(a) Durée (Dur)

(b) Phase de croissance (Growth)

Fig. I.2.21: Densités de probabilités pour la durée et la phase de croissance. Trait continu : *test set* ; trait tireté : *training set*.

Un autre critère renseignant sur la forme d'onde et utilisé par Hibert et al. [2014] est le rapport du maximum sur la moyenne de l'enveloppe. Il donne l'importance relative du maximum par rapport au reste du signal. On utilise pour cela une enveloppe lissée sur des fenêtres de 0.5 s.

Les VT se caractérisent par la présence d'un pic d'amplitude visible, ce qui n'est pas le cas des événements plus "lents" comme les EB : la valeur du rapport calculé doit donc être plus élevée pour les premiers cités (Fig. I.2.22). Les valeurs de cet attribut ont été normalisées par le logarithme pour une meilleure séparation des deux classes (Fig. I.2.22b).

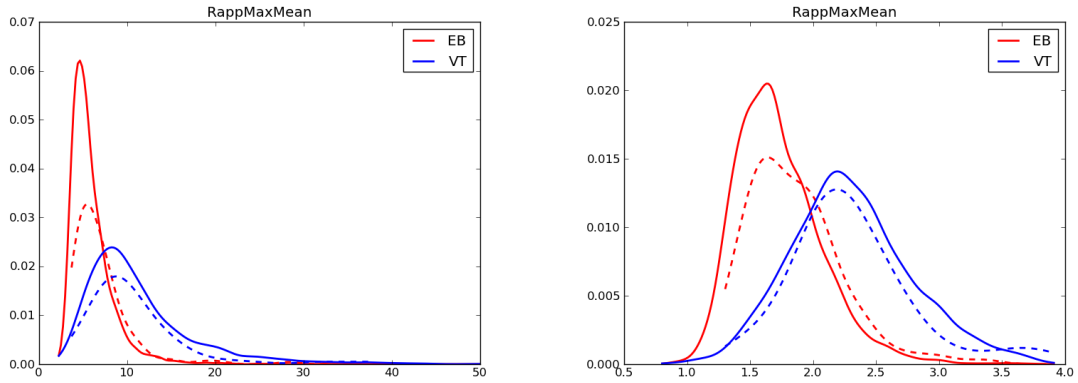


FIG. I.2.22: Densités de probabilités pour le rapport du maximum sur la moyenne de l'enveloppe non lissée, sans (a) et avec (b) la normalisation logarithmique. Trait continu : *test set* ; trait tireté : *training set*.

Distribution des amplitudes

Comme on l'avait déjà expliqué dans la section I.1.2 au début de ce manuscrit, les moments d'ordre supérieur d'une distribution constituent un bon indicateur du comportement de la variable. Si les moments d'ordre 1 et 2 (c'est-à-dire moyenne et variance) ne sont pas, *a priori*, de bons attributs puisqu'il sont directement liés à la taille de l'événement sismique, les moments d'ordre 3 et 4 (respectivement, l'asymétrie et le kurtosis) sont indépendants de la magnitude et forment ainsi de meilleurs attributs. On rappelle ici l'expression du moment d'ordre r d'une distribution $\mathbf{x}$ donnée constituée de n valeurs :

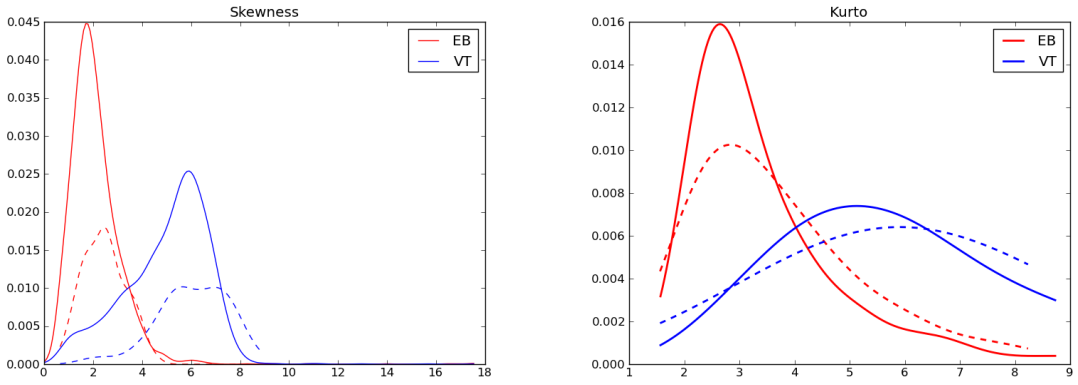
$$M_r(\mathbf{x}) = M_r(x_1 \dots x_n) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{x_i - \mu}{\sigma} \right]^r \tag{I.2.28}$$

où μ et σ sont respectivement la moyenne et la variance de la distribution.

Comme son nom l'indique, l'asymétrie (ou *skewness* en anglais) donne une mesure de l'asymétrie de la distribution. Une valeur positive indique que la distribution des amplitudes est plus longue à droite (fortes amplitudes) qu'à gauche.

Le kurtosis, quant à lui, mesure l'impulsivité du signal. Plus précisément, il mesure la pointicité (valeurs positives) ou l'aplatissement (valeurs négatives) de la distribution des amplitudes. De plus, plus les valeurs du kurtosis sont grandes, plus elles indiquent que le signal est impulsif (c'est-à-dire que la transition avec ce qui précède est « brutale »).

Les signaux des EB sont longs à émerger et relativement plats par rapport à ceux des VT qui montent vite en amplitude. Intuitivement, on s'attend donc à ce que les valeurs de l'asymétrie et du kurtosis soient plus élevées pour les VT que pour les EB. Ceci est confirmé par les PDFs calculées à partir des enveloppes lissées des signaux et visibles sur la figure I.2.23.



(a) Asymétrie (b) Kurtosis

FIG. I.2.23: Densités de probabilités pour l'asymétrie (a) et le kurtosis (b). Trait continu : *test set* ; trait tireté : *training set*.

I.2.4.3 Attributs renseignant sur le contenu fréquentiel du signal

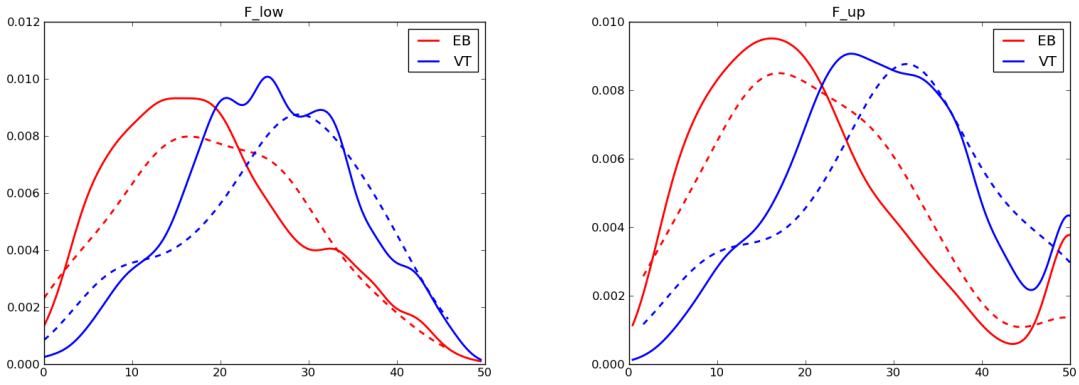
Les attributs décrits jusqu'ici donnent les premiers renseignements « basiques » sur les signaux et sont globalement représentatifs du type d'observations qu'un opérateur humain pourrait faire en visualisant les sismogrammes. De plus, ils sont faciles à calculer.

Cependant, ces caractéristiques ne renseignent pas sur les périodicités présentes dans le signal, qui ne sont d'ailleurs pas souvent facilement observables à l'oeil nu. Un passage en domaine fréquentiel est alors requis.

On peut souligner que le domaine fréquentiel permet le calcul d'une grande variabilité d'attributs. C'est d'ailleurs dans ce domaine que de nombreux auteurs ont concentré leurs investigations en recherchant soit de nouveaux attributs, soit de nouvelles méthodes de calcul.

Bande de filtrage maximisant le kurtosis

On avait expliqué en première partie (§I.1.2) la méthode du kurtogramme développée par Antoni [2007] et qui permet de récupérer les fréquences extrémales de la bande de filtrage maximisant le kurtosis. On peut légitimement supposer que ces fréquences ne seront pas les mêmes selon le type d'événement. Les PDFs de la figure I.2.24 montrent que le filtrage adapté est globalement plus basse fréquence pour les EB que pour les VT.



(a) Borne inférieure

(b) Borne supérieure

FIG. I.2.24: Densités de probabilités pour la fréquence inférieure (a) et la fréquence supérieure (b) de la bande de filtrage obtenue par analyse du kurtogramme. Trait continu : *test set* ; trait tireté : *training set*.

Analyse spectrale

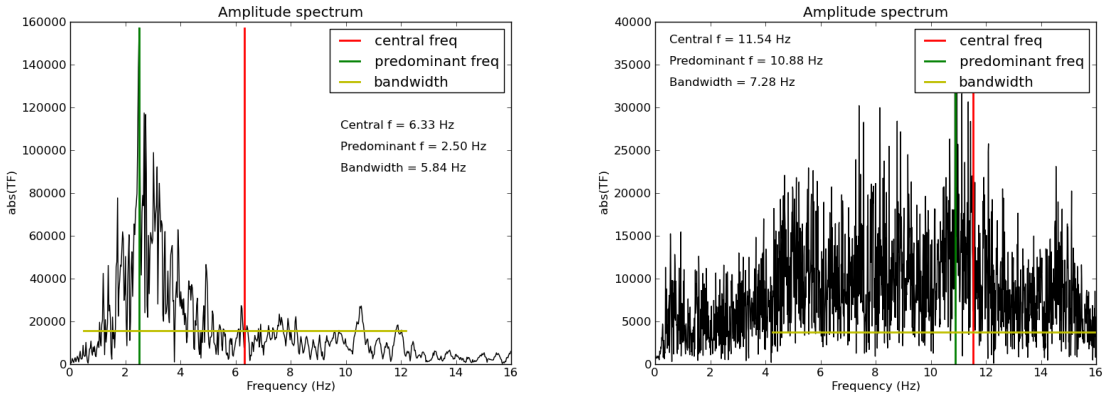
Le spectre d'un signal contient l'information sur les différentes fréquences qui sont excitées dans un signal. Il est calculé grâce à la transformée de Fourier (notée TF) :

$$TF[x(t)] = F(\nu) = \int_{-\infty}^{+\infty} x(t)e^{-2\pi i\nu t}dt \tag{I.2.29}$$

où $x(t)$ est le signal et ν est la fréquence. (En pratique, la transformée de Fourier calculée est discrète).

Deux exemples de spectres calculés pour un VT et un EB sont montrés sur la figure I.2.25 et permettent déjà de se rendre compte de leurs différences. Les amplitudes du spectre renseignent sur la prépondérance respective de chacune des fréquences présentes dans le signal. Ainsi, on observe dans l'exemple que le spectre du VT présente un pic relativement bien distinct, tandis que le spectre de l'EB est beaucoup plus étendue et occupe une large gamme de fréquences.

En se basant sur ce simple constat, on peut d'ores et déjà penser à un premier attribut qui rendra compte de la forme du spectre, comme on l'a fait dans la section précédente avec la forme d'onde, en calculant tout simplement le rapport du maximum sur la moyenne de l'enveloppe non lissée du spectre d'amplitude. D'après les observations tirées de la figure I.2.25, ce rapport devrait prendre des valeurs plus élevées pour les VT que pour les EB (FIG. I.2.26a). Or on observe l'inverse : globalement, ce sont les EB qui ont un rapport maximum sur moyenne plus élevé. Ceci semble indiquer que cet attribut ne sera pas fortement discriminant.



(a) VT (b) EB

FIG. I.2.25: Spectres d'amplitude d'un événement de type volcano-tectonique (gauche) et d'un éboulement (droite). La fenêtre de fréquence a été volontairement tronquée à 16 Hz.

L'analyse du spectre d'amplitude permet de retirer un certain nombre de caractéristiques intrinsèques au signal, telles que :

- la fréquence prédominante, c'est-à-dire la fréquence associée au maximum d'amplitude du spectre :

$$f_{dom} = \max_{\nu} |F(\nu)| \tag{I.2.30}$$

- la fréquence centrale, qui indique la fréquence moyenne à laquelle l'énergie se concentre :

$$f_c = \frac{\int \nu |F(\nu)| d\nu}{\int |F(\nu)| d\nu} \tag{I.2.31}$$

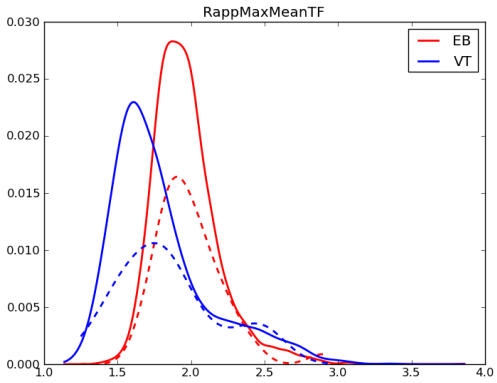
- la largeur de bande qui mesure l'étendue des fréquences contenues dans le spectre :

$$B_w^2 = \frac{\int (\nu - f_c)^2 |F(\nu)| d\nu}{\int |F(\nu)| d\nu} \tag{I.2.32}$$

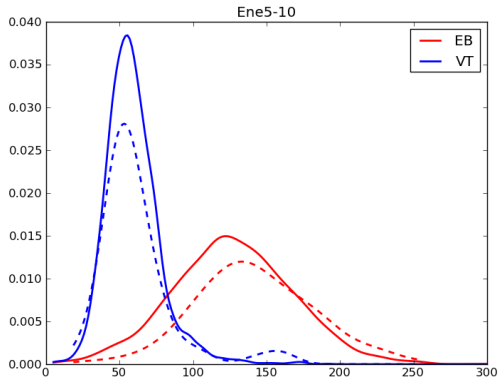
- l'énergie du signal dans une bande de fréquence (f_1, f_2) donnée :

$$E = \int_{f_1}^{f_2} |F(\nu)|^2 d\nu \tag{I.2.33}$$

Ces attributs sont régulièrement utilisés dans les autres études de discrimination [Beyreuther and Wassermann, 2008, Hammer et al., 2012], mais ne paraissent pas particulièrement discriminants au regard des PDFs calculées avec les données du Piton de la Fournaise (figures I.2.26b, I.2.27), sauf pour l'attribut énergie dans la bande de fréquence 5-10 Hz.

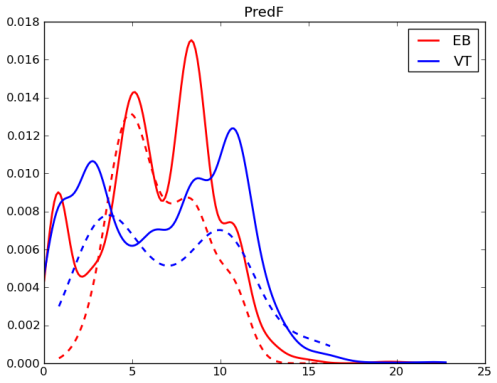


(a) Rapport max/moy du spectre

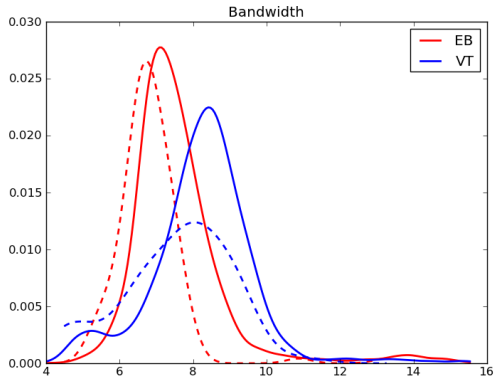


(b) Energie du signal entre 5 et 10 Hz

FIG. I.2.26: Densités de probabilités pour le rapport du maximum sur la moyenne de l'enveloppe non lissée de la TF avec normalisation logarithmique (a) et l'énergie comprise dans la bande 5-10 Hz (b). Trait continu : *test set* ; trait tireté : *training set*.



(a) Fréquence prédominante



(b) Largeur de bande

FIG. I.2.27: Densités de probabilités pour la fréquence prédominante (a) et la largeur de bande (b). Trait continu : *test set* ; trait tireté : *training set*.

Analyse temps-fréquence : le spectrogramme

Si l'analyse spectrale permet de nous renseigner sur le contenu fréquentiel du signal, elle n'est pas complète, puisqu'elle fait abstraction de la dimension temporelle. Autrement dit, on sait quelles fréquences ont été émises, mais pas quand, d'où la nécessité d'une analyse temps-fréquence via le calcul du spectrogramme. Ce dernier donne la représentation du contenu fréquentiel du signal au cours du temps. Ici, les spectrogrammes ont été calculés sur des fenêtres glissantes de 128 échantillons pour un recouvrement de 115 échantillons (voir FIG. I.2.29).

Les informations contenues dans le spectrogramme sont nombreuses, et les manières de

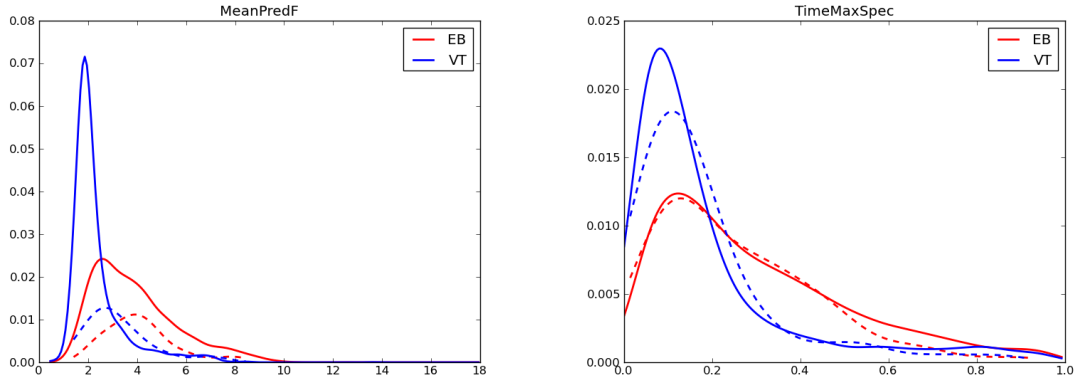
les exploiter sont multiples. Ici, on a d'abord calculé les sommes de toutes les valeurs du spectrogramme en temps et en fréquence afin d'obtenir deux courbes : l'une (le *time stack*) qui correspond grossièrement à une version lissée du spectre d'amplitude du signal ; l'autre (le *frequency stack*) qui reflète la distribution de l'énergie contenue dans le signal au cours du temps (voir Fig. I.2.29f). On a déjà parlé de l'utilité de ce dernier pour la détermination des temps de début et de fin du signal.

A partir du *time stack*, on peut ré-extraire les attributs précédemment calculés sur le spectre d'amplitude brut du signal, à savoir : la fréquence prédominante et largeur de bande. Pour cette dernière, on a simplement mesuré la largeur du pic en fixant un seuil égal à 0.1 fois l'amplitude maximale du pic principal.

Connaître l'évolution temporelle du contenu fréquentiel du signal permet aussi de calculer la moyenne de certaines caractéristiques, comme la fréquence prédominante (figure I.2.29c). Pour capturer la tendance d'évolution de la fréquence prédominante au cours du temps, on la découpe en 10 intervalles réguliers sur toute la durée du signal (de manière à pouvoir comparer les valeurs indépendamment de la longueur du signal), puis on calcule la moyenne. Ceci permet d'obtenir 10 valeurs, donc 10 attributs.

Le spectrogramme du VT (figure I.2.29b) montre un *patch* d'énergie de courte durée tandis que l'énergie de l'EB est plus étalée dans le temps. Le contenu fréquentiel diffère également, avec des fréquences prédominantes moyennes plus élevées pour l'EB (entre 2 et 10 Hz) que pour le VT (vers 2 Hz). Conformément aux spectres de la figure I.2.25, les *time stacks* montrent également que le contenu fréquentiel du VT est moins étendu que celui de l'EB.

Enfin, on a aussi choisi comme attribut le temps du maximum du spectrogramme qui indique le temps relatif pour lequel l'énergie du signal est maximale (voir Fig. I.2.28). Cet attribut est quasiment identique à la phase de croissance (Growth) définie en début de section. Sur les exemples d'illustration de la figure I.2.29, on voit que l'énergie du VT est très concentrée, d'abord parce que le signal est de courte durée, mais aussi parce que le maximum d'énergie (symbolisé par le trait jaune vertical) est atteint dès les premières secondes du signal.



(a) Fréquence prédominante moyenne

(b) Temps du maximum du spectrogramme

Fig. I.2.28: Densités de probabilités pour la fréquence prédominante moyenne (a) et le temps du maximum du spectrogramme (b). Trait continu : *test set* ; trait tireté : *training set*.

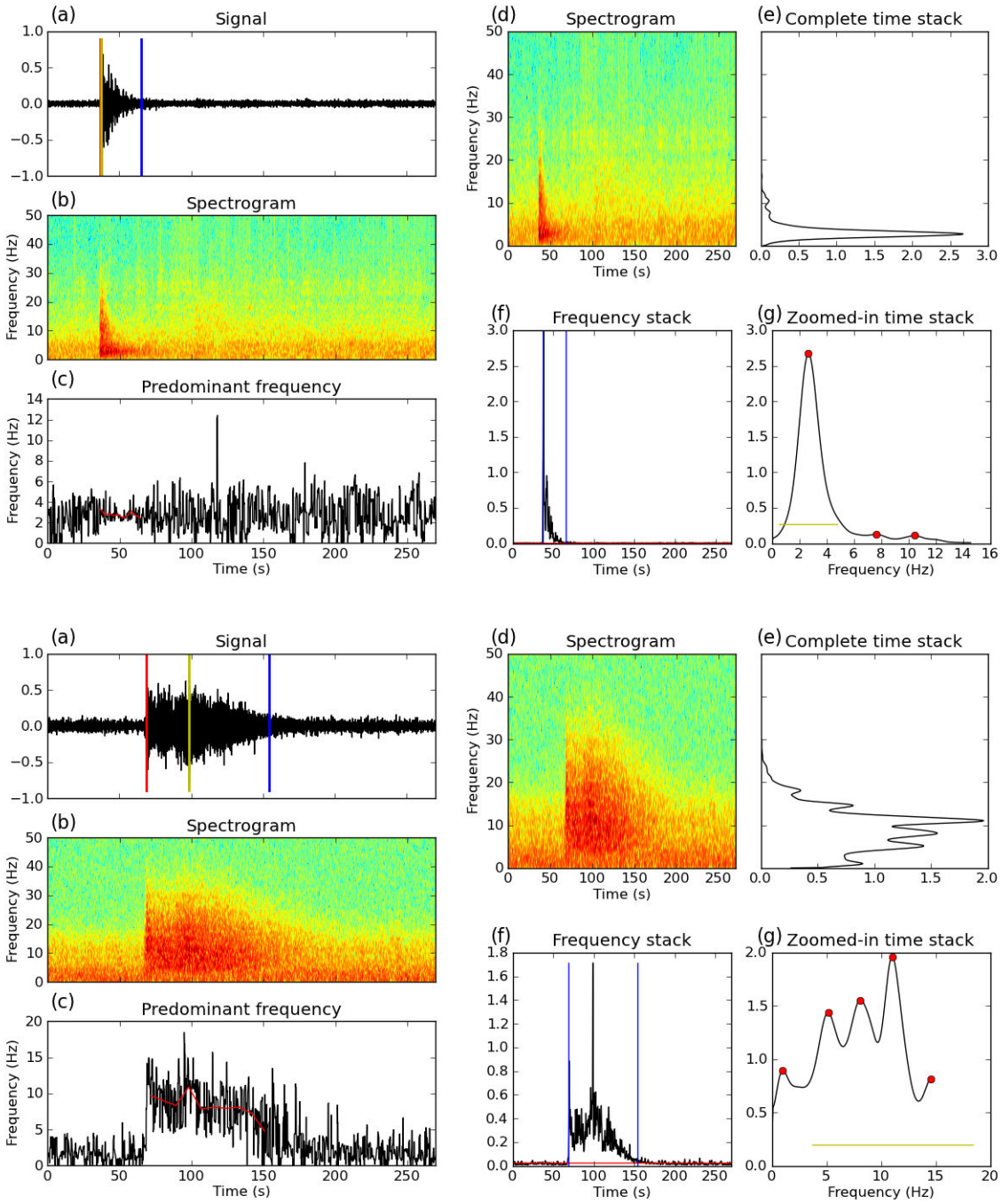


FIG. I.2.29: Spectrogrammes d'un VT (en haut) et d'un EB (en bas). (a) Signaux bruts. Le trait rouge vertical symbolise le début du signal ; le trait jaune, le temps auquel le maximum du spectrogramme est atteint ; le trait bleu, la fin de l'événement. (b) et (d) Spectrogramme. (c) Evolution de la fréquence prédominante au cours du temps. Chaque segment rouge correspond à la moyenne de la fréquence prédominante calculée dans 10 intervalles de même taille couvrant la totalité de la durée du signal. (e) *Time stack* du spectrogramme. (f) *Frequency stack* du spectrogramme. Les 2 traits bleus verticaux symbolisent le début et la fin du signal. Le trait rouge horizontal correspond au niveau moyen pris dans les premières secondes du *stack*. (g) Zoom sur la partie intéressante du *time stack*. Le trait jaune horizontal mesure la largeur du pic. Il est centré sur le maximum du pic.

Compression de l'information contenue dans le spectrogramme dans des tables de hachage

Comme on vient de le voir, l'information contenue dans le spectrogramme est très dense et l'extraction des quelques attributs est loin d'être exhaustive. Garder l'ensemble de l'information est aussi problématique, d'abord en termes de volume (le nombre de caractéristiques serait démesuré, rallongeant ainsi les temps de calcul), mais aussi et surtout en termes de complexité. Falsaperla et al. [1996] l'avait souligné en utilisant les formes d'onde brutes, la complexité de l'information ne permet pas d'optimiser l'apprentissage automatique. Il est fort probable qu'il en soit de même si l'on utilisait toute l'information du spectrogramme.

L'idée ici est de mettre en œuvre une technique, le *fingerprinting*, qui consiste à compresser l'information contenue dans une image en ne gardant que ses valeurs les plus significatives, c'est-à-dire son empreinte (*fingerprint* en anglais). Cette technique a été utilisée récemment en sismologie pour la recherche d'événements de formes d'onde similaires sur des données continues et la détection d'événements grâce à l'information contenue dans les spectrogrammes des signaux [O'Reilly et al., 2013], suivant une idée de Baluja and Covell [2006] pour la reconnaissance et l'identification de fichiers audio. Ici, l'étape de reconnaissance n'est pas celle qui nous intéresse, mais toutes les étapes la précédant, qui permettent la compression et le stockage des caractéristiques les plus représentatives de l'image, le sont.

Dans un premier temps, on explique la procédure suivie pour la compression de l'image (c'est-à-dire du spectrogramme) et le calcul de son empreinte [Baluja and Covell, 2006] :

1. on calcule les spectrogrammes (avec une normalisation en log_{10}).
2. On applique la transformée en ondelette de Haar [Stollnitz et al., 2005]. Une ondelette permet de décomposer une fonction en décrivant sa forme générale avec des détails plus ou moins fins. En prenant l'exemple d'une image 1D, la transformée en ondelette de Haar consiste à effectuer la moyenne de chaque paire de pixels : ceci dégrade la résolution de l'image « en amont », mais les coefficients contenant les détails sont stockés également « en aval ». Le processus se répète récursivement sur les valeurs moyennés, jusqu'à obtenir une seule valeur moyenne (la moyenne « globale ») suivie des coefficients de détails classés par ordre de résolution croissante (voir Fig. I.2.30). Pour une image 2D (Fig. I.2.31a), comme le spectrogramme, la transformée en ondelette de Haar se fait successivement sur les lignes, puis sur les colonnes.
3. à l'issue de la transformée en ondelette de Haar, on obtient donc une matrice de taille identique contenant l'ensemble des coefficients d'ondelette (Fig. I.2.31b). La transformée inverse de cette matrice permet de retrouver l'image initiale avec le même niveau de résolution. Pour compresser l'image, et donc perdre en résolution, on ne conserve donc qu'une certaine proportion des coefficients d'ondelette de plus grandes valeurs (Fig. I.2.31c,d). Plus la proportion gardée diminue, plus l'image reconstruite par la transformée inverse sera dégradée par rapport à l'image initiale. Ici, il s'agit de choisir la proportion qui permet de corréler l'image compressée et l'image initiale avec un bon coefficient tout en minimisant l'erreur.
4. la matrice contenant les coefficients retenus est « binarisée » : seuls les coefficients de plus grandes valeurs prennent la valeur 1 ; le reste de la matrice est mis à 0. La matrice ainsi obtenue constitue la signature (ou empreinte) du spectrogramme (Fig. I.2.31c).

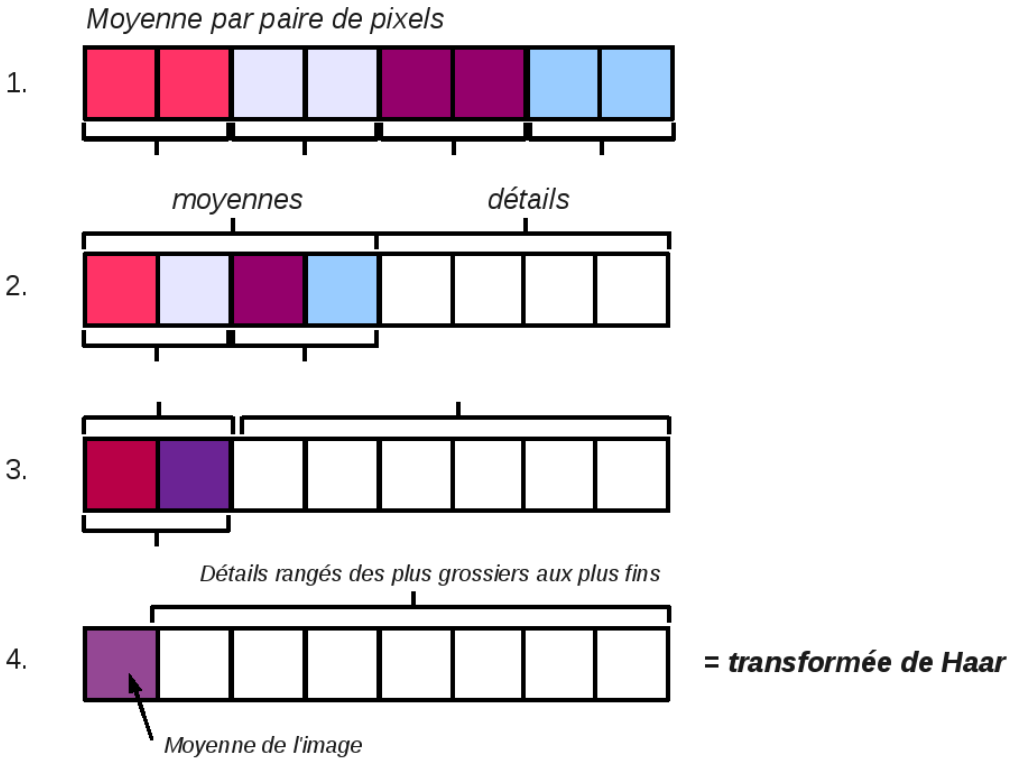


FIG. I.2.30: Schéma explicatif de la transformée de Haar sur une image 1D de 8 pixels. Voir les explications dans le texte.

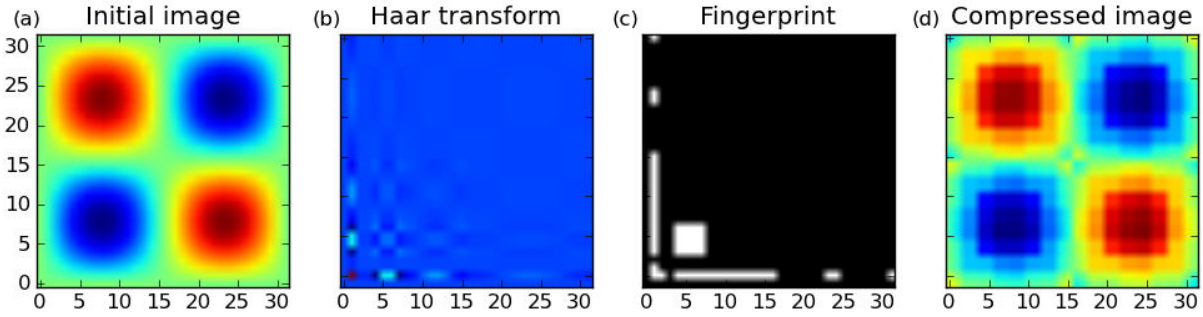
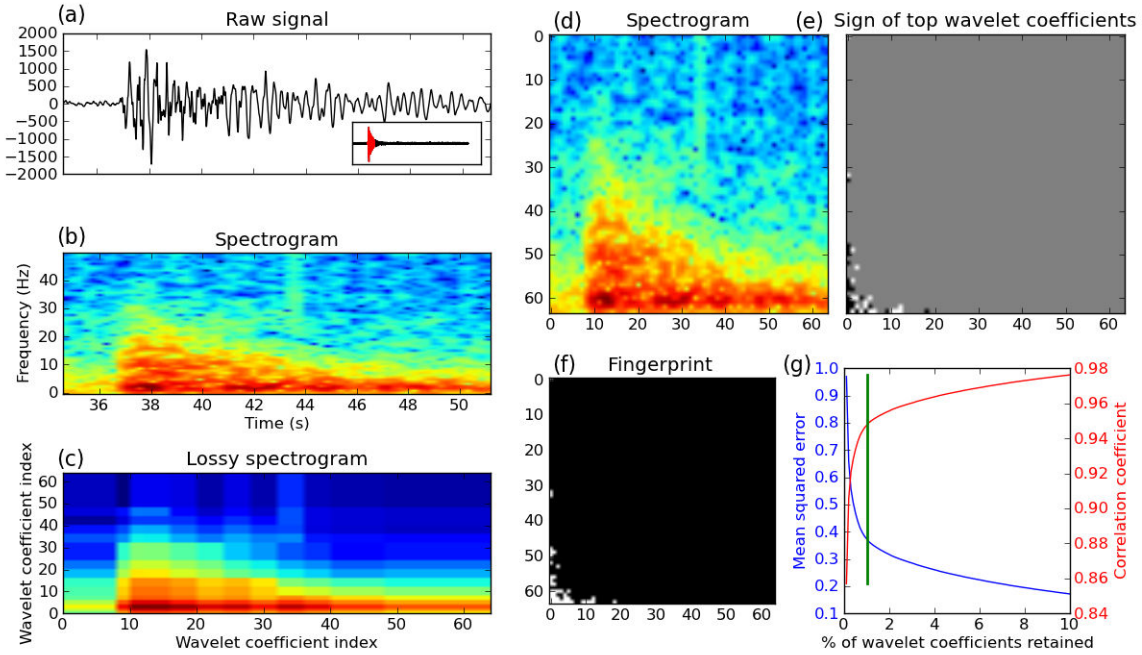


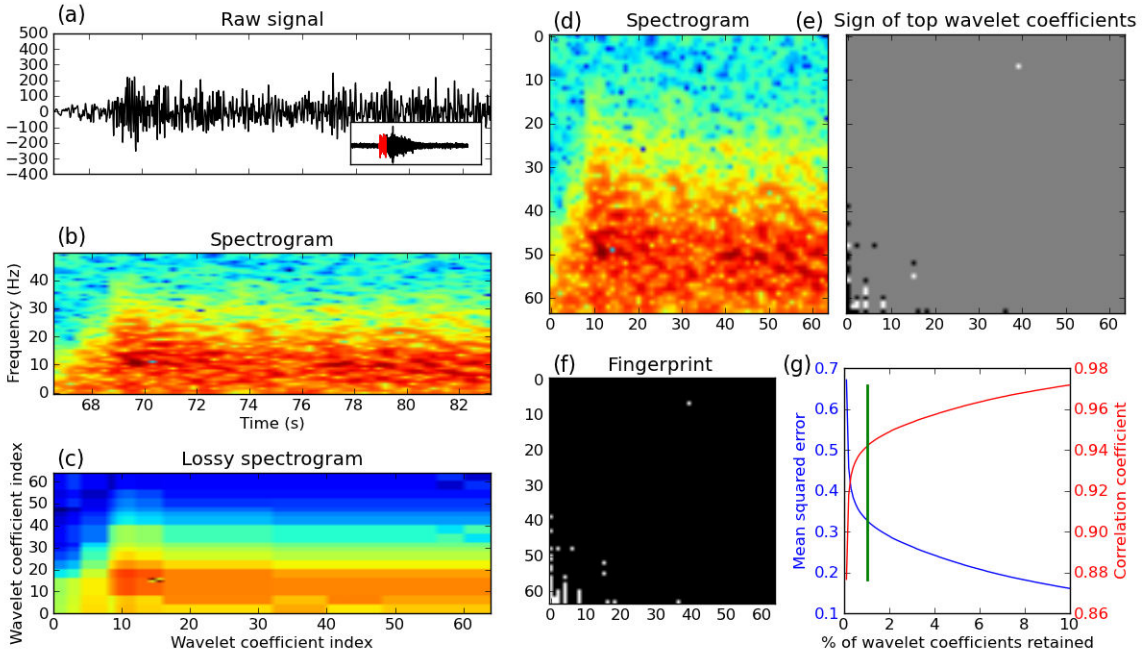
FIG. I.2.31: Exemple de la compression d'une image (a) en utilisant la transformée en ondelette d'Haar. On ne garde que 5% des coefficients maximum de la transformée (b) pour reconstruire l'image (d). L'empreinte (c) met en évidence les coefficients de plus grande valeur retenus pour faire la transformée inverse.

On présente sur la figure I.2.32 des exemples de *fingerprint* pour un VT et pour un EB. On ne traite pas le signal entier, mais seulement la partie qui est susceptible de contenir l'information importante, à savoir le début du signal. Pour compenser les erreurs éventuelles de détermination du début du signal, la fenêtre considérée commence un peu avant celui-ci. Il faut aussi noter qu'on s'arrange pour avoir une matrice de taille carrée : la longueur du signal traitée déprendra donc de la taille de la fenêtre glissante utilisée pour calculer le spectrogramme, du recouvrement entre fenêtres, et de la résolution en fréquence. On peut jouer avec ces paramètres et les adapter en fonction du résultat souhaité. Ici, les spectrogrammes ont été calculés sur des fenêtres glissantes de 1 s avec un recouvrement de 0.8 s et 64 échantillons en fréquence. Les différents tests de résolution et de taille de matrice que l'on a pu effectuer (non présentés dans ce qui suit, mais détaillé en section II.3.3.3) ont permis de montrer que choisir une haute résolution complique le problème (trop d'informations), sans compter le temps de calcul qui augmente sensiblement. Il ne faut pas tomber non plus dans l'excès inverse, avec une résolution trop basse qui fait que l'on perd l'information. Dans le même ordre d'idée, les sous-figures I.2.32g montrent qu'il est inutile de retenir un nombre de coefficients d'ondelette trop important puisque l'information peut être contenue dans seulement 1% des coefficients maximum (soit, pour une matrice de taille 64x64, une quarantaine). Le choix du pourcentage correspond approximativement à celui qui minimise l'erreur et maximise la corrélation entre l'image initiale et l'image compressée. Il est à ajuster en fonction des paramètres choisis précédemment pour le calcul des spectrogrammes.

Finalement, l'examen des empreintes obtenues montre deux motifs différents selon le type d'événement (Fig. I.2.32f) : assez "ramassé" pour le VT, beaucoup plus dispersé pour l'EB. Ceci devrait *a priori* permettre de séparer les deux types d'événements.



(a) VT



(b) EB

FIG. I.2.32: Exemples des empreintes obtenues pour un VT (haut) et un EB (bas). (a) Partie de signal brut sur lequel on calcule l'empreinte. L'encart en bas à droite montre la partie de signal retenue. (b) Logarithme du spectrogramme. (c) Spectrogramme compressé. (d) Logarithme du spectrogramme (échelle respectée sur les 2 axes). (e) Matrice signée des coefficients maximum retenus. (f) Empreinte (*fingerprint*). (g) Courbes d'erreur et de corrélation du spectrogramme initial et du spectrogramme compressé en fonction du % de coefficients retenus. La ligne verte verticale indique que l'on a retenu 1% des coefficients maximum pour construire les images de la figure.

Dans cette première étape, on a considérablement réduit l'information contenue dans les spectrogrammes. Dans un second temps, on souhaite obtenir une représentation encore plus compacte des spectrogrammes. On applique pour cela la technique du Min-Hash à l'empreinte (binaire) calculée précédemment. Elle permet de réduire la taille de l'empreinte à un vecteur de taille p et repose sur le postulat que deux signatures Min-Hash seront très similaires si et seulement si les deux signatures obtenues après la transformée en ondelette sont similaires. Le calcul de la signature Min-Hash se déroule selon les étapes suivantes :

1. on déroule la matrice de la représentation binaire obtenue précédemment et on effectue p permutations des positions des bits dans un ordre aléatoire, mais connu (FIG. I.2.33a,b,c).
2. pour chacune des permutations, on note la position du premier bit de valeur 1. A l'issue du processus, on obtient donc un vecteur de longueur p qui constitue la signature Min-Hash du spectrogramme (FIG. I.2.33c,d). Ici, on fait 500 permutations.
3. on concatène ensuite certaines valeurs de la signature Min-Hash via des fonctions de hachage, puis on les stocke dans l tables de hachage (FIG. I.2.33e,f). Une table de hachage est une structure qui permet d'associer une valeur à une clé. Ici, on a choisi de calculer 50 tables, ce qui signifie qu'on concatène les valeurs de la signature par groupes de 10. Pour la fonction de hachage, on a simplement utilisé la somme pondérée des valeurs v_i de la signature Min-Hash, c'est-à-dire $\texttt{hash value} = \sum_{i=1}^{10} i * v_i$. On peut même encore appliquer une fonction de compression après le calcul de la somme, en arrondissant la valeur à la dizaine ou centaine près, cela n'a que peu d'influence sur les résultats finaux.
4. les l valeurs de hachage ainsi calculées sont utilisées comme attributs (l valeurs = l attributs, FIG. I.2.33g).

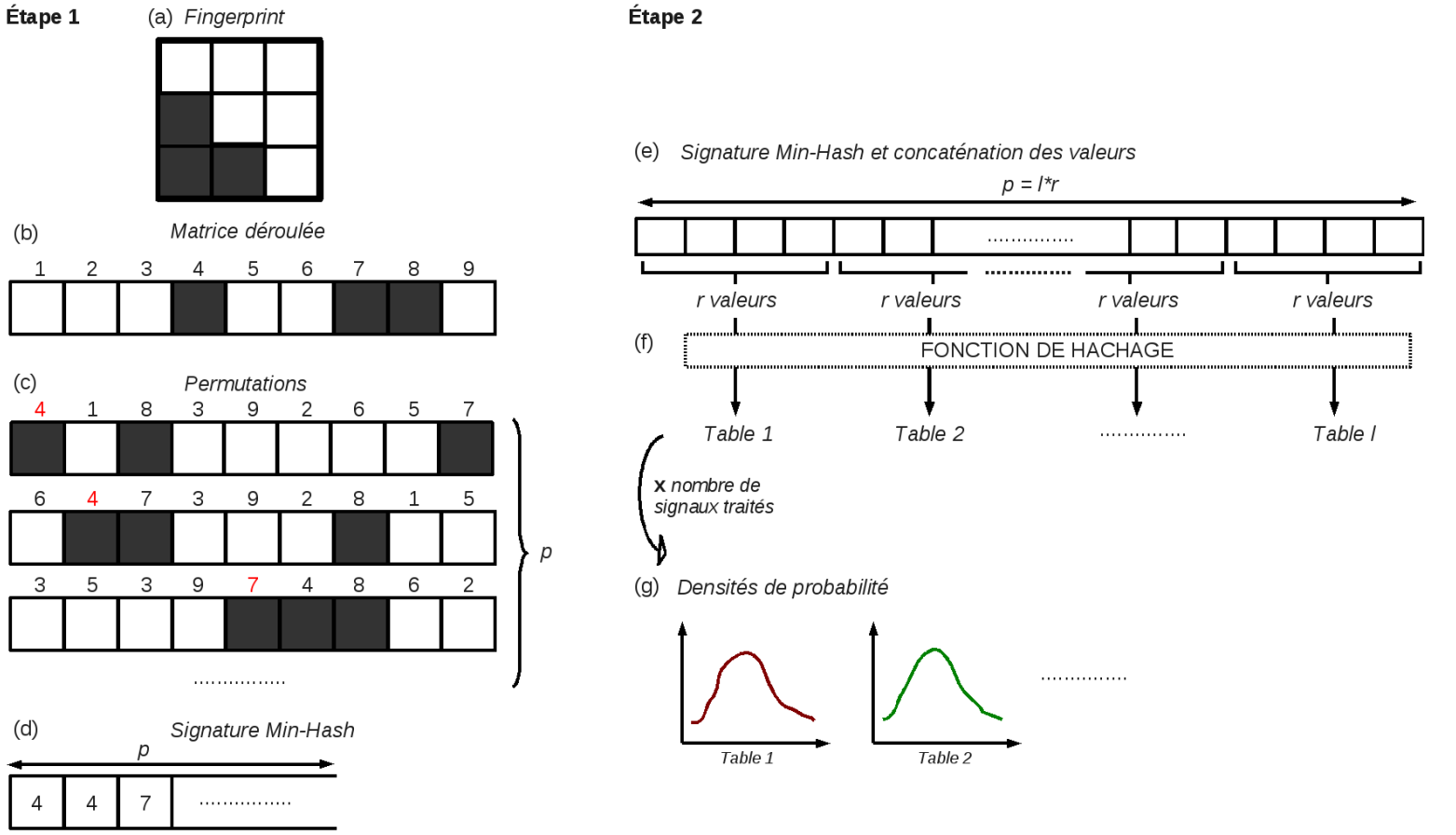


FIG. I.2.33: **(Gauche)** Schéma explicatif pour la génération de la signature Min-Hash. (a) Empreinte (matrice). Les cases noircies correspondent aux bits de valeur 1. (b) Matrice de l'empreinte déroulée. (c) p permutations de l'ordre des bits de la matrice déroulée. (d) Ecriture de la signature Min-Hash : vecteur de longueur p comprenant les positions du premier bit de valeur 1 pour chaque permutation. **(Droite)** Schéma explicatif pour la génération des tables de hachage. (e) Concaténation de la signature Min-Hash par découpage en l groupes de valeurs disjoints des autres groupes. (f) Calcul des l valeurs de hachage pour chaque table via la fonction de hachage. (g) Densités de probabilité calculées pour chaque table à partir de l'ensemble des valeurs de hachage.

I.2.4.4 Attributs basés sur le signal analytique

Les séries temporelles enregistrées par les sismogrammes peuvent être considérées comme la partie réelle d'un signal complexe : le signal analytique. En d'autres termes, si $x(t)$ est le signal enregistré, alors :

$$\hat{x}(t) = x(t) + jx_H(t) \tag{I.2.34}$$

où $\hat{x}(t)$ est le signal analytique et $x_H(t)$ est la transformée de Hilbert du signal.

L'utilité du signal analytique réside dans sa capacité à séparer les informations d'amplitude et de phase. Son analyse a notamment été développée pour l'interprétation des données de sismique réflexion [Taner et al., 1979, Barnes, 1993]. Les premiers à les avoir utilisé pour les besoins de la classification sont Beyreuther and Wassermann [2008], Hammer et al. [2012]. Pour mieux comprendre en quoi il peut nous être utile, on s'appuiera d'ailleurs sur des exemples synthétiques adaptés aux problèmes de sismique.

Connaissant le signal analytique, on définit :

- l'amplitude, ou enveloppe du signal, qui correspond à la norme du signal analytique et rend compte de la forme du signal. On la note

$$A(t) = \sqrt{x^2(t) + x_H^2(t)} \tag{I.2.35}$$

- la phase instantanée du signal, notée

$$\phi(t) = \arctan \frac{x_H(t)}{x(t)} \tag{I.2.36}$$

  dont découle la fréquence instantanée $f_i(t)$ (dérivée première par rapport au temps) :

$$f_i(t) = \frac{1}{2\pi} \frac{d\phi(t)}{dt}. \tag{I.2.37}$$

Dans les études de sismique réflexion, l'objectif est d'obtenir une image du sous-sol et de pouvoir l'interpréter du mieux possible. L'analyse des attributs du signal analytique contribue à cette interprétation de la manière suivante [Taner et al., 1979] :

- l'amplitude de l'enveloppe fournit des informations sur l'intensité des réflexions générées par les réflecteurs en profondeur.

- la phase instantanée est révélatrice des discontinuités et des irrégularités des réflecteurs présents en profondeur.

- la fréquence instantanée, dérivée de la phase, donne des indications sur les changements de phase se produisant dans le signal au cours du temps. Elle met en évidence les réflexions composites (issues de l'interférence de plusieurs réflexions). Cependant, il faut souligner deux points : (1) comme la fréquence instantanée est issue d'une dérivation, son résultat en présence de bruit devient facilement instable ; (2) lorsque plusieurs ondes se chevauchent à un instant donné, la mesure devient difficile à interpréter.

Barnes [1993] introduit également la largeur de bande instantanée $B_{w_{inst}}$:

$$B^2_{w_{inst}} = \left[\frac{\frac{dA(t)}{dt}}{2\pi A(t)} \right]^2 \tag{I.2.38}$$

qui donne une mesure des changements relatifs se produisant dans l'enveloppe. Il souligne son intérêt en remarquant le fait que, si la largeur de bande instantanée est généralement inférieure à la fréquence instantanée, la situation inverse pourrait refléter la nouvelle arrivée d'un train d'ondes.

Pour illustrer les différents attributs simplement et comprendre comment ils peuvent être utilisés, on a reproduit les exemples synthétiques créés dans son article par Barnes [1993]. Pour cela, on définit le Ricker, qui est une ondelette fréquemment utilisée dans les études de sismique pour l'élaboration de synthétiques :

$$\text{Ricker}(t) = (1 - 2\pi^2 f^2 t^2) e^{-\pi^2 f^2 t^2} \tag{I.2.39}$$

où f est la fréquence prédominante.

La figure I.2.34a montre un Ricker de fréquence 30 Hz, ainsi que son enveloppe. La fréquence et la largeur de bande instantanées sont représentées en (b). On voit que la largeur de bande instantanée s'annule au temps où l'ondelette atteint son maximum ; et que ses 2 valeurs maximales correspondent au début et à la fin du signal.

Le deuxième exemple synthétique (Fig. I.2.34c) représente des Ricker convolués avec 3 couples de dirac (-1,+1) dont l'espacement est respectivement de 40, 30 et 20 ms. La fréquence et la largeur de bande instantanées (Fig. I.2.34d) montrent :

- pour le premier couple, espacé de 40 ms, que la largeur de bande instantanée n'est pas supérieure à la fréquence instantanée quand la deuxième ondelette commence à dominer la première.

- pour le deuxième couple, espacé de 30 ms, que la largeur de bande instantanée atteint un maximum, alors que le fréquence instantanée atteint un minimum quand la deuxième ondelette commence à dominer la première.

- pour le troisième couple, l'interférence ne permet plus de séparer distinctement les deux ondelettes.

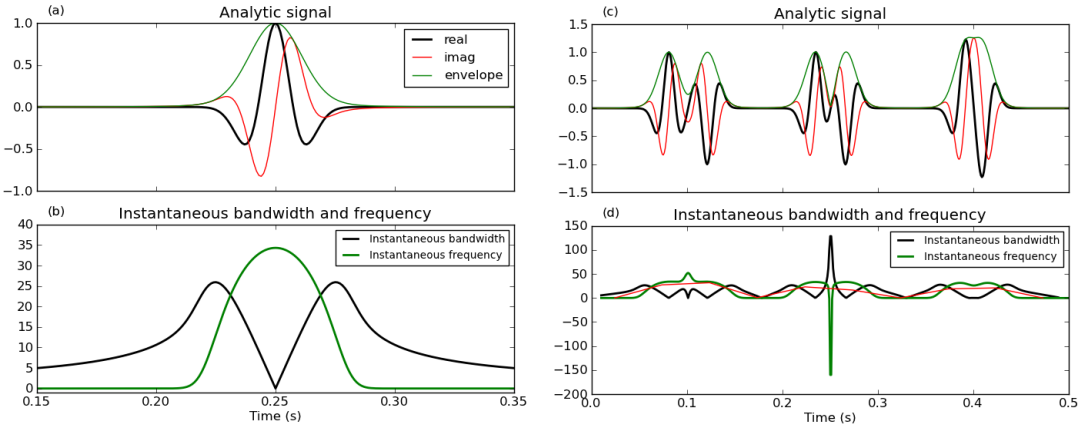


FIG. I.2.34: (a) Ricker seul de fréquence 30 Hz. (b) Largeur de bande instantanée et fréquence instantanée du Ricker. (c) Convolution d'un Ricker de fréquence 30 Hz avec 3 paires de diracs d'amplitude égale et opposée et dont l'espacement est de 40, 30 et 20 ms. (d) Largeur de bande instantanée et fréquence instantanée des interférences. Les segments rouges correspondent aux 10 valeurs moyennes de la fréquence instantanée qui seraient retenus comme attributs pour la classification automatique.

Dans notre étude, la problématique est évidemment différente de celle d'une étude de sismique réflexion, mais il faut juste retenir que si la largeur de bande instantanée et la fréquence instantanée sont complètement indépendantes (revoir les expressions mathématiques), l'étude simultanée de leurs variations peut être susceptible de contenir une information intéressante, et c'est ce qui nous intéresse.

On a donc défini et utilisé les attributs suivants :

- la pente de la phase instantanée déroulée après avoir effectué un *fit* polynomial de degré 1. Celle-ci donne la tendance d'évolution de la fréquence instantanée au cours du temps.

- la fréquence instantanée, pour laquelle on découpe le signal en 10 fenêtres de taille égale dans lesquelles on calcule les moyennes (10 attributs). La fréquence instantanée étant calculée par dérivation, elle devient facilement instable en présence de bruit, d'où l'idée de garder des valeurs moyennes.

- la largeur de bande instantanée, pour laquelle on procède de la même manière que pour la fréquence instantanée afin de faciliter leur comparaison.

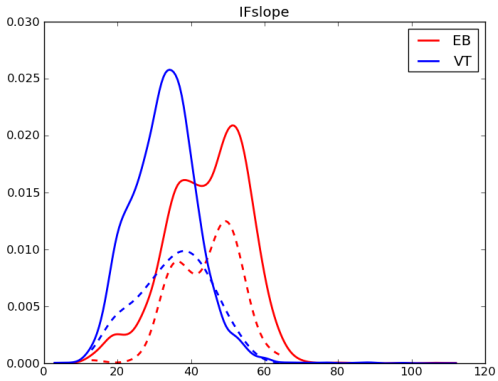
Les PDFs de ces attributs présentés sur la figure I.2.35(a-c) montrent que chaque attribut pris un à un semble assez peu discriminant, mais que les valeurs prises par les EB et VT ne se mélangent pas totalement non plus.

A ces attributs, Beyreuther and Wassermann [2008], Hammer et al. [2012] ont ajouté le temps du centroïde C qui correspond au temps auquel la moitié de la surface en-dessous de

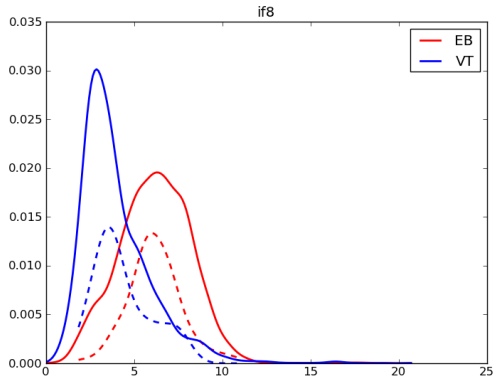
l'enveloppe $A(t)$ est atteinte. Si T est la taille de la fenêtre dans laquelle le signal est contenu, alors :

$$C = \frac{1}{T}\underset{t'}{\operatorname{argmin}} \left| \int_0^{t'} A(t)dt - 0.5 \int_0^{T} A(t)dt \right|. \tag{I.2.40}$$

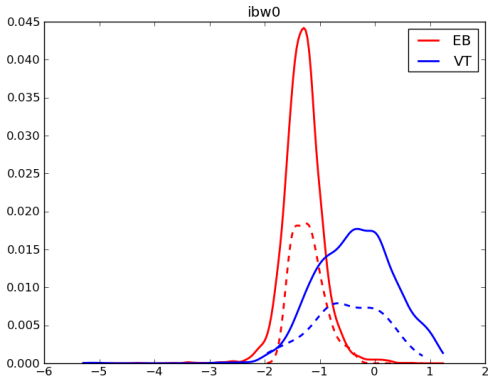
Intuitivement, on comprend que plus l'événement est long à émerger, plus son temps du centroïde sera grand. Les PDFs le confirment (Fig. I.2.35d) : les EB ont une enveloppe plus plate que les VT, d'où un temps du centroïde globalement plus élevé.



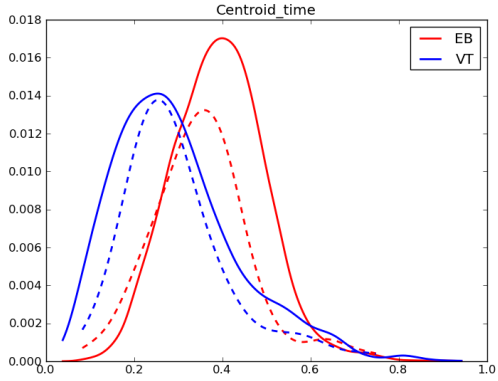
(a) Pente de la phase instantanée



(b) Fréquence instantanée



(c) Largeur de bande instantanée



(d) Temps du centroïde

Fig. I.2.35: Densités de probabilité de 4 attributs issus du signal analytique. Trait continu : *test set* ; trait tireté : *training set*.

I.2.4.5 Attributs issus de l'analyse de polarisation des données 3C

Lorsque les stations ont 3 composantes (une verticale, deux horizontales), il est possible de mener une analyse de polarité [Jurkevics, 1988, Ohrnberger, 2001]. Notons $x_z(t) = [x_{1z}, \ldots, x_{Nz}]$, $x_n(t) = [x_{1n}, \ldots, x_{Nn}]$ et $x_e(t) = [x_{1e}, \ldots, x_{Ne}]$ les sismogrammes enregistrés respectivement sur les composantes verticale (Z) et horizontales (N,E) du sismomètre (où N

est le nombre d'échantillons du signal retenu), alors on peut définir la matrice des données $\mathbf{X}$ suivante :

$$\mathbf{X} = \begin{bmatrix} \mathbf{x_z} \\ \mathbf{x_n} \\ \mathbf{x_e} \end{bmatrix} = \begin{bmatrix} x_{1z} & x_{2z} & \dots & x_{Nz} \\ x_{1n} & x_{2n} & \dots & x_{Nn} \\ x_{1e} & x_{2e} & \dots & x_{Ne} \end{bmatrix} \tag{I.2.41}$$

à partir de laquelle on peut calculer la matrice de covariance S :

$$\mathbf{S} = \frac{\mathbf{X}^T\mathbf{X}}{N} \text{ soit } S_{ij} = \frac{1}{N}\sum_{k=1}^{N} x_{ik}x_{kj} \tag{I.2.42}$$

Les coefficients de $\mathbf{S}$ décrivent l'équation d'un ellipsoïde dont les directions et longueurs des axes peuvent être déterminés par décompositon de la matrice en vecteurs et valeurs propres. Les trois vecteurs propres sont notés $(\mathbf{u_1}, \mathbf{u_2}, \mathbf{u_3})$, et les trois valeurs propres $(\lambda_1, \lambda_2, \lambda_3)$.

Quatre nouveaux attributs permettant de caractériser le signal en découlent :

- la rectilinéarité, qui quantifie le degré de linéarité du mouvement de la particule :

$$\text{rect} = 1 - \frac{\lambda_2 + \lambda_3}{2\lambda_1} \tag{I.2.43}$$

  Lorsque le mouvement est purement linéaire, ce qui est le cas théoriquement pour les ondes P et S, la rectilinéarité vaut 1. Lorsque le mouvement de la particule n'a pas de direction préférentielle, la valeur tend à se rapprocher de 0.

- la planarité, qui indique si le mouvement de particule est polarisé dans un plan (valeur égale à 1) ou s'il n'existe pas de polarisation préférentielle (valeur tendant vers 0) :

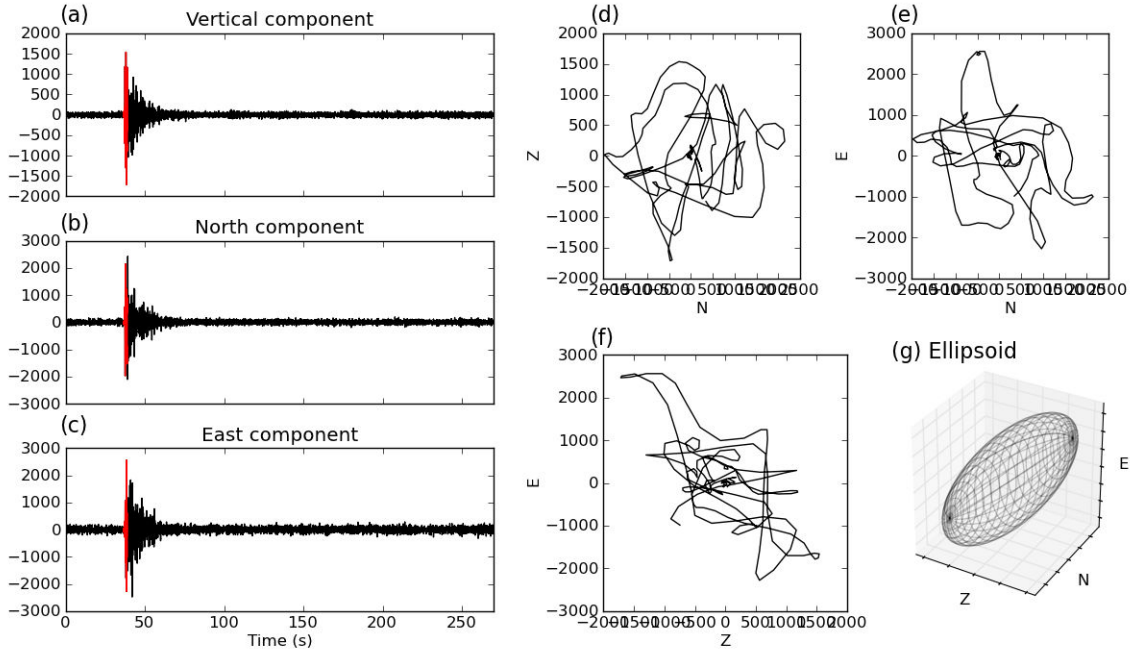
$$\text{plan} = 1 - \frac{2\lambda_3}{\lambda_1 + \lambda_2} \tag{I.2.44}$$

- l'azimut ϕ_P et l'angle d'incidence θ_P de l'onde P, définis par :

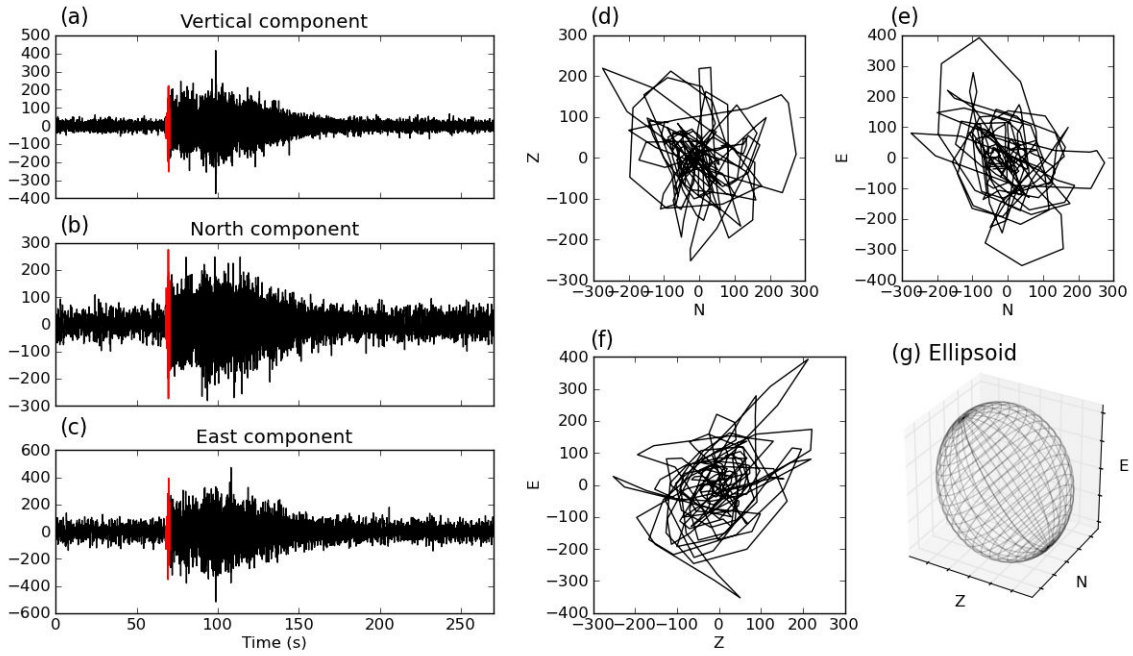
$$\phi_P = \arctan\frac{u_{21}\,\text{sign}(u_{11})}{u_{31}\,\text{sign}(u_{11})} \text{ et } \theta_P = \arccos(u_{11}) \tag{I.2.45}$$

Les PDFs calculées pour les données du *training set* et du *test set* montrent que les VT ont une rectilinéarité et une planarité en moyenne plus élevée que les EB, bien qu'une proportion non négligeable des EB du *test set* prenne des valeurs de planarité comparables à celles des VT.

Globalement, les attributs basés sur l'analyse de polarité n'ont pas l'air discriminants du tout (voir Fig. I.2.36).

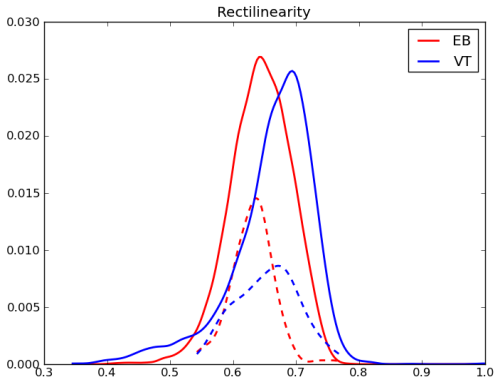


(a) VT

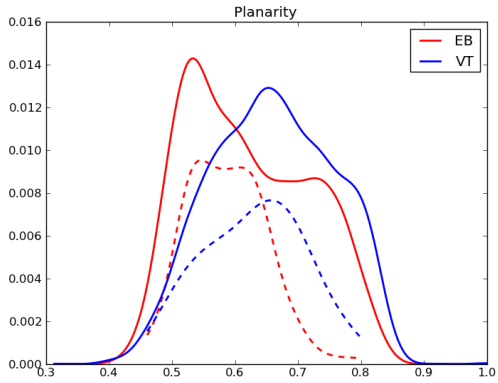


(b) EB

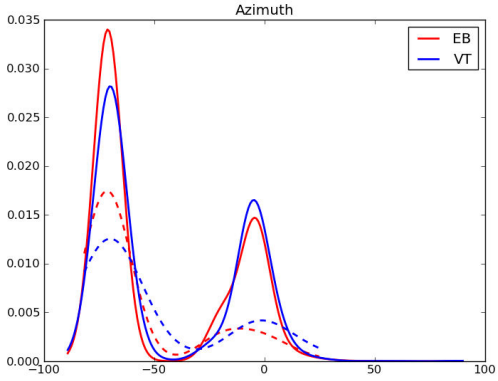
FIG. I.2.36: Analyse de polarité pout un VT (haut) et un EB (bas). (a-c) Sismogrammes enregistrés sur les 3 composantes dans l'ordre Z,N,E. En rouge, la partie du signal retenue pour l'étude de polarisation. (d-f) Mouvement des particules dans les trois plans de l'espace. (g) Ellipsoïde.



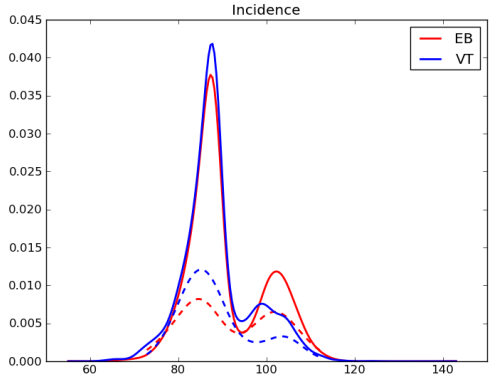
(a) Rectilinéarité



(b) Planarité



(c) Azimut



(d) Angle d'incidence

FIG. I.2.37: Densités de probabilité des quatre attributs issus de l'analyse de polarité. Trait continu : *test set* ; trait tireté : *training set*.

I.2.5 Récapitulatif

Dans les deux sections précédentes, on a présenté les techniques utilisées pour la classification automatique (non spécifiques à la sismologie), puis les attributs sismiques définis pour caractériser les signaux sismologiques. Au total, ce sont plus d'une cinquantaine d'attributs qui sont calculés pour chaque événement, hors tables de hachage. L'ensemble est résumé dans le tableau I.2.2.

Le tableau I.2.3 réprend également l'ensemble des termes définis dans cette section et qui sont régulièrement ré-employés dans la suite de ce travail.

Le code de régression logistique a été entièrement écrit dans le cadre de cette thèse. Pour les algorithmes de la SVM et des K-moyennes, le package Scikit-learn de Python a été utilisé [Pedregosa et al., 2011].

Attribut	Nom	Expression mathématique	Bibliographie
Domaine temporel			
Durée	Dur	$t_f - t_i$	Ibs-von Seht [2008], Hibert et al. [2014]
AsDec	AsDec	$\frac{t_{max} - t_i}{t_f - t_{max}}$	Hibert et al. [2014]
Croissance	Growth	$\frac{t_{max} - t_i}{t_f - t_i}$	Hibert et al. [2014]
Asymétrie (r=3)	Skewness	$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i - \mu}{\sigma}\right]^r$	Curilem et al. [2009]
Kurtosis (r=4)	Kurto		Hibert et al. [2014]
Maximum sur moyenne	RappMaxMean	$\frac{\max A(t)}{\overline{A(t)}}$	Hibert et al. [2014]
Domaine spectral			
Kurtogramme	F_{up} et F_{low}	-	-
Maximum sur moyenne TF	RappMaxMeanTF	$\frac{\max \lvert F(\nu)\rvert}{\overline{\lvert F(\nu)\rvert}}$	-
Fréquence dominante	PredF	$\max_{\nu}\lvert F(\nu)\rvert$	Ibs-von Seht [2008], Hammer et al. [2012]
Fréquence dominante (*time stack*)	sPredF	$\max_{\nu}\lvert time\ stack\rvert$	
Fréquence dominante moyenne	MeanPredF	-	
Fréquence dominante (évolution)	$v_{1...10}$	-	
Fréquence centrale	CentralF	$\frac{\int \nu\lvert F(\nu)\rvert d\nu}{\int \lvert F(\nu)\rvert d\nu}$	idem
Largeur de bande	Bandwidth	$\frac{\int (\nu - f_c)^2\lvert F(\nu)\rvert d\nu}{\int \lvert F(\nu)\rvert d\nu}$	idem
Largeur de bande (*time stack*)	sWidth		-
Energie dans $[f_1, f_2]$	Ene$f_1 - f_2$	$\int_{f_1}^{f_2} \lvert F(\nu)\rvert^2 d\nu$	Curilem et al. [2009], Hibert et al. [2014]
Temps du maximum	TimeMaxSpec	$\max_t \text{spectrogram}(\nu, t)$	-
Tables de hachage	1 à 50	-	-
Signal analytique			
Pente de la phase instantanée	IFslope	$\arctan \frac{x_H(t)}{x(t)}$	-
Fréquence instantanée	$if_{1...10}$	$\frac{1}{2\pi}\frac{d\phi(t)}{dt}$	Hammer et al. [2012]
Largeur de bande instantanée	$ibw_{1...10}$	$(2\pi A(t))^{-1}\frac{dA(t)}{dt}$	Hammer et al. [2012]
Temps du centroïde	Centroid_time	$\frac{1}{T}\text{argmin}_{t'}\left\lvert\int_0^{t'} A(t)dt - 0.5\int_0^{T} A(t)dt\right\rvert$	Hammer et al. [2012]
Polarisation			
Planarité	Planarity	$1 - \frac{2\lambda_3}{\lambda_1 + \lambda_2}$	Ohrnberger [2001], Beyreuther et al. [2008], Hammer et al. [2012]
Rectilinéarité	Rectilinearity	$1 - \frac{\lambda_2 + \lambda_3}{2\lambda_1}$	
Azimut	Azimuth	$\arctan \frac{u_{21}\,\text{sign}(u_{11})}{u_{31}\,\text{sign}(u_{11})}$	
Angle d'incidence	Incidence	$\arccos(u_{11})$	

TAB. I.2.2: Tableau récapitulatif des attributs sismiques calculés. On rappelle pour chacun d'entre eux leur formulation mathématique ainsi que le nom utilisé. La colonne bibliographie précise pour chaque attribut les études de classification automatique récentes dans lesquelles ils ont déjà été utilisés (liste non exhaustive).

Terme employé	Définition
Attribut	Caractéristique mesurée sur le signal sismique. A un attribut est associée une valeur pour un événement donné.
Densité de probabilité (PDF)	Donne la loi de probabilité d'une variable (i.e. un attribut) donnée.
Apprentissage supervisé	Qui cherche à classer des données dont on connaît déjà la structure.
Apprentissage non-supervisé	Qui cherche à classer des données dont on ne sait rien de la structure.
Training set	Jeu de données sur lequel le système d'apprentissage supervisé s'entraîne.
Test set	Jeu de données complet qu'on cherche à classer.
Surface de décision	Hyperplan séparateur recherché pour séparer les données.
Régression logistique (LR) - p.31	Méthode d'apprentissage supervisé qui cherche les paramètres définissant la surface de décision.
Machine à vecteurs de support (SVM) - p.41	Méthode d'apprentissage supervisé qui cherche les paramètres définissant la surface de décision maximisant la marge entre les éléments les plus proches.
K-moyennes - p.48	Méthode d'apprentissage non supervisé qui regroupe les éléments dans les groupes dont la moyenne est la plus proche.
Matrice de confusion - p.47	Manière de présenter les résultats de la classification entre observations et prédictions.
Sur-apprentissage - p.38	Situation dans laquelle l'algorithme d'apprentissage supervisé s'est trop bien adapté aux données du *training set*. La variance est élevée, mais l'erreur est faible (Tab. I.2.1).
Sous-apprentissage - p.38	Situation dans laquelle l'algorithme d'apprentissage supervisé ne s'est pas bien adapté aux données du *training set*. La variance est faible, mais l'erreur est élevée (Tab. I.2.1).

Tab. I.2.3: Définitions des termes fréquemment utilisés dans la suite du manuscrit.

Partie II

Traitement des données du Piton de la Fournaise, La Réunion

Sommaire

CHAPITRE II.1

Présentation générale

Le volcan du Piton de la Fournaise, culminant à 2631 m d'altitude, est situé sur l'île de la Réunion dans l'océan Indien. Cette île a vu le jour grâce à un volcanisme de point chaud qui a donné naissance à trois volcans. Aujourd'hui, seul le Piton de la Fournaise est encore en activité. Il donne régulièrement lieu à des éruptions de type effusif (une par an en moyenne depuis 1998), ce qui en fait l'un des volcans les plus actifs du monde. Cette activité volcanique intense s'accompagne d'une activité sismique importante.

L'enregistrement et l'analyse de la sismicité dans un tel contexte s'avère primordiale :

- d'abord dans un but de surveillance et d'observation du volcan. Une recrudescence de la sismicité constitue généralement un signe précurseur d'une éruption. Elle est liée le plus souvent à des mouvements de magma en profondeur. L'Observatoire Volcanologique du Piton de la Fournaise (OVPF) possède ainsi une dizaine de stations sismologiques permanentes enregistrant en continu l'activité sismique du volcan.

- ensuite dans un but de meilleure compréhension des processus se produisant dans l'édifice volcanique.

C'est dans ce cadre que le projet ANR UnderVolc a eu lieu [Brenguier et al., 2012]. 21 stations large-bandes ont pu être installées sur le volcan pour compléter le réseau pré-existant pendant 3 ans (2009-2012) (voir TAB. II.1.1). Comme le relief autour du volcan est relativement accidenté, le réseau présente l'avantage d'avoir une géométrie 3D très intéressante pour la localisation de la sismicité et compensera le fait qu'on n'utilise que l'information sur la première arrivée dans notre processus de localisation (§I.1). Les stations utilisées ont un pas d'échantillonnage de 100 Hz.

Divers types de signaux sismiques sont enregistrés sur le Piton de la Fournaise. Parmi ceux qui sont directement liés au volcan, on peut citer les événements volcano-tectoniques (notés VT) et les éboulements (EB) (voir FIG. II.1.1).

Si une activité sismique régulière est enregistrée sur le volcan, il existe des périodes où l'activité s'intensifie. Ces périodes, pouvant durer plusieurs heures, sont appelées **crises** et

sont caractérisées par une forte hausse du nombre de VT (généralement plusieurs centaines).

Ces crises sismiques ne sont pas nécessairement suivies par des éruptions dans le cas du Piton de la Fournaise. On distingue ainsi deux types d'essaims sismiques : les essaims dits **pré-éruptifs**, qui précèdent une éruption ; et les essaims dits **intrusifs**, qui ne donnent pas lieu à une éruption. Notons que le terme *intrusif* demeure relativement impropre car les éruptions sont aussi précédées de phases intrusives. Cependant, pour des raisons de clarté, nous continuerons d'utiliser ces deux termes.

Station	Latitude	Longitude	Altitude	X UTM (km)	Y UTM (km)	Instrument
UV01	-21.24366	55.65286	2378 m	360.212	7650.290	Taurus
UV02	-21.27387	55.77934	471 m	373.364	7647.052	Taurus
UV03	-21.22336	55.75783	998 m	371.088	7652.626	Taurus
UV04	-21.26741	55.76162	1004 m	371.520	7647.753	Taurus
UV05	-21.24862	55.71409	2523 m	366.571	7649.794	Taurus
UV06	-21.23979	55.75247	1413 m	370.546	7650.803	Taurus
UV07	-21.22886	55.69158	2214 m	364.217	7651.962	Taurus
UV08	-21.24642	55.68451	2193 m	363.499	7650.012	Taurus
UV09	-21.21110	55.72021	1961 m	367.173	7653.952	Taurus
UV10	-21.28373	55.72497	1806 m	367.732	7645.916	Taurus
UV11	-21.23971	55.70921	2552 m	366.057	7650.776	Taurus
UV12	-21.25534	55.72472	2075 m	367.680	7649.059	Taurus
UV13	-21.29164	55.70792	2085 m	365.970	7645.026	Taurus
UV14	-21.20184	55.69540	1781 m	364.589	7654.956	Taurus
UV15	-21.24510	55.70884	2579 m	366.023	7650.179	Trident
SNE	-21.23911500	55.71790000	2505 m	366.958	7650.849	Q330
FJS	-21.22949333	55.72229000	2123 m	367.405	7650.849	Q330
RVL	-21.25589444	55.70052222	2110 m	365.170	7648.977	Q330
FOR	-21.26192500	55.71870000	2049 m	367.061	7648.325	Q330
FLR	-21.24074167	55.73287167	1947 m	368.513	7650.682	Q330
HDL	-21.25072667	55.79059000	242 m	374.512	7649.623	Q330

TAB. II.1.1: Caractéristiques des stations du réseau temporaire utilisé sur le Piton de la Fournaise dans le cadre du projet ANR UnderVolc.

Au cours du projet UnderVolc (2009-2012), 12 crises sismiques d'importance significative ont eu lieu sur le volcan (TAB. II.2.2). On dispose des données sismologiques continues correspondant à ces crises pour 20 stations.

On s'attachera dans le chapitre suivant à détecter et localiser les événements sismiques lors de ces crises avec Waveloc. On a en plus la chance de disposer des localisations manuelles effectuées par A. Schmid [Schmid, 2011] au cours de sa thèse pour la crise du 14 octobre 2010 : ceci permettra un étalonnage et un ajustement des paramètres de localisation de Waveloc, et surtout, une évaluation de la robustesse de l'algorithme.

Dans le deuxième chapitre de cette partie, on s'intéressera à la classification de deux des types de signaux observés sur le Piton de la Fournaise : les événements volcano-tectoniques (notés VT) et les éboulements (notés EB) (FIG. II.1.1). L'étude des EB sur le Piton de la Fournaise a été motivée par l'effondrement du cratère principal du volcan (le cratère Dolomieu) lors de l'éruption d'avril 2007. Les EB générés suite à cet effondrement sont nombreux et sont bien enregistrés par le réseau de sismomètres. Leur étude a fait l'objet de la thèse de Hibert [2012]. Le jeu de données utilisé pour la partie classification nous a été fourni par Hibert [2012],

Hibert et al. [2014] (plus de détails seront donnés dans la partie consacrée).

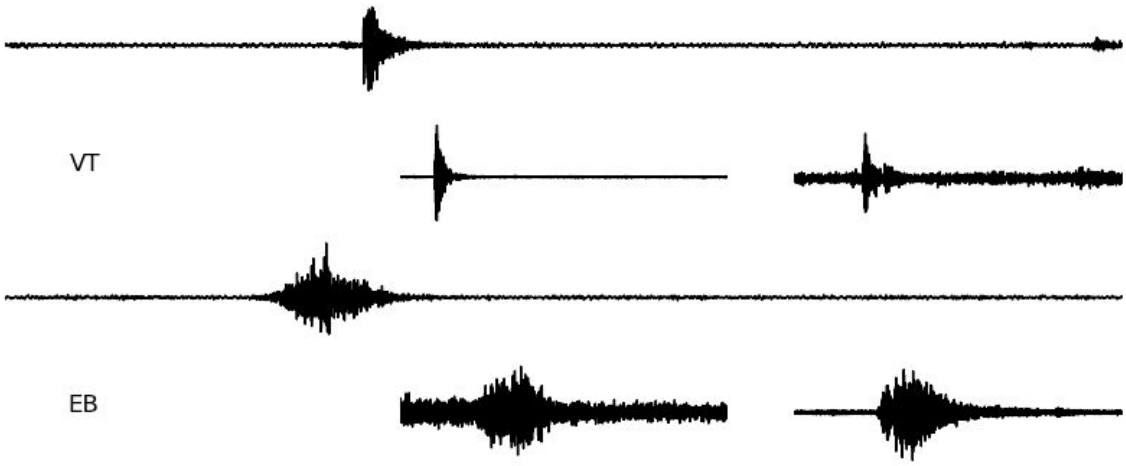


FIG. II.1.1: Formes d'ondes de trois événements volcano-tectoniques (VT) et de trois éboulements (EB) enregistrés sur la composante verticale de la station BOR et tirés aléatoirement. On peut déjà noter des différences dans l'allure générale des sismogrammes (durée du signal, impulsivité...)

II.2.1 Tests de résolution avec Waveloc

Le réseau de stations sismologiques installé sur le volcan du Piton de la Fournaise a une extension d'environ 14 km dans la direction Est-Ouest et 10 km dans la direction Nord-Sud. Du fait de la topographie du volcan, les stations sont réparties entre 250 et 2500 m au-dessus du niveau de la mer (voir TAB. II.1.1).

Le choix de la grille s'avère primordial : dans l'idéal, il faut une grille qui permette d'avoir une résolution suffisante sur les localisations tout en n'augmentant pas démesurément le temps de calcul, puisque l'on souhaite aussi être efficace.

Comme la sismicité apparaît très localisée dans une zone autour du cratère, nous avons finalement limité l'espace d'étude à des dimensions de 8x6x3 km [Taisne et al., 2011], avec pour origine X=362 km, Y=7647 km et Z=500 m de profondeur dans le système de projection cartographique UTM (*Universal Transverse Mercator*). Le pas d'échantillonnage est de 250 m dans les trois directions de l'espace. Il constitue un bon intermédiaire entre un échantillonnage très large (500 m par exemple) ou plus resserré (100 m par exemple) qui demande un temps de calcul beaucoup plus long.

La grille qu'on utilise contient donc 32x24x12 = 9216 points, soit autant de sources sismiques potentielles à explorer lors de l'étape de migration (§I.1). Ici, on dispose des données de 20 stations échantillonnées à 100 Hz, d'où, pour une seule heure de données, plusieurs milliards de valeurs à stocker : ceci justifie pleinement la nécessité de simplifier le problème en ne gardant que le maximum absolu des sommations pour chaque échantillon en temps dans la trace $S_{max}(t)$.

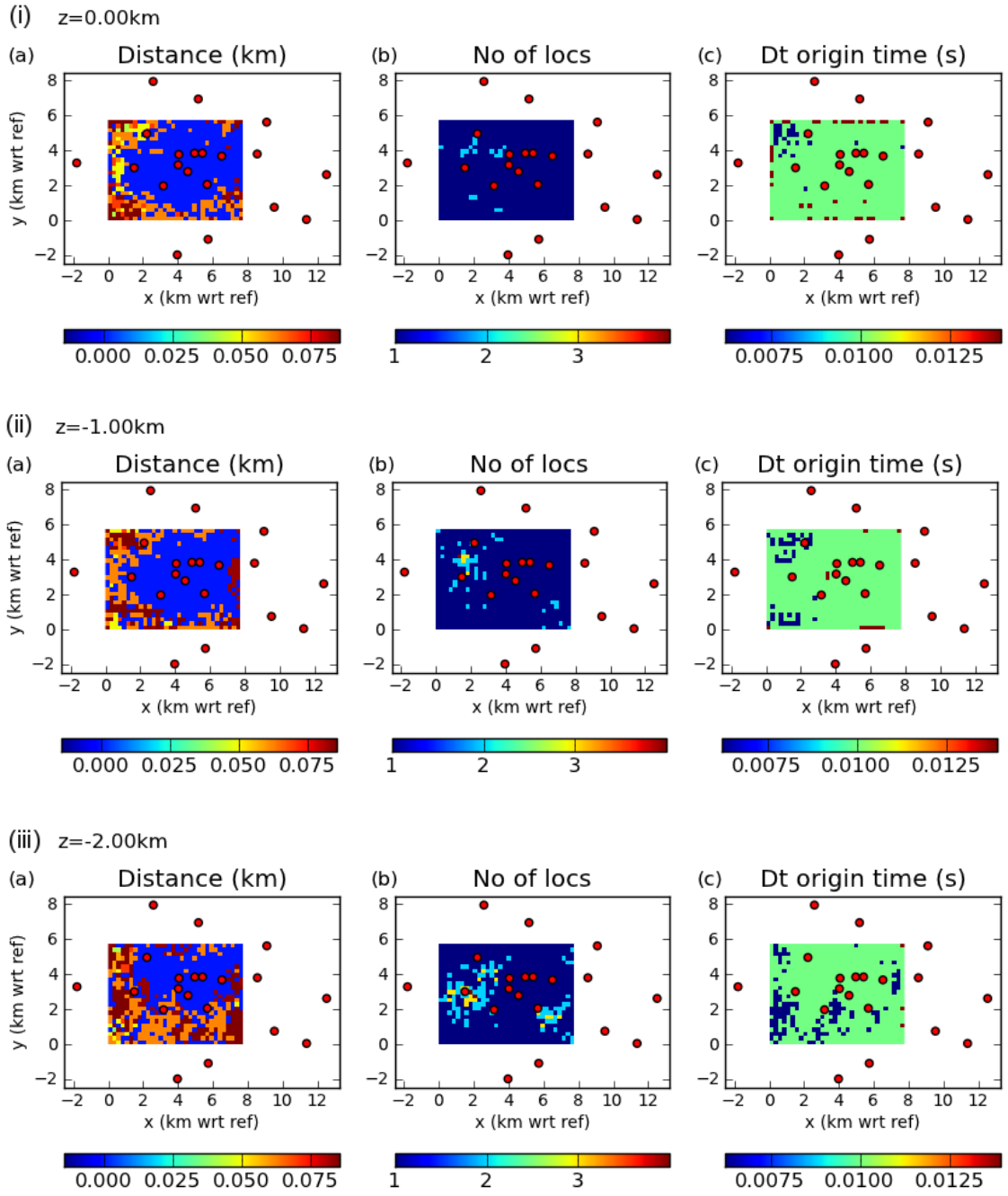


FIG. II.2.1: Tests de résolution effectués sur le réseau de 20 stations du Piton de la Fournaise pour différentes profondeurs : (i) 0 km (niveau de la mer), (ii) -1 km et (iii) -2 km. L'axe des z est positif vers le bas. Chaque test cartographie les erreurs sur la localisation (a), le nombre de localisations de Waveloc (b) et les erreurs sur les temps origine (c). Les ronds rouges correspondent aux stations. La grille utilisée est de taille réduite par rapport à l'étendue du réseau de stations. L'origine en x correspond à la coordonnée UTM de 362 km ; celle en y à 7647 km.

Les tests de résolution sont menés de la manière suivante : pour chaque point de la grille, on crée des traces synthétiques (non bruitées) telles qu'elles seraient enregistrées en chacune des stations du réseau. On cherche ensuite à retrouver la bonne localisation suivant les étapes de Waveloc. Pour chaque point étudié, on calcule la distance qui sépare la localisation trouvée par Waveloc de la localisation initiale ; le nombre de localisations faites par Waveloc (sachant qu'il peut y avoir plusieurs maxima qui dépassent le niveau de détection dans $S_{max}(t)$) ; et l'écart sur les temps origine.

On a choisi de construire les synthétiques dans des traces de 20 s de long échantillonnées à 100 Hz. Le pulse triangulaire correspondant à l'événement se produit au temps origine 6 s avec une amplitude égale à 1. Le niveau de détection est fixé à la moitié du nombre de stations, soit 10.

La figure II.2.1 présente les résultats des tests de résolution pour trois profondeurs différentes (axe des z positif vers le bas ; voir les tests pour d'autres profondeurs en annexes A3).

On voit qu'étant donné la bonne couverture du réseau, les tests de résolution sont plutôt bons : l'erreur sur les localisations est généralement inférieure à 100 m. Elle se dégrade sur les bords et lorsque l'altitude augmente, mais la zone située autour du cratère (autour de x=366 km et y=7651 km) reste toujours bien résolue.

Les écarts sur les temps origine sont minimes (de l'ordre de 0.01 s) sur l'ensemble de la grille quelle que soit la profondeur considérée.

Enfin, le nombre de localisations nous indique que le risque d'avoir des localisations multiples (donc potentiellement très éloignées de la vraie localisation) augmente lorsque l'on s'élève en altitude.

II.2.2 Ajustement des paramètres de Waveloc : l'exemple de la crise du 14 octobre 2010

Dans cette section, nous détaillerons précisément la manière de procéder que nous avons adoptée pour localiser automatiquement les VT de la crise du 14 octobre 2010. Le choix de cette crise n'a pas été fait au hasard, puisqu'en effet nous disposons des localisations manuelles réalisées par A. Schmid [Schmid, 2011] : 447 événements ont été localisés en utilisant le logiciel NonLinLoc [Lomax, 2011] et avec le modèle de vitesse 3D des ondes P fourni par Prono et al. [2009].

Le même modèle de vitesse sera utilisé dans notre étude, ce qui facilite grandement la comparaison entre les localisations automatiques de Waveloc et les localisations manuelles. De plus, cela permettra d'ajuster les paramètres de Waveloc plus finement avant d'évaluer la robustesse et la fiabilité de l'algorithme.

La figure II.2.2 donne un exemple de résultat obtenu avec Waveloc (se référer à la partie I.1 pour la méthode). Les formes d'ondes de l'événement considéré sont visibles sur la figure II.2.3. Celles-ci montrent que, même si la qualité des données est variable selon les stations, les premières arrivées sont émergentes en plusieurs stations. Les coupes dans les 3 plans de l'espace (II.2.2a,d,f) montrent une plus grande complexité par rapport au test synthétique (I.1.6a,d,f). Le pic de $S_{max}(t)$ est aussi plus large. Cette complexité peut s'expliquer par la présence de bruit dans les données et par la diversité des valeurs prises par les maxima des traces du gra-

dient du kurtosis $\dot{K}_+$ (II.2.3b). Dans cet exemple, l'incertitude mesurée sur le temps origine est de ± 0.1 s (rappelons que cette valeur correspond à l'intervalle formé à 95% de l'amplitude maximale du pic). Les incertitudes en x, y et z sont quant à elles déterminées par simple calcul de l'écart-type dans cet intervalle : dans cet exemple, elles sont respectivement de ± 0.2, ± 0.1 et ± 0.2 km.

En donnant à Waveloc l'ensemble des données continues disponibles pour la crise du 14 octobre 2010, on obtient un catalogue de sismicité à partir duquel on va pouvoir effectuer des comparaisons avec le catalogue manuel.

Pour les comparaisons, on considère qu'un événement manuel et qu'un événement automatique correspondent au même événement lorsque la différence entre leur temps origine est inférieure à 1 s. Ceci va ensuite permettre de calculer les différences entre les paramètres hypocentraux automatiques et manuels et d'en regarder la répartition sous forme d'histogrammes. Ces derniers vont non seulement nous donner une idée de la précision maximale que l'on peut espérer atteindre avec Waveloc, mais vont aussi montrer à quel point le choix des paramètres (bande de filtrage, choix de la fonction caractéristique et méthode de calcul...) est capital. Les résultats obtenus lors des différents tests sont résumés dans le tableau II.2.1.

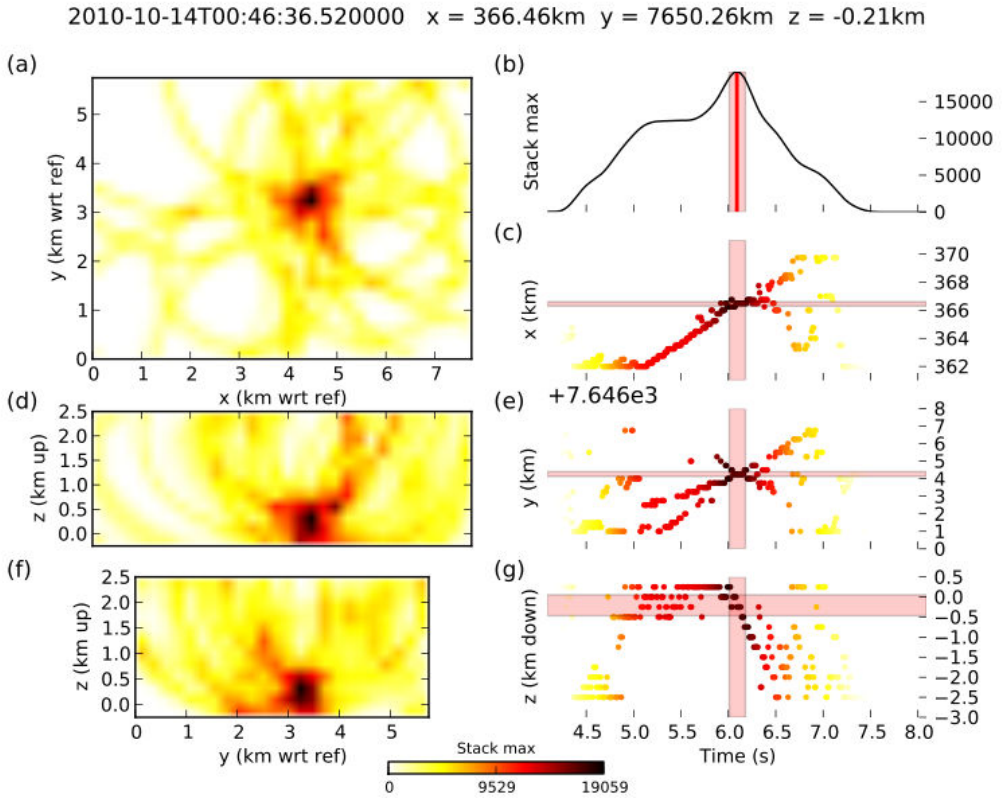


FIG. II.2.2: Migration et localisation d'un VT sur le Piton de la Fournaise. Les zones colorées en rouge donnent les incertitudes sur les 4 paramètres hypocentraux. Pour le détail concernant la description des sous-figures, se référer à la figure I.1.6.

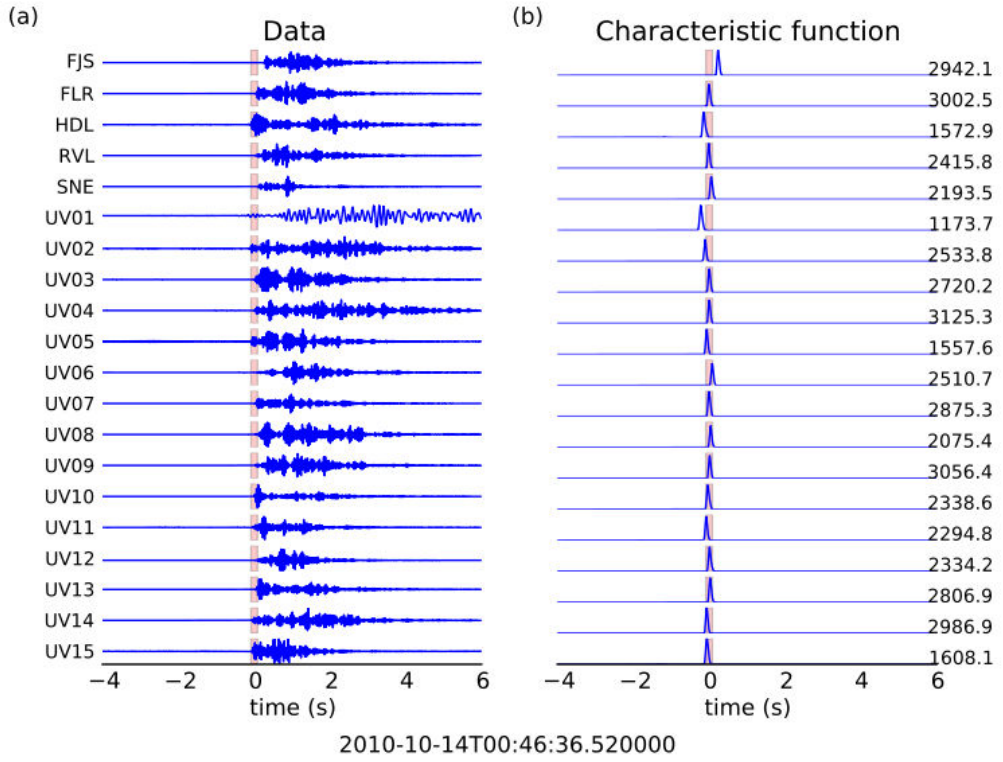


FIG. II.2.3: (a) Formes d'ondes brutes migrées ; (b) traces du gradient du kurtosis $\dot{K}_+$ migrées. La valeur du maximum de chaque trace $\dot{K}_+$ est notée à droite. Toutes les traces sont alignées par rapport au temps de trajet théorique à l'hypocentre. La zone colorée en rouge donne l'incertitude sur le temps origine.

Dans le traitement standard de l'OVPF, les données sont filtrées entre 4 et 10 Hz. C'est donc tout naturellement que nous avons commencé les premiers tests en filtrant les données dans cette bande de fréquence. Les résultats des localisations sont inscrits dans la première ligne du tableau II.2.1 (test 1). On a initialement utilisé les traces de kurtosis calculées de manière non-récursive. La taille de la fenêtre de calcul des kurtosis et le seuil de détection avaient été fixés de manière à retrouver un nombre d'événements similaires à ceux de la localisation manuelle (autour de 450). La taille de fenêtre qui semblait le mieux convenir était "moyenne" (3.0 s), ni trop large, ni trop petite.

En diminuant uniquement la taille de cette fenêtre de calcul (test 2, 1.0 s), on constate que les paramètres hypocentraux sont déterminés avec une précision accrue et que plus de 90% des événements localisés le sont aussi manuellement. En revanche, on est limité en termes de nombre d'événements localisés. Le seuil de détection étant déjà relativement bas, il est difficile de l'abaisser davantage, au risque de localiser des signaux qui n'ont plus rien à voir avec des VT...

Le calcul des kurtosis étant relativement coûteux en temps, il est beaucoup plus intéressant d'utliser une méthode de calcul récursive qui permet d'accélérer le processus. Cependant, comme évoqué en §I.1, cela change l'aspect des kurtosis et la taille de la fenêtre de calcul doit être réadaptée. Ici, dans le test 3, on garde une fenêtre de 1.0 s. Les résultats montrent

une grosse perte dans la précision des localisations bien que leur nombre augmente. On peut expliquer ces observations par le fait que les kurtosis récursifs ont un temps de décroissance beaucoup plus lent que les kurtosis standards : ils sont donc plus larges, plus étalés dans le temps. De ce fait, les chances d'avoir des sommations constructives augmentent au détriment de la précision. Rappelons aussi que les kurtosis récursifs ont une amplitude plus élevée que les standards. Comme le seuil de détection est resté le même pour les tests 2 et 3, il n'est pas impossible qu'un certain nombre d'événements localisés ne soient pas des VT.

Enfin, le test 4 permet de montrer qu'il est bien sûr possible d'obtenir une meilleure précision sur les localisations (comparables à celles du test 2), à condition de diminuer fortement la taille de la fenêtre de calcul. Toutefois, comme on l'a vu précédemment, prendre une fenêtre trop petite limite aussi le nombre de détections.

La comparaison des résultats des tests 1 et 5 permet, quant à elle, de voir l'effet d'un changement du seuil de détection : plus le seuil est élevé, moins il y a de détections. Par contre, quasiment toutes les localisations automatiques correspondent à des événements également classés manuellement (on peut supposer qu'il s'agit des plus "évidents"). La précision sur les paramètres est aussi améliorée.

	Paramètres					**Statistiques**					
	Fonction	Calcul	w	Filtre	K_t	W	C	Δt	Δx	Δy	Δz
1	$K(t)$	Stand.	3.0 s	4-10 Hz	50	452	333	0.47 s	-0.22 km	0.12 km	0.79 km
2			1.0 s			229	215	0.29 s	-0.01 km	-0.03 km	0.58 km
3		Rec.	1.0 s			345	253	0.58 s	-0.41 km	-0.26 km	0.77 km
4			0.25 s			184	166	0.32 s	0.01 km	-0.09 km	0.49 km
5	$K(t)$	Stand.	3.0 s	4-10 Hz	200	143	138	0.34 s	-0.03 km	-0.02 km	0.75 km
6				20-35 Hz		514	368	0.29 s	0.03 km	-0.17 km	0.44 km
7	$\dot{K}_+(t)$				2000	527	366	0.11 s	-0.05 km	0.01 km	0.42 km
8	$G(t)$					638	386	0.14 s	-0.02 km	-0.02 km	0.17 km
9	$K(t)$	Rec.	1.0 s	20-35 Hz	300	646	382	0.33 s	-0.08 km	-0.25 km	0.32 km
10	$\dot{K}_+(t)$				2000	716	386	0.10 s	0.06 km	0.03 km	0.26 km
11	$G(t)$					764	390	0.11 s	-0.04 km	0.01 km	0.09 km

TAB. II.2.1: Récapitulatif de quelques résultats obtenus pour différentes combinaisons de paramètres de Waveloc. La fonction caractéristique utilisée est indiquée dans la première colonne : $K(t)$ pour le kurtosis, $\dot{K}_+(t)$ pour le gradient du kurtosis, $G(t)$ pour les gaussiennes. La méthode de calcul est soit récursive (Rec.), soit standard (Stand.). w désigne la taille de la fenêtre sur laquelle la fonction caractéristique a été calculée. K_t désigne le seuil de détection utilisé ; W donne le nombre d'événements total localisés par Waveloc ; C est le nombre d'événements communs avec la localisation manuelle ; Δ* est l'erreur moyenne sur le paramètre hypocentral considéré.

Dans la partie I.1, on avait évoqué l'importance de bien choisir les paramètres de filtrage des données afin que ceux-ci maximisent les valeurs de kurtosis. On utilise pour cela la méthode d'analyse du kurtogramme. Dans notre cas, cette méthode a montré que la meilleure bande de filtrage pour nos données est 20-35 Hz (sauf pour la station UV01 pour laquelle on a conservé le filtrage entre 4 et 10 Hz). La comparaison des tests 5 et 6 confirme l'importance du choix du filtre : non seulement la précision sur les paramètres hypocentraux est améliorée, mais en plus le nombre de localisations est triplé, d'où une plus grande efficacité. Les histogrammes correspondants à ces 2 tests sont visibles sur la figure II.2.4 (en bleu : test 5 ; en vert : test 6).

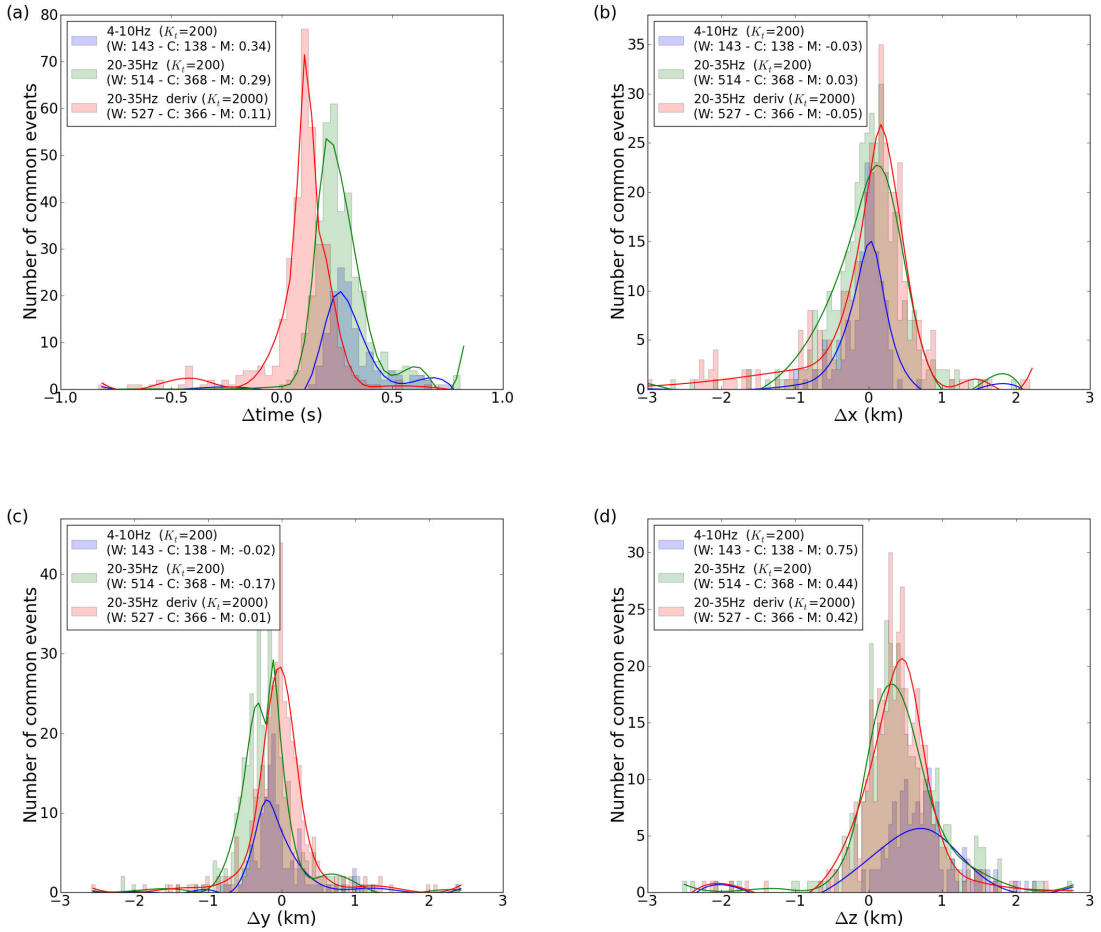


FIG. II.2.4: Histogrammes des événements communs montrant la précision sur chacun des paramètres hypocentraux et l'influence des paramètres de filtrage et du choix de la fonction caractéristique. Les kurtosis sont calculés sur des fenêtres glissantes de 3 s. Les barres bleues correspondent aux données filtrées entre 4 et 10 Hz ; Les barres vertes, aux données filtrées entre 20 et 35 Hz ; les barres rouges, aux données filtrées entre 20 et 35 Hz et migrées en utilisant le gradient du kurtosis comme fonction caractéristique. Pour comprendre les légendes, se référer au tableau II.2.1.

Malgré les premières petites améliorations constatées, la précision des paramètres hypocentraux, et plus particulièrement celle sur le temps origine, reste limitée. Dans la partie I.1, on avait mis en évidence le retard systématique du maximum du kurtosis par rapport au vrai début du signal. L'idée était donc d'utiliser, non plus les kurtosis, mais les gradients des kurtosis qui vont atteindre leur maximum à l'endroit de la rupture de pente liée au kurtosis. Le test 7 (à comparer avec le test 6) prouve à quel point la précision sur les temps origine est améliorée : elle est presque triplée (on passe d'une erreur de 0.29 s à 0.11 s). Remarquons une nouvelle fois la nécessité d'apdapter le seuil de détection : comme le gradient des kurtosis atteint des amplitudes beaucoup plus importantes que le kurtosis seul, la valeur de K_t a été

multipliée par 10. Ce choix a aussi été fait de manière à ce que les résultats des tests 6 et 7 soient facilement comparables, particulièrement en terme de nombre d'événements localisés. Les histogrammes correspondants au test 7 sont représentés sur la figure II.2.4 (couleur rouge).

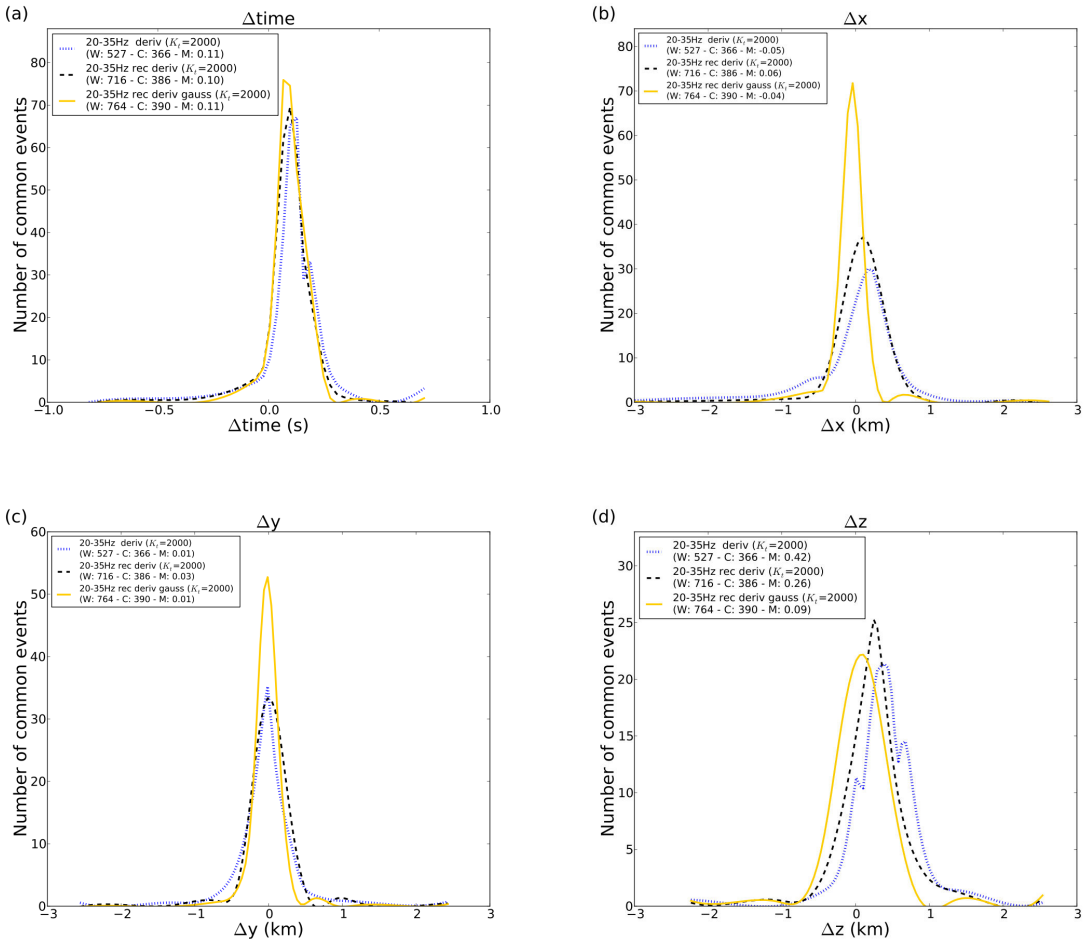


FIG. II.2.5: Histogrammes des événements communs montrant la précision sur chacun des paramètres hypocentraux et l'influence du kurtosis récursif et de la convolution gaussienne. Les données sont filtrées entre 20 et 35 Hz et la fonction caractéristique utilisée est le gradient des kurtosis $\dot{K}_+(t)$. Pour comprendre les légendes, se référer au tableau II.2.1.

Enfin, une fois de plus, on a vu qu'on peut considérablement accélérer le temps de calcul des kurtosis en utilisant la méthode récursive. Les fonctions caractéristiques des tests 9 à 11 du tableau II.2.1 utilisent la récursivité. Les histogrammes correspondants sont visibles dans la figure II.2.5.

Le test 9 est à comparer avec le test 6. On a choisi une taille de fenêtre raisonnable (1.0 s - ni trop petite, pour ne pas limiter excessivement le nombre de détections possibles ; ni trop grande) de manière à obtenir des précisions comparables à celles du test 6. Il a été nécessaire d'élever un peu le seuil de détection afin que le nombre de localisations ne soit pas trop grand (ceci étant dû au fait que les kurtosis récursifs atteignent des amplitudes supérieures).

En gardant maintenant cette même taille de fenêtre, on utilise comme fonction caractéristique les gradients de kurtosis (test 10), qui permettent d'obtenir une bien meilleure précision sur les temps origine (comparer les tests 9 et 10). La comparaison avec le test 7 (même chose, mais avec les kurtosis standards) montre une amélioration notable de la précision sur les profondeurs (presque doublée).

Dans la partie I.1 de ce manuscrit, on avait aussi discuté de la forme asymétrique des kurtosis et, dans une moindre mesure, de leur gradient. Cette asymétrie peut s'avérer gênante lors de la sommation. Un moyen d'y remédier est donc de remplacer les maxima locaux de la fonction caractéristique par une suite de diracs (en conservant les amplitudes), puis de convoluer le tout avec une gaussienne. Les résultats obtenus correspondent au test 11. L'amélioration la plus notable concerne la précision sur les profondeurs, qui atteint désormais une centaine de mètres. Le nombre de localisations augmente d'une cinquantaine d'événements (par rapport au test 10) et le nombre de localisations communes aux manuelles est le plus grand jamais atteint (plus de 85%). Précisons que le choix de la largeur de la gaussienne a aussi son importance : ici, on a pris une demi-largeur de 0.1 s, qui correspond approximativement à la largeur mesurée à la base des gradients de kurtosis initiaux ; ceci signifie donc que les pics de la nouvelle fonction caractéristique sont deux fois plus larges. Mais comme ils sont centrés sur le maximum, ils sont en comparaison plus larges à gauche et moins à droite... En choisissant une largeur strictement équivalente (soit une demi-largeur égale à 0.05 s), on s'est aperçu qu'on réduisait le nombre total d'événements localisés, et surtout, qu'on n'améliorait pas autant la précision sur les profondeurs.

Bilan

La série de tests réalisée a permis d'affiner les paramètres de Waveloc et de les ajuster afin d'obtenir le maximum d'informations contenues dans les données et ce, avec la meilleure précision possible. Les paramètres retenus au final pour la crise du 14 octobre 2010 sont ceux du onzième et dernier test. On constate qu'en terme de nombre d'événements, la méthode automatique localise environ 300 événements en plus par rapport à ce qui a été fait manuellement. Tous ces événements ont été vérifiés visuellement afin de s'assurer qu'ils correspondaient réellement à des VT. Il n'est d'ailleurs pas impossible que le catalogue manuel fourni soit incomplet. Plus de 85% des événements manuels localisés l'ont aussi été avec Waveloc. Les 15% d'événements qui n'ont pas été localisés peuvent se classer en deux catégories :

- les événements qui sont détectés, mais dont le nombre de traces ayant un rapport signal-sur-bruit assez bas est trop élevé. Le critère du nombre de stations minimum requis pour détecter un événement n'est donc pas rempli.

- plus rarement, les événements qui ne sont pas du tout détectés car l'amplitude du pic de $S_{max}(t)$ est inférieure à celle du seuil de détection.

Dans la suite, pour localiser l'ensemble des événements des crises dont nous disposons, nous utiliserons les paramètres suivants :

- filtrage des données brutes entre 20-35 Hz.
- calcul des kurtosis de manière récursive sur des fenêtres glissantes de 1.0 s.
- migration avec les traces gaussiennes (issues de la convolution avec les gradients des

kurtosis précédentes).
- seuil de détection ajustable en fonction des crises. En effet, le nombre de stations les ayant enregistrées est variable. De manière générale, on a remarqué qu'un seuil fixé à 100 fois le nombre de stations disponibles était correct (ceci n'est évidemment vrai qu'avec les choix de calcul énoncés dans les points précédents).
- nombre de détections minimum égal ou supérieur à la moitié du nombre de stations disponibles pour que la localisation soit acceptée.

II.2.3 Analyse de la sismicité

On a aussi à disposition les catalogues de sismicité de l'OVPF correspondants aux différentes crises. Ceux-ci sont construits à partir du logiciel Earthworm (détection et pointé automatiques) [Johnson et al., 1995]. Les localisations sont calculées avec Hypo71 [Lee and Lahr, 1975]. Seules 81 événements ont été localisés dans les catalogues. De plus, certains catalogues semblent incomplets (crises du 9 décembre 2010 notamment), voire sont manquants (crise du 2 janvier 2010). En ce qui concerne les détections, les résultats sont consignés dans le tableau II.2.2.

Globalement, le nombre de détections de Waveloc est toujours supérieur à celui de l'OVPF (sauf pour la crise du 14 décembre 2009). Comme dit dans la section précédente, on a fixé un seuil de détection égal à 100 fois le nombre de stations disponibles, ce qui devrait assurer une certaine "constance" dans les résultats, et, dans tous les cas, faciliter les comparaisons.

Toutefois, on observe une grande disparité dans le nombre de détections de Waveloc : il y a par exemple un nombre de détections qui semble exagéré pour les crises du 14 et 18 octobre 2009, où le nombre de stations disponibles est restreint. En effet, comme le seuil de détection choisi est fixé à 100 fois le nombre de stations, il est abaissé, d'où un plus grand nombre de détections. Mais toutes ces détections ne correspondent pas à des événements et ne donnent donc pas lieu à des localisations (10 et 15% seulement de localisations par rapport aux détections). On n'observe pas d'aussi fortes distensions entre nombre de détections et nombre de localisations pour les autres crises du catalogue, ce qui paraît beaucoup plus raisonnable.

On a considéré, comme pour les localisations de la section précédente, que 2 détections étaient communes lorsque la différence observée sur les temps origine était inférieure à 1 s. Même si l'écart dans le nombre des détections est de taille variable entre Waveloc et l'OVPF, on observe en général que plus de 60% des détections de l'OVPF sont identiques à celles de Waveloc.

Enfin, la comparaison des quelques localisations disponibles pour l'ensemble des crises avec celles de Waveloc a permis de mettre en évidence 72 événements communs (on rappelle que le nombre de localisations diffère du nombre de détections car il dépend de paramètres tels que le seuil de détection et les SNR des signaux ; tous les événements détectés par Waveloc ne sont donc pas localisés). Les comparaisons des paramètres hypocentraux (Fig. II.2.6) montrent que les coordonnées x et y sont similaires (pas plus de 200 m de différence en moyenne). Les différences en t et en z, en revanche, sont un plus marquées : les temps origine trouvés par l'OVPF sont en moyenne plus tardifs que ceux de Waveloc et les profondeurs sont plus superficielles (ce qui est cohérent compte-tenu du trade-off entre t et z). Ceci peut en partie s'expliquer par le fait que les temps origine calculés par l'OVPF ont une précision à la seconde seulement (contre le centième de seconde avec Waveloc).

Crise	Type	Stations	Détect. Waveloc	Détect. OVPF	Détections communes	Loc. Waveloc
14 octobre 2009	intrusive	7	1486	338	131 (39%)	140
18 octobre 2009	intrusive	9	2212	488	309 (63%)	321
29 octobre 2009	intrusive	15	781	322	300 (93%)	509
5 novembre 2009	pré-éruptive	17	439	181	109 (60%)	322
14 décembre 2009	pré-éruptive	13	445	557	314 (56%)	348
29 décembre 2009	intrusive	17	196	188	74 (39%)	186
2 janvier 2010	pré-éruptive	17	586	-	-	515
23 septembre 2010	intrusive	20	904	863	597 (69%)	646
14 octobre 2010	pré-éruptive	20	1168	929	628 (68%)	764
9 décembre 2010	intrusive	21	1386	161	130 (80%)	305
9 décembre 2010	pré-éruptive					685
2 février 2011	intrusive	19	652	643	481 (75%)	525

TAB. II.2.2: Crises sismiques enregistrées sur le Piton de la Fournaise entre 2009 et 2011. Le nombre de stations disponibles est indiqué dans la colonne stations. Le tableau donne aussi le nombre de détections et de localisations obtenues par Waveloc, et, à titre comparatif, le nombre de détections de l'OVPF.

Si les catalogues de l'OVPF ne sont pas complets en ce qui concerne les localisations, ils contiennent néanmoins l'information sur les magnitudes. Les magnitudes sont des magnitudes de durée calculées à partir des données des 4 stations sommitales (SNE, UV05, UV11 et UV15) et d'après la définition de Lee et al. [1972] :

$$M_d = -0.87 + 2\log\tau + 0.0035\Delta \tag{II.2.1}$$

où τ est la durée du séisme et Δ est la distance hypocentrale.

Les magnitudes calculées par Waveloc sont des magnitudes locales (voir §I.1.5). Le calcul des magnitudes avec Waveloc a aussi été limité aux données des 3 composantes des stations sommitales. Les distributions de magnitude pour l'ensemble des crises sont présentées dans la figure II.2.7.

Le nombre d'événements pour chacun des deux catalogues reste comparable (4700 événements pour l'OVPF ; 5300 pour Waveloc). La forme des distributions est relativement similaire si on excepte l'échantillonnage irrégulier des magnitudes de l'OVPF. On voit cependant qu'il y a plus d'événements de magnitude négative avec Waveloc et qu'en contrepartie, il "manque" des événements de magnitude positive proche de 0. Graphiquement, on constate d'ailleurs que la loi de Gutenberg-Richter calculée avec les résultats de Waveloc est légèrement courbe alors que celle de l'OVPF est une ligne droite.

Les b-values calculées à partir des deux catalogues donnent des valeurs de 1 pour Waveloc et 1.1 pour l'OVPF. Ces valeurs sont conformes à celles observées à l'échelle globale [Rierola, 2000, McNutt, 2002].

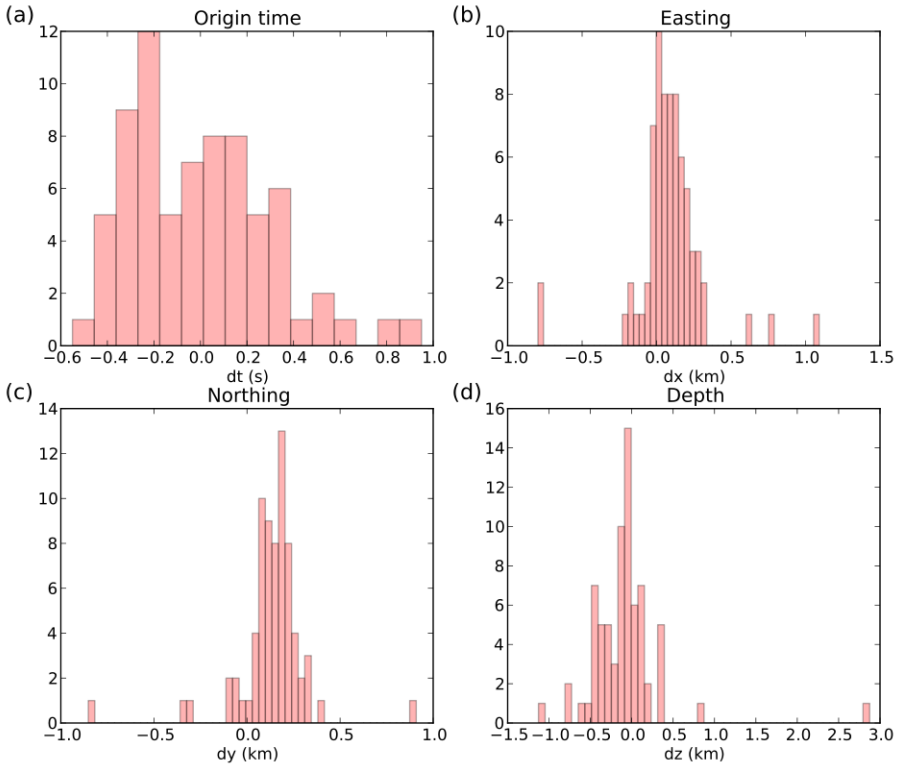


FIG. II.2.6: Comparaison des paramètres hypocentraux de Waveloc et de l'OVPF pour les 72 localisations communes.

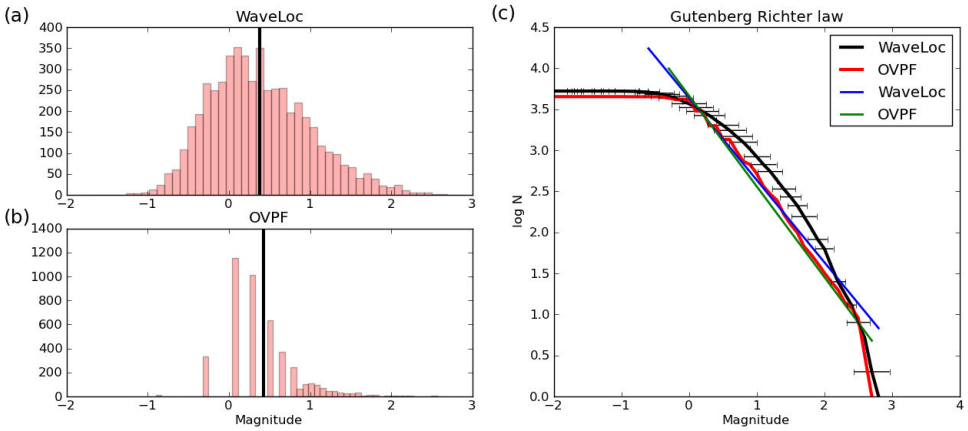


FIG. II.2.7: Distribution des magnitudes sur l'ensemble des crises. (a) Magnitudes locales calculées avec Waveloc. (b) Magnitudes de durée calculées par l'OVPF. La barre verticale noire symbolise la magnitude moyenne dans les deux cas. (c) Lois de Gutenberg-Richter et régressions linéaires associées. Les barres horizontales représentent les incertitudes mesurées pour chaque intervalle de magnitude.

De manière générale, on observe que plus un événement a une forte magnitude, plus il est visible sur les sismogrammes car les amplitudes sont plus élevées. On pourrait donc légitimement penser que les amplitudes de $S_{max}(t)$ donnent l'information sur les magnitudes directement. Or la figure II.2.8 montre clairement qu'il n'existe pas de corrélation entre la magnitude et les amplitudes résultant de la sommation des fonctions caractéristiques migrées, mais qu'il existe une limite inférieure : pour une valeur du pic de $S_{max}(t)$ donnée, la magnitude de l'événement ne peut descendre en-dessous d'une certaine valeur.

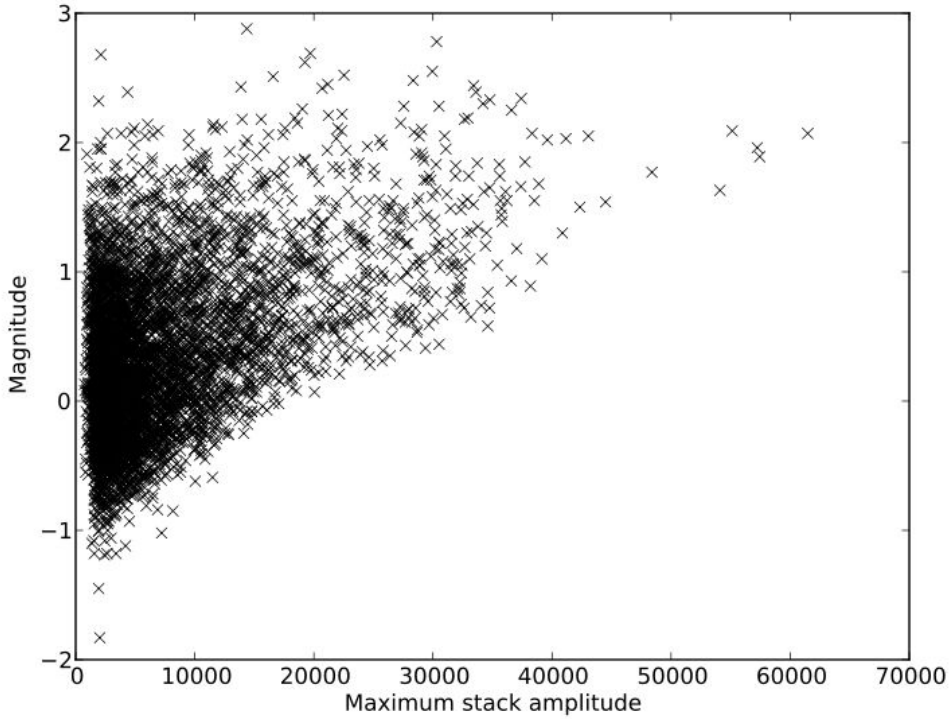


FIG. II.2.8: Magnitudes en fonction des amplitudes de $S_{max}(t)$.

Les résultats des localisations pour la crise du 14 octobre 2010 sont présentées dans la figure II.2.9. Cette crise a été suivie d'une éruption. La répartition des événements sismiques localisés par Waveloc (a) montre un essaim sismique principal situé à l'aplomb du cratère du volcan. Les événements sont pour la plupart localisés au-dessus du niveau de la mer et n'atteignent pas la surface (ils ne dépassent pas les 1500 m d'altitude). Ceci est cohérent avec une autre étude menée par Battaglia and Brenguier [2011], Battaglia [2012]. L'orientation de l'essaim en profondeur semble très légèrement nord-sud.

La comparaison avec les localisations manuelles (b) montre une dispersion plus importante de la sismicité avec Waveloc. Ceci est très probablement dû aux 300 événements supplémentaires qui ont été localisés automatiquement. En effet, lorsque l'on compare strictement les événements communs aux deux catalogues (FIG. II.2.10), on n'observe pas de différence majeure entre les 2 essaims : la sismicité est toujours concentrée dans la même zone (autour de x=366 km et y=7650 km). On peut aussi noter la présence de quelques événements se produisant dans une petite zone au sud de l'essaim principal.

La dernière remarque intéressante que l'on peut faire concerne la répartition des magni-

tudes au sein de l'essaim : il semble en effet que les événements de plus forte magnitude se produisent préférentiellement en profondeur et qu'ils soient entourés d'événements de plus faible magnitude.

La figure II.2.9 montre qu'il existe une certaine dispersion de la sismicité autour de l'essaim principal. Il doit donc être possible de localiser plus précisément les événements en utilisant la méthode de double-différence détaillée en I.1.6. La relocalisation se fait en 3 étapes :

- la recherche des multiplets, c'est-à-dire des événements de formes d'onde similaires par inter-corrélation. Les signaux ont été corrélés sur des fenêtres de 4.5 s (0.5 s avant et 4.0 s après le temps origine) et les délais ont été calculés en domaine de Fourier dès lors que le coefficient de corrélation était supérieur à 0.7. Un exemple de deux événements quasi-parfaitement corrélés est donné dans la figure II.2.11a.

- la formation de familles de multiplets. Pour cela, on a considéré que deux événements appartenaient à un même cluster lorsque leur coefficient de corrélation était supérieur à 0.7 en au moins 8 stations. Les relations entre événements au sein d'un cluster sont complexes et il est toujours intéressant de constater que deux événements peuvent appartenir à une même famille uniquement parce qu'ils sont partagent des simitudes avec un autre événement commun. Un exemple de cluster est représenté dans la figure II.2.11b.

- la relocalisation à proprement parler, par la méthode de double-différence. L'exemple donné dans la figure II.2.11c montre bien un rapprochement des événements composants le cluster à l'issue de la relocalisation. La comparaison des relocalisations avec les localisations manuelles de la crise du 14 octobre 2010 donne une idée de la précision obtenue sur chacun des paramètres hypocentraux (voir Fig. II.2.12).

La figure II.2.13 est une carte de sismicité de la crise pré-éruptive du 14 octobre 2010 après relocalisation. On constate que le nombre d'événements est décimé (seuls 400 événements ont pu être relocalisés). On obtient un essaim sismique bien moins dispersé que sans la relocalisation et plus fin. On confirme également que la principale zone sismogène se situe entre 0 et 1000 m.

Un exemple de résultats de localisations pour une crise dite intrusive est donné en figure II.2.14 (crise du 23 septembre 2010). La zone sismogène principale ne semble pas différer du cas précédent (crise pré-éruptive), ce qui indique que la distinction entre les deux types de crises *a priori* n'est pas évidente et qu'il faut affiner davantage l'analyse de la sismicité [Battaglia and Brenguier, 2011]. L'essaim se trouve toujours compris entre 0 et 1000 m au-dessus du niveau de la mer. Les magnitudes sont toujours réparties selon le même motif « concentrique », avec les plus fortes magnitudes uniquement au centre de l'essaim et les plus faibles allant jusqu'en périphérie.

Les résultats pour l'ensemble des autres crises sont présentés en annexe A2. La figure II.2.15 montre, pour chacune des crises, l'évolution de la sismicité au cours du temps. On observe très nettement que le nombre d'événements augmente à l'apex de la crise, augmentation également marquée par une hausse générale des magnitudes.

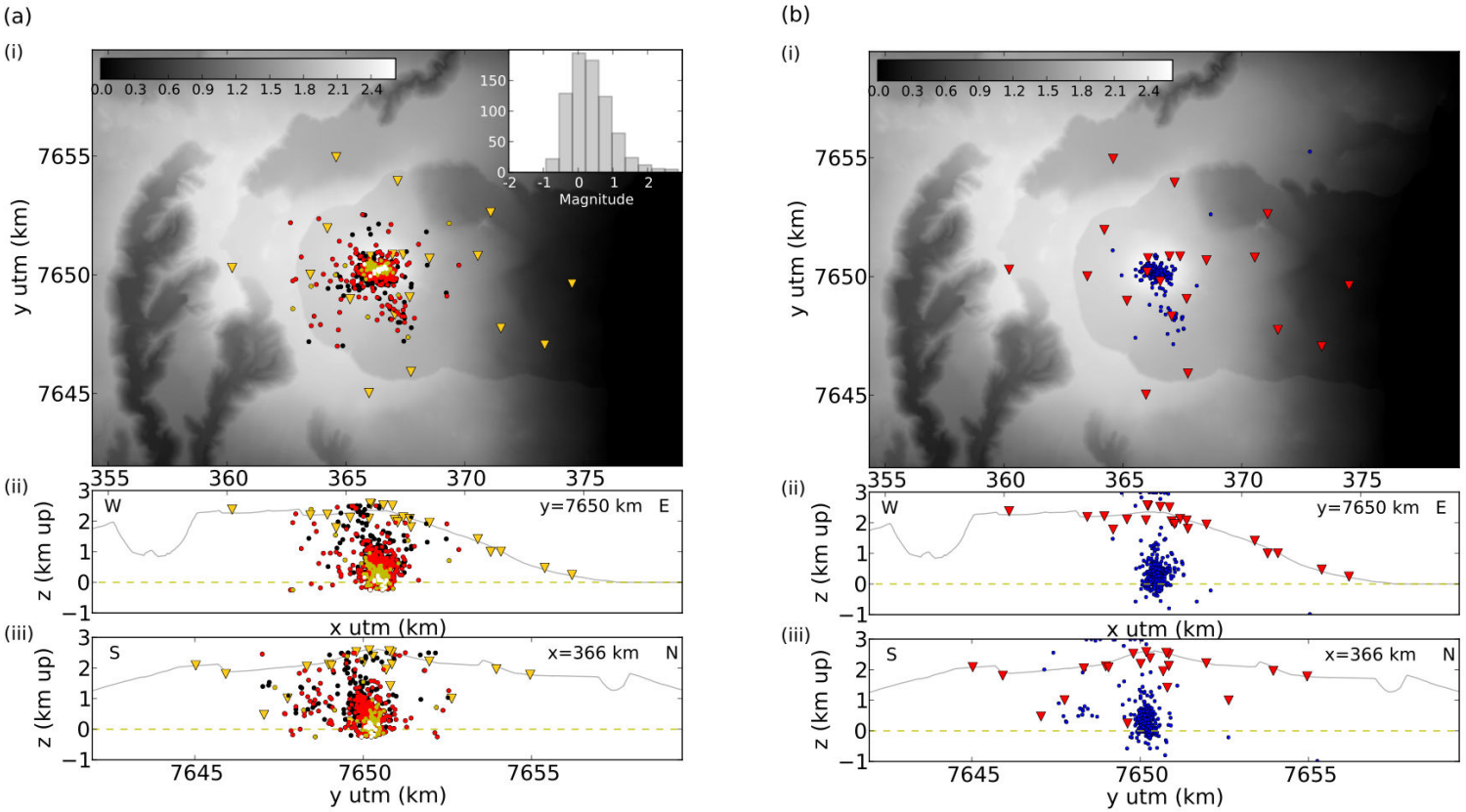


FIG. II.2.9: Cartes de localisation de la sismicité pour la crise pré-éruptive du 14 octobre 2010 dans les trois plans cartésiens ((i)-(iii)). (a) Localisations automatiques de Waveloc. (b) Localisations manuelles [Schmid, 2011]. Les stations du réseau sont représentées par les triangles inversés. La ligne pointillée sur les coupes (ii) et (iii) correspond au niveau de la mer. La coloration en niveaux de gris en (i) donne l'altitude en (km). L'échelle de couleurs utilisée pour représenter les événements en (a) dépend de l'intervalle de magnitude auquel ils appartiennent : noir pour $M_L < 0$, rouge pour $0 \leq M_L < 1$, jaune pour $1 \leq M_L < 2$ et blanc pour $M_L \geq 2$. L'histogramme de répartition des magnitudes est également affiché dans l'encart en haut à droite.

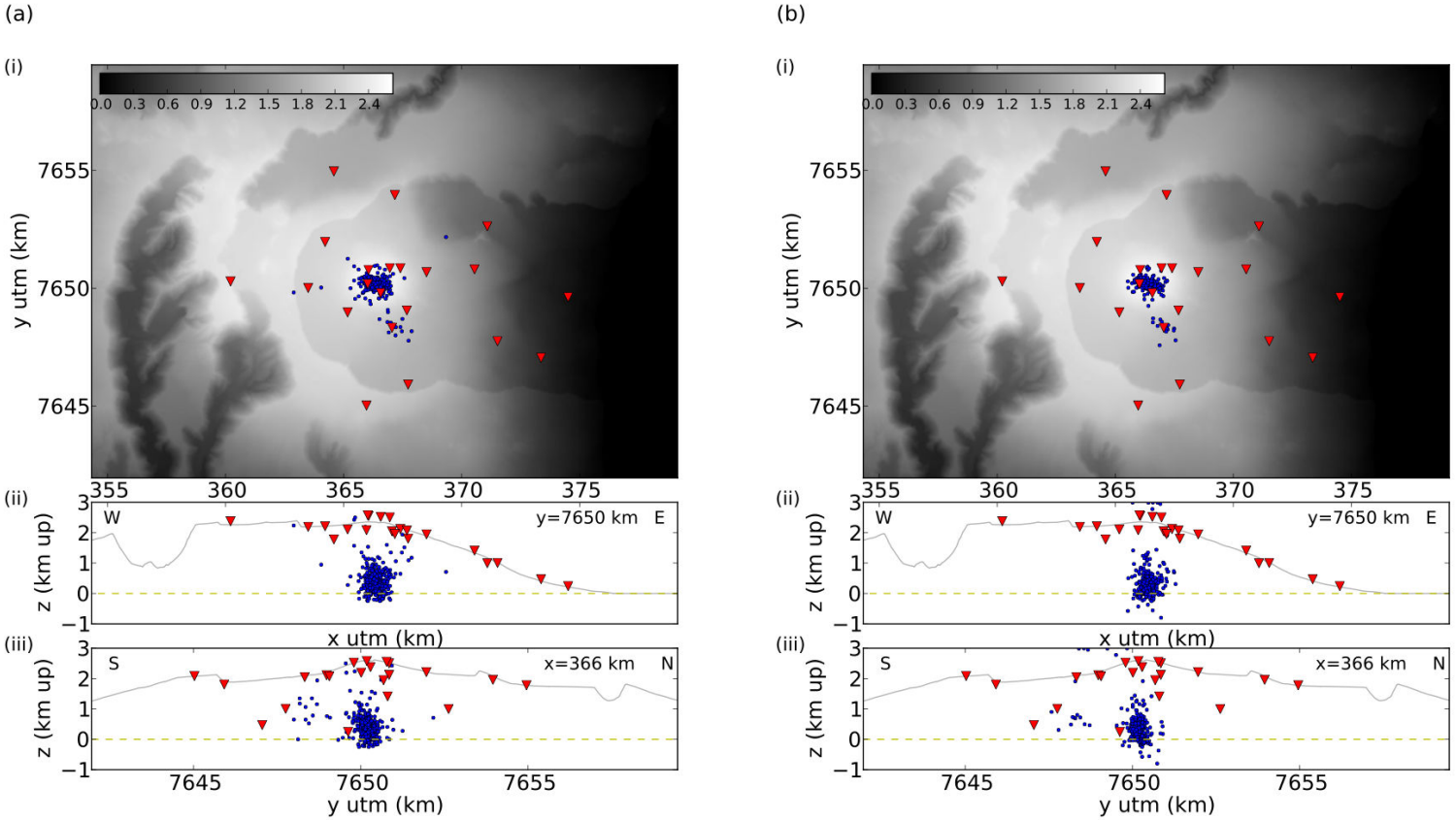


FIG. II.2.10: Cartes de localisation de la sismicité pour la crise du 14 octobre 2010 dans les trois plans cartésiens ((i)-(iii)). (a) Evénements communs au catalogue manuel représentés selon leurs localisations automatiques. (b) Evénements communs au catalogue manuel représentés selon leurs localisations manuelles. Description des sous-figures identique à la figure II.2.9.

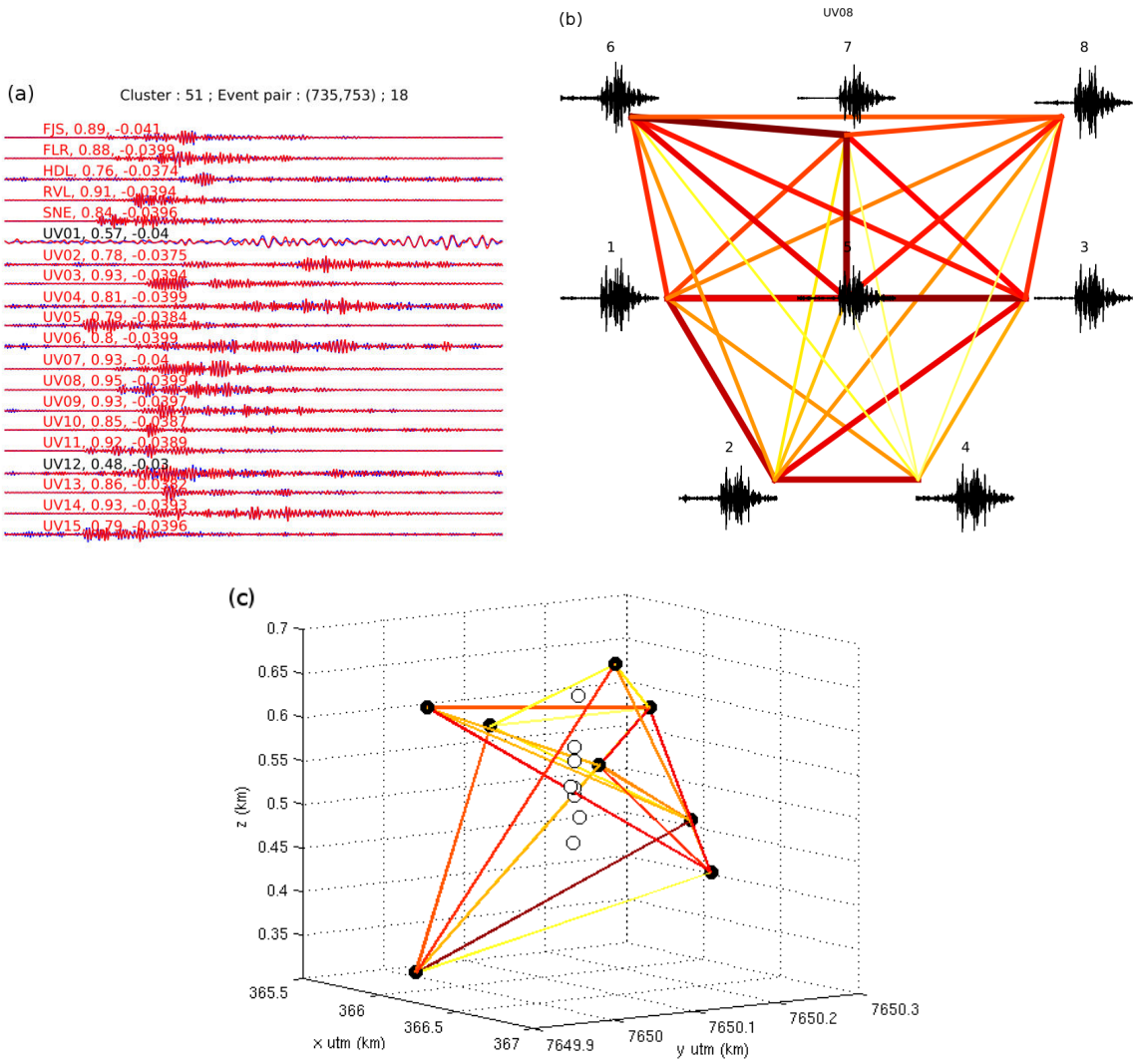


FIG. II.2.11: (a) Illustration de la corrélation entre deux événements de formes d'onde très similaires. Les signaux des deux événements sont superposés dans une fenêtre de 8 s. Les valeurs du coefficient de corrélation et du délai (en s) sont inscrites au-dessus de chaque trace. Le coefficient de corrélation dépasse la valeur 0.7 pour 18 stations sur 20. (b) Exemple d'un cluster de 8 événements et des relations internes entre-eux pour une station donnée. Les traces durent 7 s. La couleur et l'épaisseur de chacun des traits reliant un événement à un autre est fonction de leur degré de corrélation. Plus les événements sont corrélés, plus le lien est épais et de couleur foncée. L'échelle de couleur commence à la valeur 0.7. (c) Relocalisation des événements d'un même cluster. Les point noirs correspondent aux localisations initiales de Waveloc ; les points blancs aux relocalisations. La couleur des traits reliant les événements entre-eux dépend du nombre de stations pour lesquelles le coefficient de corrélation dépasse 0.7 : plus elle est foncée, plus ce nombre est élevé (et plus la corrélation est forte). L'échelle de couleur commence à 8 (stations).

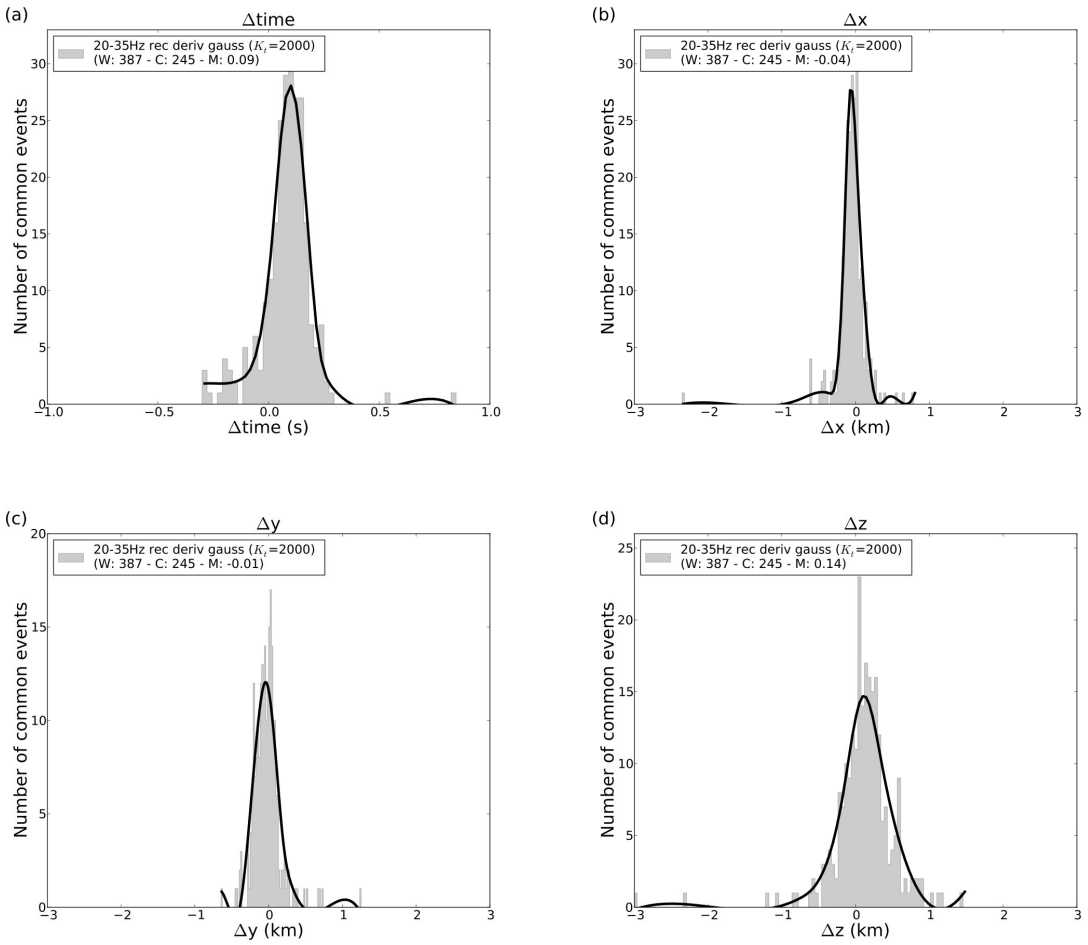


FIG. II.2.12: Histogrammes des événements communs montrant la précision sur chacun des paramètres hypocentraux après la relocalisation effectuée pour la crise éruptive du 14 octobre 2010.

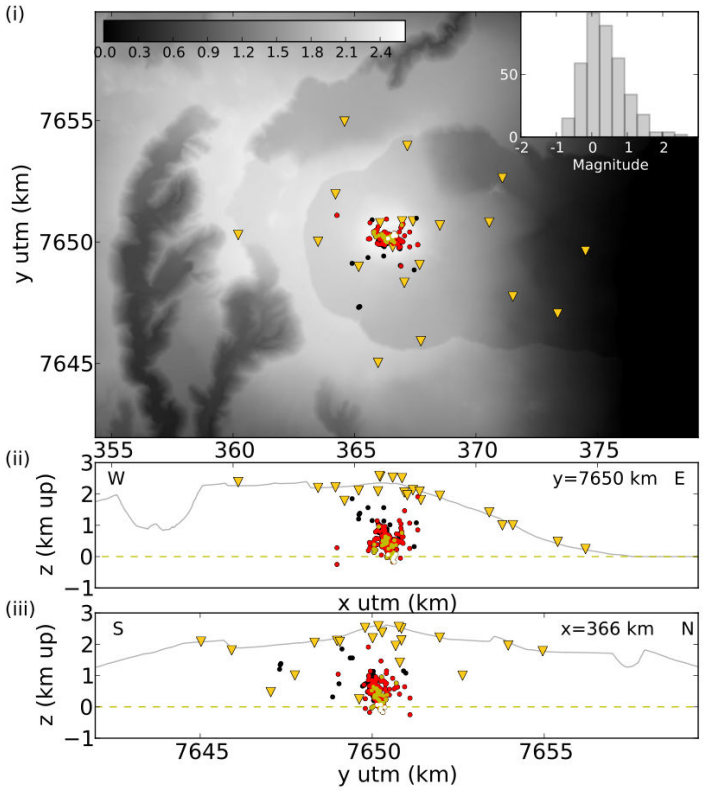


FIG. II.2.13: Carte de localisation de la sismicité pour la crise du 14 octobre 2010 après relocalisation. Description des sous-figures identique à la figure II.2.9.

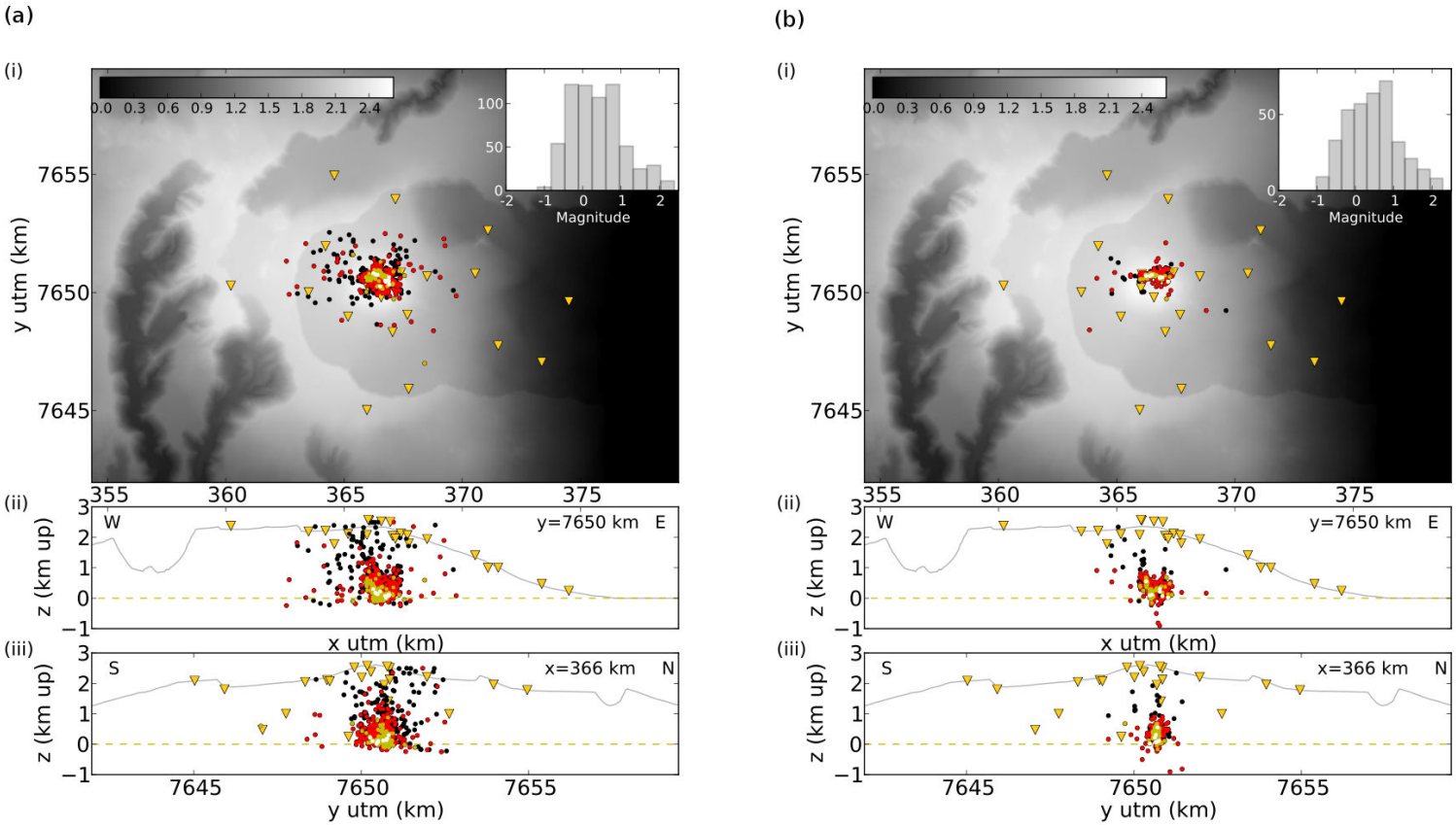


FIG. II.2.14: Cartes de localisation de la sismicité pour la crise intrusive du 23 septembre 2010. (a) Localisations automatiques "brutes". (b) Localisations après relocalisation. Description des sous-figures identique à la figure II.2.9.

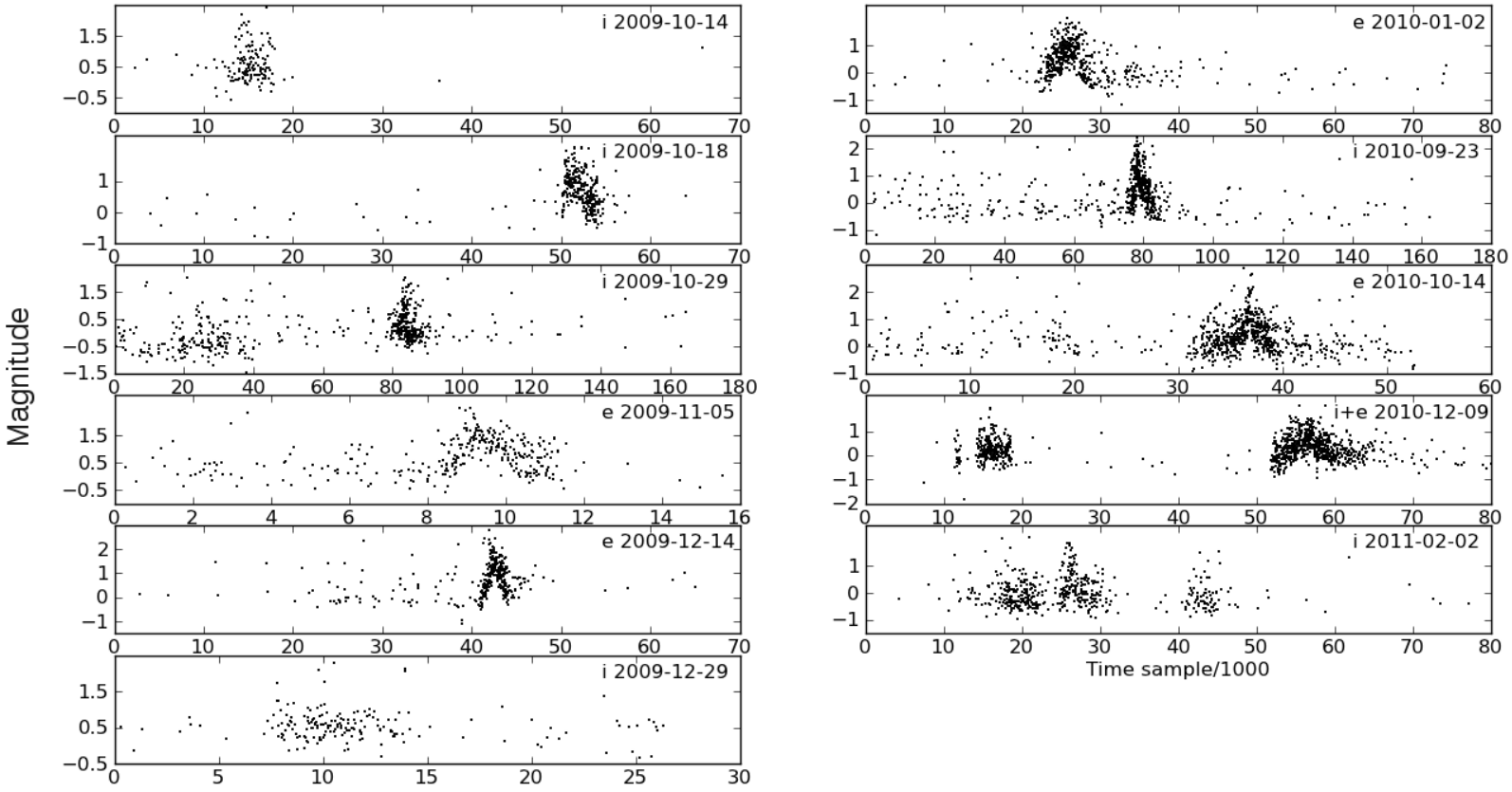


FIG. II.2.15: Evolution de la sismicité et des magnitudes au cours du temps pour chacune des crises.

II.2.4 Conclusion et discussion

Cet exemple d'application de Waveloc sur des données de microsismicité volcanique a permis de démontrer :

- la fiabilité de Waveloc : la comparaison avec les localisations manuelles disponibles a montré que la précision atteinte par les localisations automatiques était bonne (0.1 s sur les temps origine, quelques centaines de m sur les paramètres hypocentraux restants). Un soin particulier doit être apporté aux choix des différents paramètres de traitement des données (bande de filtrage, taille des fenêtres de calcul de la fonction caractéristique...) et de calcul de Waveloc (seuil de détection, critères de localisation - SNR...). Une inspection visuelle des événements localisés sur un échantillon de données permet un ajustement rapide. La comparaison avec des localisations manuelles constitue un plus indéniable. Les localisations peuvent encore être améliorées en procédant à une analyse plus détaillée de la sismicité par recherche des multiplets et relocalisation par double-différence.

- l'efficacité de Waveloc, en termes de temps de calcul notamment. Pour la crise du 14 octobre 2010, 16 h de données échantillonnées à 100 Hz sur 20 stations étaient disponibles. Les temps de calcul requis pour chacune des étapes sur un processeur de 2.67 GHz et 8 Go de RAM se décomposent de la manière suivante : le calcul des kurtosis non-récursifs sur une fenêtre glissante de 3 s nécessite 10 mn par station ; pour les kurtosis récursifs avec une fenêtre de 1 s, ce temps est réduit à 1 mn par station. L'étape de migration sur la grille contenant 9216 points dure environ 3 h. En comparant ce temps avec celui mesuré pour les données des autres crises, on estime que la migration de 5 h de données échantillonnées à 100 Hz prend environ 1 h. L'étape de localisation a une durée variable dépendant du seuil de détection choisi : plus celui-ci est bas, plus le temps s'allonge. Pour la crise du 14 octobre 2010, avec un seuil de détection fixé à 2000 permettant la détection d'un peu plus de 1200 événements, le calcul a pris 8 mn. La recherche des multiplets par inter-corrélation est évidemment fonction de la taille du catalogue considéré ($\frac{N!}{2!(N-2)!}$ paires où N est le nombre d'événements) et du nombre de stations du réseau : on estime, pour un catalogue avoisinant les 800 événements, un coût de 2 mn par station. Seules quelques minutes sont nécessaires pour les deux dernières étapes de clustering et relocalisation par double-différence. On arrive finalement à un temps de calcul de 7 h pour traiter l'ensemble des données de la crise du 14 octobre 2010. Ce temps est même réduit à 4h30 lorsque le calcul des kurtosis est mené récursivement. Ces temps sont à comparer avec ceux nécessaires à la localisation manuelle : le temps requis pour pointer environ 800 événements et localiser 450 d'entre-eux a été évalué à 2 ou 3 semaines. Dans les observatoires, où les données sont enregistrées en continu, un tel outil constituerait un gain de temps indéniable au traitement des données et à l'obtention de cartes de sismicités préalables. Des analyses plus rigoureuses de la sismicité sont toujours possibles par la suite...

Concernant la sismicité sur le Piton de la Fournaise, les résultats obtenus pour les différentes crises montrent clairement une zone sismogène principale située entre 0 et 1000 m d'altitude à l'aplomb du cratère (X=366 km,Y=7650 km). Cette zone est systématiquement activée lors des épisodes de crises, que cela donne lieu à une éruption ou non. Les magnitudes

locales montrent aussi que des événements de plus forte magnitude se produisent lors des crises et qu'ils se localisent toujours au centre de l'essaim.

CHAPITRE II.3

Classification automatique

II.3.1 Présentation du jeu de données

Dans l'introduction de cette partie, on avait mentionné les deux types principaux d'événements sismiques enregistrés sur le Piton de la Fournaise : les volcano-tectoniques (ou VT) et les éboulements (ou EB). Ces derniers ont fait l'objet d'une étude plus approfondie menée par C. Hibert dans le cadre de sa thèse [Hibert, 2012]. Entre autres, une classification automatique des VT et des EB grâce à une technique de logique floue a déjà été proposée et appliquée avec succès.

Dans ce chapitre, on travaillera sur le même jeu de données que celui utilisé par C. Hibert. Celui-ci est constitué de deux jeux :

- le *training set*, composé de 191 événements dont 124 EB et 67 VT.
- le *test set*, composé de 7482 événements dont 4133 EB et 3349 VT.

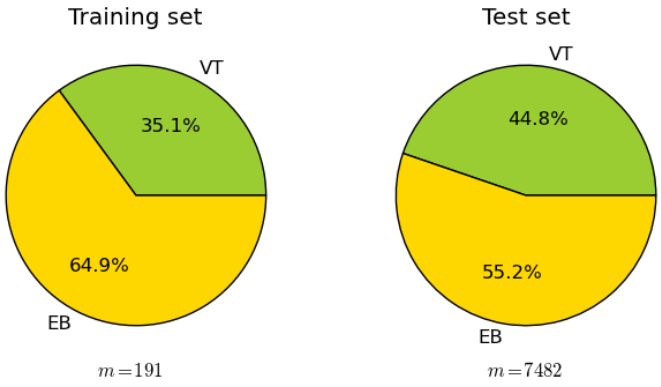


FIG. II.3.1: Diagrammes de répartition du *training set* et du *test set*.

Les signaux correspondants à chaque événement ont été enregistrés sur l'une des stations permanentes de l'OVPF, la station BOR (3 composantes), et ont une durée de 270 s pour le *training set* et 200 s pour le *test set*.

Il s'agit d'un très bon jeu de données. Le *training set* a été créé soigneusement. Cependant, les PDFs des attributs montrés dans les figures de la section I.2.4 ont été calculées sur ce jeu de données et ont mis en évidence quelques disparités entre le *test set* et le *training set* : les PDFs ne sont pas toujours superposables. De plus, on remarque que les EB sont présents en plus grand nombre dans le *training set* (en proportion de la taille du jeu de données). On verra par la suite que ce *training set* permet tout de même d'obtenir d'excellents résultats ; c'est donc celui qui sera utilisé tout au long de cette étude.

Ce jeu de données fourni comprend aussi les valeurs des 5 attributs calculés par C. Hibert [Hibert et al., 2014] :
- le kurtosis de l'enveloppe (noté Kurto) ;
- la durée du signal (notée Dur) ;
- le rapport des durées des phases de croissance et de décroissance de l'enveloppe (noté AsDec) ;
- le rapport du maximum sur la moyenne de l'enveloppe (noté RappMaxMean) ;
- l'énergie du signal dans la bande 20-30 Hz (noté Ene).

La description des attributs a été donnée dans la partie I.2.4. Ici, ils ont tous été normalisés par le logarithme, sauf la durée.

Dans ce chapitre, on présentera d'abord les résultats obtenus avec la régression logistique et la SVM en utilisant les 5 attributs sus-cités ; puis on les comparera avec ceux de la méthode de logique floue mise en œuvre par Hibert et al. [2014]. On testera ensuite nos deux algorithmes en recalculant les attributs utilisés par Hibert et al. [2014], et en en ajoutant de nouveaux (voir leur description en partie I.2.4 de ce manuscrit ; Tab. I.2.2).

II.3.2 Résultats avec les 5 attributs fournis par Hibert (2014)

Dans cette section, le nombre d'attributs étant limité, on s'attachera à bien présenter les résultats obtenus pour un attribut seul ou pour plusieurs combinaisons d'attributs.

II.3.2.1 Avec un seul attribut

La classification effectuée avec un seul attribut va permettre d'identifier le ou les attributs qui ont un bon pouvoir séparateur. Les résultats sont présentés sous forme de figures. La sous-figure (a) des figures II.3.2a, II.3.2b, II.3.3a représente l'histogramme des valeurs de l'attribut considéré :
- en vert, les valeurs correspondant aux VT du *training set* ;
- en bleu, les valeurs correspondant aux EB du *training set* ;
- en tiretés noirs, les valeurs correspondant aux événements bien classés du *test set* ;
- en tiretés rouges, les valeurs correspondant aux événements mal classés du *test set*.

Les pourcentages de classification associés à chaque histogramme sont inscrits en haut à gauche de la figure et correspondent aux résultats de la régression logistique. La fonction hypothèse dont les paramètres sont déterminés lors de la régression logistique est tracée en jaune ; la fonction hypothèse « équivalente » pour la SVM est tracée en magenta. Les traits verticaux orange (régression logistique) et violet (SVM) correspondent aux séparateurs. Les

sous-figures (b) et (c) montrent les matrices de confusion obtenues pour le *test set*, respectivement avec la régression logistique et la SVM. On rappelle qu'une matrice de confusion donne les pourcentages d'événements bien classés pour chaque classe sur la diagonale et la répartition des événements mal classés dans les autres classes (§I.2.2.4). Pour la figure II.3.3b, on a uniquement représenté l'histogramme et les PDFs de l'attribut en distinguant les EB (bleu) et les VT (vert) du *training set* et les EB (noir) et les VT (blanc) du *test set*.

Les histogrammes des EB et VT pour AsDec (Fig. II.3.2a) montrent que les valeurs prises par les VT sont entièrement superposées par celles des EB. Pour les EB, les valeurs occupent un plus large intervalle. Elles sont aussi plus élevées que pour les VT, ce qui s'accorde avec le fait que les EB sont moins impulsifs que les VT. La régression logistique permet tout de même de classer 80% des événements correctement et ce, dans des proportions équivalentes (80% pour les VT et 80% pour les EB). La majorité des événements mal classés (histogramme rouge tireté) se situe dans la zone de superposition observée.

Les valeurs de durée prises par les EB et les VT montrent un comportement différent (Fig. II.3.2b) :

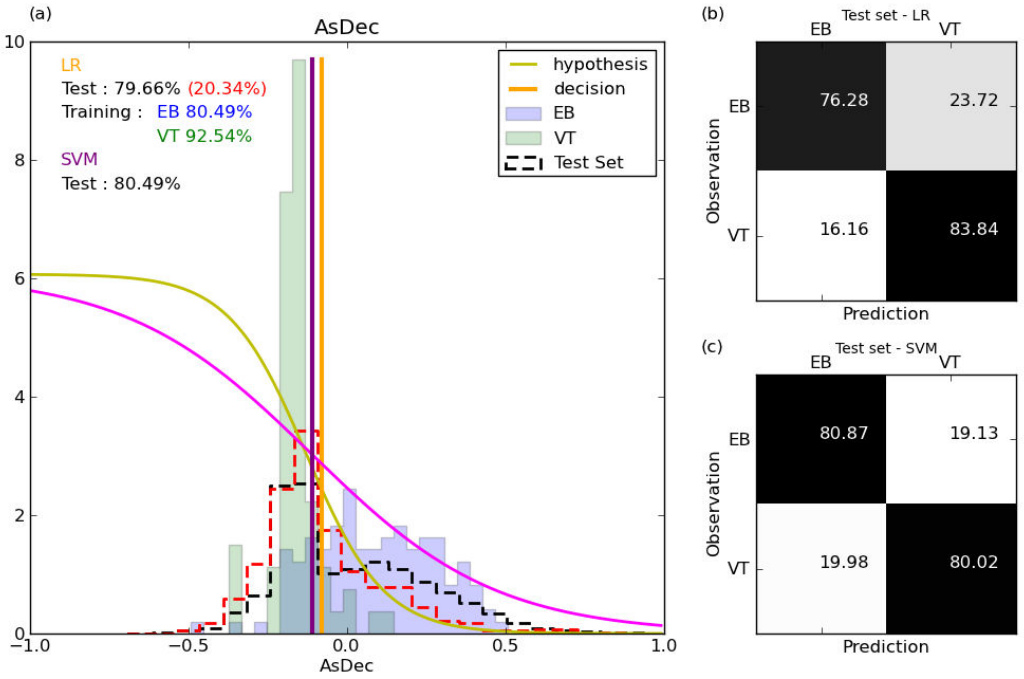
- les VT sont de courte durée et l'étendue des valeurs prises pour l'ensemble des VT est relativement restreinte.
- les EB durent plus longtemps et prennent des valeurs plus variables.

Comme pour AsDec, on note que les valeurs des durées des VT se recoupent entièrement avec celles des EB, ce qui n'empêche cependant pas de classer correctement les trois-quarts du *test set*. Sans surprise, l'essentiel des événements mal classés correspond à de faibles valeurs de durée. Les VT sont légèrement mieux classés que les EB (72% et 79% respectivement).

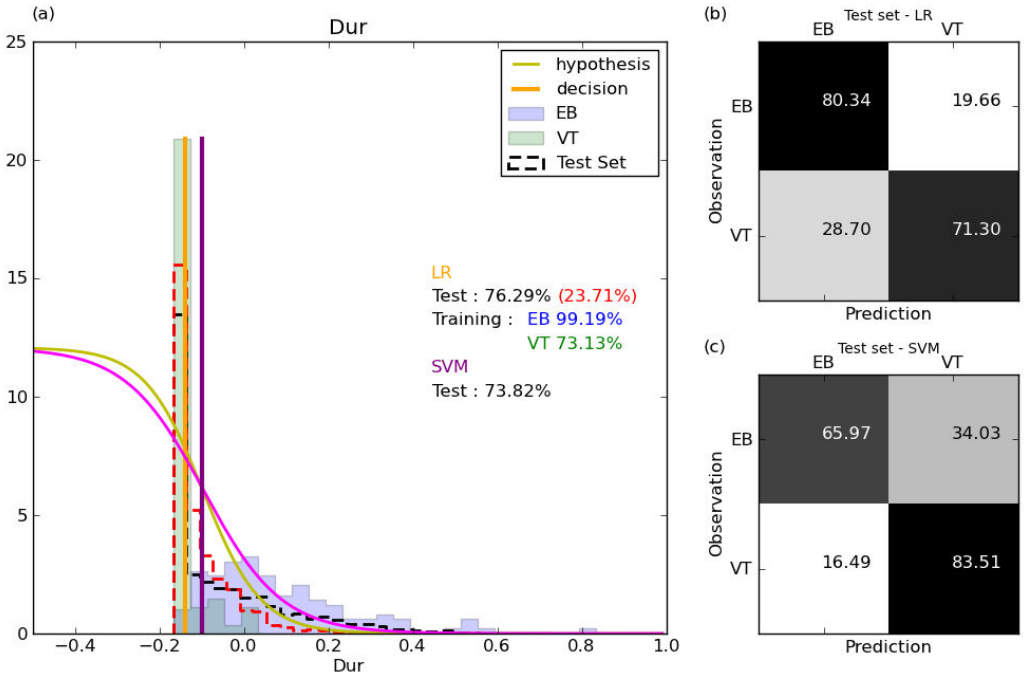
L'examen des variations de chaque attribut associé à un autre a permis de mettre en évidence la corrélation entre deux des attributs : le kurtosis et le rapport du maximum sur la moyenne de l'enveloppe (Fig. II.3.4), qui donnent tous les deux une indication sur la forme du signal. Dans ce qui suit, on utilisera désormais un seul et unique attribut résultant d'une combinaison linéaire entre les deux attributs d'origine. Il sera noté KRapp.

Les résultats obtenus pour ce nouvel attribut (Fig. II.3.3a) montrent que la séparation entre EB et VT se fait très bien : plus de 90% des événements sont bien classés et la majorité des événements mal classés sont situés dans la "zone de transition" entre les histogrammes des EB et des VT.

On ne présente pas les résultats obtenus pour l'énergie entre 10 et 30 Hz, car cet attribut n'est pas suffisamment discriminant pris seul. Sur la figure II.3.3b, on voit que les valeurs prises par les EB sont toujours confondues avec celles des VT. Pour les VT, on observe 2 "pics" principaux dans la distribution. On remarque également que les valeurs prises par le *test set* sont limitées à un intervalle différent par rapport à celle du *training set*, ce qui explique en partie pourquoi il n'est pas possible de bien classer les événements avec ce seul attribut (problème de représentativité). De plus, compte tenu de l'aspect des distributions, il faudrait plutôt chercher à calculer une fonction hypothèse de degré polynomial supérieur à 1.

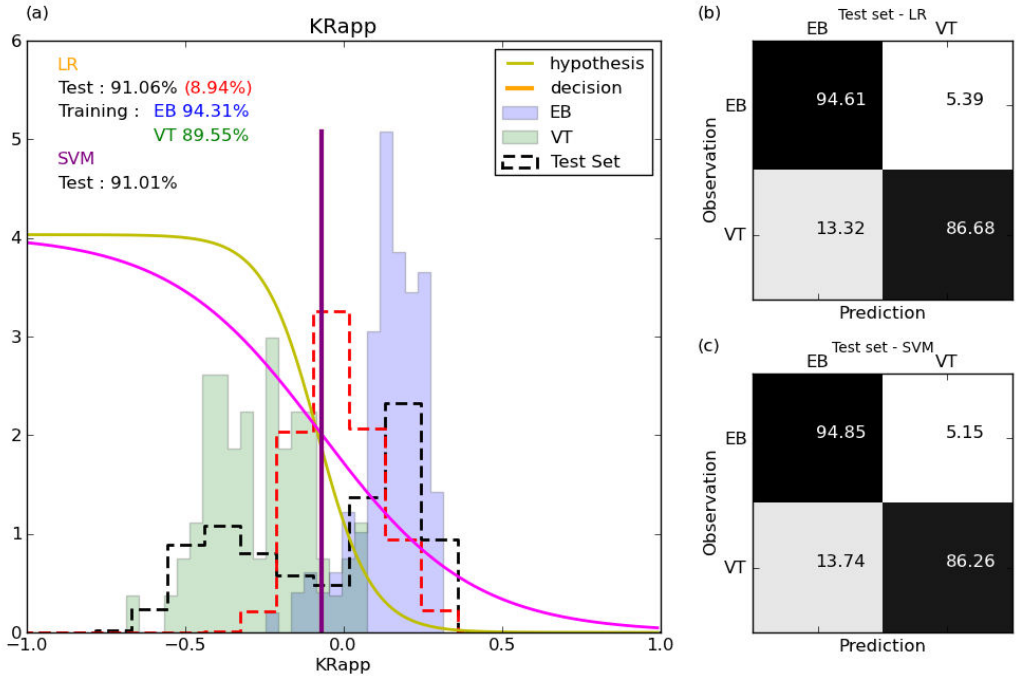


(a) AsDec

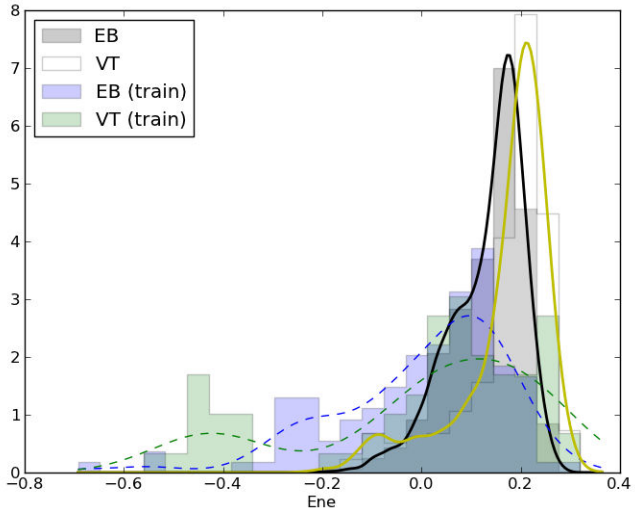


(b) Durée

FIG. II.3.2: Histogrammes et résultats de la classification pour AsDec et Dur. Détails des sous-figures dans le texte.



(a) KRapp



(b) Energie entre 10 et 30 Hz

FIG. II.3.3: Histogrammes et résultats de la classification pour KRapp et Ene. Détails des sous-figures dans le texte.

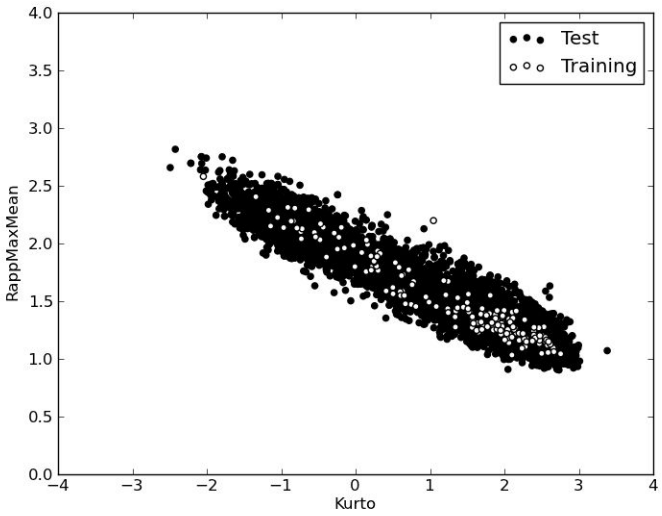


FIG. II.3.4: Corrélation entre deux attributs : kurtosis et rapport du maximum sur la moyenne de l'enveloppe.

Discussion : choix du seuil

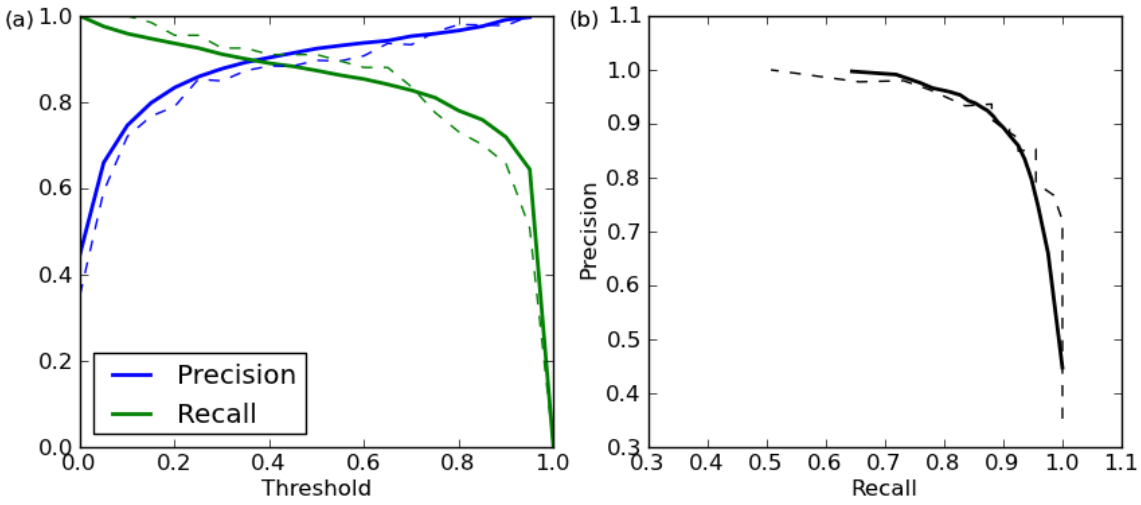


FIG. II.3.5: (a) Précision (bleu) et rappel (vert) pour le *test set* (courbes pleines) et le *training set* (courbes pointillées) en fonction du seuil fixé. (b) Courbe de précision en fonction du rappel. Courbes calculées pour la classification avec l'attribut KRapp.

Dans la section I.2.2.1, on avait mentionné l'importance du choix du seuil afin de trouver le meilleur équilibre possible entre la précision et le rappel. La figure II.3.5 est un exemple pris lors de la classification effectuée avec le seul attribut KRapp. En (a), pour les données du *training set*, on voit que le seuil optimal doit être proche de 0.5 (intersection des 2 courbes de

précision et de rappel). On voit qu'il est un peu moins élevé si l'on s'en tient uniquement aux données du *test set* (autour de 0.4).

Dans notre problème de classification, un seuil élevé revient à classer plus d'EB au détriment des VT et inversement. Le seuil est calculé automatiquement à partir des données du *cross-validation set* (qui représente 20 % du *training set*).

Bilan

Les résultats présentés dans cette section montrent que KRapp est l'attribut le plus discriminant et qu'il permet déjà de fournir une excellente classification. On peut espérer améliorer encore un peu le pourcentage de réussite en utilisant plusieurs attributs.

II.3.2.2 Avec diverses combinaisons d'attributs

Dans cette section, on utilise des combinaisons d'attributs. Il est notamment intéressant de voir comment la classification évolue en combinant les attributs les moins discriminants a priori (AsDec, Dur, Ene). Pour KRapp, on s'attend à une amélioration légère des résultats. Les résultats sont consignés dans le tableau II.3.1. Ils sont donnés pour la régression logistique et la SVM linéaire. Il est inutile d'appliquer la SVM non-linéaire : les résultats sont quasiment identiques à ceux de la régression.

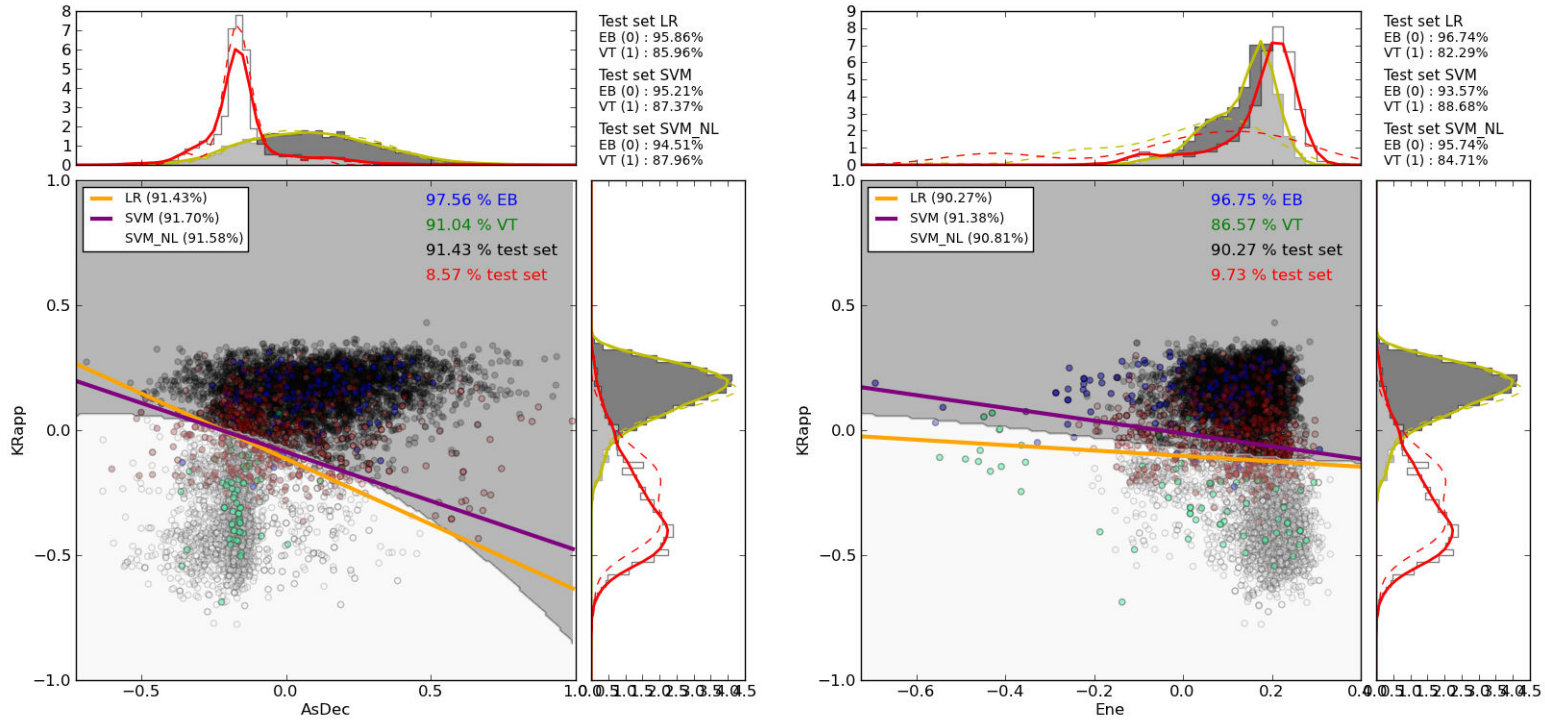
	Attribut(s)	Régression logistique		SVM linéaire	
		% training	% test set	% training	% test set
1 attribut	AsDec	84±0%	79±1%	≈ 84%	≈ 80%
	Dur	89±5%	74±3%	≈ 87%	≈ 74%
	Ene	35%	45%	≈ 65%	≈ 55%
	Kurto	92±1%	90±1%	≈93%	≈91%
	RappMaxMean	91±1%	88±1%	≈91%	≈90%
	KRapp	92±2%	90±1%	≈ 93%	≈91%
2 attributs	AsDec + Dur	95±3%	84±2%	≈97%	≈ 86%
	AsDec + Ene	84±1%	80±1%	≈ 85%	≈ 79%
	AsDec + KRapp	94±1%	91±1%	≈ 94%	≈ 92%
	Dur + Ene	88±5%	74±3%	≈ 84%	≈ 71%
	Dur + KRapp	96±2%	90±2%	≈ 99%	≈ 92%
	Ene + KRapp	93±1%	90±1%	≈ 92%	≈ 90%
3 attributs	AsDec + Dur + Ene	97±1%	86±1%	≈ 98%	≈ 87%
	AsDec + Dur + KRapp	98±1%	90±2%	100%	≈ 92%
	AsDec + Ene + KRapp	95±1%	91±0%	≈ 96%	≈ 91%
	Dur + Ene + KRapp	97±1%	92±0%	≈ 99%	≈ 93%
4 attributs	AsDec + Dur + Ene + KRapp	98±1%	92±1%	100%	≈ 94%

TAB. II.3.1: Résultats de la classification obtenus pour différentes combinaisons d'attributs pour la régression logistique et la SVM linéaire. Les pourcentages de classification par classe sont donnés dans les matrices de confusion.

Combinaison de 2 attributs

Sans surprise, les meilleurs résultats de classification ($> 90\%$ sur le *test set*) sont obtenus pour les trois combinaisons d'attributs impliquant KRapp ; les deux moins bonnes combinaisons ($< 80\%$) impliquent Ene.
La comparaison des résultats avec ceux obtenus pour les attributs seuls permet de voir que, dans tous les cas, si Ene n'intervient pas, les combinaisons d'attributs améliorent la classification. Ceci est particulièrement visible pour AsDec et Dur, qui sont les deux attributs "moyens" (ni médiocres comme Ene, ni excellents comme KRapp) : on passe de moins de 80% à 84%.
Afin de mieux visualiser le séparateur, on présente les résultats obtenus dans les figures II.3.6 et II.3.7. Chaque sous-figure représente les valeurs du premier attribut en fonction des valeurs du deuxième attribut. Le code couleur n'a pas changé par rapport à celui de la section précédente (histogrammes). La droite orange symbolise le séparateur dont les paramètres ont été déterminés grâce à la régression logistique ; la violette à celui de la SVM. Les pourcentages de classification affichés et la couleur de fond qui cartographie la séparation des deux classes sont ceux de la régression logistique.

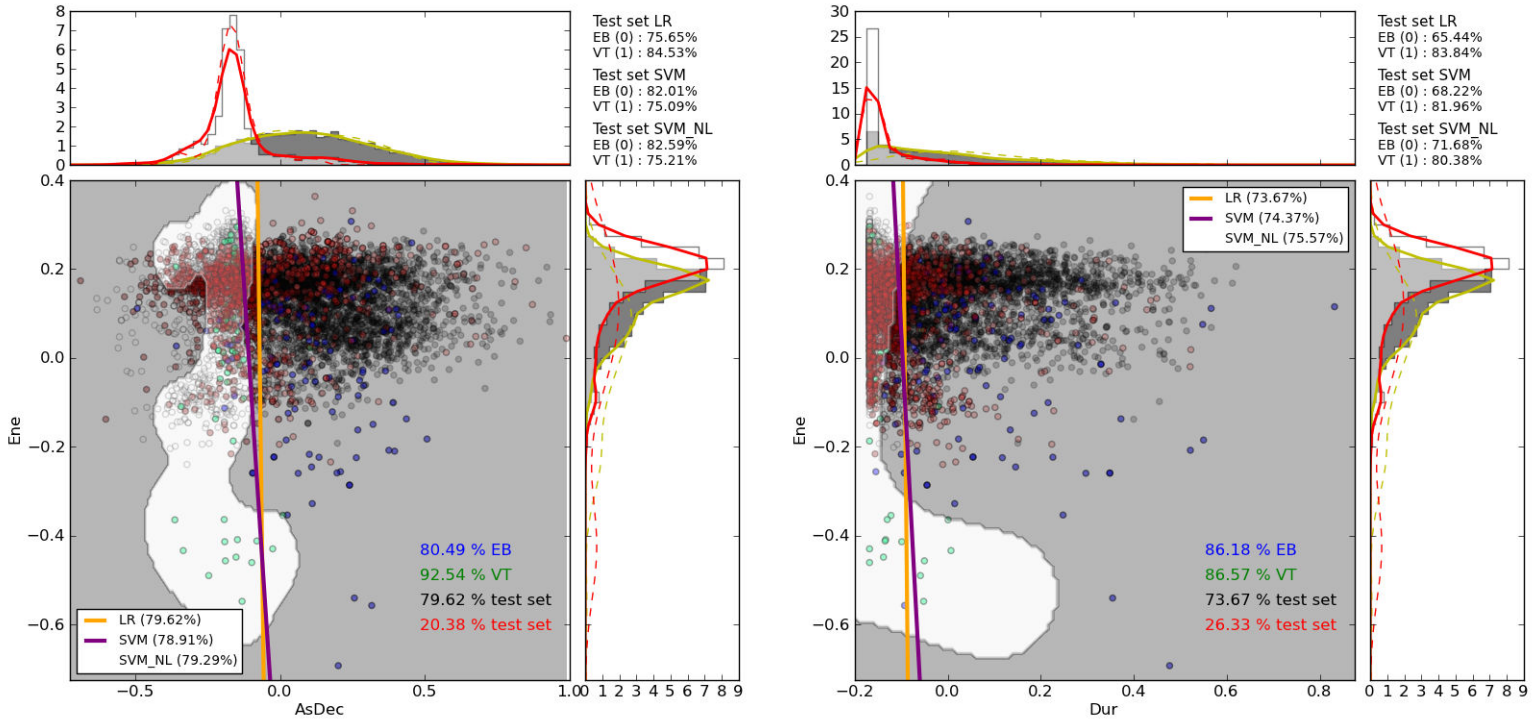
On voit sur la figure II.3.6 que la séparation des EB et des VT dans le *training set* est très nette et ne pose pas de difficultés. L'essentiel des événements du *test set* mal classés par la suite se situe dans la zone de transition entre les 2 classes. Sur la figure II.3.7, en revanche, la distinction entre EB et VT s'avère beaucoup moins aisée, en raison notamment du recouvrement des PDFs pour les attributs des différentes classes. Il est beaucoup plus difficile dans ces conditions de trouver un séparateur optimal : une grande partie des EB est classée à tort dans la classe des VT.



(a) AsDec et KRapp

(b) Ene et KRapp

FIG. II.3.6: Séparateurs calculés pour les deux meilleures combinaisons d'attributs. La figure centrale représente les valeurs prises par chaque attribut pour le *training set* (VT en vert, EB en bleu) et le *test set* (classification correcte en noir pour les EB et blanc pour les VT ; mauvaise classification en rouge). Les pourcentages de classification obtenus pour la LR sont inscrits en bleu et vert, repectivement pour les EB et les VT du *training set* et en noir pour l'ensemble du *test set*. Les surfaces de décision de la LR et de la SVM sont représentées par les droites continues, respectivement en orange et violet. Les pourcentages de classification correcte par classe pour le *test set* et pour les deux méthodes sont inscrits dans le coin en haut à droite. La couleur de fond cartographie la séparation des VT (blanc) et des EB (noir) en fonction des valeurs prises par les deux attributs d'après les résultats de la SVM non-linéaire. Les histogrammes et PDFs de chacun des attributs sont représentées de part et d'autre de la figure.



(a) AsDec et Ene

(b) Dur et Ene

FIG. II.3.7: Séparateurs calculés pour diverses les deux moins bonnes combinaisons d'attributs. Pour la description, voir la figure II.3.6

Combinaison de 3 attributs

Lorsque l'on combine les attributs par 3, on note une plus grande constance dans les résultats (rien en-dessous de 85%). La meilleure combinaison s'avère être celle impliquant Dur, Ene et KRapp. Le fait que AsDec n'en fasse pas partie peut paraître surprenant car c'est un meilleur attribut que Ene pris seul. Mais on avait déjà vu dans le paragraphe précédent, pour les paires d'attributs, que la combinaison de KRapp et Ene fonctionnait bien...
Globalement, la combinaison AsDec et Dur avec un autre attribut est celle qui marche le moins bien. On voit toutefois que le problème est surtout liée à la généralisation au *test set* : si on s'en tient aux pourcentages de réussite obtenus sur le *training set*, les résultats sont supérieurs à ceux de la combinaison AsDec, Ene, KRapp (ce qui n'est plus le cas pour le *test set*). Le problème se posait déjà pour les paires d'attributs, voire pour les attributs seuls (mauvaise généralisation pour Dur par rapport à AsDec).
Ces observations tendent à montrer que les combinaisons d'attributs sont très importantes, et que ce ne sont pas forcément les meilleurs attributs combinés ensemble qui donneront les meilleurs résultats. Ils montrent également que le *training set* doit être le plus représentatif possible du *test set* pour éviter les problèmes de généralisation.
La figure II.3.8 présente les résultats de la combinaison de AsDec, Dur et KRapp. Le plan trouvé grâce à la régression logistique sépare bien les données du *training set*.

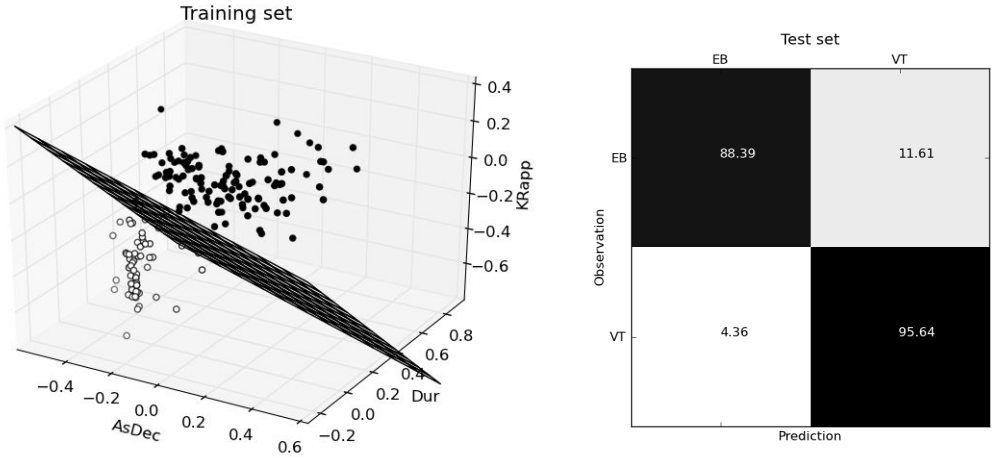
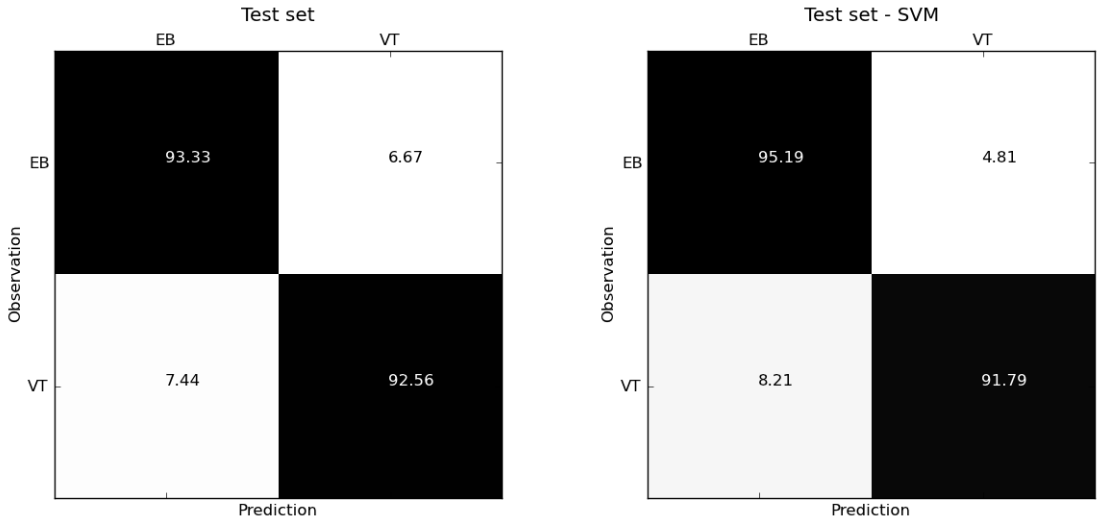


FIG. II.3.8: Résultats pour Asdec, Dur et KRapp. **(Gauche)** Plan séparateur trouvé avec les données du *training set*. **(Droite)** Matrice de confusion associée (données du *test set*).

Avec les 4 attributs

On combine maintenant l'ensemble des attributs disponibles, soit 4 au total. Le tableau II.3.1 nous apprend que c'est la meilleure combinaison possible, que ce soit pour le *training set* ou le *test set*, puisqu'on atteint respectivement les 98% et 92% de réussite pour la régression logistique ; et 100% et 94% pour la SVM.
Ce résultat paraît logique puisque l'on essaie de classer les événements en gardant le maximum d'information les concernant. Cependant, il aurait pu en être autrement car on a vu que les combinaisons d'attributs n'apportent pas nécessairement des améliorations.

Les matrice de confusion obtenues pour les deux méthodes de discrimination sont représentées sur la figure Fig. II.3.9. Elles montrent que les EB sont légèrement mieux classés que les VT.



(a) Régression logistique

(b) SVM linéaire

Fig. II.3.9: Matrices de confusion obtenues lorsque l'on utilise les 4 attributs disponibles pour la régression logistique **(gauche)** et la SVM linéaire **(droite)**.

II.3.2.3 Conclusion

La méthode de régression logistique pour la séparation des EB et des VT paraît très bien fonctionner. La comparaison des classifications obtenues avec celles de l'OVPF montrent qu'on est en mesure de classer correctement 92% des événements (et au maximum 94%).

Si l'on compare nos résultats avec ceux de Hibert et al. [2014] sur le même jeu de données (Tab. II.3.2), on remarque qu'ils sont très légèrement supérieurs à ceux de la logique floue (92 vs 91%).

On constate également que globalement, pour la régression logistique, les taux de classification des EB et des VT sont toujours proches et leur différence n'excède pas les 10%, ce qui n'est pas le cas de la logique floue et de la SVM où les écarts de classification entre les deux classes peuvent atteindre 20%.

Finalement, ces premiers résultats montrent que la régression logistique et la SVM s'avèrent être des méthodes de discrimination satisfaisantes.

Une autre chose importante à souligner est le fait qu'il est difficile de "prévoir" les résultats d'une combinaison d'attributs. On l'a vu avec l'attribut énergétique : il est peu discriminatoire seul, mais certaines de ses associations sont bonnes, voire très bonnes. Ceci est important pour la suite, où le nombre d'attributs considéré est plus important et où il devient moins évident de tout regarder de manière aussi détaillée.

Enfin, on remarque qu'ici, la taille du *training set* est relativement restreinte par rapport à celui du *test set*. Comme le problème de discrimination se limite à deux classes facilement

séparables, cela ne semble pas poser de problème.

Attribut(s)	Logique floue			Régression logistique			SVM linéaire		
	VT	EB	Total	VT	EB	Total	VT	EB	Total
AsDec	82%	77%	80%	79±6%	80±3%	79±1%	80%	81%	80%
Kurto	80%	84%	82%	86±5%	93±5%	90±1%	86%	95%	91%
RappMaxMean	77%	97%	88%	91±4%	86±5%	88±1%	81%	94%	88%
Durée	84%	65%	73%	79±7%	71±11%	74±3%	85%	65%	74%
Tous	89%	92%	91%	95±2%	86±6%	90±2%	94%	90%	92%

TAB. II.3.2: Résultats obtenus par C. Hibert sur le même jeu de données que nous, en utilisant une technique de logique floue.

Dans la suite, on va s'intéresser aux résultats que l'on peut obtenir en utilisant un plus grand nombre d'attributs calculés automatiquement (voir §I.2.4, TAB. I.2.2) pour le même jeu de données.

II.3.3 Résultats avec les attributs décrits en I.2.4, p. 49

Les résultats sont consignés dans le tableau II.3.3. Les démarches effectuées pour obtenir ces résultats sont détaillées dans les sections qui suivent.

Comme on effectue des découpages aléatoires dans le *training set* en *sub-training set* (60%), *cross-validation set* (40%) et *test set* (20%)pour la régression logistique (voir §I.2.2.1, p.38), on introduit inéluctablement au moins une petite variabilité sur les résultats. Pour la quantifier, on effectue dix tirages aléatoires au sein du *training set* initial : les résultats dans le tableau correspondent donc à la moyenne des pourcentages obtenus pour les dix tirages suivi de l'écart-type (±). On étudiera aussi un peu plus loin dans cette section l'influence de cette variabilité sur l'aspect des PDFs.

	Méthode	Training set	Test set		
			Global	EB	VT
1	Attributs de Hibert (2014) recalculés				
	LR	95±1%	88±1%	88±4%	87±3%
	SVM linéaire	≈90%	≈90%	≈88%	≈92%
2	Tous les attributs sans tables de hachage				
	LR	96±2%	88±2%	84±5%	93±3%
	SVM linéaire	100%	≈88%	≈91%	≈84%
	SVM non linéaire	100%	≈88.5%	≈91%	≈86%
3	Tous les attributs, sauf ceux de Hibert (2014) et tables de hachage				
	LR	93±3%	87±2%	81±5%	94±2%
	SVM linéaire	≈96%	≈91%	≈90%	≈92%
4	Tables de hachage seules				
	LR	89±4%	83±2%	82±8%	84±6%
	SVM linéaire	≈93%	≈85%	≈88%	≈81%
5	Tables de hachage + Dur + KRapp				
	LR	96±1%	86±1%	80±3%	93±2%
	SVM linéaire	≈95%	≈89%	≈91%	≈88%
6	Tous les attributs et tables de hachage				
	LR	97±1%	90±1%	86±3%	94±2%
	SVM linéaire	≈98%	≈92%	≈91%	≈92%

TAB. II.3.3: Résultats de la classification pour différents attributs et combinaisons d'attributs.

II.3.3.1 Calcul des 5 attributs définis par Hibert (2014)

Dans un premier temps, on a recalculé les 5 attributs utilisés par Hibert et al. [2014] à partir des formes d'ondes fournies, ce qui devrait permettre de valider (ou pas) les attributs et les manières de les calculer (voir §I.2.4 pour les détails).

La comparaison des PDFs (FIG. II.3.10(a-e)) montre quelques différences qui méritent d'être soulignées :

- pour AsDec (a), les intervalles de valeurs occupés par les EB et les VT correspondent globalement bien, mais la séparation est moins nette avec nos calculs car les PDFs sont plus étendues. En essayant avec une enveloppe plus lissée, on a constaté que les PDFs étaient plus étroites, mais quasiment identiques pour VT et EB.

- pour Dur (b), les PDFs sont très proches.

- pour Ene (c), on voit que la séparation entre les valeurs des EB et des VT est toujours délicate. Les gammes de valeurs ne sont pas tout à fait les mêmes (décalage vers des valeurs plus grandes), mais surtout, on n'observe pas le même comportement pour les EB et les VT... : dans un cas, les VT sont plus énergétiques dans la bande 10-30 Hz que les EB ; dans l'autre, c'est l'inverse. Comme l'attribut n'est pas très discriminant à la base, ce n'est pas vraiment problématique... Le calcul de l'énergie dans une autre

bande de fréquence (5-10 Hz, voir (f)) se révèle avoir un très bon pouvoir discriminant, avec des EB contenant plus d'énergie dans cette bande basse-fréquence que les VT (voir les spectrogrammes de la figure I.2.29, p.60).

- pour Kurto (d), on observe des comportements très différents, avec dans un cas [Hibert et al., 2014] des valeurs de kurtosis plus élevées pour les EB que les VT ; et l'inverse dans l'autre cas... Intuitivement, les EB générant des signaux plus émergents que les VT, on s'attend à obtenir des valeurs plus élevées pour les VT. Le pouvoir discriminatoire de l'attribut recalculé est moins bon que celui de Hibert et al. [2014]. Pour essayer de comprendre d'où pouvait provenir la divergence observée, on a testé l'influence du lissage de l'enveloppe (qui a un effet direct sur l'écart-type des amplitudes) sur la valeur du kurtosis (Fig. II.3.11). On remarque que les kurtosis prennent des valeurs globalement moins élevées lorsque le lissage de l'enveloppe est fort (ce qui est normal, puisqu'on atténue les variations d'amplitude), et que ceci résulte en une distinction moins nette entre VT et EB, voire à une inversion de la tendance des PDFs : plus la fenêtre de lissage choisie est grande, plus les kurtosis des EB sont grandes par rapport à celles des VT. Par défaut, on avait choisi une fenêtre de lissage de 0.5 s.

- pour RappMaxMean (e), les PDFs sont semblables, quoique légèrement moins séparables avec nos calculs.

Les divergences observées sont imputables à plusieurs facteurs, le premier étant bien évidemment la manière de calculer les attributs. Les calculs des attributs dépend directement de la détermination du début et de la fin du signal, laquelle conditionne la taille de la fenêtre dans laquelle on travaille. Pour pallier aux éventuelles erreurs, on a choisi une fenêtre "élargie" de 0.2 s avant et après le signal effectif. De plus, l'enveloppe a été lissée sur des fenêtres de 0.5 s. Les tests effectués avec un lissage plus grand (jusqu'à 10 fois) n'ont pas permis une amélioration du pouvoir séparateur des attributs.

Ceci étant dit, on ne devrait tout de même pas observer des variations inverses des PDFs en fonction des EB et des VT (Ene, Kurto).

Résultats

Les résultats des classifications effectuées avec les différents attributs et combinaisons d'attributs sont consignés dans le tableau II.3.4 et sont à comparer avec ceux du tableau II.3.1 (p.113). Ils nous apprennent que :

- globalement, les résultats sont légèrement en-deçà de ceux obtenus avec les attributs calculés par Hibert et al. [2014] lui-même. On arrive quand même à une classification de 90% des événements dans le meilleur des cas (SVM linéaire avec 4 attributs ou la meilleure combinaison de 3 attributs).

- le moins bon attribut pris seul est désormais AsDec. Il est d'ailleurs intéressant de noter que, comme pour Ene dans la section précédente, le comportement de la SVM et de la LR est différent : si la SVM a tendance à tout classer en EB (qui correspond à la classe majoritaire dans le *training set*), la LR agit à l'inverse en classant tout en VT. De manière générale, on remarque que la SVM a tendance à sous-classer les VT au profit des EB lorsque les attributs ou les combinaisons d'attribut ne sont pas très bonnes, ce

qui permet d'assurer une majorité de classifications correctes lors du passage au *test set* : un taux de réussite global supérieur à 80% correspond généralement à un bon taux de classification des VT.

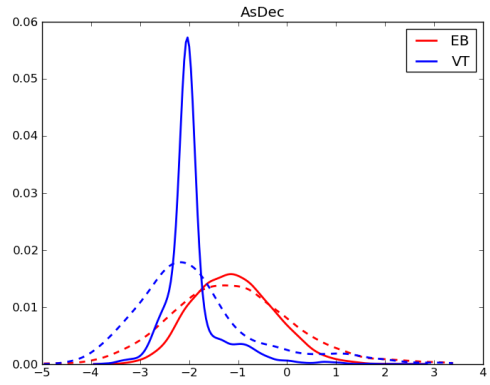
- le meilleur attribut pris seul est KRapp (suivi de Dur) pour la LR et Dur pour la SVM (suivi de KRapp).

- la meilleure combinaison de deux attributs est Dur + KRapp (soit la combinaison des deux meilleurs attributs seuls). AsDec étant le moins bon attribut pris seul, on s'attendait à ne pas retrouver les résultats de la section précédente où l'association Dur + AsDec était très bonne. Néanmoins, on peut souligner une fois de plus le fait que le pouvoir plus ou moins séparateur d'un attribut pris seul ne se répercute pas obligatoirement dans les combinaisons d'attributs. Par exemple, l'association AsDec + Dur est bien meilleure que Dur seul.

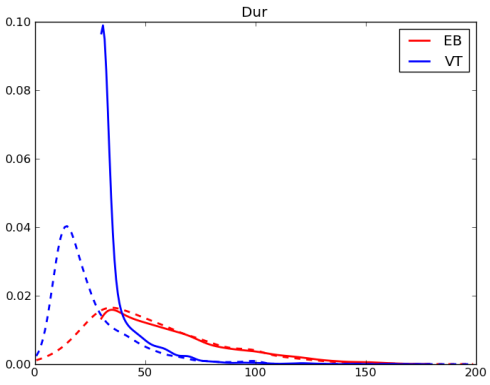
- l'association des 3 meilleurs attributs est, comme dans le cas précédent, Dur + Ene + KRapp et donne des résultats équivalents à ceux issus de l'association de tous les attributs. On note cependant, pour la LR, que l'on classe de manière plus homogène EB et VT lorsque tous les attributs sont pris en compte ; la situation inverse se produit pour la SVM linéaire.

	Attribut(s)	**Régression logistique**			**SVM linéaire**		
		VT	**EB**	**Total**	**VT**	**EB**	**Total**
1 attribut	AsDec	≈98%	≈2%	≈35%	≈13%	≈96%	≈58%
	Dur	79±6%	67±12%	73±4%	≈71%	≈80%	≈76%
	Ene	84±14%	55±11%	68±3%	≈13%	≈93%	≈57%
	Kurto	69±13%	81±9%	76±2%	≈51%	≈91%	≈73%
	RappMaxMean	72±15%	76±19%	74±5%	≈47%	≈94%	≈73%
	KRapp	76±11%	79±9%	77±1%	≈53%	≈92%	≈75%
2 attributs	AsDec + Dur	85±3%	78±9%	81±3%	≈86%	≈77%	≈81%
	AsDec + Ene	65±26%	70±17%	68±4%	≈27%	≈93%	≈63%
	AsDec + KRapp	67±12%	82±8%	76±1%	≈52%	≈92%	≈74%
	Dur + Ene	87±3%	69±4%	77±1%	≈84%	≈73%	≈78%
	Dur + KRapp	92±4%	85±8%	88±3%	≈92%	≈85%	≈88%
	Ene + KRapp	75±11%	81±8%	78±1%	≈54%	≈91%	≈ 75%
3 attributs	AsDec + Dur + Ene	88±3%	76±8%	81±3%	≈87%	≈82%	≈84%
	AsDec + Dur + KRapp	87±5%	89±6%	88±2%	≈97%	≈62%	≈78%
	AsDec + Ene + KRapp	63±19%	86±8%	76±5%	≈52%	≈92%	≈74%
	Dur + Ene + KRapp	91±4%	85±5%	88±2%	≈93%	≈87%	≈90%
4 attributs	AsDec + Dur + Ene + KRapp	87±3%	88±4%	88±1%	≈92%	≈88%	≈90%

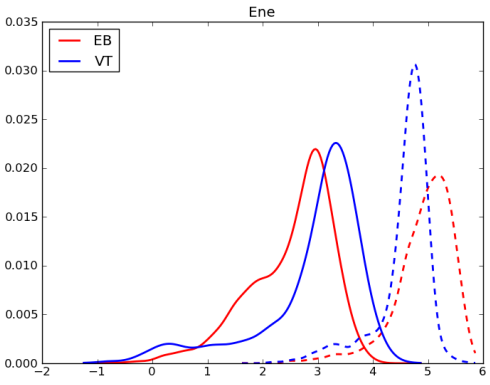
TAB. II.3.4: Résultats obtenus avec la régression logistique et la SVM linéaire en recalculant les attributs utilisés par Hibert et al. [2014]. Tous sont normalisés par le logarithme, sauf Dur et KRapp (qui résulte de la corrélation de Kurto et RappMaxMean déjà normalisés). A comparer avec le tableau II.3.1.



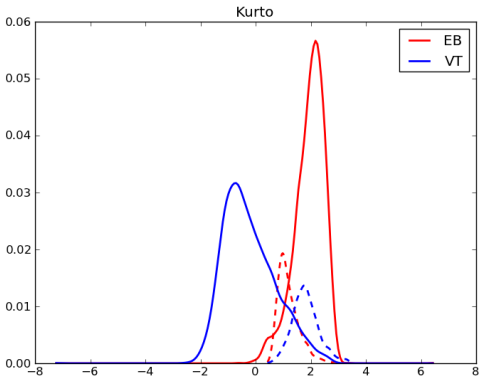
(a) AsDec



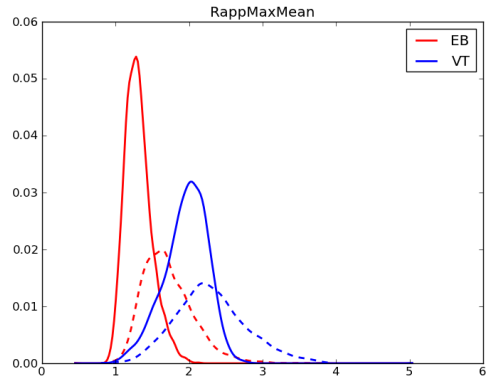
(b) Durée (Dur)



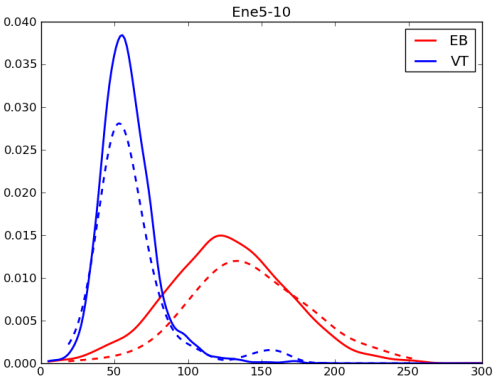
(c) Energie entre 10 et 30 Hz (Ene)



(d) Kurtosis (Kurto)



(e) RappMaxMean



(f) Energie entre 5 et 10 Hz

FIG. II.3.10: (a-e) Comparaison des densités de probabilité pour les 5 attributs définis par Hibert et al. [2014]. Les courbes continues correspondent aux attributs calculés par Hibert et al. [2014] ; les courbes tiretées aux attributs recalculés (voir §I.2.4). Tous ces attributs ont été normalisés par le logarithme, sauf la durée. (f) Densité de probabilité de l'énergie entre 5 et 10 Hz (sans normalisation). Courbes continues : *test set* ; courbes tiretées : *training set*.

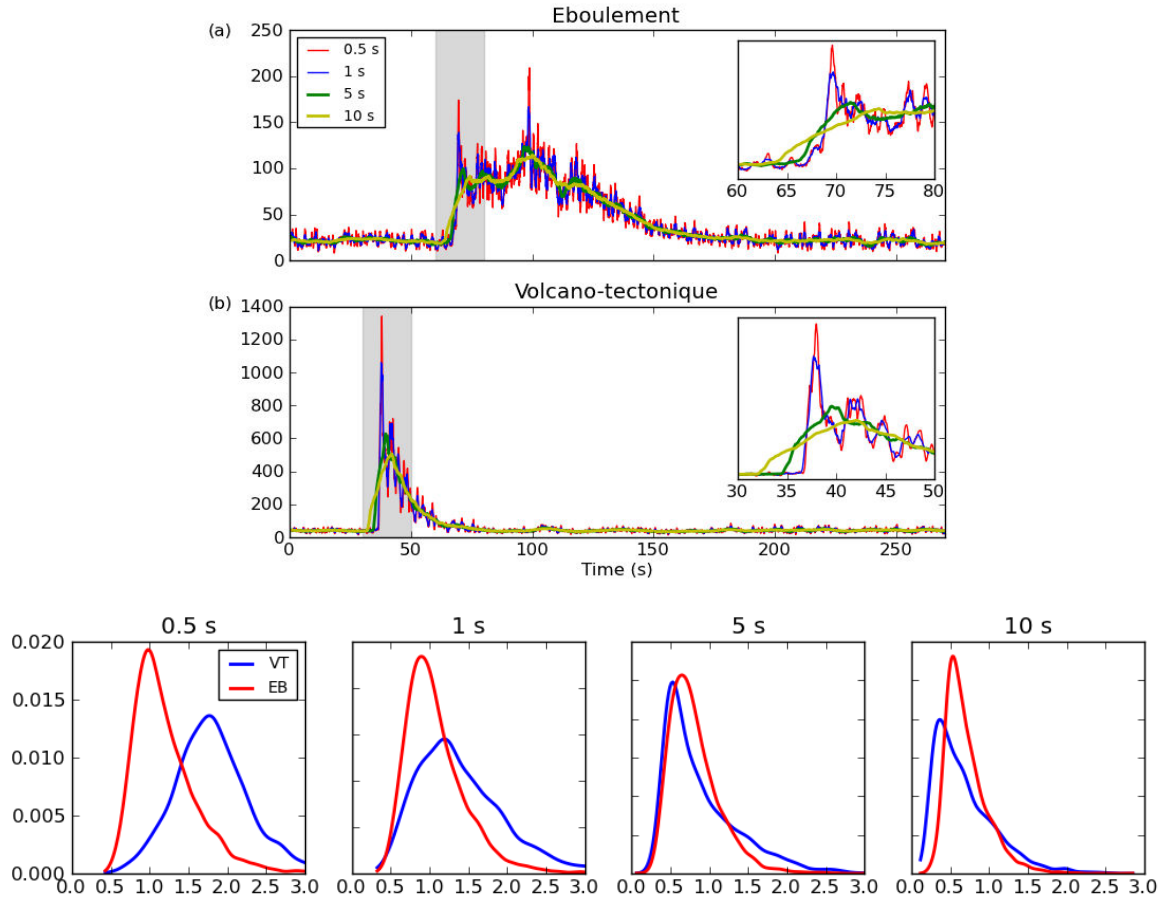


FIG. II.3.11: **(Haut)** Enveloppes d'un EB (a) et d'un VT (b) lissées sur des fenêtres de 0.5, 1, 5 et 10 s. Les encarts zooment sur le début du signal. **(Bas)** PDFs de l'attribut de kurtosis selon le lissage de l'enveloppe.

Variabilité du *training set*

On observe dans le tableau II.3.4 une forte dispersion des résultats de la LR pour certains attributs pour les classes des EB et des VT (de l'ordre de 10%). Il est intéressant de noter que cette dispersion ne se répercute souvent pas sur le résultat final, où la dispersion ne dépasse jamais les 5%. Autrement dit, avec un attribut (ou une combinaison d'attribut) donné, on classe toujours bien une même proportion d'événements, mais la répartition des bonnes classifications au sein des deux classes des VT et des EB peut varier.

Les figures II.3.12 et II.3.13 donnent un exemple pour l'attribut RappMaxMean (rapport du maximum sur la moyenne de l'enveloppe). Les comparaisons des PDFs pour les différents *subsets* montre que la variabilité liée aux tirages est minime, mais pas nulle non plus (l'aspect des PDFs change légèrement). D'autres facteurs, comme le choix du seuil de probabilité qui maximise le score F_1 (revoir §I.2.7, p.36), pourraient aussi avoir une influence. Toutefois, ceci montre qu'une variation, même très faible, dans les PDFs du *training set* a une forte influence sur les résultats de la classification. Cette variation se fera d'autant plus ressentir que le

training set est de petite taille, ce qui est le cas ici.

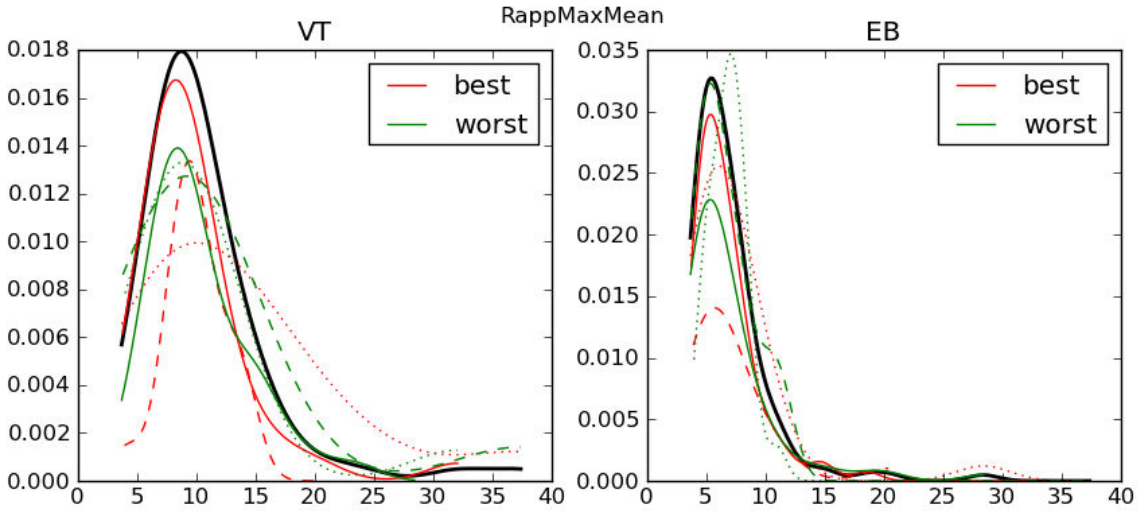


FIG. II.3.12: Densités de probabilités séparées pour les VT **(gauche)** et les EB **(droite)** pour RappMaxMean. Sur chaque figure, la courbe noire est la densité de probabilité calculée sur l'ensemble du *training set* ; les courbes rouges correspondent au tirage donnant les meilleurs résultats de classification ; les courbes vertes, aux moins bons. Pour ces deux couleurs, les courbes pleines représentent le *sub-training set* (60%), les courbes tiretées au *cross-validation set* (20%), et les courbes pointillées au *sub-test set* (20%).

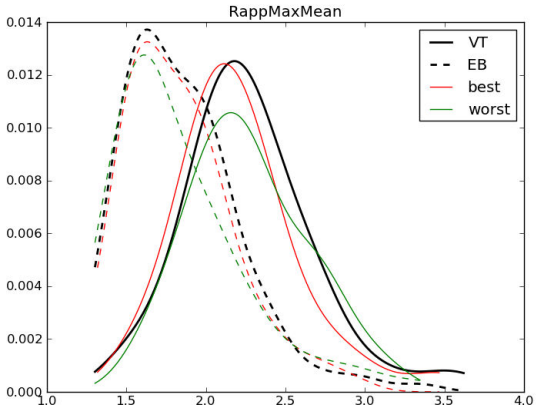


FIG. II.3.13: Densités de probabilités pour RappMaxMean pour les VT (courbes pleines) et les EB (courbes tiretées). La couleur noire correspond aux densités de probabilité calculées sur l'ensemble du *training set* (100%) ; la couleur rouge, à celles calculées pour le tirage (*sub-training set* 60%) ayant permis d'obtenir la meilleure classification ; la couleur verte, à celles du moins bon tirage.

II.3.3.2 Résultats avec tous les attributs hors tables de hachage

Un récapitulatif de l'ensemble des attributs utilisés est proposé dans le tableau I.2.2, p. 74. Les résultats de classification sont visibles dans le tableau II.3.3(2) et sur la figure II.3.14 pour la LR, la SVM linéaire et la SVM non-linéaire. Ceux-ci montrent que :

- on arrive au maximum à 88% de bonne classification, ce qui est équivalent (voire légèrement moins bon pour la SVM) à ce qu'on obtenait avec les 4 attributs étudiés dans les sections précédentes.

- la SVM non-linéaire n'apporte rien de plus car le problème est suffisamment "simple".

- on est en situation de sur-apprentissage, i.e. le taux de classification du *training set* est largement supérieur à celui du *test set* (de l'ordre d'une dizaine de %). On avait vu en section I.2.2.1 que dans ce cas, il vaut mieux réduire le nombre d'attributs ou augmenter la taille du *training set*. Ici, la taille du *training set* est en effet relativement réduite par rapport à celle du *test set* (seulement de l'ordre de 2% !), donc lorsqu'on utilise un trop grand nombre d'attributs, on laisse moins de "degré de liberté" au système qui va avoir tendance à trop bien ajuster les données.

- la SVM classe mieux les EB (classe majoritaire) que les VT, ce qui n'est pas le cas de la LR.

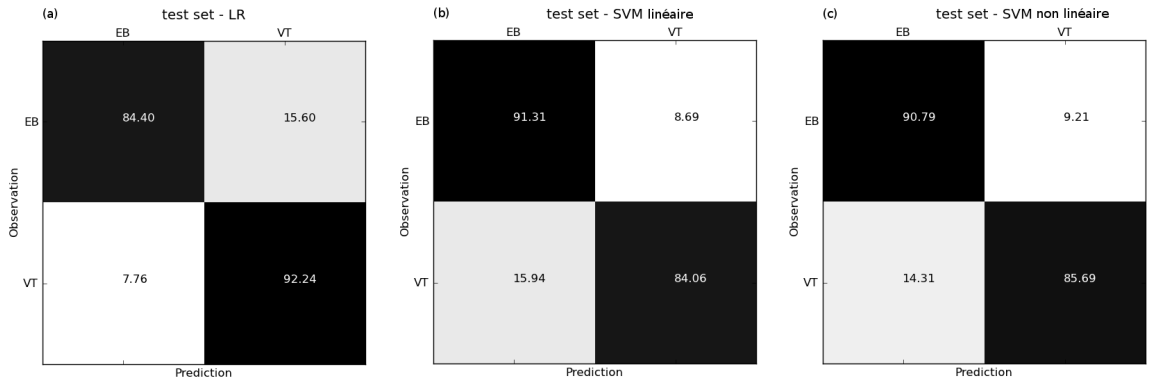


FIG. II.3.14: Matrices de confusion obtenues avec plus de 50 attributs pour (a) la régression logistique, (b) la SVM linéaire et (c) la SVM non-linéaire.

Comme on sait aussi déjà que la combinaison de AsDec, Dur, Ene et KRapp donne de bons résultats, on regarde ce qu'il en est si on supprime ces 4 attributs (voir TAB. II.3.3(3)). Curieusement, les résultats de la LR sont légèrement inférieurs à la situation précédente (-1%), alors que ceux de la SVM sont meilleurs (+3%, on atteint 91%). Les VT sont mieux classés avec cette combinaison d'attributs (94% et 92% respectivement pour la LR et la SVM, ce qui correspond aux taux maximum d'extraction des VT pour tous les tests effectués).

Bilan

Ces résultats servent à montrer que ce n'est pas le nombre d'attributs qui importe, ni leur pouvoir séparateur pris seul, mais leur combinaison.

II.3.3.3 Résultats avec les tables de hachage seules

Paramètre	Valeur	% training	% test	Ecart	Commentaires
Taille de la fenêtre de bruit avant le P-onset	0 s	85±5%	74±4%	11%	nlap = 0.6
	1 s	91±4%	81±3%	10%	NFFT = 256
	2 s	88±4%	78±3%	10%	longueur du signal ≈ 16 s
	3 s	89±5%	82±3%	7%	
	4 s	89±3%	78±3%	11%	
	5 s	88±4%	77±3%	11%	
	10 s	88±5%	77±2%	11%	
	15 s	82±6%	68±2%	14%	
	20 s	59±8%	50±2%	9%	
Recouvrement des fenêtres (nlap)	0.0	95±1%	82±3%	13%	NFFT = 256
	0.1	89±5%	76±3%	13%	résolution : 64x64
	0.2	93±1%	78±2%	15%	longueur du signal variable
	0.3	91±2%	79±2%	12%	
	0.4	89±2%	79±2%	10%	
	0.5	90±2%	80±3%	10%	
	0.6	91±3%	81±3%	10%	
	0.7	90±3%	77±3%	13%	
	0.8	89±3%	77±3%	12%	
	0.9	85±5%	70±2%	15%	
NFFT	128	93±2%	83±3%	10%	nlap = 0.8 ; signal de ≈16 s
	256	91±3%	82±3%	9%	nlap = 0.6 ; signal de ≈16 s
	512	90±3%	79±3%	11%	nlap = 0.2 ; signal de ≈16 s
Résolution	128x128	73±7%	61±2%	12%	nlap = 0.6 ; NFFT = 512
	64x64	93±2%	83±3%	10%	nlap = 0.8 ; NFFT = 128
	32x32	92±2%	82±2%	10%	nlap = 0.2 ; NFFT = 256
	16x16	89±2%	76±2%	13%	nlap = 0 ; NFFT = 512 ; signal de 4.8 s ; 5% coeff.
	8x8	83±4%	73±3%	10%	nlap = 0 ; NFFT = 1024 ; signal de 1 s ; 10 coeff.
Nombre de permutations	100	89±4%	51±4%	38%	nlap = 0.6
	200	89±4%	78±2%	11%	NFFT = 256
	300	92±3%	82±2%	10%	
	400	89±3%	79±3%	10%	
	500	93±1%	83±3%	10%	
	600	86±4%	78±3%	8%	
	700	89±2%	80±3%	9%	
	800	89±4%	77±2%	12%	
	900	95±2%	85±1%	10%	
	1000	90±3%	78±4%	12%	
Nombre de tables	10	82±4%	80±2%	2%	nlap = 0.6
	25	82±4%	77±4%	5%	NFFT = 256
	50	91±2%	82±2%	9%	
	100	94±1%	84±1%	10%	

TAB. II.3.5: Résultats de la classification obtenus pour divers calculs des tables de hachage et avec la méthode de régression logistique. Description du tableau dans le texte.

Dans la section §I.2.4.3, on a détaillé une méthode de compression de l'information contenue dans les spectrogrammes via les tables de hachage. On a partiellement évoqué les paramètres susceptibles d'influencer les valeurs stockées dans ces tables, comme notamment les paramètres de calcul des spectrogrammes, ou encore le nombre de permutations utilisées pour le calcul de la signature Min-Hash...

Dans le tableau II.3.5 sont consignés les résultats d'un certain nombre de tests qui ont été effectués sur le jeu de données du Piton de la Fournaise. La première colonne indique le paramètre que l'on a fait varier. La deuxième colonne donne la valeur de ce paramètre et les deux colonnes suivantes donnent les pourcentages de classification du *training set* et du *test set* pour cette valeur de paramètre. La colonne "écart" indique l'écart de pourcentage qui existe entre *training set* et *test set*, fournissant ainsi une idée sur la généralisation de l'un à l'autre.

On avait précisé dans la section §I.2.4.3 les paramètres finaux retenus pour le calcul des tables de hachage, à savoir : on calcule un spectrogramme de taille **64x64** sur des fenêtres glissantes proches de 1 s (soit **NFFT=128** échantillons) avec un taux de recouvrement de **0.8**, et en considérant le signal **2 s** avant la première arrivée (ceci correspond au final à un signal de durée $\approx$ 16 s). On retient **1%** des coefficients maximum de l'empreinte. On effectue ensuite **500** permutations de la matrice de fingerprint et on calcule **50** tables de hachage. Ces paramètres sont surlignés en jaune dans la deuxième colonne du tableau II.3.5. Sauf mention contraire dans la colonne "commentaires", ce sont les paramètres utilisés par défaut dans le calcul des différents tests.

Les trois autres couleurs surlignent respectivement, pour chaque paramètre considéré, le pourcentage maximum de classification pour le *training set* (bleu) et pour le *test set* (vert), ainsi que l'écart minimum entre ces deux pourcentages (rouge) ; ceci afin de faciliter la lecture du tableau.

Sans rentrer dans les détails, le tableau II.3.5 nous apprend que :

- sans surprise, si la taille de la fenêtre de bruit considérée avant la première arrivée augmente, les résultats de la classification sont dégradés (la fenêtre de 20 s ne contient même que du bruit). Ceci étant, il semble quand même nécessaire de considérer le signal quelques secondes avant son début, notamment pour pallier aux erreurs et aux imprécisions éventuelles de détermination du P-onset.

- lorsque l'on fait varier le taux de recouvrement entre fenêtres glissantes, on fait intrinsèquement varier la longueur du signal considéré puisque la taille de la matrice finale reste fixe (64x64). L'idée était de savoir si une partie du signal en particulier pouvait contenir un maximum d'information (par exemple, si le tout début du signal était suffisant ou pas...). Les résultats montrent globalement qu'il n'existe pas une très forte variabilité (on reste autour de 80% de classification pour le *test set*), sauf pour un recouvrement de 0.9. Dans ce dernier cas, le signal retenu est de petite taille (quelques secondes seulement) : l'information importante ne semble donc pas contenue dans les premières secondes du signal. Il faut au contraire garder un maximum de signal (les meilleurs résultats sont d'ailleurs obtenus sans recouvrement).

- lorsque le nombre de points retenus pour le calcul de chaque spectre augmente, les taux de classification ont tendance à diminuer (même s'ils restent relativement proches les uns des autres). Remarque : pour pouvoir comparer pleinement les résultats et conserver

la même longueur totale de signal, on a aussi fait varier le taux de recouvrement.

- une résolution trop élevée (128x128), en plus d'être coûteuse en temps de calcul, ne permet pas une discrimination satisfaisante des EB et des VT. En choisissant une résolution trop basse (16x16 ou 8x8), on perd également en taux de bonne classification. Il y a peu de différences entre les matrices 64x64 et 32x32. Remarque : afin de conserver une longueur de signal identique, on a aussi dû faire varier le taux de recouvrement et le nombre d'échantillons contenu dans chaque fenêtre de calcul.

- un nombre trop insuffisant de permutations (100) pour le calcul de la signature Min-Hash ne permet pas une bonne classification. Comme le nombre de tables de hachage calculées est fixe et vaut 50, c'est donc le nombre de valeurs de hachage concaténées dans une seule table qui varie : pour 100 permutations, on concatène des paires de valeurs ; pour 200, des groupes de 4... Ici, on remarque que lorsque le nombre de permutations dépasse 200, les résultats de la classification sont semblables. Pour alléger le temps de calcul, on a donc choisi un nombre de permutations intermédiaire (500).

- il semble préférable d'utiliser un plus grand nombre de tables de hachage (mais pas trop non plus, sinon on retourne dans la situation précédente où le nombre de valeurs de la signature à concaténer est trop faible). On remarque cependant que la généralisation entre le *training set* et le *test set* est bien meilleure lorsque le nombre de tables de hachage est moins élevé (ce qui est cohérent, puisque l'on avait vu en §I.2.2.1 qu'il suffisait de réduire le nombre d'attributs pour diminuer le sur-apprentissage).

Bilan

L'ensemble des résultats présentés dans le tableau II.3.5 montre qu'il est possible de classer avec sûreté entre 80 et 85% des VT et des EB du jeu de données avec la LR. Le tableau II.3.3(4) donne aussi les résultats de la SVM, qui sont similaires à ceux de la LR. Ces pourcentages sont inférieurs à ceux obtenus avec les seuls attributs de Hibert [2012].

Néanmoins, la méthode de *fingerprinting* utilisée reste intéressante en termes de compression de l'information et montre qu'il est possible de l'utiliser seule pour la classification en atteignant des taux de réussite honorables. De plus, on remarque dans le tableau II.3.3(4) que, même si le nombre d'attributs reste élevé, on est moins en situation de sur-apprentissage qu'avec tous les autres attributs sismiques « classiques » (l'écart entre le *training-set* et le *test set* est inférieur à 10%).

Une piste à explorer pour des développements futurs serait la recherche d'une très bonne fonction de hachage. Ici, on s'est contenté d'une somme pondérée des valeurs concaténées de la signature Min-Hash. Il existe sûrement des fonctions plus complexes qui devraient permettre un meilleur stockage de l'information.

II.3.3.4 Combinaison des tables de hachage avec d'autres attributs

Les tables de hachage calculées précédemment sont issues de la compression de l'information contenue dans les spectrogrammes : elle ne classe donc les événements que sur leur contenu fréquentiel et temporel et permet déjà d'obtenir de très bons résultats.

La combinaison de ces tables de hachage avec des attributs temporels (qui prennent en compte l'ensemble du signal et non pas une petite partie) ou des indicateurs de forme des signaux devrait permettre d'améliorer encore un peu la classification. Ainsi, pour obtenir les résultats présentés dans le tableau II.3.3(5), on a ajouté l'attribut de durée (Dur) et l'attribut KRapp (combinaison du kurtosis et du rapport du maximum sur la moyenne de l'enveloppe, qui renseignent tous les deux sur la forme et l'impulsivité du signal).

Ces résultats montrent une nette amélioration par rapport aux tables de hachage seules (+4% pour la SVM, +3% pour la LR).

Enfin, on a également essayé une combinaison prenant en compte tous les attributs « classiques » et toutes les tables de hachage (Tab. II.3.3(6)), soit au total plus d'une centaine d'attributs. Les résultats obtenus sont excellents (90 et 92% pour la LR et la SVM, respectivement) et comparables à ceux obtenus avec les 4 attributs fournis par C. Hibert (Tab. II.3.2), voire meilleurs en terme d'équilibre de taux de classification entre VT et EB.

II.3.4 Résumé et conclusion

Le but de cette partie était de classer automatiquement les EB et les VT sur le volcan du Piton de la Fournaise. Une telle automatisation a déjà été mise en œuvre par Hibert [2012], Hibert et al. [2014] grâce à une technique de logique floue. En travaillant avec le même jeu de données, on a souhaité expérimenter deux autres méthodes de classification automatisée : la LR et la SVM. Les points qui méritent d'être soulignés à l'issue de cette étude sont finalement les suivants :

- la LR et la SVM sont des méthodes de classification très satisfaisantes qui donnent des résultats similaires à ceux de la logique floue avec cet exemple d'application.

- on atteint au mieux 90% de réussite avec la LR et 92% avec la SVM. La SVM est donc légèrement meilleure que la LR. Quelles que soient les combinaisons d'attributs considérées, on n'arrive pas à dépasser ces pourcentages, ce qui est probablement dû à une limitation du jeu de données.

- presque n'importe quelle combinaison de plus de 4 attributs permet d'atteindre entre 80 et 85% de classifications correctes, et ceci indépendamment du pouvoir plus ou moins discriminatoire d'un attribut donné.

- c'est la combinaison des attributs, et non leur nombre, qui permet une bonne classification. Quelques attributs judicieusement associés peuvent donner de meilleurs résultats qu'un grand nombre d'attributs.

- le mode de calcul des attributs est important.

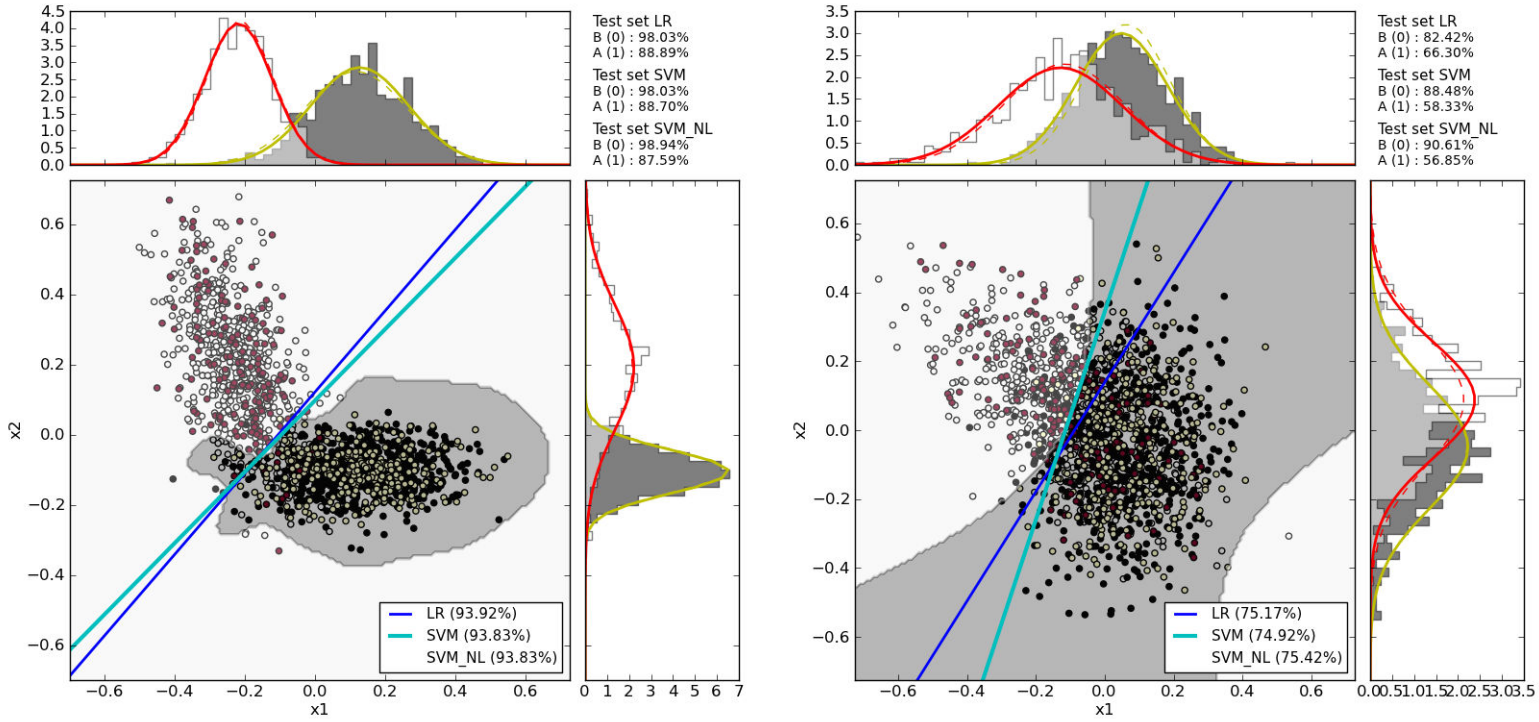
A propos du jeu de données, il est aussi important de remarquer que :

- il s'agit d'un très bon jeu de données. Les VT et les EB sont facilement identifiables. Le problème de classification binaire ne pose pas de difficultés particulières. Les méthodes

linéaires suffisent.

- le *training set* et le *test set* sont légèrement déséquilibrés et comptent plus d'EB que de VT. Globalement, on a noté que la SVM avait tendance à classer plus facilement les EB que les VT, à l'inverse de la LR. Les tests synthétiques de la figure II.3.15, construits selon les proportions du jeu de données du Piton de la Fournaise (voir FIG. II.3.1), confirment cette observation : que la séparation entre les 2 classes soit facile ou difficile, la SVM sera toujours plus prompte à classer les éléments de la classe majoritaire. Ceci s'explique par le fait qu'elle cherche à maximiser la marge avec les éléments les plus proches. La LR, quant à elle, trouve généralement le séparateur qui équilibre au mieux les classes, i.e. qui classe bien le plus d'EB possible ET le plus de VT possible (calcul du score F_1).

- le *training set* est de très petite taille par rapport au *test set*. Ici, cela ne pose pas de problème car les EB et les VT se séparent facilement. Il ne semble pas y avoir de gros problème de représentativité (sauf pour certains attributs pour lesquels les PDFs du *training set* et du *test set* ne sont pas superposables) : les pourcentages de classifications correctes restent élevés. Néanmoins, on est très souvent en situation de sur-apprentissage (le pourcentage du *training set* est supérieur à celui du *test set*), ce qui suggère un nombre d'éléments insuffisant dans le *training set*.



(a) 2 classes facilement séparables

(b) 2 classes difficilement séparables

FIG. II.3.15: Tests synthétiques respectant les proportions du *training set* et du *test set* du jeu du Piton de la Fournaise (FIG. II.3.1). La classe A (blanche) est minoritaire par rapport à la classe B (noire). La SVM (droite bleu clair) favorise toujours la classe majoritaire par rapport à la LR (droite bleu foncé).

Partie III

Traitement des données du Kawah Ijen, Indonésie

Sommaire

CHAPITRE III.1

Présentation générale

III.1.1 Contexte

Le volcan du Kawah Ijen, culminant à 2386 m d'altitude, est situé sur l'île indonésienne de Java. C'est un volcan de type explosif, encore actif, qui fait partie de la caldeira de l'Ijen. Il est caractérisé et connu pour la présence d'un lac acide dans le cratère sommital, ainsi que pour sa solfatare d'où est extrait le minerai de soufre. De nombreuses études ont été effectuées sur le volcan, notamment en ce qui concerne le lac acide (mesures de température, du niveau ; observations : explosion de bulles, changements de couleurs de la surface ; analyse géochimique), mais relativement peu concernant les enregistrements sismiques (un seul sismomètre jusqu'en 2010), contrairement à d'autres volcans indonésiens tels que le Merapi [Ohrnberger, 2001]. Les observations et enregistrements sont traités dans un observatoire sur place (*Kawah Ijen Observatory*). On ne s'intéressera ici qu'aux signaux sismologiques.

Développer un outil de détection, localisation et classification automatique serait intéressant pour l'observatoire : le nombre de signaux enregistrés étant relativement important, cela peut constituer une aide précieuse au traitement manuel des données et un gain de temps non négligeable. Cela permettrait aussi une meilleure compréhension des phénomènes sismologiques se produisant sur le volcan, et pourrait, à terme, aider à mieux appréhender les signaux précurseurs des éruptions (types et nombre d'événements...).

De plus, un tel outil pourrait aussi accroître les connaissances sur le volcan et aider à la corrélation avec d'autres observations géophysiques (phénomènes hydrothermaux par exemple). En effet, la multiplicité des signaux enregistrés par les sismomètres sur le Kawah Ijen augure la complexité du volcan (huit classes d'événements principales ont été identifiées et définies par le personnel de l'observatoire sur place).

Jusqu'en 2010, une seule station sismologique enregistrait l'activité sismique du Kawah Ijen : la station IJEN, appartenant au CVGHM (*Center for Volcanology and Geohazards Mitigation*, Indonésie), dont les données sont filtrées systématiquement entre 1 et 10 Hz.

En 2010 ont commencé les installations d'un certain nombre de stations sismologiques autour du volcan, menées par deux équipes différentes (ROB - *Royal Observatory of Belgium* et USGS - *United States Geological Survey*), permettant de passer d'une seule à une quinzaine de

stations. Le choix des sites s'est fait selon plusieurs critères : le bruit ambiant (il faut le limiter au maximum en se plaçant loin des lieux d'activité humaine, des arbres...), l'accessibilité (le terrain étant relativement accidenté, il faut pouvoir accéder aux données facilement) et la proximité au volcan (plus on s'éloigne, moins cela présente d'intérêt pour étudier les phénomènes intrinsèques au volcan). Deux des trois stations large-bandes appartenant au ROB ont été installées sur le cratère actif du volcan afin de favoriser l'étude des signaux liés au système hydrothermal.

Le tableau III.1.1 fournit quelques informations sur les 16 stations utilisées dans cette étude et la figure III.1.1 montre leur répartition autour du volcan. La figure III.1.2 illustre la disponibilité des données pour chacune des stations. Les disponibilités sont assez disparates : seules les données de la station IJEN, au sommet du volcan, sont disponibles en continu pour la période considérée. Pour la station POS, les données sont disponibles uniquement pour certaines périodes. Enfin, un certain nombre de stations a été installé de manière temporaire fin 2011 (IBLW, IMLB...). Ceci coïncide avec la crise qui a eu lieu en 2011-2012 sur le volcan (crise caractérisée par un premier essaim sismique en mai 2011, une forte augmentation de la sismicité d'octobre 2011 jusqu'à mi-janvier 2012, une reprise début mars 2012, puis fin juin 2012).

Le catalogue de sismicité dont on dispose couvre une période allant de début février 2011 à fin juin 2012 et prend donc en compte les crises mentionnées précédemment. Ceci est particulièrement intéressant pour l'étude du volcan et des interactions de la sismicité avec les phénomènes hydrothermaux. Néanmoins, l'augmentation du nombre de séismes enregistrés risque de compliquer la tâche de classification.

L'ensemble des données continues utilisées dans cette thèse nous ont été fournies par le ROB sous forme de fichiers miniseed découpés par heure ou par jour selon les stations.

Station	Longitude	Latitude	Altitude	Instrument	Composantes	Fréquence	Appartenance
DAM	114.2362	-8.0586	2208 m	Trillium 120P	Z-N-E	200 et 100 Hz	ROB
IBLW	114.1672	-7.9933	1021 m	LE-3Dlite	Z-N-E	125 Hz	ROB
IGEN	114.1648	-8.0603	1500 m	LE-3Dlite	Z-N-E	125 Hz	ROB
IJEN	114.2395	-8.0622	2330 m	L4	Z	100 Hz	USGS
IMLB	114.1003	-8.0253	1530 m	LE-3Dlite	Z-N-E	125 Hz	ROB
IPAL	114.2057	-8.0598	1600 m	LE-3Dlite	Z-N-E	125 Hz	ROB
IPLA	114.1948	-8.0370	1450 m	LE-3Dlite	Z-N-E	125 Hz	ROB
IPSW	114.2792	-7.9915	651 m	LE-3Dlite	Z-N-E	125 Hz	ROB
KWUI	114.2371	-8.0526	2140 m	L4	Z	100 Hz	USGS
MLLR	114.1195	-8.1531	1370 m	2sa180	Z-N-E	100 Hz	USGS
POS	114.2463	-8.0543	2379 m	Trillium 120P	Z-N-E	200 et 100 Hz	ROB
POSI	114.2570	-8.1468	730 m	L4	Z	100 Hz	USGS
PSG	114.2276	-8.0725	1853 m	Trillium 120P	Z-N-E	100 Hz	ROB
PUN	114.2395	-8.0622	2330 m	Trillium 120 P	Z-N-E	100 Hz	ROB
RAUN	114.1195	-8.1531	1370 m	L4	Z	100 Hz	USGS
TRWI	114.2395	-8.0622	2330 m	L22D	Z-N-E	100 Hz	USGS

TAB. III.1.1: Caractéristiques des stations utilisées dans l'étude.

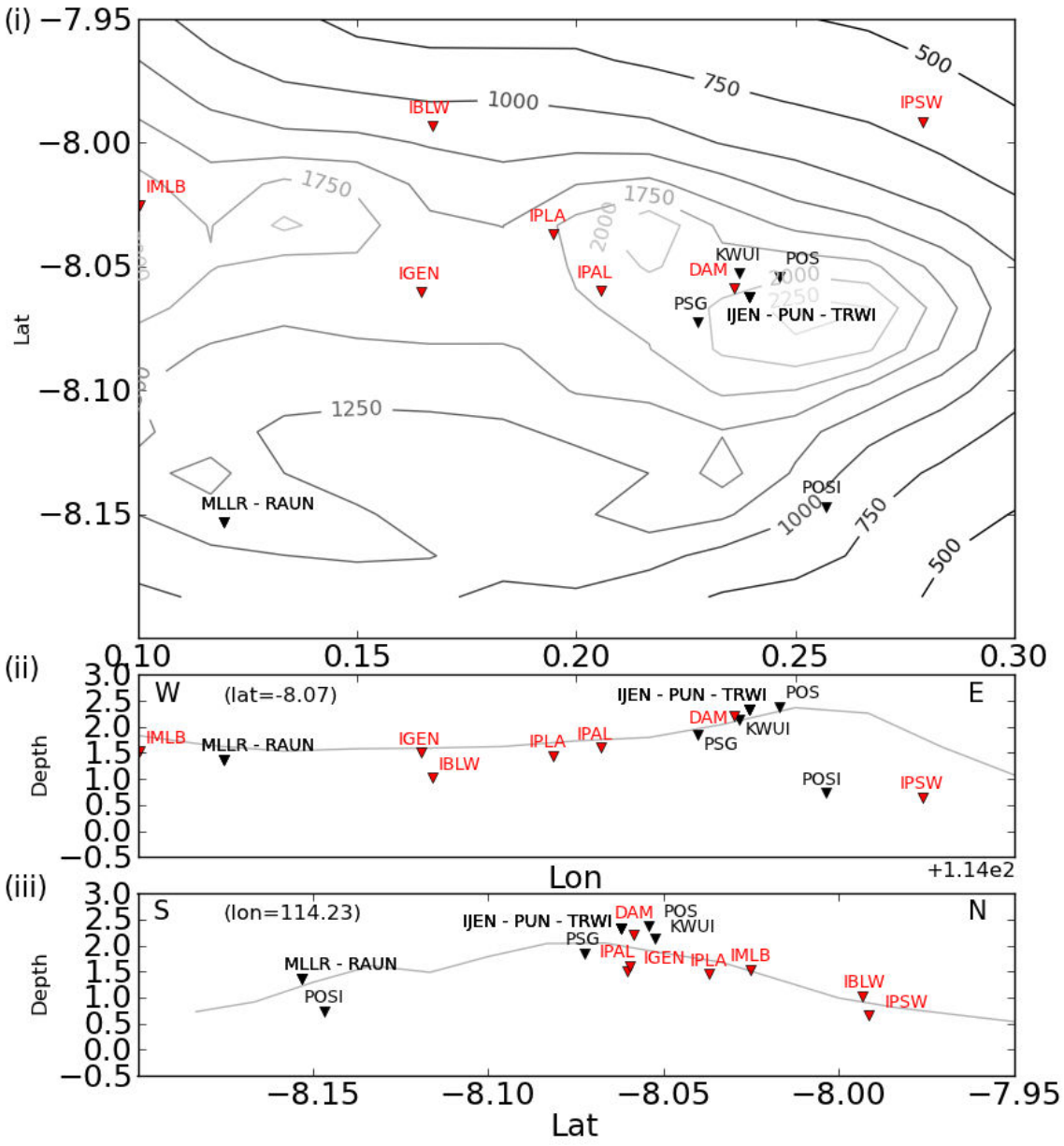


FIG. III.1.1: Réseau sismologique installé autour du volcan. Les stations colorées en noir corrspondent à celles qui ont été retenues pour la localisation avec Waveloc.

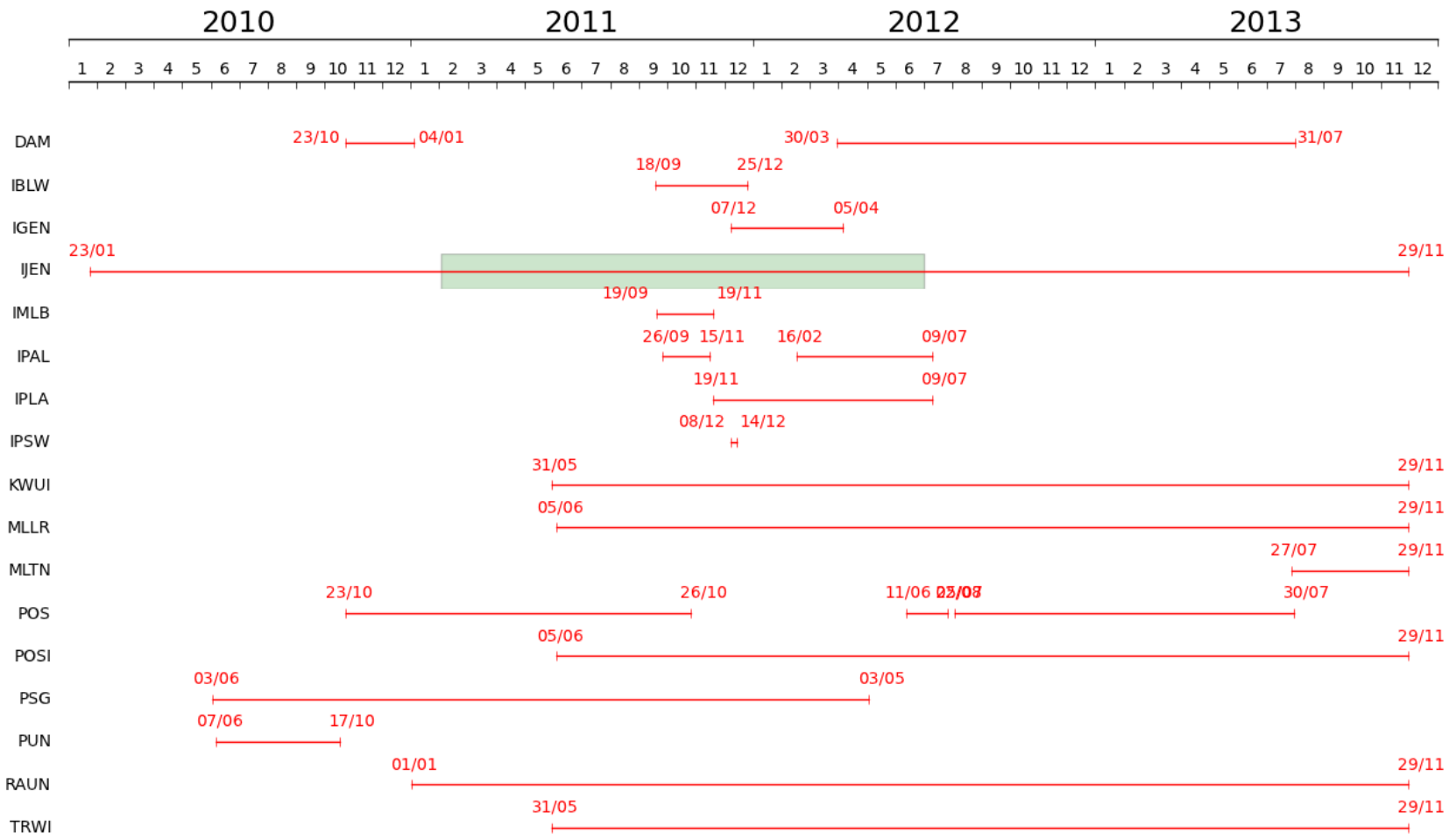


FIG. III.1.2: Disponibilité des données sismologiques par station pour la période 2010-2013. La zone colorée correspond au catalogue utilisé pour la classification et la station à partir de laquelle il a été élaboré.

III.1.2 Classification des événements sur le Kawah Ijen

III.1.2.1 Introduction aux différents types d'événements enregistrés en milieu volcanique

Les signaux enregistrés sur les volcans en général sont de nature très diverse et traduisent la multiplicité des processus physiques qui les génèrent. Ces processus sont complexes et résultent le plus souvent des interactions entre gaz, liquide et solide [Chouet and Matoza, 2013].

Tous les volcans ne sont pas identiques et présentent des systèmes et des mécanismes différents. De ce fait, il n'existe pas de classification uniforme d'événements sismiques pour l'ensemble des volcans du globe (on pourrait probablement établir une classification spécifique pour chacun d'entre-eux), mais certains événements caractéristiques se retrouvent d'un volcan à un autre [McNutt, 1986, 2002, Lahr et al., 1994]. La classification "de base" la plus communément utilisée est celle établie par Minakami [1960, 1974] à partir de l'observation des formes d'ondes enregistrées sur quelques volcans japonais. Quatre types d'événements différents sont ainsi définis :

- les événements haute-fréquence (HF - *High Frequency* en anglais) ou de type A ;
- les événements basse-fréquence (LF - *Low Frequency* en anglais) ou de type B ;
- les explosions ;
- les tremors volcaniques.

A ces événements, on peut ajouter d'autres types, comme ceux qui se produisent plus en surface et qui sont le plus souvent spécifiques d'un volcan donné (séismes glaciaires, glissements de terrain, éboulements, coulées pyroclastiques...).

Enfin, deux nouveaux types ont émergé ces dernières années :

- les hybrides, qui présentent à la fois des caractéristiques des HF et des LF.
- les très longue période (VLP - *Very Long-Period* en anglais), qui ont été mis en évidence depuis l'avènement des sismomètres large-bandes.

Les principales caractéristiques de tous ces types d'événements sont reportées dans le tableau III.1.2.

Nom	Sismogramme	Fréquence	Origine probable
Haute fréquence (A)	P et S visibles	5-15 Hz	cisaillement/glissement sur les failles
Basse fréquence (B)	P émergente ; pas de S	1-5 Hz	lié aux fluides
Explosion	onde de choc aérienne	-	propagation des ondes à la fois dans le sol et dans l'air
Tremor volcanique	longue durée	1-5 Hz	série de LF
Hybride	début HF ; coda LF	-	séisme se produisant à proximité d'une cavité remplie de fluide et la mettant en résonance
Très longue période	-	< 1 Hz	mouvements de magma
Autres	glissements de terrain, éboulements, glaciers...		

Tab. III.1.2: Classification de base des événements enregistrés en milieu volcanique. D'après McNutt [2002].

III.1.2.2 Classification manuelle et présentation du catalogue

Une classification des événements se produisant sur le Kawah Ijen est effectuée manuellement par le personnel de l'observatoire sur place. Elle est essentiellement basée sur l'aspect de la forme d'onde, selon la classification de Minakami [1974] introduite dans la section précédente, et plus spécifiquement sur l'identification des ondes P et S.

Huit types d'événements ont ainsi pu être définis, avec quelques différences notables par rapport à la classification "usuelle" : les tectoniques (avec une distinction entre les tectoniques lointains ($>$ 50 km) et les locaux ($\approx$20 km)) ; les volcano-tectoniques de type A ou de type B selon leur profondeur (les "A" sont plus profonds), les explosions, les tremors et les événements moins fréquents, tels que les LF, les glissements de terrain et les hybrides.

Les divergences que l'on peut souligner sont les suivantes :
- les types A et B ne se distinguent pas par leur contenu fréquentiel (voir Tab. III.1.2), mais par leur profondeur : les A sont supposés avoir lieu proche de la base du volcan, alors que les B sont supposés avoir lieu proche du sommet (voir Tab. III.1.3) ;
- la classe des LF est distincte de celle des types B. Elle désigne typiquement des tremors de courte durée.
- les hembusans (équivalent aux explosions) ne se caractérisent pas par une onde de choc aérienne. Leur définition demeure encore relativement floue...
- les tremors désignent ici des *tremors harmoniques monochromatiques*, c'est-à-dire dont la fréquence dominante représente 80 à 90% des amplitudes spectrales. La fréquence dominante associée à ces événements se situe autour 1.3 Hz pour le Kawah Ijen.

On a répertorié dans le tableau III.1.3 l'ensemble des caractéristiques définissant les différents types d'événements enregistrés sur le Kawah Ijen.

Nom		Sismogramme	Durée	Fréquence	Remarques
Tectonique	lointain	S-P $>$ 10 s	-	-	peu de locaux enregistrés
	local	4 $<$ S-P $<$ 10 s			
VT de type A		S-P $<$ 4 s ; impulsif	$>$ 10 s	1-12 Hz	peu d'enregistrements
VT de type B		pas de S visible ; P émergente	$<$ 10 s	-	nombreux pendant les crises
Explosion		pas de S visible	-	large spectre	pas de formes d'onde spécifiques
Tremor		pas de S visible	variable	1-2 Hz	nombreux pendant les crises
Basse fréquence		pas de S visible ; P émergente	8-20 s	$<$ 5 Hz	tremor de courte durée
Glissement de terrain		pas de S visible	-	-	peu observés
Hybride		pas de S visible	-	-	début du signal HF ; coda BF

Tab. III.1.3: Classification des événements sur le Kawah Ijen d'après les informations recueillies dans la thèse de Caudron [2013]. Jusqu'en juillet 2011, les critères de classification étaient exclusivement visuels ; ensuite, les critères spectraux ont aussi été pris en compte.

Il faut aussi noter qu'en juillet 2011 a été introduit le logiciel SWARM, développé par l'*Alaska Volcano Observatory* et l'USGS. Il permet d'analyser les formes d'onde en temps réel en fournissant notamment les critères spectraux à prendre en compte pour la classification.

Ceci a créé une certaine confusion dans les classifications qui ont suivies, car le personnel de l'observatoire a dû être formé pour intégrer ces nouveaux paramètres.

Le logiciel SWARM a permis la mise en évidence des trois derniers types évoqués (à savoir : les LF, les hybrides et les glissements), c'est pourquoi leur classification a été plus systématique qu'auparavant, mais elle a aussi pu engendrer des confusions avec les types pré-existants.

Le catalogue utilisé dans cette étude sous le nom de « **catalogue brut** » comprend quasiment 8400 événements répartis dans les huit classes définies précédemment et ayant eu lieu entre début février 2011 et fin juin 2012. La classification a d'abord été effectuée dans l'observatoire sur place à partir des enregistrements de la station IJEN (filtrée entre 1 et 10 Hz), puis révisée par C. Caudron dans le cadre de sa thèse [Caudron, 2013], avec, entre autres, l'aide du logiciel *Seismo_volcanalysis* developpé par Lesage [2009].

La figure III.1.3 donne une idée de l'aspect et de la diversité des signaux enregistrés sur le volcan. Chaque événement a été tiré aléatoirement à partir du catalogue manuel. On voit d'ores et déjà que la classification automatique ne va pas s'annoncer facile compte tenu de la diversité des signaux au sein d'une même classe d'événements (longsorans, tectoniques, hembusans et hybrides) et de la similitude des signaux entre classes (tectoniques et tremors ; BF et volcaniques de type A, voire type B).

Type d'événement	Nombre	% jeu	Numéro	Nom court	Terminologie indonésienne
Volcanique de type B	3514	42%	0	VB	Vulkanik B
Tectonique	1789	21%	1	Tecto	Tektonik
Tremor harmonique	1238	15%	2	Tr	Tremor
Explosion de gaz	697	8%	3	Hem	Hembusan
Volcanique de type A	677	8%	4	VA	Vulkanik A
Basse fréquence	329	4%	5	LF	LF
Glissement de terrain	95	1%	6	Eb	Longsoran
Hybride	32	1%	7	Hy	Hibrid

Tab. III.1.4: Répartition des différentes classes au sein du jeu de données (8371 événements)

La composition du jeu de données (voir tableau III.1.4) est très hétérogène : 3 classes représentent plus de 75% du jeu (volcaniques de type B, tectoniques et tremors), et la répartition au sein des classes restantes est inégale (de quelques dizaines à quelques centaines d'événements). Une distribution hétérogène peut poser problème lors de la classification, notamment pour la classification des petites classes. Les tests synthétiques de la figure III.1.4 pour un jeu de données constitué de 3 classes facilement séparables, mais inéquitablement réparties (une classe occupant 50% du jeu, une autre 40% et la troisième 10%), en sont une illustration. On remarque que pour la classe de plus petite taille (blanche), il y a une légère différence entre les PDFs du *training set* et du *test set*. En effet, moins il y a d'éléments disponibles, plus il sera difficile d'avoir un échantillonnage représentatif pour créer le *training set*. De plus, la régression logistique ne classe que très peu d'événements de la petite classe, contrairement à la SVM qui est plus efficace. Pourtant, en termes de pourcentages de réussite globaux, l'écart entre les deux n'est que de 2%. Ceci permet d'insister sur le fait qu'il ne faut pas se fier uniquement

au pourcentage global de classification, qui peut être trompeur, et qu'il vaut mieux accorder de l'importance aux pourcentages de réussite par classe.

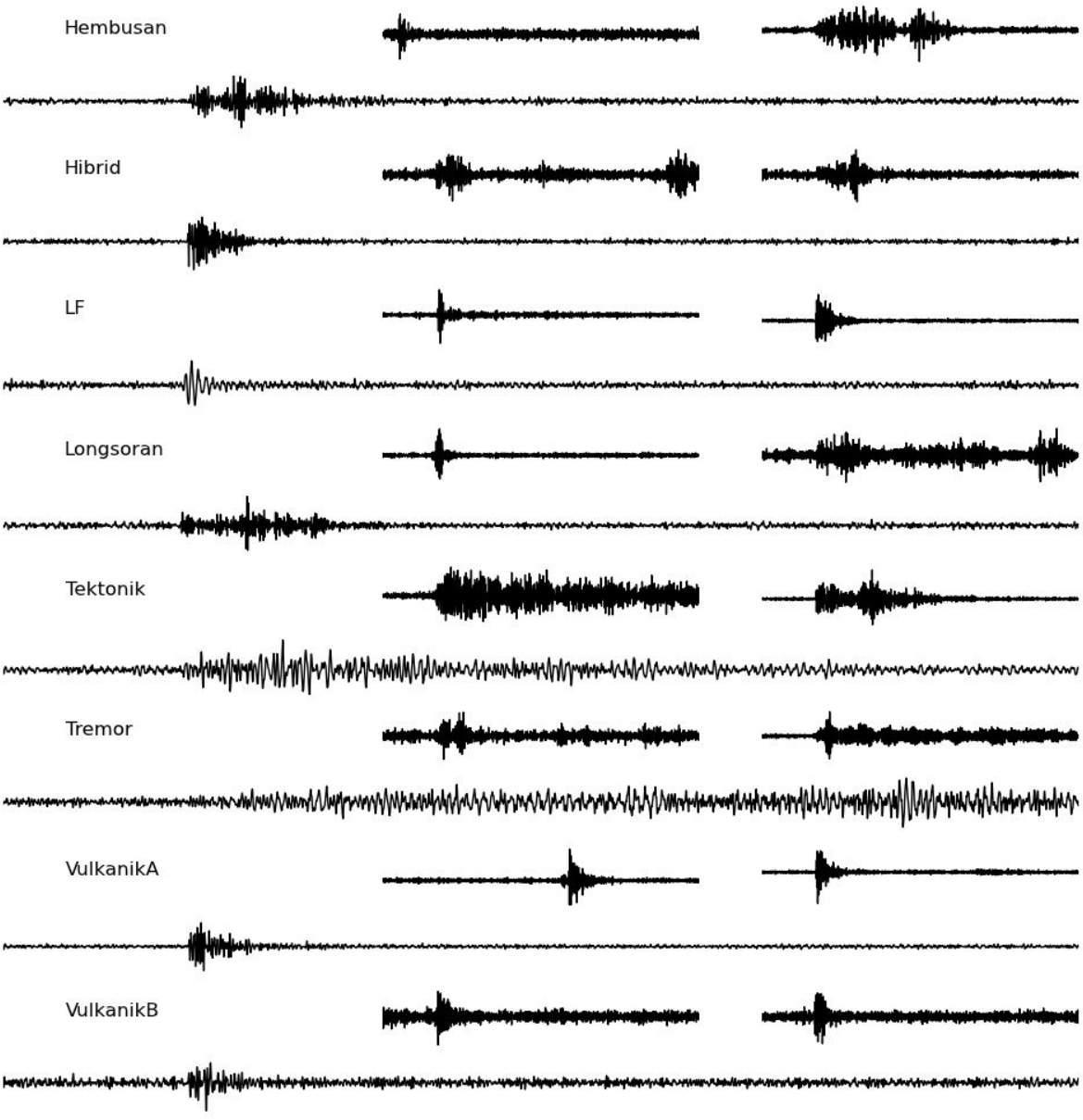
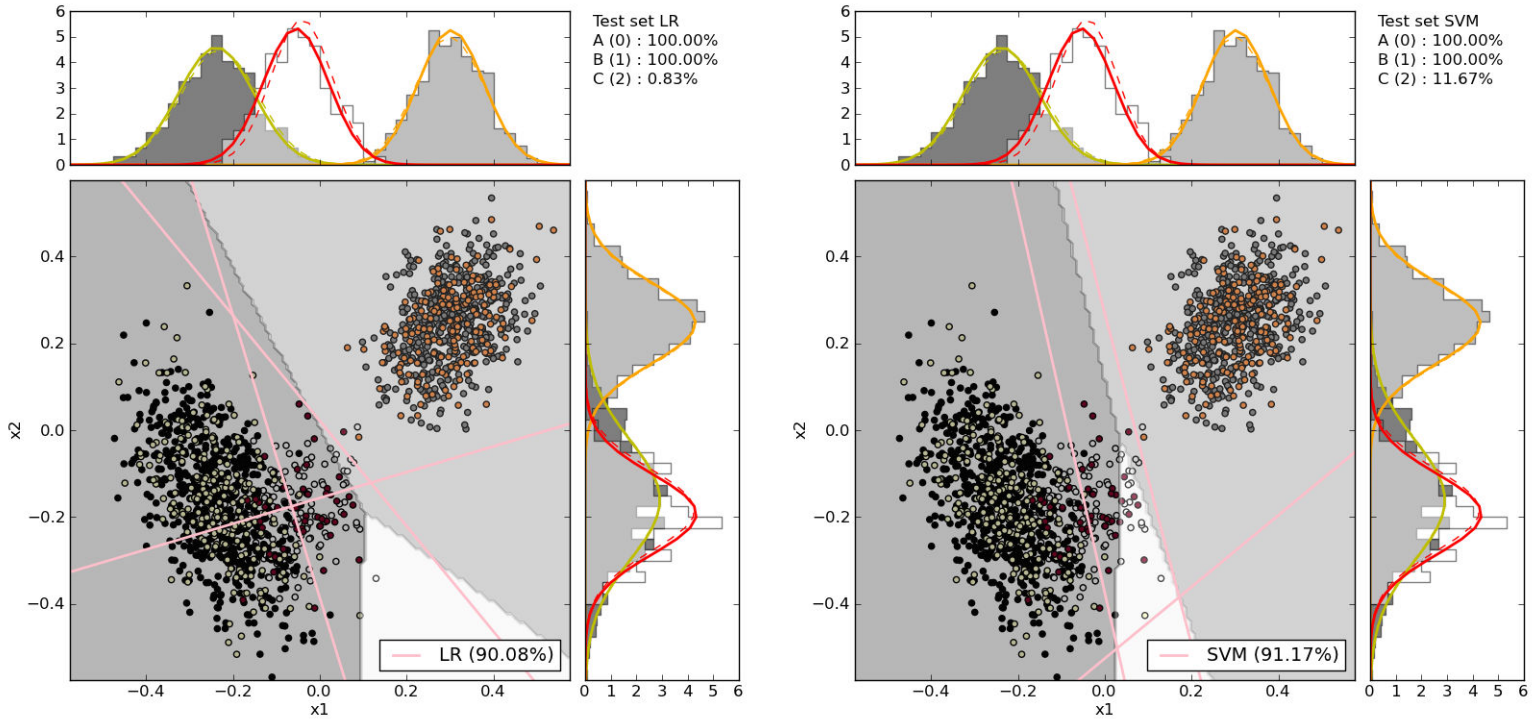


FIG. III.1.3: Sismogrammes des 8 types d'événements principaux enregistrés sur le volcan du Kawah Ijen. Les événements ont été tirés aléatoirement à partir du catalogue brut manuel.



(a) Régression logistique

(b) SVM

FIG. III.1.4: Résultats de la classification pour un jeu de données synthétiques comportant 3 classes inégalement réparties pour la régression logistique (a) et la SVM linéaire (b) . La classe noire constitue 50% des données ; la grise 40% et la blanche 10%. Les éléments du *training set* sont également représentés (en jaune, orange et rouge respectivement). Les densités de probabilité associées aux deux caractéristiques x_1 et x_2 sont visibles de part et d'autre de la figure, avec les courbes continues pour le *test set* et les courbes tiretées pour le *training set*.

CHAPITRE III.2

Classification automatique sur le volcan du Kawah Ijen

III.2.1 Premiers résultats

On essaie, à partir du jeu de données complet, de retrouver la classification manuelle grâce aux méthodes d'apprentissage supervisées décrites en §I.2.2. Ces méthodes nécessitent une phase d'entraînement du système qui permet de déterminer le séparateur optimal à partir d'un jeu de données réduit, mais suffisamment représentatif. Contrairement au cas du Piton de la Fournaise, où le *training set* et le *test set* nous étaient fournis, on ne dispose pas ici d'un *training set*. Il a donc fallu en créer un à partir de l'ensemble des données du *test set* (Fig. III.2.1).

Génération du *training set*

Le *training set* est composé aléatoirement en retenant 40% des événements de chaque classe afin que les proportions de chacune des classes soient respectées. Le *training set* est ensuite lui-même décomposé en trois groupes pour le calcul des courbes d'apprentissage (voir §I.2.2.1 : 60% en *sub-training set*, 20% en *cross- validation set* et 20% en *sub-test set*). L'apprentissage en lui-même se fait donc sur 60% des 40% du jeu de données entier, ce qui devrait déjà permettre sa bonne représentativité. Afin de s'assurer que l'ensemble des données du *test set* peut potentiellement être utilisée dans le *training set*, 10 tirages aléatoires différents ont été effectués. Nous verrons plus tard que la composition du *training set* peut s'avérer primordiale dans les résultats de la classification.

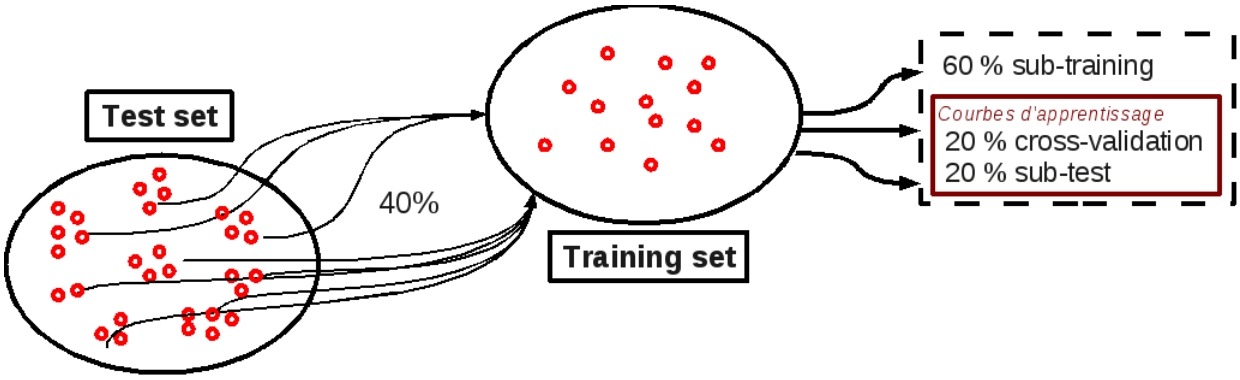


FIG. III.2.1: Génération du *training set* à partir du *test set*.

On n'a pas à disposition toutes les formes d'ondes correspondants aux événements du catalogue (données de trop mauvaise qualité, trous...). Selon les stations considérées, on a de 500 à 6000 formes d'ondes. Les résultats qui suivent sont ceux obtenus pour la station IJEN uniquement, station sur laquelle la classification manuelle a été réalisée, sauf si spécifié différemment. On a choisi cette station, d'une part parce que le catalogue a été effectué à partir de ses données ; d'autre part parce que les données étaient globalement de bonne qualité pour l'ensemble du jeu. Il est bien sûr possible de classer les événements grâce aux données des autres stations, mais les résultats ne seront pas présentés ici.

Le nombre d'événements disponibles à la station IJEN est d'un peu plus de 5900 (au lieu des 8400 du catalogue). La répartition du jeu de données n'est cependant pas modifiée en termes de pourcentages.

Classification multi-station

On présentera tout de même quelques résultats de classification multi-station pour laquelle on a procédé comme suit :

- on réalise la classification pour chaque composante de chaque station : pour une station et une composante données, on crée un *training set* (40% des données disponibles). Durant le processus d'apprentissage, le séparateur est déterminé et permet de prédire les classifications du *test set* pour le même couple (station, composante).
- ensuite, pour chaque événement du catalogue, on réunit l'ensemble des classifications trouvées (leur nombre peut varier selon le nombre de couples (station, composante) où l'événement a été enregistré) et on garde celle qui apparaît le plus souvent. Si plusieurs classes apparaissent un même nombre de fois, on décide de ne pas classer l'événement ; l'idéal étant bien sûr d'avoir une seule et unique classe associée à l'événement.

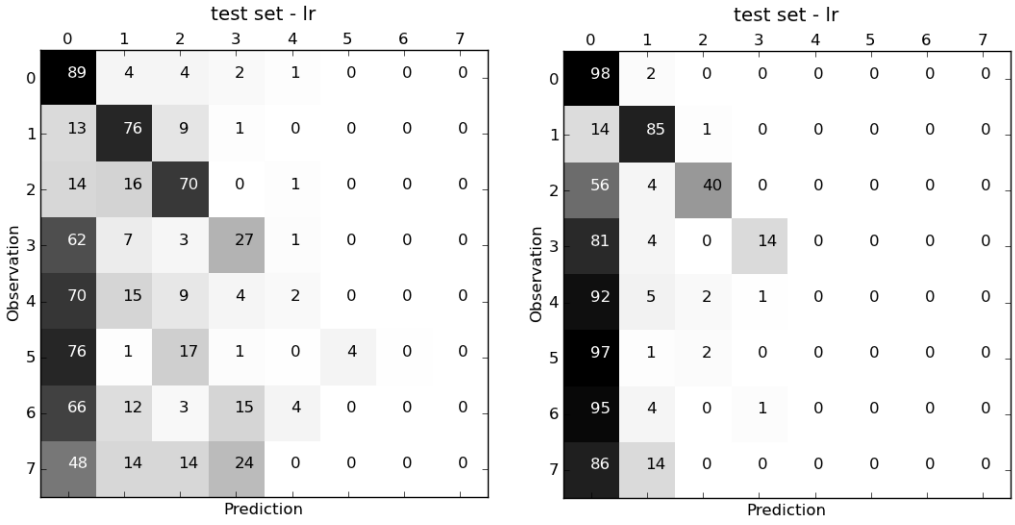
On précise également que, compte tenu de la complexité du jeu de données, avec huit classes, on a préféré utiliser la SVM non-linéaire avec un noyau gaussien (§I.2.2.2) plutôt que la SVM linéaire.

III.2.1.1 Classification sur le catalogue brut

Dans un premier temps, on essaie de distinguer l'ensemble des classes d'événements en utilisant toutes les caractéristiques disponibles (une cinquantaine au total, voir TAB. I.2.2).

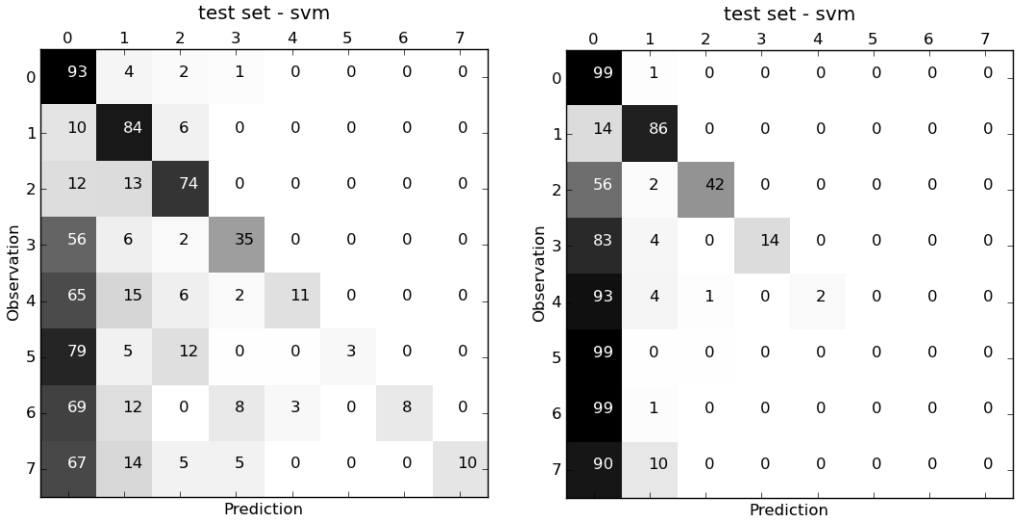
On pourra réduire par la suite le nombre d'attributs utilisés en cherchant à ne garder que ceux qui sont les plus discriminants. Les résultats sont présentés pour la régression logistique (Fig. III.2.2) et la SVM (Fig. III.2.3). On rappelle, pour une meilleure compréhension des résultats, que les matrices de confusion donnent, pour une ligne i donnée, les pourcentages de répartition de la classe i observée dans les différentes classes prédites (voir §I.2.2.4).

Afin d'être sûr que la manière de créer le *training set* était la meilleure qui soit dans notre cas, on a aussi classé les événements en utilisant un *training set* avec un nombre fixe d'événements dans chaque classe (Fig. III.2.4). Afin de conserver une taille de *training set* comparable à la précédente, on a choisi de mettre au plus 400 événements de chaque classe dans le *training set* (comme il existe des classes où le nombre d'événements ne dépasse pas 400, ceux-ci sont présents à la fois dans le *training set* et dans le *test set*, ce qui biaisera les résultats des petites classes). Le maintien d'une taille de *training set* identique était importante car plus celui-ci est grand, plus la prise en compte de la variabilité "naturelle" des événements au sein d'une classe est facile.



(a) Station IJEN (b) Toutes les stations

Fig. III.2.2: Matrices de confusion obtenues après régression logistique. La signification des numéros de classe est donnée dans le tableau III.1.4.



(a) Station IJEN (b) Toutes les stations

FIG. III.2.3: Matrices de confusion obtenues après SVM. La signification des numéros de classe est donnée dans le tableau III.1.4.

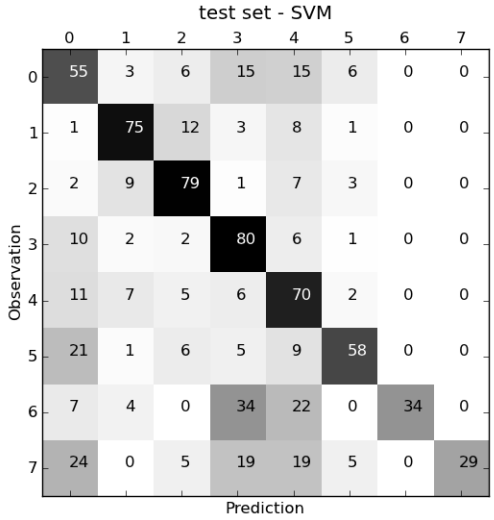


FIG. III.2.4: Matrice de confusion obtenue après SVM à la station IJEN et en utilisant tous les attributs disponibles. Le nombre d'événements de chaque classe dans le *training set* n'est pas proportionnel à la taille de la classe dans le *test set*, mais est fixé **à 400**.

L'analyse des résultats nous apprend que :

- à partir de la station IJEN uniquement (Fig. III.2.2a et Fig. III.2.3a), la classification des événements est correcte pour les 3 classes de plus grande taille qui composent les trois-quarts du jeu de données (VB, tectoniques et tremors), et mauvaise pour les 5 classes restantes. La SVM (non-linéaire) permet d'améliorer les taux de bonne classification de quasiment toutes les classes.

- à partir de toutes les stations disponibles (Fig. III.2.2b et Fig. III.2.3b), la classification des VB et des tectoniques est renforcée au détriment des autres classes. Pour mieux comprendre comment fonctionne la classification multi-classe, on a réprésenté sur la figure III.2.5a l'histogramme du pourcentage de classification identique entre stations pour un événement donné, c'est-à-dire le rapport du nombre de stations où l'événement a été "bien" classé sur le nombre total de stations. On rappelle que la classe de l'événement correspond à la plus représentée. L'histogramme III.2.5b donne la répartition du nombre de stations disponibles sur l'ensemble du catalogue.

- l'immense majorité des événements qui ne sont classés manuellement ni en VB, ni en tectoniques, et, dans une moindre mesure, ni en tremors, est classée en VB. Ceci est encore plus vrai lorsque l'on utilise les résultats obtenus sur l'ensemble des stations.

- lorsque le *training set* utilisé n'est pas proportionnel au *test set* (Fig. III.2.4), on voit que les pourcentages d'extraction sont meilleurs pour les petites classes. En revanche, pour les deux classes principales (VB et tectoniques), la classification se fait moins bien. Ceci est particulièrement visible pour les VB, pour lesquels on atteint seulement un peu plus que les 50% de réussite, ce qui semble indiquer que la classe est sous-représentée dans le *test set*. De plus, il faut souligner que les 3 classes de plus petite taille sont représentées par moins de 400 événements dans le jeu de données complet : les échantillons dans le *training set* et dans le *test set* sont donc identiques, ce qui peut expliquer l'augmentation observée en termes de pourcentages.

Ces premiers résultats montrent qu'il s'avère difficile de classer des événements provenant de classes sous-représentées dans le jeu de données, notamment si le pouvoir discriminatoire de leurs attributs n'est pas très élevé. On a vu aussi qu'il était possible d'améliorer la classification de ces petites classes, en choisissant de générer un *training set* uniforme qui ne respecte pas les proportions de chaque classe ; mais cette amélioration se fait au détriment des grandes classes car le *training set* ne constitue plus un échantillon représentatif du *test set*. Ceci confirme que l'entraînement d'un algorithme automatique a toujours tendance à favoriser le classement dans la classe majoritaire, mais prouve également qu'avec les attributs dont nous disposons il devrait *a priori* être possible de réussir à discriminer les événements des petites classes. Par conséquent, on peut essayer d'améliorer la classification en appliquant différentes stratégies, et plus particulièrement en faisant des extractions successives de chaque classe.

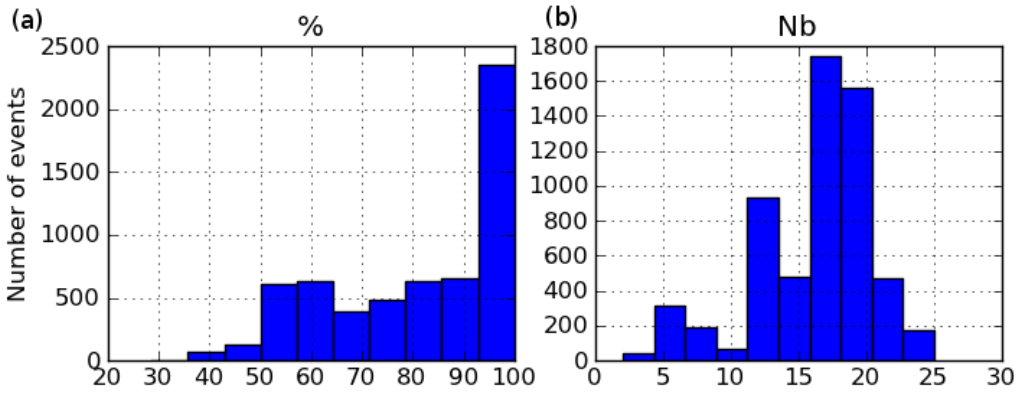


FIG. III.2.5: Histogrammes (a) des pourcentages de classification identique pour un événement donné, (b) du nombre de stations disponibles pour la classification. Ils ont été calculés pour la SVM multi-stations avec toutes les stations et tous les attributs.

III.2.1.2 Stratégie mise en place : les extracteurs

Les extracteurs, comme leur nom l'indique, permettent d'extraire chaque classe une à une et devraient donc assurer la bonne classification des éléments les plus caractéristiques. A chaque extraction, le problème de classification multi-classe est ramené à un problème binaire, ce qui le simplifie fortement.

Deux types d'extracteurs ont été mis en place :

- l'extracteur « **un contre tous** » (FIG. III.2.6a) : on extrait chaque classe une à une en considérant toujours l'ensemble du jeu de données. Le *training set* et le *test set* restent identiques pendant tout le processus. Il est donc possible qu'un événement soit classé dans plusieurs classes ou qu'il ne soit jamais classé.

- l'extracteur « **classe par classe** » (FIG. III.2.6b) : on extrait chaque classe une à une, en commençant par celle de plus grande taille, en supprimant à la fin de chaque extraction les événements déjà classés. La taille du *test set* est donc réduite à l'issue de chaque extraction, ainsi que celle du *training set*. A la fin de toutes les extractions, il existe un certain nombre d'événements qui n'ont pas été classés. Il n'est pas possible qu'un événement soit classé plusieurs fois.

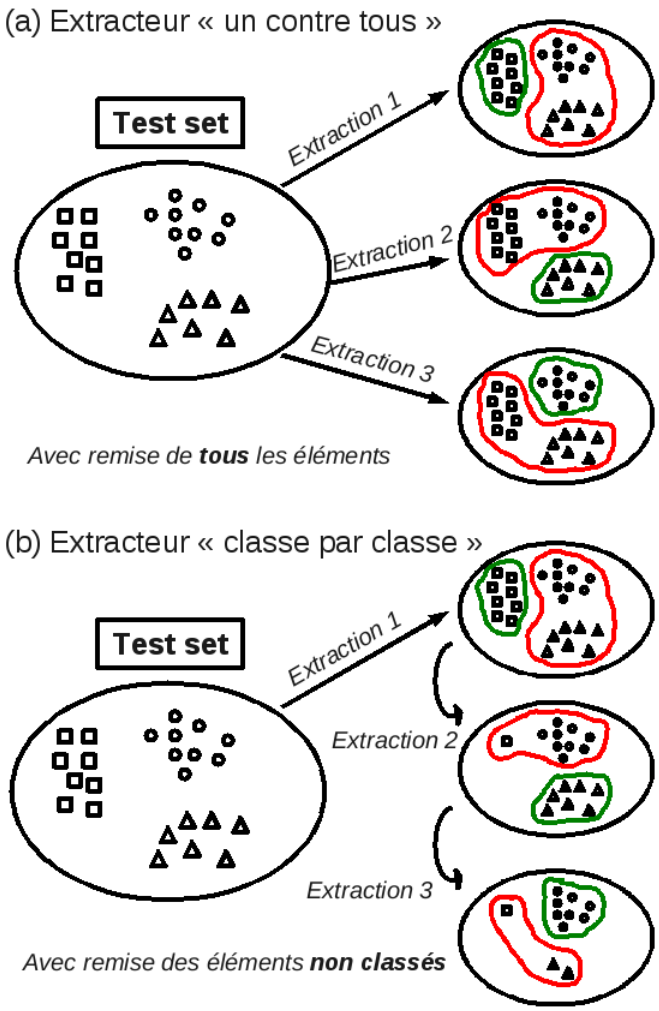


FIG. III.2.6: Représentations schématiques des deux extracteurs. (a) « Un contre tous » : à chaque extraction, le problème devient binaire (la classe à extraire (vert) contre l'ensemble des autres classes (rouge)). Le jeu de données contient toujours le même nombre d'événements. (b) « Classe par classe » : à l'issue de chaque extraction, on supprime les événements qui ont été classés dans la classe à extraire, y compris ceux qui n'appartiennent pas à cette classe d'après la classification manuelle. Les événements appartenant à cette classe mais n'étant pas extraits sont, quant à eux, conservés.

On ne présente dans cette section que les résultats obtenus avec les 2 extracteurs avec la méthode SVM non linéaire et seulement avec les 8 attributs sismiques les plus discriminants, principalement pour des raisons de temps de calcul. Les mêmes études ont été réalisées avec la méthode de régression logistique, mais n'apportent pas d'informations complémentaires.

Le choix de ces 8 attributs sur la base de leur pouvoir discriminant sera explicité un peu plus loin au cours de cette étude.

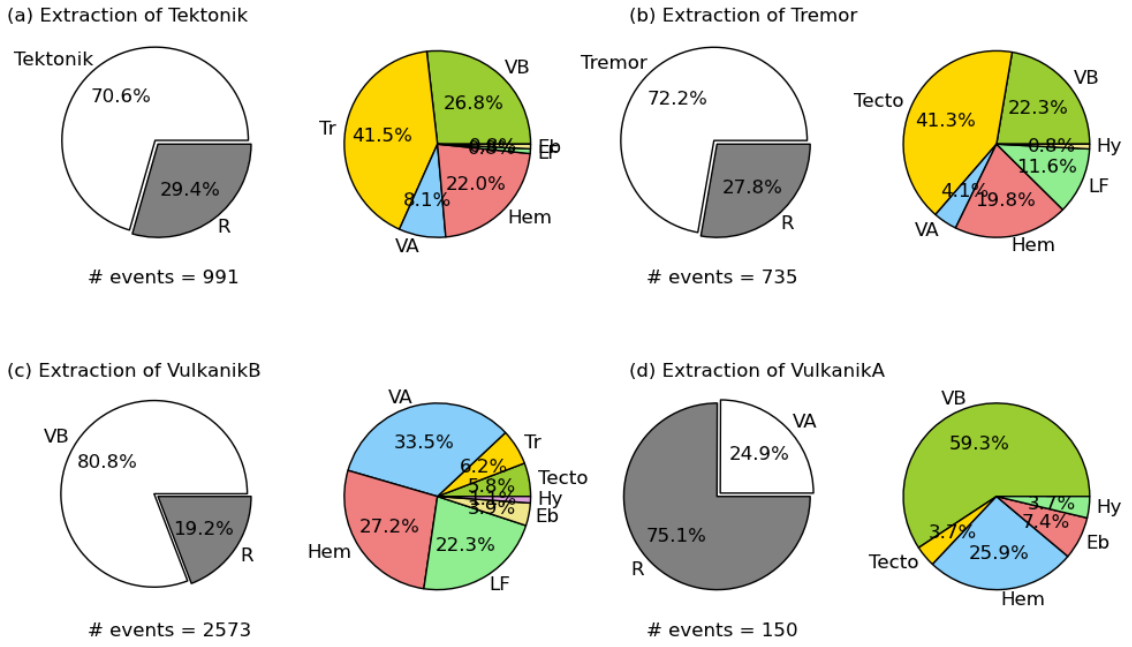
Les résultats obtenus grâce aux 2 extracteurs à partir du catalogue brut et sur les données de la station IJEN sont présentés dans la figure III.2.7. On remarque que les taux d'extraction

des 3 classes principales (VB, tectoniques et tremors) sont compris entre 70 et 80% et restent semblables quel que soit l'extracteur utilisé. Pour la classe des tectoniques, on remarque que la majeure partie des événements mal classés est de type tremor, puis de type VB, puis de type hembusan. Pour la classe des tremors, une majorité d'événements mal classés provient de la classe des tectoniques, puis des hembusans, des VB et des basses fréquences. Pour les VB, les confusions se font pour un tiers avec les VA, pour un "petit" tiers avec les hembusans et pour un "petit" quart avec les basses fréquences. Enfin, pour la classe des VA, qui est d'importance moindre en proportion du catalogue, on note une nette amélioration du taux d'extraction (de 25 à 45%) en fonction du type d'extracteur utilisé. On s'attendait à un tel résultat puisque l'extracteur « classe par classe » réduit le nombre d'événements au fur et à mesure des extractions, limitant ainsi les confusions avec d'autres classes : on remarque que la confusion avec les VB est réduite (de 60 à 40%), ainsi que celle avec les hembusans. Les confusions avec les autres types d'événements, en revanche, augmentent (tectoniques et éboulements surtout). Il est important de souligner que les résultats sont donnés ici en proportion de ce qui a été extrait : autrement dit, il est probable que l'extracteur « classe par classe » ait extrait moins de VA que l'extracteur « un contre tous » puisqu'un certain nombre de VA a déjà été mal classé lors des extractions des classes précédentes (VB, tectoniques, tremors). Si on réduit le nombre d'événements classés par l'extracteur, les chances que ces événements appartiennent effectivement à la classe extraite augmentent mécaniquement (Fig. III.2.8).

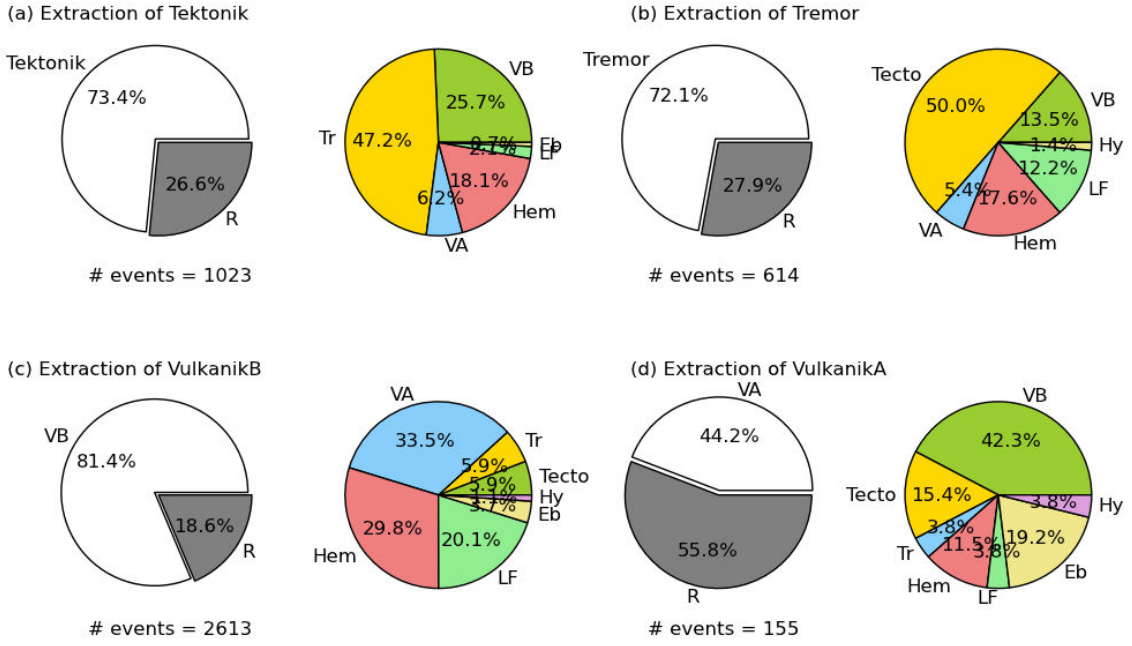
On constate donc qu'il faut manipuler les résultats obtenus avec précaution : la manière dont on les représente revêt une importance capitale et peut s'avérer trompeuse si on n'y prête pas attention. Par exemple, la représentation sous forme de diagrammes de pourcentages que l'on a choisie ici ne rend pas compte des nombres d'événements effectifs extraits dans chaque classe. Conclure que l'extracteur « classe par classe » est meilleur que le « un contre tous » pour la classe des VA pourrait ainsi s'avérer trop hâtif.

La figure III.2.9 représente les diagrammes de répartition de chacune des classes à l'issue des extractions sur l'ensemble du jeu de données. Ceux-ci montrent que les deux extracteurs donnent des résultats très semblables, avec presque la moitié des événements classés en VB et un quart de non classés. On peut alors conclure, grâce à cette figure, que l'extracteur « classe par classe » semble meilleur que le « un contre tous » en ce qui concerne la classe des VA, puisque 44.2% des VA extraits sont bien classés (contre seulement 24.9% pour l'autre extracteur) et que la proportion de VA totale classée est de l'ordre de 2.5% du jeu total pour les deux extracteurs.

La comparaison de ces diagrammes avec le diagramme de répartition du jeu de données initial (Fig. III.2.12a) montre que les proportions des trois plus grandes classes (VB, tectoniques et tremors) sont plutôt bien respectées (dans une limite de ±5%). Pour les autres classes, en revanche, les extractions sont beaucoup plus difficiles : c'est particulièrement notable pour les hembusans, qui ne sont pas extraits du tout alors qu'ils représentent 8% du jeu de données complet. Cette incapacité à extraire correctement les classes de petite taille se traduit par une forte proportion d'événements non classés.



(a) Extracteur « un contre tous »



(b) Extracteur « classe par classe »

FIG. III.2.7: Exemple d'extraction des 4 classes majoritaires d'événements enregistrés à la station IJEN par les 2 types d'extracteurs avec la méthode SVM non linéaire. A chaque classe correspond une paire de diagrammes. Le premier diagramme représente, pour l'ensemble des éléments extraits dans la classe considérée, le pourcentage de bonne classification (partie blanche). La répartition des élements extraits qui ne proviennent pas a priori de la classe considérée (partie grise) est détaillée dans le second diagramme.

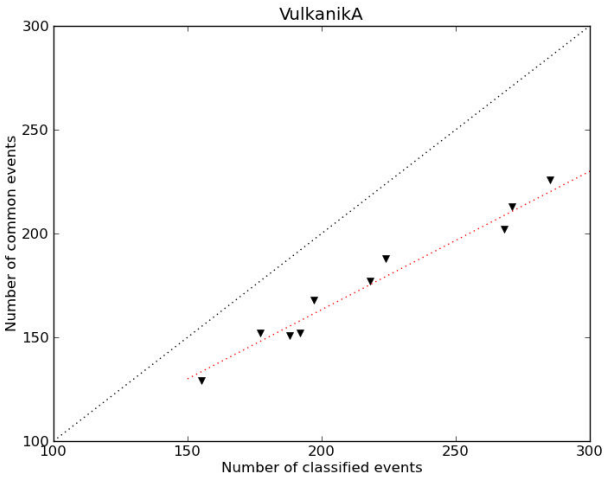
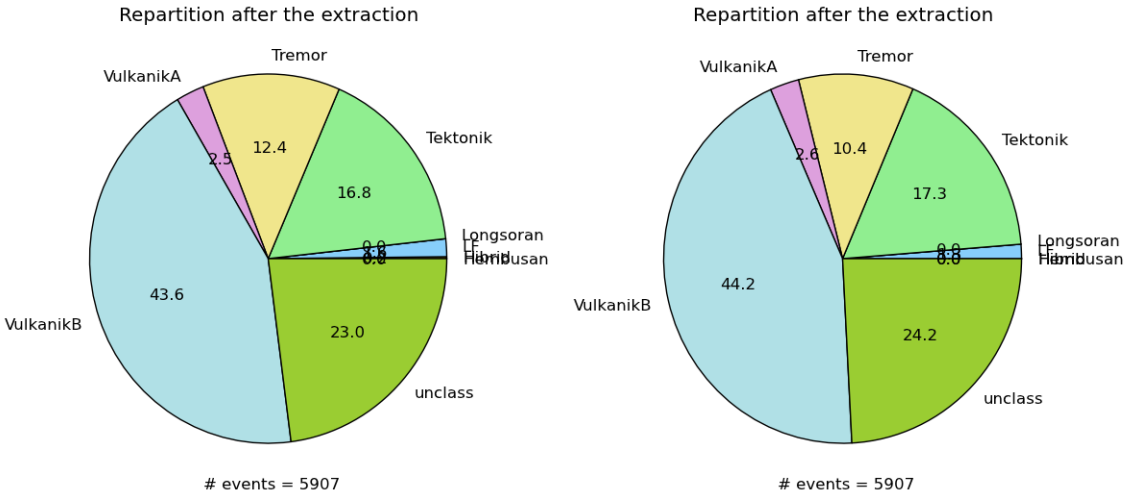


FIG. III.2.8: Evolution du nombre d'événements communs au catalogue manuel en fonction du nombre d'événements classés dans la classe des volcaniques de type A. La ligne pointillée noire symbolise l'égalité entre les deux. La ligne pointillée rouge symbolise la tendance observée sur 10 extractions différentes : elle s'éloigne de la ligne noire et croît plus lentement, ce qui signifie que plus le nombre d'événements classés devient grand, plus il est difficile de retrouver tous les événements communs (d'où une baisse du taux de réussite).



(a) Extracteur « un contre tous »

(b) Extracteur « classe par classe »

FIG. III.2.9: Diagrammes de répartition des classes à la fin de chacune des extractions.

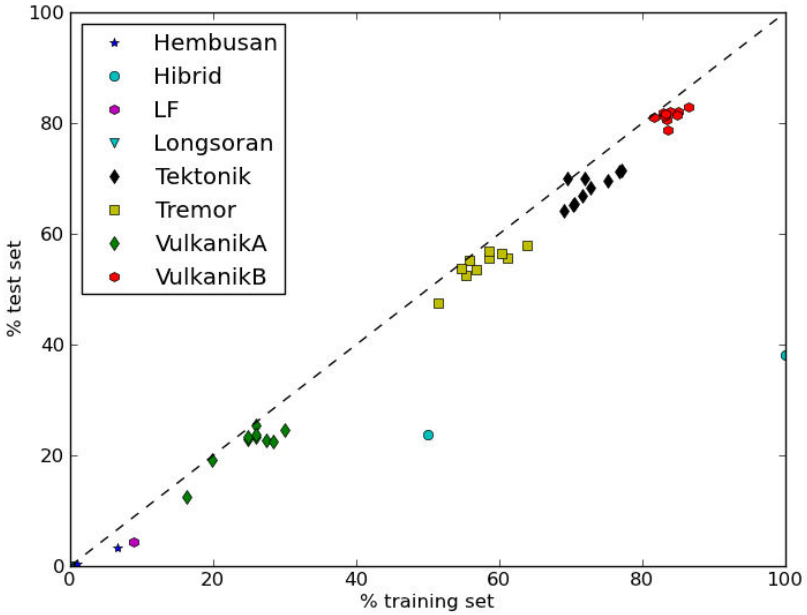
Remarquons également que les résultats de la figure III.2.7 sont présentés pour un tirage de *training set* donné (III.2.1). L'inspection des résultats pour d'autres *training sets* a montré que :

- les résultats sont stables pour les trois classes principales (VB, tectoniques, tremors) au sein d'un même extracteur.

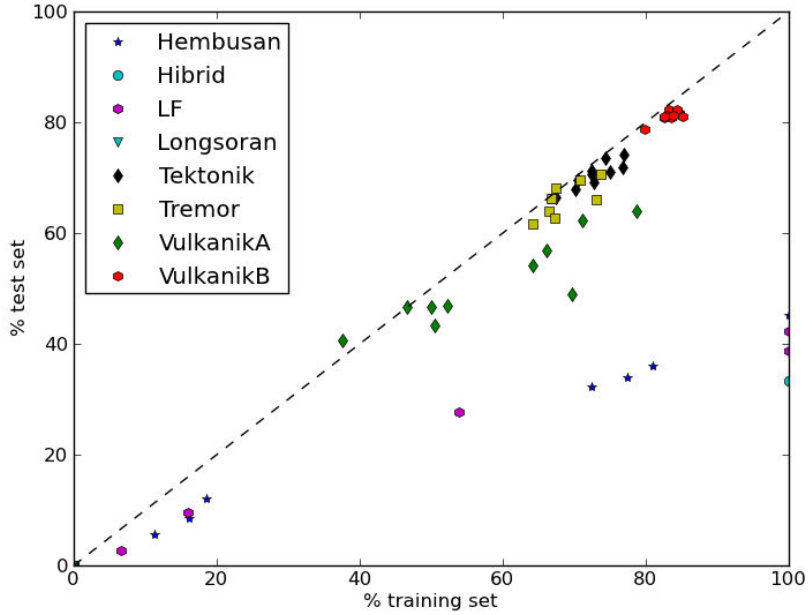
- les résultats sont semblables pour les trois classes principales quel que soit l'extracteur considéré.

- il existe une plus grande disparité dans les résultats pour la classe des VA. Dans le cas de l'extracteur « un contre tous », le taux de bonne classification oscille entre 20 et 30% ; dans le cas de l'extracteur « classe par classe », il est compris entre 45 et 70%. La grande différence observée dans le second cas peut s'expliquer par la forte dépendance qui existe avec les résultats des extractions précédentes. De plus, la distribution des classes avec lesquelles les VA sont confondus montre que, dans le cas de l'extraction « un contre tous », les VB sont la principale source de mauvaise classification (les proportions variant généralement de 40 à 75%), suivis des hembusans (20 à 40%), puis des tectoniques et des glissements de terrain (dans des proportions assez variables), les autres classes étant assez rarement représentées. Dans le cas de l'extracteur « classe par classe », une grande majorité de VB, tectoniques et tremors est déjà classée avant l'extraction des VA : les confusions possibles avec les autres classes, de plus petites tailles, vont s'accroître, notamment avec celle des basses fréquences et celle des longsorans. On note également plus de confusions avec les tectoniques.

On utilise la représentation suivante pour essayer d'affiner la compréhension que l'on a de nos résultats : les taux de réussite de classification du *test set* en fonction de ceux du *training set*. Afin d'étudier l'influence que pourrait avoir la composition du *training set* sur les résultats, on effectue 10 fois le processus de classification avec génération d'un nouveau *training set*. Pour que les comparaisons soient valables, les 10 *training sets* ont été conservés et sont identiques quelle que soit la méthode utilisée. La ligne pointillée noire permet de voir si la généralisation de la fonction hypothèse trouvée grâce au *training set* est correcte : plus on s'en approche, mieux c'est. Lorsque le pourcentage du *test set* est inférieur à celui du *training set*, alors on est en situation de sur-apprentissage : l'erreur par rapport au *training set* est minimisée, mais la généralisation n'est pas parfaite et la variance est élevée (voir §I.2.2.1).



(a) Extracteur « un contre tous »



(b) Extracteur « classe par classe »

FIG. III.2.10: Pourcentages de bonne classification du *test set* vs celui du *training set* pour chaque classe d'événements. La méthode SVM non linéaire a été utilisée pour les extractions.

Les résultats des extractions pour 10 tirages de *training set* différents et avec la méthode SVM non linéaire sont présentés sur la figure III.2.10 et montrent que :

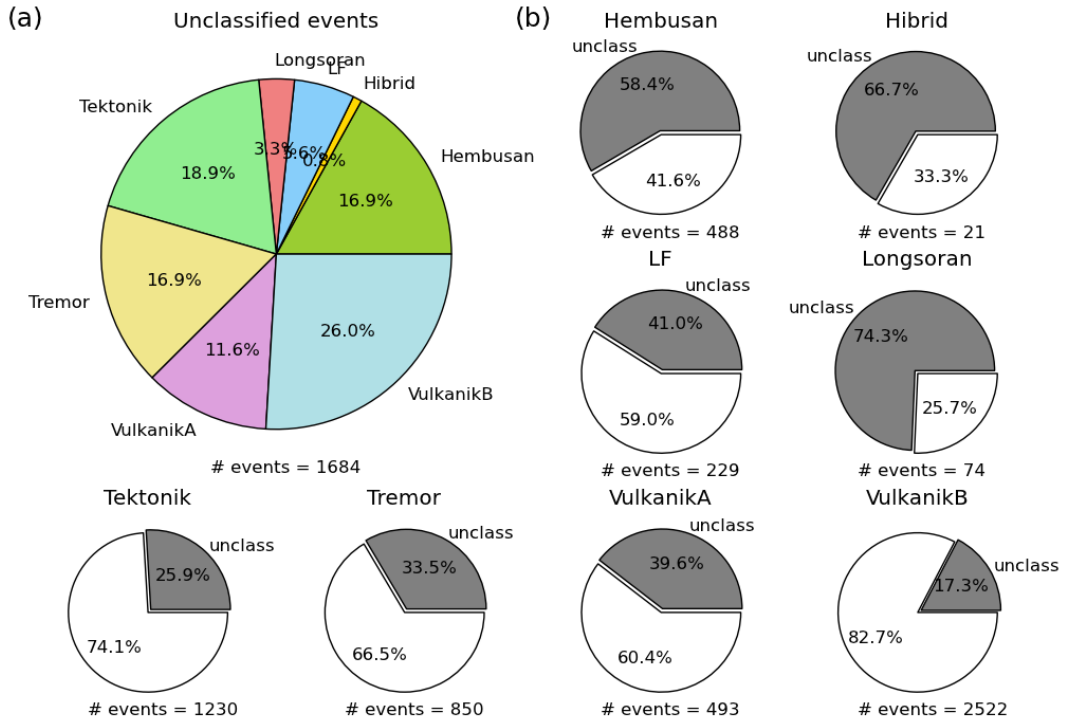
- trois classes seulement sont bien classées : les VB (environ 80%), les tectoniques (autour de 70%) et les tremors (de 50 à plus de 60%).

- l'extracteur « classe par classe » permet d'améliorer sensiblement l'extraction des classes, notamment celle des tremors et des VA. Ceci s'explique par le fait que les événements déjà classés soient supprimés au cours des extractions : il y a donc moins de confusions possibles. Dans le cas des VA, si le taux d'extraction peut approcher des 60%, on note que la généralisation au *test set* diminue (sur-apprentissage). On note également une forte dispersion des résultats.

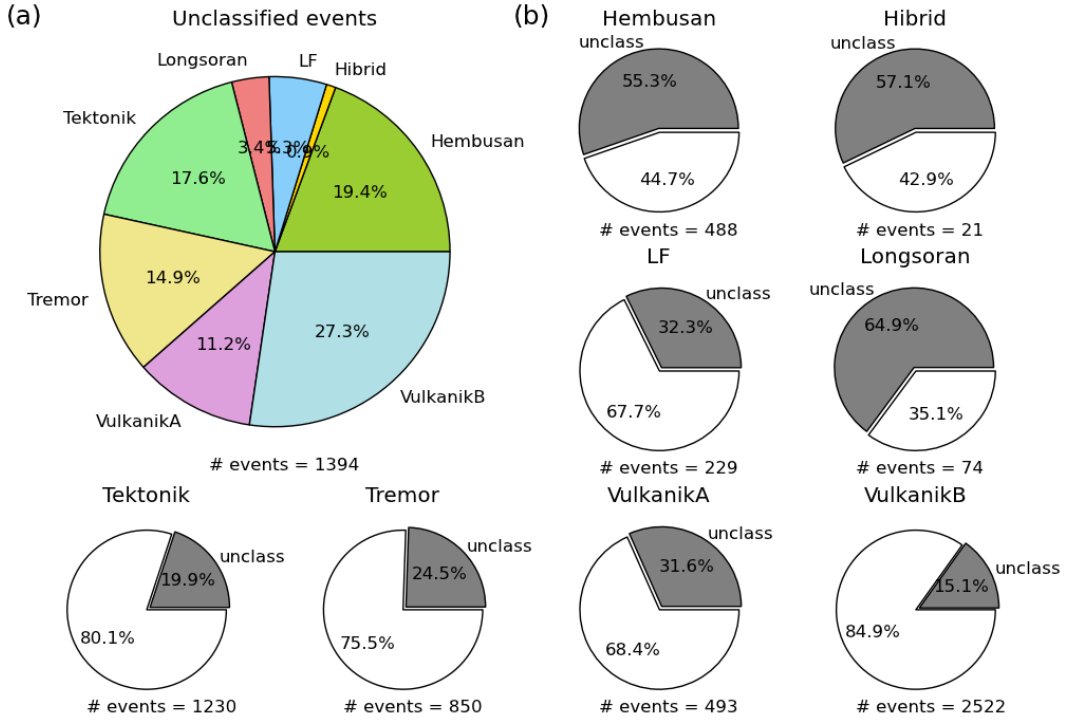
- on n'arrive pas à classer correctement les événements appartenant aux classes suivantes : hembusans, hybrides, BF et glissements de terrain.

- la dispersion des résultats due au *training set* est inférieure à une dizaine de % pour les 3 classes principales. La dispersion des résultats est plus importante dans le cas de l'extracteur « classe par classe » car on effectue un nouveau tirage de *training set* à chaque étape : ce tirage fait aléatoirement dépend en plus des résultats de l'extraction précédente, ce qui induit une forte variabilité.

A l'issue des extractions, il est possible que des événements ne soient classés dans aucune classe. La figure III.2.11 détaille, pour un tirage donné, la composition des événements non classés ainsi que leurs proportions pour chaque classe. On remarque que les résultats obtenus sont similaires pour les 2 extracteurs, bien que, globalement, les pourcentages d'événements non classés diminuent avec l'extracteur « classe par classe ». Ceci n'est pas surprenant dans la mesure où les événements déjà classés sont éliminés au fur et à mesure des extractions, réduisant ainsi le nombre d'événements à classer et augmentant les chances de classer les événements restants. Un quart des événements non classés appartient à la classe des VB, mais ceci ne représente que 15% de la classe entière des VB. Un quart des événements tectoniques et un quart à un tiers des tremors ne sont pas classés. Pour les classes de plus petite taille, assez peu représentées dans le jeu de données initial comme les les hybrides et les glissements de terrain, environ 2/3 des événements ne sont pas classés. Enfin, plus de la moitié des hembusans n'est pas classée non plus, bien qu'ils représentent presque 10% du jeu de données.



(a) Un contre tous



(b) Classe par classe

FIG. III.2.11: Pour chacune des sous figures : (a) Diagramme de répartition des événements non classés à l'issue des extractions. (b) Diagrammes de répartition des événements classés et non classés pour chaque classe.

La figure III.2.12 complète les observations précédentes en présentant le diagramme de répartition des événements qui ne sont systématiquement pas classés après les extractions après 10 tirages de *training set* différents. Idéalement, pour un jeu de données avec des classes bien séparables, si le classifieur est bien entrainé et bien équilibré par rapport aux différentes classes, il aurait été logique que le diagramme des événements non-classés conserve les mêmes proportions que le diagramme du jeu de données en entier, voire qu'il n'y ait pas d'inclassés du tout. Or on remarque que la part des VB non classés est beaucoup moins importante que la part des VB dans le jeu de données, ce qui signifie que l'on classe plutôt bien les VB. En revanche, la classe des hembusans constitue plus d'un tiers des non classés : il est donc plus difficile de classer ces événements. De même, presque 10% des événements jamais classés appartiennent à la classe des glissements de terrain.

L'extracteur « classe par classe » fournit des résultats très similaires à l'extracteur « un contre tous ». On ne présente donc pas ses résultats.

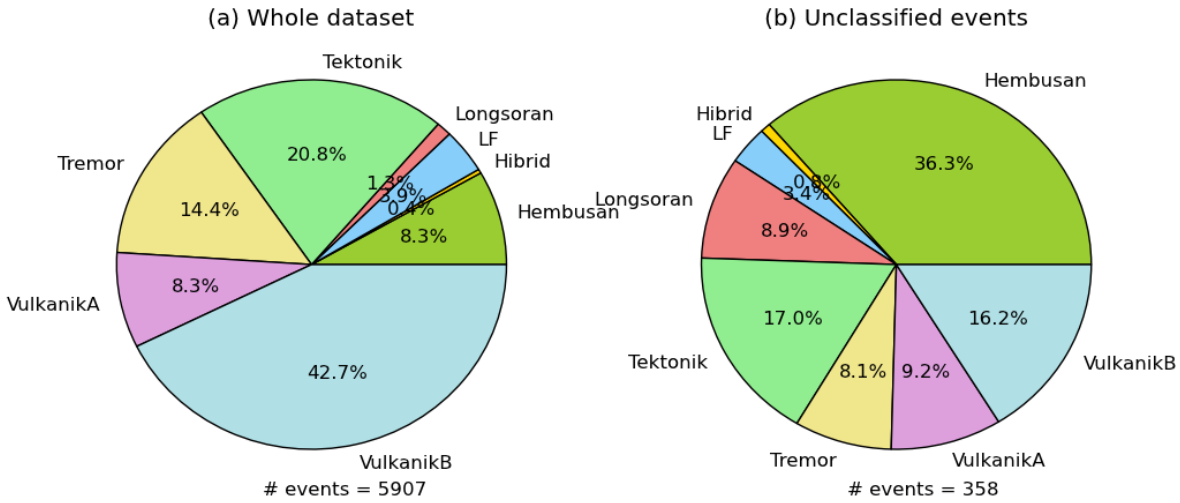


FIG. III.2.12: Diagrammes de répartition : (a) du jeu de données ; (b) des événements jamais classés après 10 tirages de *training set* pour l'extracteur « un contre tous ».

III.2.1.3 Bilan pour le catalogue brut

Les résultats mettent en évidence que les trois classes principales (VB, tectoniques et tremors.) se séparent relativement bien. Il est plus difficile de classer les événements des autres classes, comme les hembusans dont un certain nombre n'est jamais classé au cours des extractions. On peut aussi conclure que la source de confusion pour les VB est constituée par les VA et les hembusans et qu'il existe également une confusion importante entre tectoniques et tremors.

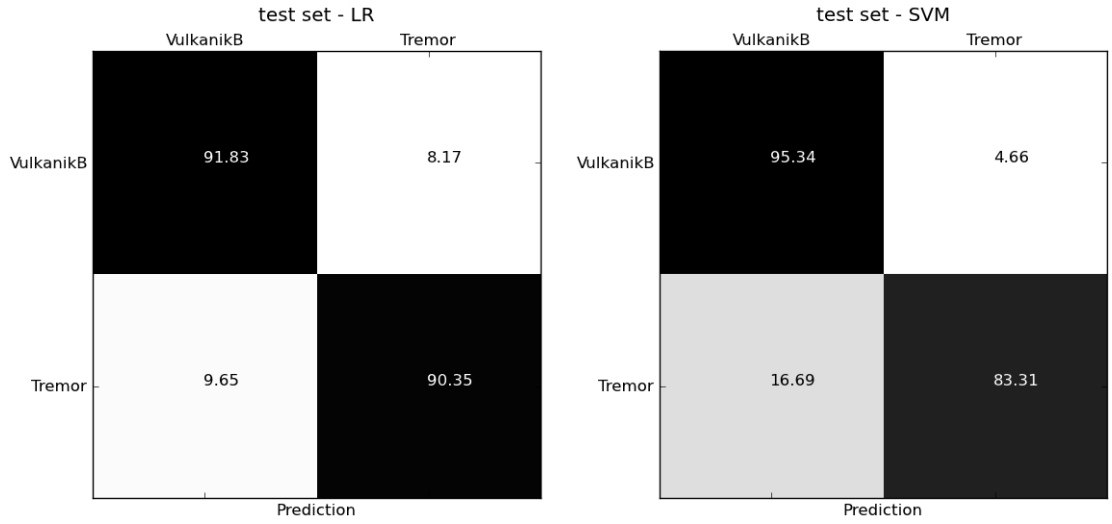
Si les confusions avec les VA et les hembusans peuvent sembler "normales" (les VA et les VB ne diffèrent que par leur profondeur supposée, et les hembusans regroupent des événements qui n'ont pas de "forme d'onde spécifique"), celle avec les tectoniques pose plus de problème, puisque ceux-ci sont généralement très facilement identifiables (avec une onde S bien visible).

De plus, ceci est gênant dans la mesure où les tectoniques ne sont pas impliqués dans les phénomènes liés au volcan ; ce ne sont donc pas les événements qui nous intéressent le plus.

Les résultats suggèrent donc dans un premier temps une simplification et/ou une reclassification du catalogue. En effet, le catalogue est fortement déséquilibré et cela pose problème pour la classification : regrouper des classes "proches", partageant des caractéristiques similaires et difficilement discernables, peut permettre de simplifier le processus de classification. De plus, l'examen des formes d'ondes (Fig. III.1.3) et des densités de probabilité calculées pour tous les attributs (Tab. I.2.2) montre clairement que la fiabilité du catalogue de départ peut être remise en question. Une reclassification est donc la solution envisageable.

Dans ce qui suit, on se focalisera donc plus spécifiquement sur les VB et les tremors, la classe des tectoniques n'apparaissant pas forcément fiable en terme de classification manuelle (Fig. III.1.3).

III.2.2 Résultats pour le catalogue brut restreint à 2 classes



(a) Régression logistique (b) SVM

Fig. III.2.13: Matrices de confusion obtenues après classification des VB et des tremors uniquement, pour la station IJEN avec tous les attributs.

On repart du catalogue brut en ne gardant que les VB et les tremors. Les résultats présentés sur la figure III.2.13 montrent que la séparation entre les deux classes est très bonne (supérieure à 90%). La comparaison entre la régression logistique et la SVM met en évidence que la méthode linéaire suffit pour classer ces 2 classes. Il est même intéressant de noter que la classification des tremors est bien meilleure avec la régression logistique. Les résultats obtenus pour une SVM linéaire sont presque identiques à ceux obtenus avec la régression linéaire. On peut expliquer cette supériorité des méthodes linéaires par le fait que les deux classes se séparent bien naturellement. La méthode non-linéaire va avoir tendance à trop s'ajuster aux

données du *training set*, ne permettant pas ainsi une bonne généralisation lors du passage au *test set* (Fig. III.2.14). On peut d'ailleurs supposer que si les résultats sont moins bons pour les tremors avec la SVM non-linéaire, c'est que le système s'est trop bien ajusté aux données des VB, présents en plus grand nombre, lors de la phase d'apprentissage.

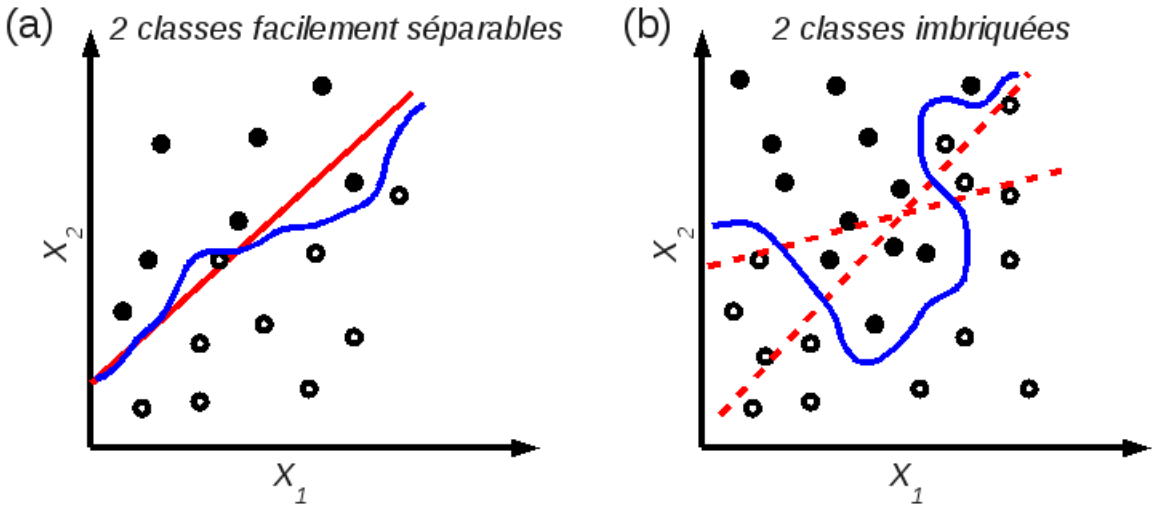


Fig. III.2.14: Représentation schématique de deux cas de figure : (a) Séparation de deux classes bien distinctes : le séparateur linéaire (rouge) est meilleur que le non-linéaire (bleu) car il permettra une meilleure généralisation au *test set*. (b) Séparation de deux classes "mélangées" : aucun séparateur linéaire ne paraît convenir (lignes tiretées rouges), le séparateur non linéaire est préférable.

La figure III.2.15 montre l'influence du choix des caractéristiques utilisées. Les résultats sont présentés pour la régression logistique et la SVM non linéaire pour tous les attributs (indiqué dans les figures comme 50 attributs), une sélection de 30 attributs et puis les 8 meilleurs attributs (voir §I.2.4). Ces derniers ont été choisis après l'examen des densités de probabilité, en retenant ceux qui prennent des valeurs permettant de bien séparer les 2 types d'événements assez distinctement. Les 8 attributs les plus discriminants ici semblent être : le temps du centroïde, la durée, l'énergie du signal dans la bande 0-5 Hz, la fréquence supérieure déterminée par l'analyse du kurtogramme, le kurtosis, le rapport du maximum sur la moyenne de l'enveloppe du signal, l'asymétrie et le temps relatif correspondant au maximum du spectrogramme. Lorsqu'on garde les 30 meilleurs attributs, on ajoute notamment à la liste précédente les valeurs de la fréquence instantanée et de la fréquence prédominante au cours du temps.

Avec tous les attributs, les résultats sont supérieurs à 92% quelle que soit la méthode utilisée. Ils sont légèrement meilleurs en apprentissage pour la SVM que pour la régression logistique, mais avec une généralisation moins bonne. Lorsque l'on réduit le nombre de caractéristiques à 30, les résultats obtenus avec la régression logistique sont comparables aux précédents. Pour la SVM, on perd un peu en taux de bonne classification, mais le problème de sur-apprentissage évoqué précédemment est moins flagrant et la généralisation du *training* au *test set* est meilleure. Lorsque l'on réduit encore plus drastiquement le nombre d'attributs,

en ne gardant que les 8 "meilleurs" selon les probabilités de densité calculées, on voit que les résultats de la SVM ne sont pas tellement affectés par rapport au cas avec 30 attributs. En revanche, les résultats de la régression logistique sont fortement affectés (les % sont désormais compris entre 87 et 92%). On note également une plus forte dispersion des résultats en fonction du *training set* considéré, ce qui semble montrer qu'il faut essayer de trouver un bon équilibre entre le nombre d'attributs considéré et le nombre d'échantillons du *training set*. En effet, s'il y a une variabilité naturelle à l'intérieur d'une classe et si on abaisse la taille du *training set*, on accroît la variabilité des résultats finaux, car le *training set* sera moins "représentatif".

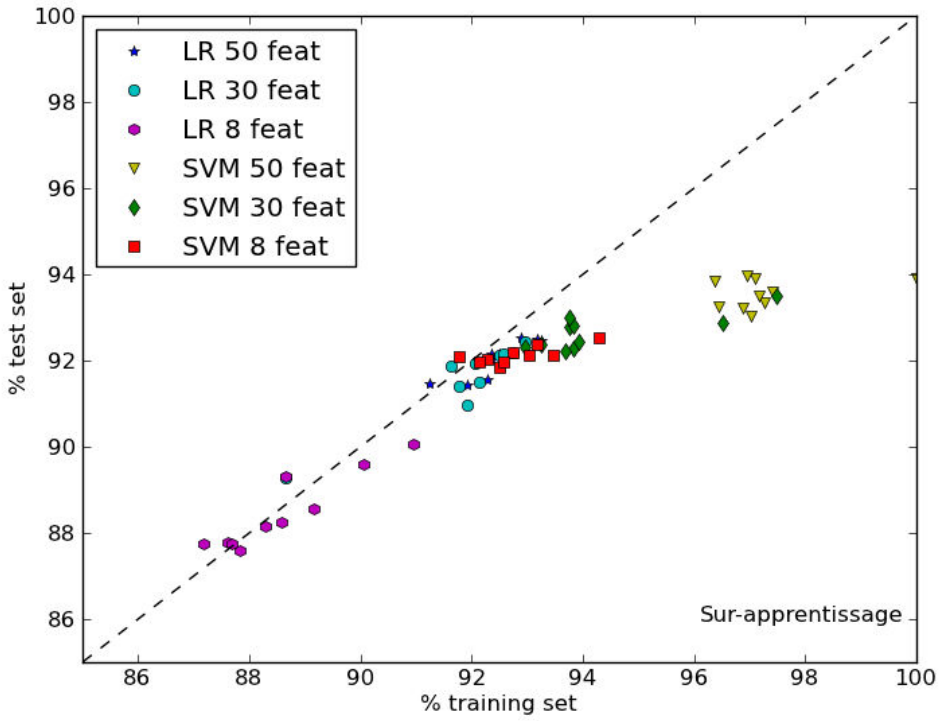


FIG. III.2.15: Résultats obtenus pour séparer les VB et les tremors seulement avec la régression logistique et la SVM selon le nombre d'attributs utilisés. Dix *training sets* différents ont été générés à chaque essai.

Ces résultats permettent de tirer les conclusions suivantes :

- pour la SVM non-linéaire, il est plus intéressant d'utiliser un nombre d'attributs relativement faible (8). Ceci permet de pallier aux problèmes de sur-apprentissage et n'affecte que peu les résultats finaux. La non-linéarité de l'apprentissage compense le faible nombre d'attributs.

- pour la régression logistique, il ne faut pas réduire le nombre d'attributs de manière trop importante, de manière à conserver un taux de réussite assez élevé, car le processus d'apprentissage n'est que linéaire.

On voit donc ici que la SVM non-linéaire avec un nombre de caractéristiques réduit est équivalente à la régression logistique avec plus de caractéristiques. En effet, la SVM non-linéaire transforme l'espace des données en un espace de plus grande dimension (voir §I.2.2.2), introduisant ainsi plus de variables d'ajustement.

III.2.3 Reclassification

On a vu en section III.2.1 qu'essayer de séparer les 8 classes d'origine à partir du catalogue brut donnait des résultats peu probants, surtout pour le classement des événements appartenant aux petites classes. On a aussi vu dans la section précédente qu'on n'avait pas de difficultés particulières à séparer les deux classes principales, les VB et les tremors, quand elles sont les seules présentes dans le catalogue. On s'intéresse dans cette section à la séparation de ces deux classes au sein d'un jeu de données contenant aussi d'autres types d'événements.

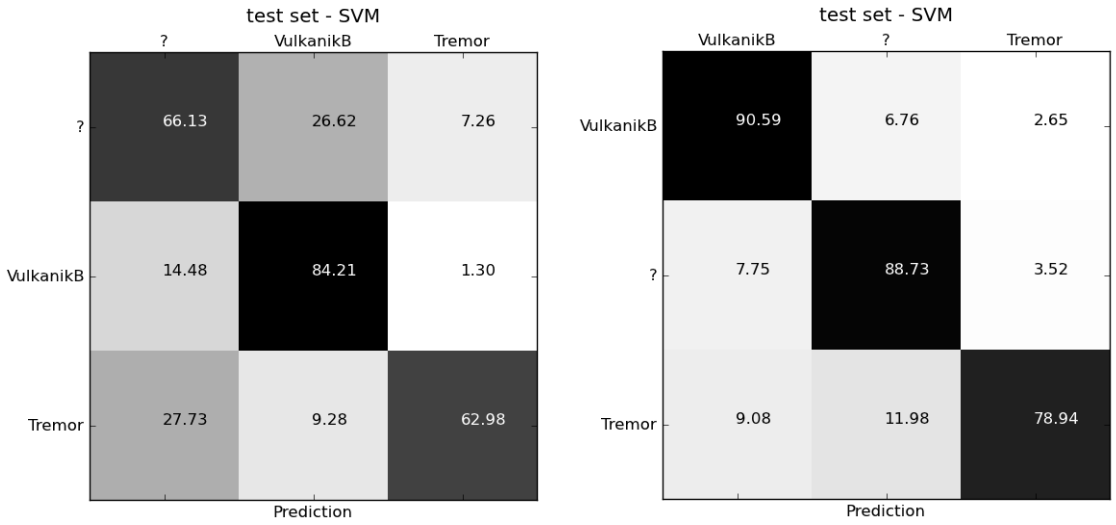
III.2.3.1 Catalogue brut ramené à 3 classes

On introduit, donc, une troisième classe qui contient l'ensemble des événements des autres classes, sans distinction entre eux. On la nommera « classe des indéterminés ». Le tableau III.2.1 présente la répartition du catalogue brut dans ces trois classes : deux classes, les VB et les indéterminés, sont de taille identique, et la troisième, celle des tremors, est plus petite (un tiers environ des deux autres classes). On appellera ce catalogue « **catalogue simplifié** ».

Classe	Catalogue simplifié		Catalogue amélioré	
	Nombre	% du catalogue	Nombre	% du catalogue
Volcanique de type B	2534	43%	3209	54%
Tremor	851	14%	1035	18%
Indéterminé (?)	2536	43%	1677	28%

Tab. III.2.1: Répartition du catalogue brut (5921 événements) ramené à 3 classes (colonnes 2 et 3), puis amélioré par un premier passage de la SVM (colonnes 4 et 5).

La matrice de confusion obtenue avec la méthode SVM avec 8 attributs est présentée sur la figure III.2.16a. Elle montre que si la classification des VB est bonne (80%), elle l'est nettement moins pour les tremors et les indéterminés (autour de 60%).
La première ligne de la matrice de confusion nous apprend qu'un quart des événements indéterminés est classé automatiquement dans la classe des VB. Comme cela représente un nombre d'événements non négligeable, on peut se poser une nouvelle fois la question de la validité du catalogue. Il serait donc intéressant de savoir si les événements indéterminés classés en VB ne sont pas réellement des VB. Pour cela, on crée un nouveau catalogue, qu'on nommera « catalogue amélioré », dans lequel l'ensemble des événements indéterminés classés automatiquement en VB et tremors est classé respectivement en VB et tremors (voir Tab. III.2.1) : on fait donc confiance au classement effectué par le premier passage de l'algorithme de classification. Les résultats obtenus avec ce catalogue amélioré en utilisant la SVM sont visibles sur la figure III.2.16b.



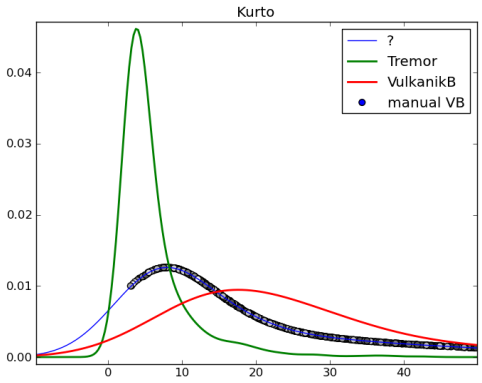
(a) Catalogue brut ramené à 3 classes (b) Catalogue brut ramené à 3 classes amélioré

FIG. III.2.16: Matrices de confusion obtenues avec la SVM sur le catalogue brut ramené à 3 classes puis amélioré après un passage de la SVM en utilisant 8 attributs.

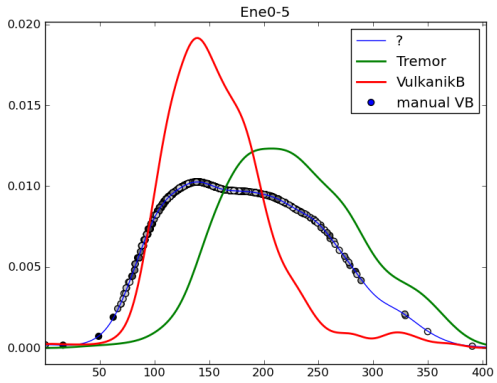
Dans le catalogue amélioré par les premiers résultats de la SVM, la proportion d'événements indéterminés a été fortement amoindrie au profit des VB (voir TAB. III.2.1), ce qui déséquilibre encore plus le jeu de données. La matrice de confusion obtenue à partir du catalogue amélioré montre que l'on retrouve maintenant 90% des VB par rapport aux 66% en utilisant catalogue brut ; les indéterminés et les tremors sont aussi bien mieux classés. Cette amélioration dans les résultats est naturelle et attendue, car on a simplement conforté l'apprentissage de l'algorithme de classification. Mais est-ce qu'on ce choix était-il justifié ? Est-ce que les événements que le premier passage du SVM a reclassé en VB ou tremors le sont vraiment ?

Pour vérifier cela, on regarde la valeur des attributs des événements reclassés et on les compare aux densités de probabilité calculées pour les VB et les tremors (FIG. III.2.17). On voit, dans le cas des tremors (sous-figures c et d), que les valeurs prises pour le kurtosis et l'énergie entre 0 et 5 Hz correspondent bien à celles prises par des tremors (densité de probabilité en vert). Pour les VB, en revanche, la séparation est un peu moins nette car les valeurs des attributs pour les événements reclassés couvrent une large gamme, correspondant parfois plus aux fortes densités de probabilité pour les tremors plutôt que pour les VB. De plus, les densités de probabilités pour ces deux attributs se recouvrent sur un partie relativement large de la gamme de valeurs possibles, ce qui ne facilite pas cette analyse simpliste en attributs uniques, qui ne peut pas prendre en compte la combinaison des attributs.

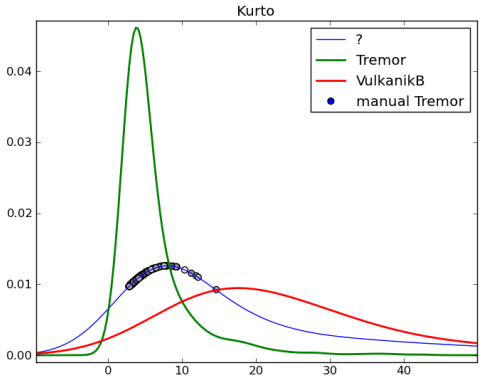
Si on compare les densités de probabilité pour ces deux attributs pour le catalogue brut et pour celui qui a été amélioré, on ne note aucune différence (FIG. III.2.18), ce qui nous rassure sur le fait que la reclassification automatique a bien suivi l'information disponible dans le jeu de données d'entraînement.



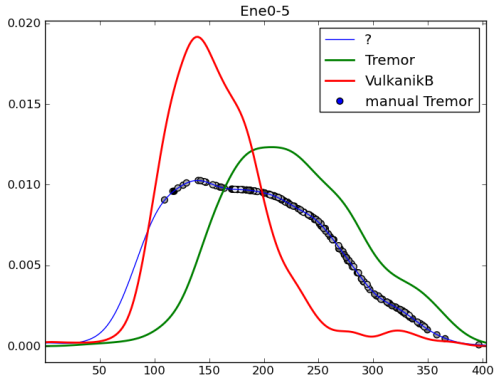
(a) VT - Kurtosis



(b) VT - Energie entre 0 et 5 Hz

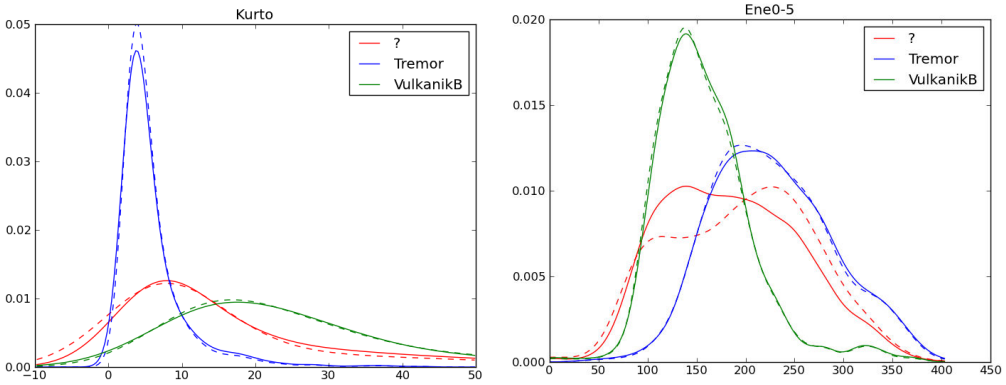


(c) Tremor - Kurtosis



(d) Tremor - Energie entre 0 et 5 Hz

FIG. III.2.17: Valeurs de certaines caractéristiques prises par les événements non classés manuellement mais classés automatiquement en VT ou tremors. Les courbes de densité de probabilité (traits pleins) sont calculées sur l'ensemble du jeu de données. Les points correspondent aux valeurs des attributs prises par les événements classés dans les indéterminés manuellement, mais classés en VT (a,b) ou en tremors (c,d) à l'issue de la SVM. Leur couleur est fonction de leur probabilité de classement : plus ils sont foncés, plus la probabilité est proche de 1.



(a) Kurtosis (b) Energie entre 0 et 5 Hz

FIG. III.2.18: Comparaison des densités de probabilité de deux attributs pour le catalogue brut simplifié (3 classes, courbes continues) et le catalogue amélioré (courbes tiretées).

Bilan

Les résultats obtenus dans cette section montrent que si l'on crée un catalogue prenant en compte les prédictions de l'algorithme, on améliore sensiblement les résultats de la classification. Cependant, cela ne paraît pas forcément correct d'un point de vue purement scientifique car cette amélioration peut être causée simplement par le fait d'avoir donné raison à l'algorithme *a priori*. De plus, les densités de probabilité des 8 attributs retenus ne montrent pas toutes une séparation nette des événements "mal" classés en premier lieu : sans revenir aux sismogrammes initiaux, on ne peut donc pas être sûrs de la reclassification de ces événements, même si certains attributs sont plus clairement discriminants que d'autres.

Finalement, comme on nourrit toujours des doutes sur la qualité du catalogue brut de départ, et comme on ne peux pas faire aveuglément confiance aux améliorations automatiques du catalogue par un premier passage de SVM, on envisage de reclasser le catalogue brut nous mêmes, manuellement. Ce catalogue et les résultats obtenus en l'utilisant sont présentés dans la prochaine section.

III.2.3.2 Catalogue reclassifié

Ayant de sérieux doutes sur la qualité de la classification manuelle initiale (voir FIG. III.1.3), il nous a semblé intéressant de procéder à une nouvelle classification manuelle dans le cadre de cet étude. Nous avons donc produit un nouveau catalogue, appelé « **catalogue reclassifié** » en classant les événements du catalogue brut en 3 classes (VT, tremors et indéterminés) sur la base de critères visuels (le moyen utilisé par les opérateurs de l'observatoire volcanologique du Kawah Ijen pour produire le catalogue brut initial). Des événements extraits aléatoirement de chacune de ces trois classes sont montrés en FIG. III.2.19.

Les éléments visuels que nous avons pris en considération pour la classification manuelle sont les suivants :

- les VT sont les événements relativement impulsifs, qui sortent bien par rapport au bruit environnant, et qui sont plutôt de courte durée (quelques dizaines de secondes) ;
- les tremors, quant à eux, sont les événements beaucoup plus émergents, de longue durée (plusieurs dizaines, voire centaines de secondes) ;
- les indéterminés sont les événements pour qui l'appartenance à l'une des deux classes sus-citées n'apparaît pas évidente.

On obtient un jeu de données assez déséquilibré, une majorité d'événements n'ayant pas pu être classée avec certitude dans l'une des deux classes étudiées (voir répartition entre les classes en Tab. III.2.2).

Classe	Catalogue reclassifié		Catalogue reclassifié amélioré	
	Nombre	% du catalogue	Nombre	% du catalogue
Volcanique de type B	1800	30%	2149	36%
Tremor	777	13%	947	16%
Indéterminé (?)	3362	57%	2825	48%

Tab. III.2.2: Répartition des différentes classes au sein du catalogue après reclassifications manuelle et automatique (5939 événements).

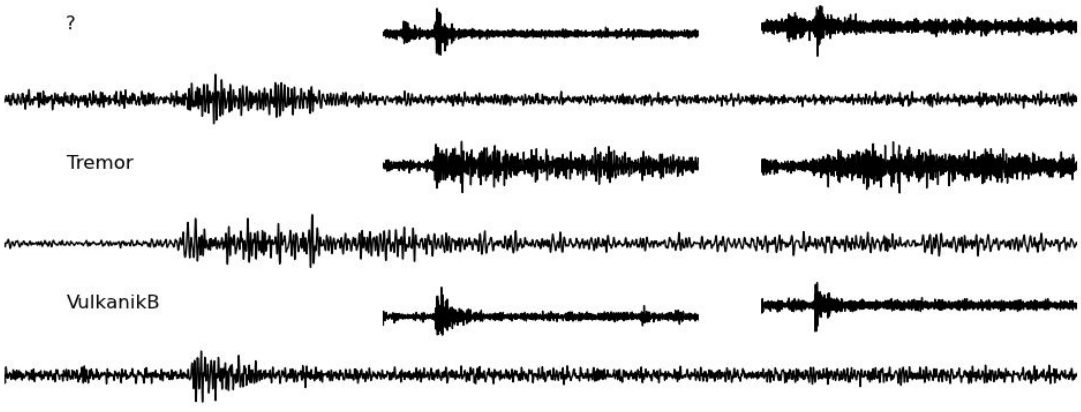
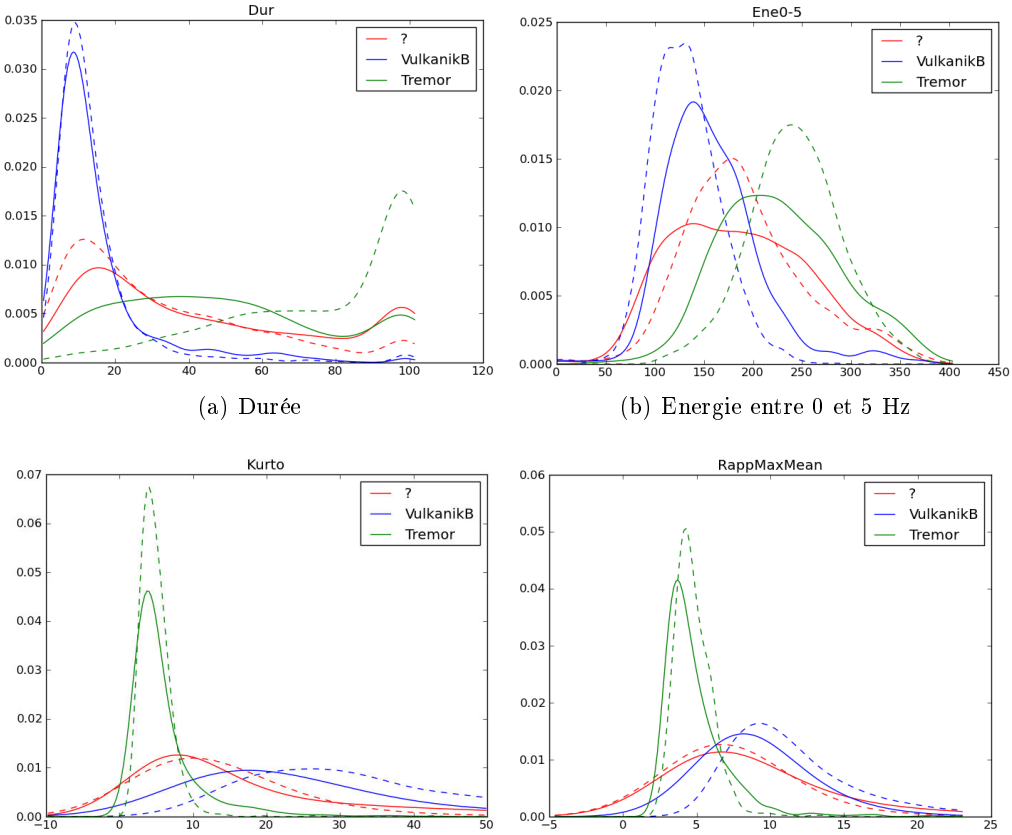


Fig. III.2.19: Formes d'ondes de quelques événements choisis au hasard dans le catalogue après la reclassification manuelle.

On a comparé sur la figure III.2.20 les densités de probabilités de certains attributs sismiques pour les événements du catalogue brut et du catalogue reclassé. On voit ainsi que la séparation entre VT et tremors s'accentue avec le nouveau catalogue pour la durée et l'énergie.

En revanche, pour les deux autres attributs présentés en exemple, les densités de probabilités n'apparaissent pas tellement modifiées.



(a) Durée

(b) Energie entre 0 et 5 Hz

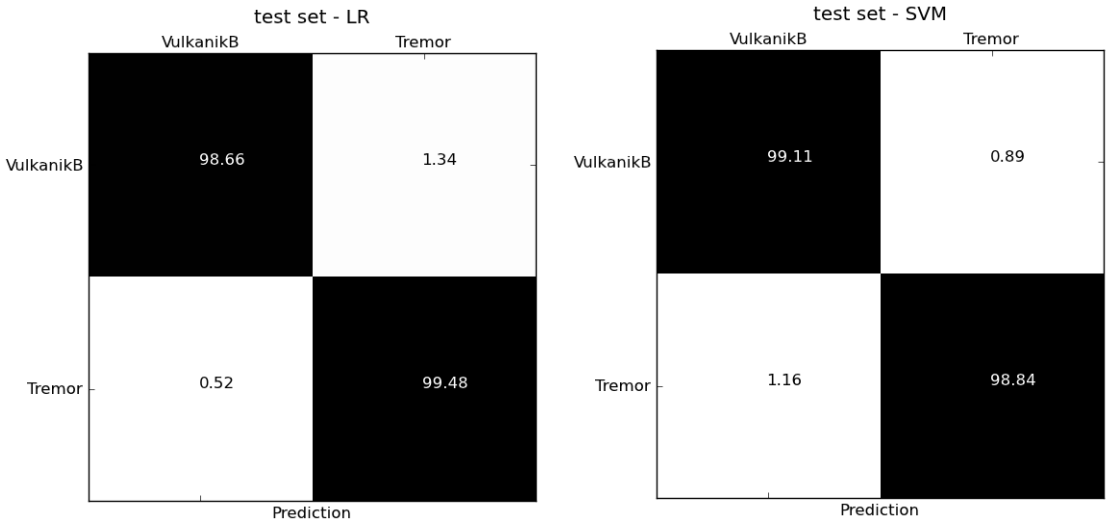
(c) Kurtosis

(d) Rapport du maximum sur la moyenne de l'enveloppe

FIG. III.2.20: Comparaison des densités de probabilités de quelques attributs sismiques pour le catalogue brut (trait plein) et le catalogue reclassé (trait tireté).

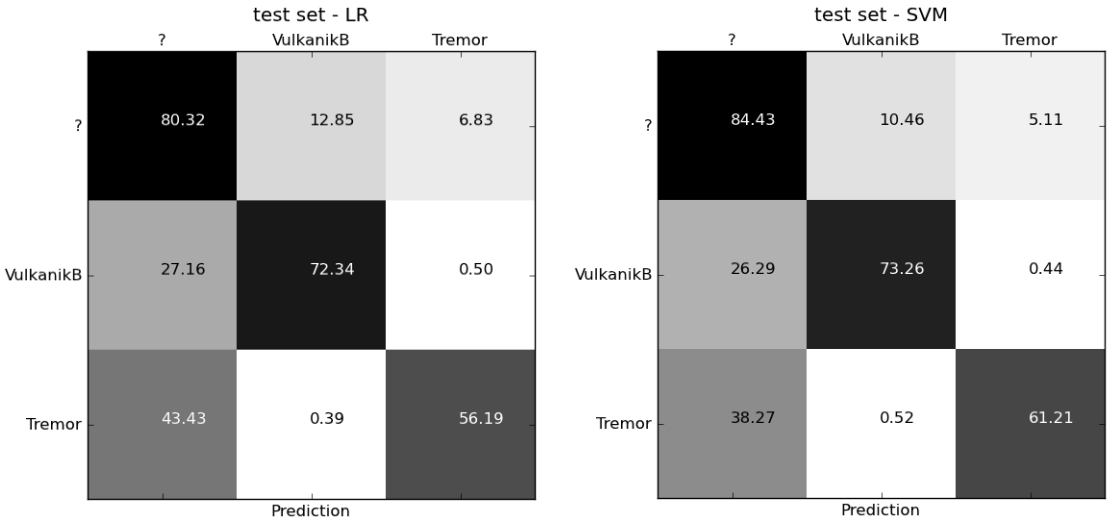
On présente les résultats obtenus pour la classification des VB et des tremors seulement (figure III.2.21), puis en tenant compte des événements non classés (indéterminés, figure III.2.22) avec tous les attributs (TAB. I.2.2, p.74).

Sans surprise, la distinction entre les VT et les tremors ne pose pas de problème (voir FIG. III.2.21). On a augmenté le taux de réussite (autour de 99%), ce qui valide la reclassification effectuée, bien que le nombre d'événements dans ces deux classes soit inférieur par rapport au catalogue brut.



(a) Régression logistique (b) SVM

FIG. III.2.21: Matrices de confusion obtenues après reclassification pour les VB et les tremors.



(a) Régression logistique (b) SVM

FIG. III.2.22: Matrices de confusion obtenues après reclassification, événements indéterminés compris.

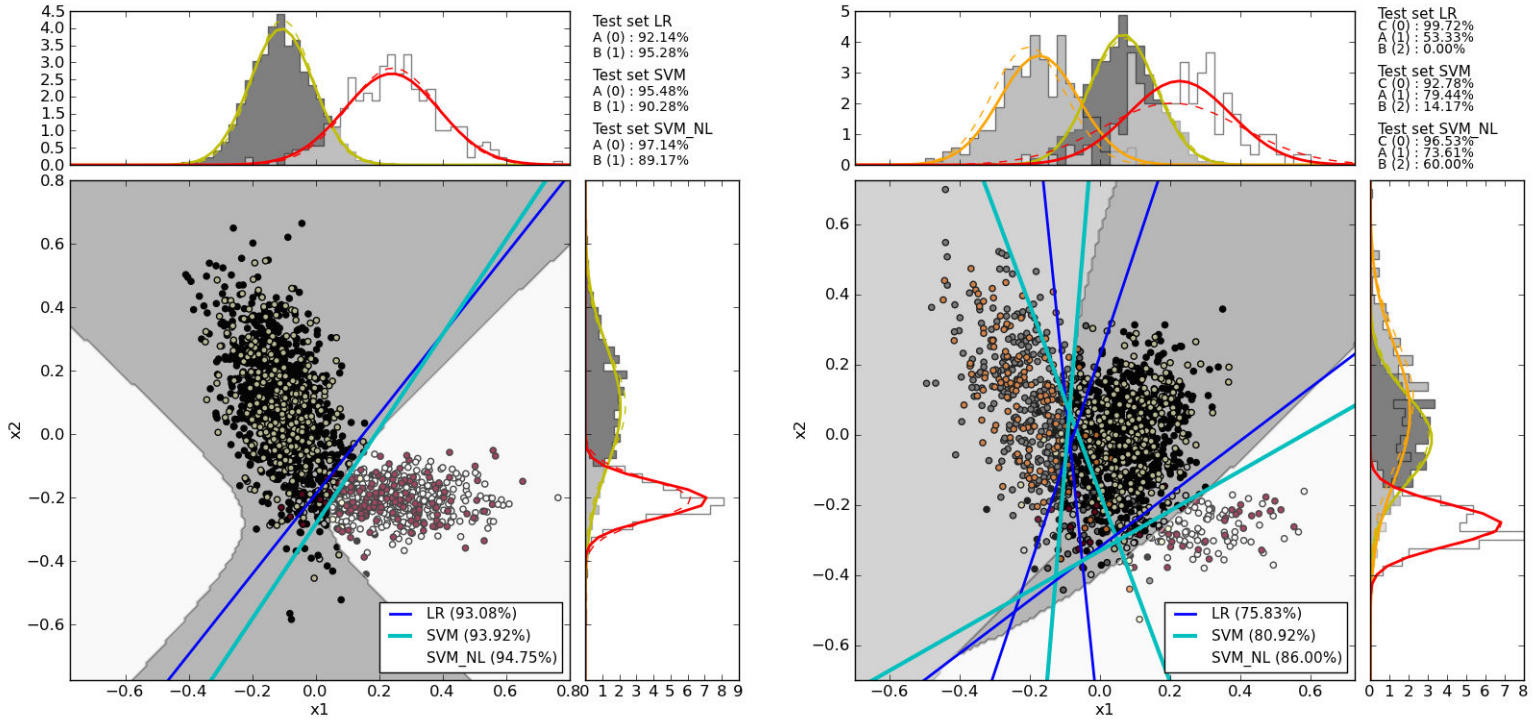
Lorsque l'on introduit les événements non classés dans le processus de classification, les résultats sont nettement moins bons (FIG. III.2.21a). La classe des indéterminés représentant plus de la moitié du jeu de données et étant constituée d'événements qu'il n'était pas évident de classer visuellement, le risque de confusion avec les événements des 2 autres classes est grand. Ainsi, un quart des VB et plus de 40% des tremors sont classés automatiquement en indéterminés.

Ces observations permettent de confirmer le fait que la taille des classes au sein du jeu de données est un paramètre qui influence fortement la classification finale. Les tests synthétiques de la figure III.2.23 permettent de comprendre plus précisément ce qu'il se passe. Dans le cas (a), on a 2 classes qui se séparent relativement bien. On a respecté les proportions des données réelles avec une classe représentant 70% du jeu, et l'autre, 30%. Dans le cas (b), on a ajouté une troisième classe, qui est "intermédiaire" par rapport aux 2 classes précédentes. De surcroît, cette nouvelle classe occupe 60% du jeu ; les 2 autres classes étant réduites à 30% et 10% respectivement. Les résultats de la régression logistique et de la SVM sont fortement dégradés entre les cas (a) et (b) (une dizaine de %). La classification se fait au détriment des classes de petite taille. Conformément à ce qui avait déjà été observé sur la figure III.1.4, la SVM donne tout de même des résultats supérieurs à ceux de la régression logistique. La SVM non-linéaire, surtout, est largement supérieure à la LR et à la SVM linéaire dans ce cas.

Comme dans la section précédente, on s'intéresse alors aux événements qu'on n'a pas classé manuellement, mais qui appartiennent à la classe des VT ou des tremors à l'issue de la classification automatique. En regardant les valeurs prises par certaines caractéristiques de ces événements (Fig. III.2.24), on s'aperçoit que les événements indéterminés manuels classés en tremors et en VT prennent bien les valeurs spécifiques des tremors et des VT, respectivement. Attention cependant, comme avec le catalogue brut, ceci n'est pas valable pour tous les attributs. Même si ce n'est pas toujours très clair, on observe globalement une légère augmentation des probabilités de classement lorsque la valeur de l'attribut ne peut plus être confondue avec celle des autres classes. Ceci est particulièrement visible sur la sous-figure III.2.24c.

L'inspection visuelle des sismogrammes de certains de ces événements permet de confirmer leur appartenance aux classes trouvées automatiquement (voir Fig. III.2.25). On peut donc tenter de générer un nouveau catalogue, en ajoutant au catalogue reclassifié manuellement l'information obtenue grâce à la classification automatique (Tab. III.2.2).

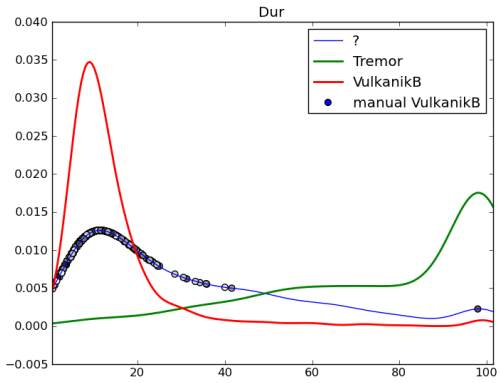
Les résultats avec le catalogue reclassifié amélioré sont présentés en figure III.2.26b. Les taux de réussite de chaque classe augmentent en moyenne de 10%. Il reste néanmoins encore un grand nombre de VT (15%) et de tremors (25%) qui sont classés à tort dans la classe des indéterminés.



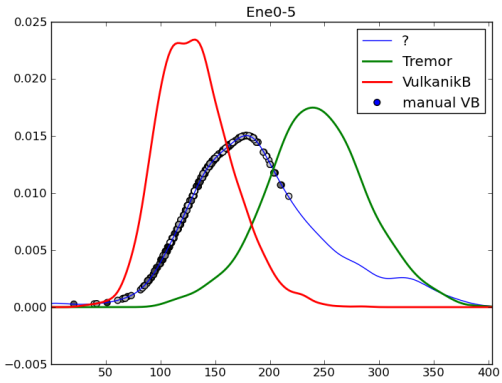
(a) 2 classes bien séparées

(b) 3 classes "mélangées"

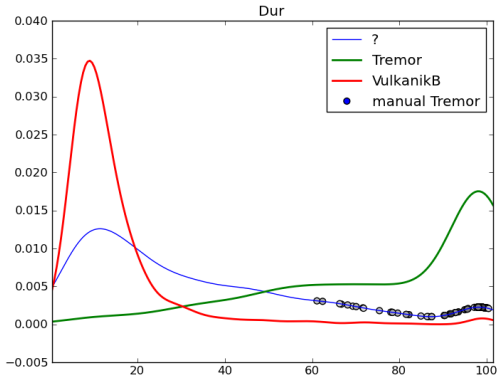
FIG. III.2.23: (a) Classification pour un jeu de données synthétiques comportant 2 classes facilement séparables, mais inégalement réparties (70% "noir", 30% "blanc"). (b) Classification pour le même jeu de données synthétiques avec l'ajout d'une troisième classe qui prend des valeurs intermédiaires. De plus, cette classe occupe 60% du jeu de données (noir) ; et les deux autres occupent désormais 30% (gris) et 10% (blanc). Les séparateurs linéaires déterminés par la LR et la SVM (linéaire) sont représentés respectivement par les droites bleu foncé et bleu clair. La couleur de fond correspond à la classification de la SVM non linéaire.



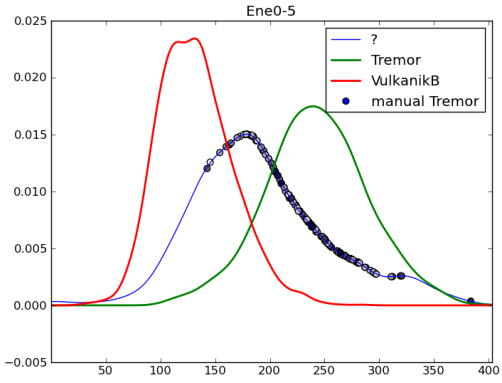
(a) VT - Durée



(b) VT - Energie



(c) Tremor - Durée



(d) Tremor - Energie

FIG. III.2.24: Valeurs de certaines caractéristiques prises par les événements non classés manuellement mais classés automatiquement. Les courbes de densité de probabilité (traits pleins) sont calculées sur l'ensemble du jeu de données. Les points correspondent aux valeurs des attributs prises par les événements classés dans les indéterminés manuellement, mais en VT (a,b) ou en tremors (c,d). Leur couleur est fonction de la probabilité de classement de l'événement : plus elle est foncée, plus la probabilité est proche de 1.

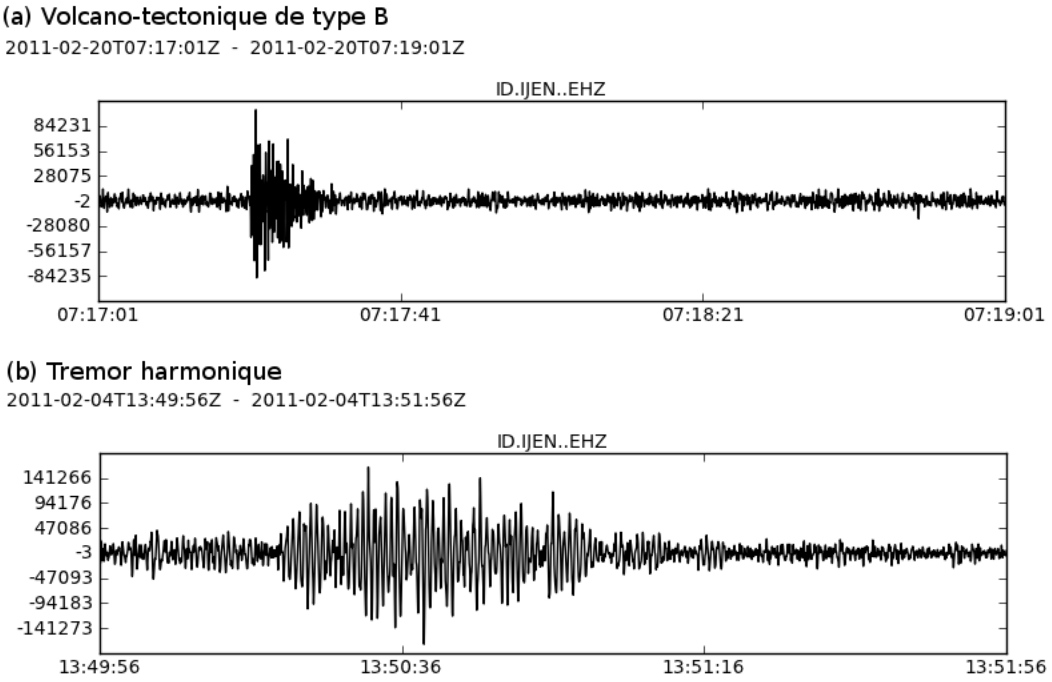
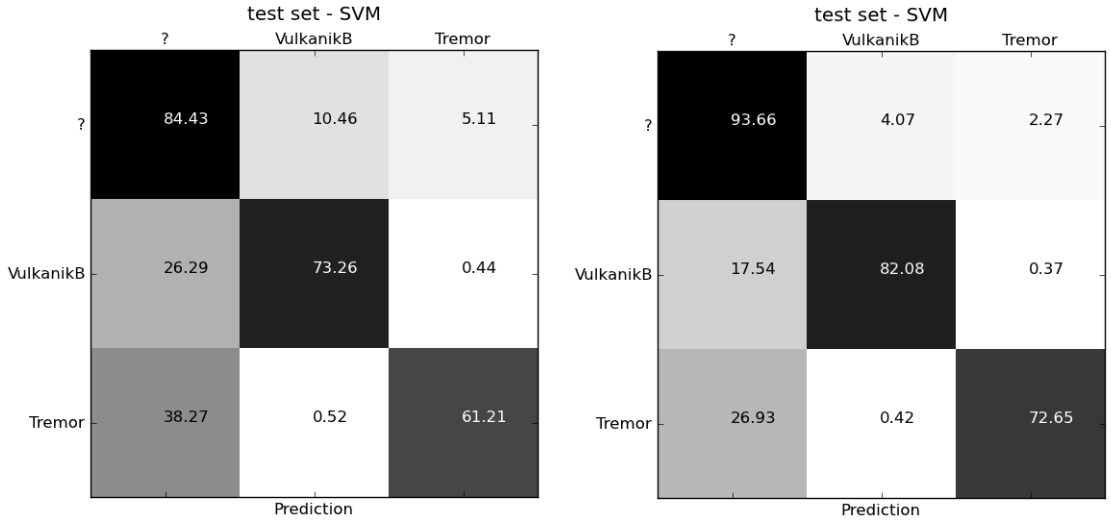


FIG. III.2.25: Formes d'ondes d'un VB (a) et d'un tremor (b) qui avaient été classés manuellement en indéterminés.



(a) Catalogue reclassé (b) Catalogue reclassé et amélioré

FIG. III.2.26: Résultats de la SVM pour le catalogue reclassé (a) puis le catalogue amélioré après la première SVM (b).

La figure III.2.27 récapitule les résultats obtenus par la SVM avec les différents catalogues détaillés précédemment. On observe une première amélioration entre le catalogue brut ramené à 3 classes et le catalogue reclassé : on passe de 72% à 76% de réussite et on réduit légèrement le sur-apprentissage. Lorsqu'on réinjecte dans de nouveaux catalogues les événements non classés manuellement mais classés automatiquement, on augmente de manière très nette le taux de bonne classification (autour d'une dizaine de %). Il faut noter que ces résultats sont assez similaires indépendamment du catalogue de départ. En conclusion, le catalogue reclassé avant réinjection semble légèrement meilleur que le catalogue brut. Après réinjection, les résultats sont identiques.

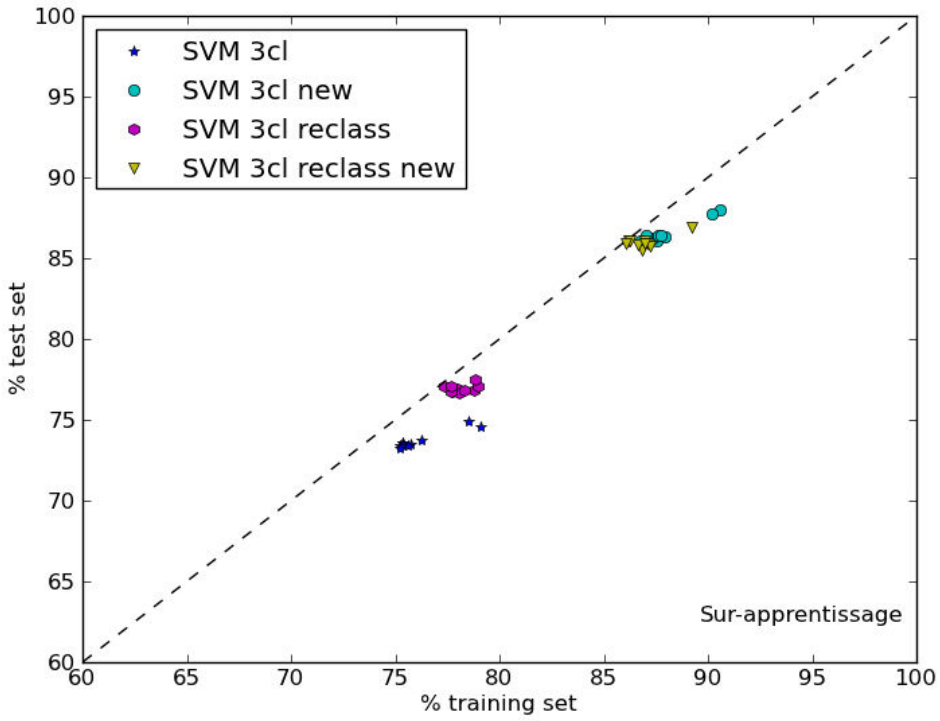


FIG. III.2.27: Comparaison des résultats obtenus pour 10 *training sets* différents pour la SVM avec 8 attributs. (Bleu) Catalogue brut ramené à 3 classes. (Cyan) Idem, avec réinjection des résultats de la SVM. (Magenta) Catalogue reclassé visuellement. (Jaune) Idem, avec réinjection des résultats de la SVM.

La figure III.2.28 montre l'amélioration des résultats obtenus après reclassification pour la SVM et la régression logistique. Dans le cas de la régression logistique, on gagne une dizaine de % de bonne classification en passant de 65 à 75%. Dans le cas de la SVM, en utilisant tous les attributs, on atteint quasiment 80% après reclassification, mais on est en sur-apprentissage. La réduction du nombre d'attributs permet cependant d'éviter ce problème et les résultats sont légèrement supérieurs à ceux de la régression logistique. Enfin, le catalogue généré à l'issue de la reclassification et en réinjectant les résultats de la SVM permet d'atteindre les 85% de bonne classification.

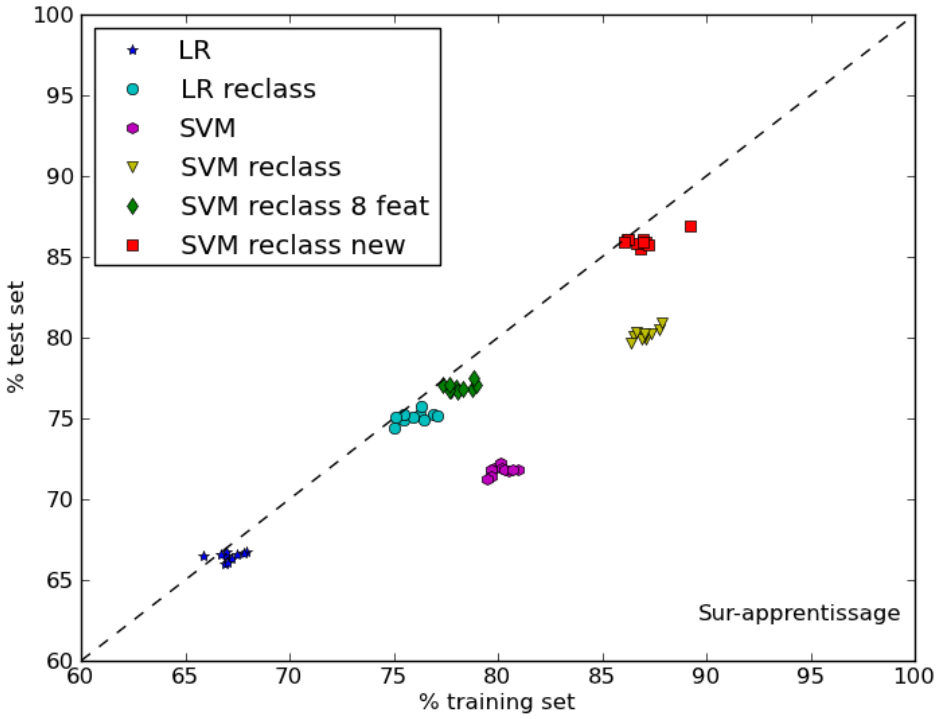


FIG. III.2.28: Comparaison des résultats obtenus pour 10 *training sets* différents pour la régression logistique (LR) et la SVM en appliquant différentes stratégies. (Etoiles bleues) Régression logistique sur le catalogue brut. (Ronds cyan) Régression logistique sur le catalogue reclassé. (Hexagones magenta) SVM sur le catalogue brut. (Triangles jaunes) SVM sur le catalogue reclassé. (Losanges verts) SVM sur le catalogue reclassé avec 8 attributs. (Carrés rouges) SVM sur le catalogue reclassé avec 8 attributs.

Bilan

La reclassification du catalogue effectuée améliore les résultats finaux de la classification. Cependant, une grande partie des événements est indéterminée : on a gardé *a priori* les tremors et VB les plus facilement séparables en laissant de côté les cas plus délicats...donc on a plus de chances de bien classer ces événements, et d'augmenter le taux de réussite.

La réinjection des événements manuels indéterminés classés automatiquement dans les 2 autres classes dans le catalogue permet d'améliorer sensiblement la classification. Néanmoins, la validité d'une telle manipulation ne peut être établie scientifiquement : la séparation des événements n'est nette que pour un nombre réduit d'attributs et le fait de donner raison au système automatique dès l'entrée ne peut qu'accroître la réussite. De plus, on peut remarquer que les résultats avec la réinjection ne semblent pas dépendre du catalogue utilisé (FIG. III.2.27).

III.2.4 Conclusions sur la classification supervisée

Les résultats présentés dans les sections précédentes permettent de classer de manière sûre les VT et les tremors lorsque seules ces 2 classes sont présentes dans le jeu de données. La

classification se dégrade dès lors qu'on introduit une classe supplémentaire. Ceci peut être dû au fait que la classe des indéterminés est susceptible de contenir des événements partageant des caractéristiques communes avec les VT et les tremors.
De manière plus générale, on a mis en évidence quelques points importants pour le bon déroulement de la classification automatique :

- la fiabilité du catalogue manuel sur lequel on s'appuie pour entraîner le système est essentielle.
- la taille des classes dans le jeu de données joue un rôle décisif dans le processus de classification. Plus une classe est grande, plus le classifieur aura tendance à classer les éléments des autres classes dans cette classe "par défaut".
- le choix des attributs qui permettent de discriminer au mieux les différents types d'éléments est également très important. Il est nécessaire et utile d'ajuster le choix et le nombre des attributs au problème considéré afin de tirer le meilleur parti possible du classifieur.

III.2.5 Classification non supervisée

Comme les résultats de la classification supervisée montrent que la séparation des différentes classes est loin d'être évidente, on choisit de regarder ce que contient le jeu de données sans classification a priori des éléments : on applique pour cela une méthode de classification non-supervisée, les K-moyennes, qui consiste à regrouper les éléments d'un jeu de données en K groupes de même variance, K étant spécifié par l'utilisateur (description de la méthode en I.2.3, p. 48).

On présente dans un premier temps les résultats obtenus pour la séparation en 8 classes du jeu de données en prenant en compte tous les attributs disponibles (Fig. III.2.29) ; puis les résultats pour 3 classes avec tous les attributs (Fig. III.2.30a) et un nombre d'attributs réduit selon les mêmes critères que pour la classification supervisée (Fig. III.2.30b). Les figures donnent à chaque fois la répartition des classes selon la classification manuelle (a) et selon la classification non-supervisée (b) ainsi que la répartition des événements au sein de chaque classe automatique (c).

Sur le catalogue brut avec 8 classes

Dans le cas à 8 classes, on voit que 3 classes automatiques sont de tailles négligeables ; une classe occupe 10% du jeu de données ; 2 classes occupent respectivement 20% et les deux dernières occupent respectivement 25%. La répartition des classes manuelles au sein des classes 0 et 7 (de taille 25%) montre qu'elles contiennent plus de la moitié de VB. Les VA constituent presque un quart de la classe 0 ; ils sont beaucoup moins nombreux dans la classe 7 qui contient en contrepartie un peu plus de hembusans et de tremors. Concernant les classes de taille 20%, on observe des diagrammes très différents : pour la classe 2, la classe majoritaire est celle des tectoniques (presque 60%), suivie de celle des tremors (presque 30%), les autres classes se partageant le reste. Pour la classe 5, en revanche, la classe majoritaire est une fois de plus celle des VB (55%) et la répartition des autres classes ne met pas en évidence une classe dominante. La classe 1, de taille 10%, est, quant à elle, composée de tectoniques et de tremors dans des proportions équivalentes (environ 35% chacune) et de 20% de VB.

Tout ceci nous apprend que :
- il n'est possible de ne séparer que 5 classes sur les 8 définies a priori.
- globalement, dans les classes où les VB sont majoritaires, les tremors sont présents en faibles proportions (et réciproquement), ce qui confirme nos observations précédentes selon lesquelles la séparation entre ces 2 classes se fait bien.
- il existe une classe comprenant presque tous les VA.
- il existe 2 classes où les classes manuelles dominantes sont les tectoniques et les tremors. Or on avait déjà vu que la classification manuelle des tectoniques semblait peu fiable et que les confusions avec les tremors étaient fréquentes : ceux-ci pourraient appartenir à une seule et même classe.

Sur le catalogue reclassifié avec 3 classes

Dans le cas où l'on cherche à séparer le jeu de données en 3 classes, on observe un comportement différent selon le nombre d'attributs utilisés. Lorsque tous les attributs sont utilisés, seules 2 classes se séparent réellement (Fig. III.2.30a). La classe qui occupe un tiers du jeu est composée essentiellement d'indéterminés et de tremors. La classe qui occupe les deux tiers restants contient presque autant d'indéterminés que de VB. On constate donc une fois de plus que la séparation entre tremors et VB se fait aisément.

Lorsque l'on utilise un nombre restreint d'attributs, 3 classes sont définies (Fig. III.2.30b) : 2 classes occupent respectivement 40% du jeu de données, et la classe restante, 20%. La répartition des classes manuelles au sein de ces classes montre qu'on obtient une classe où il n'y a que des VB et des indéterminés (avec une majorité de VB) ; une classe où il n'y a que des tremors et des indéterminés (avec une majorité d'indéterminés) ; une classe composée aux trois-quarts d'indéterminés.

La figure III.2.31 représente les densités de probabilité de la durée et de l'énergie pour les classes manuelles et pour les classes automatiques. Les courbes pour l'énergie montrent clairement la séparation entre les 3 classes : celles obtenues pour les classes déterminées automatiquement sont globalement moins larges que pour les manuelles. Ceci s'accorde bien avec les diagrammes de la figure III.2.30b et permet de conclure que :
- une classe est composée de tremors et d'indéterminés et semble correspondre à la classe des tremors plus les indéterminés qui partagent des caractéristiques communes aux tremors (classe 0).
- une classe est composée de VB et d'indéterminés et semble correspondre à la classe des VB plus les indéterminés qui partagent des caractéristiques communes aux VB (classe 1).
- une classe est composée essentiellement d'indéterminés et semble correspondre à la classe des indéterminés plus les VB et tremors qui ont des caractéristiques "limites" qui peuvent aussi bien correspondre à la classe des indéterminés (classe 2).

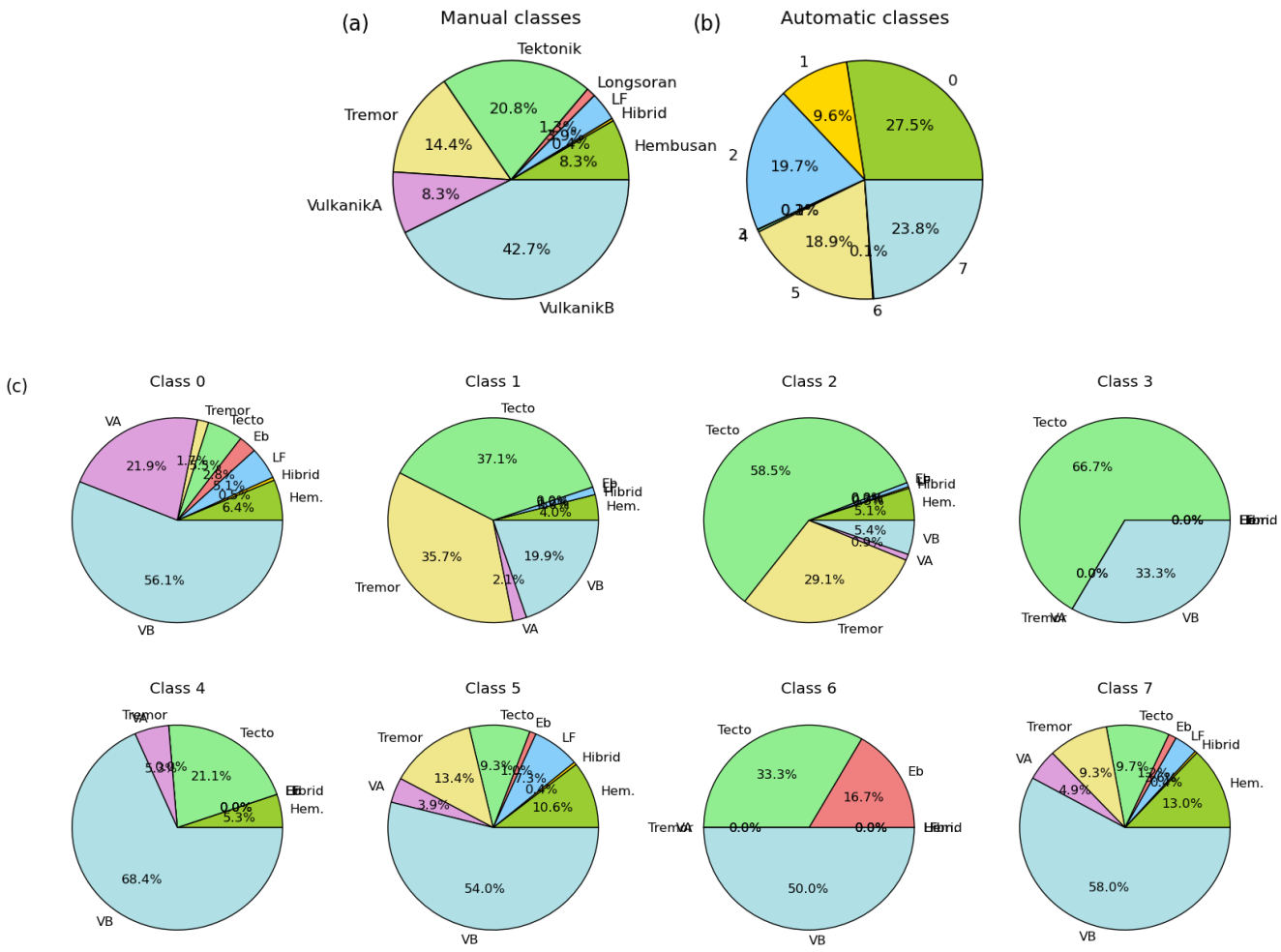
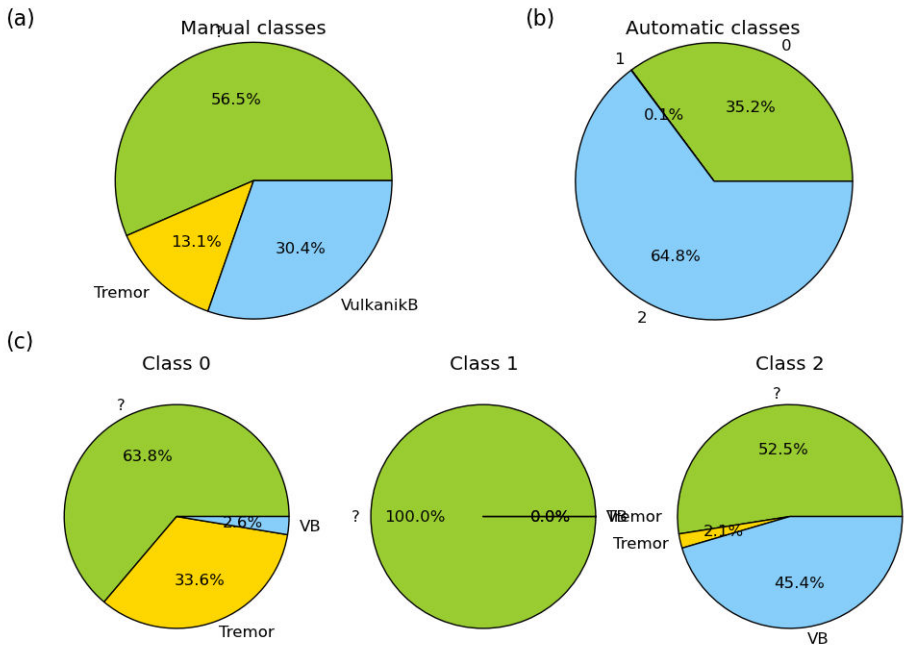
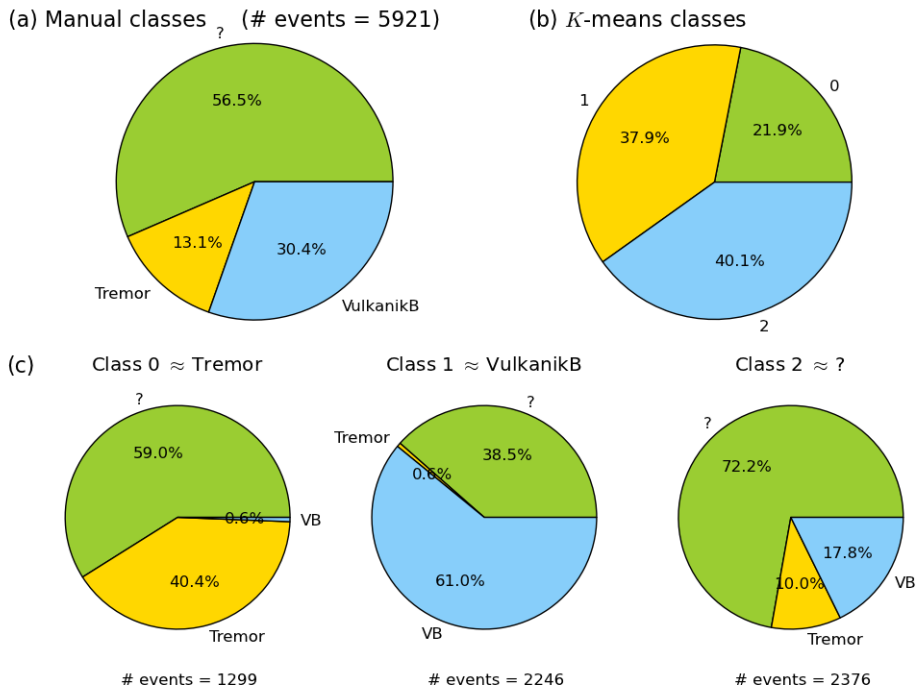


FIG. III.2.29: Résultats de la classification non supervisée appliquée au catalogue brut avec plus de 50 attributs. (a) Diagramme de répartition du catalogue brut (classes manuelles). (b) Diagramme de répartition des classes trouvées par l'algorithme des K-moyennes (classes automatiques). (c) Diagrammes de répartition de chaque classe automatique en termes de classes manuelles.

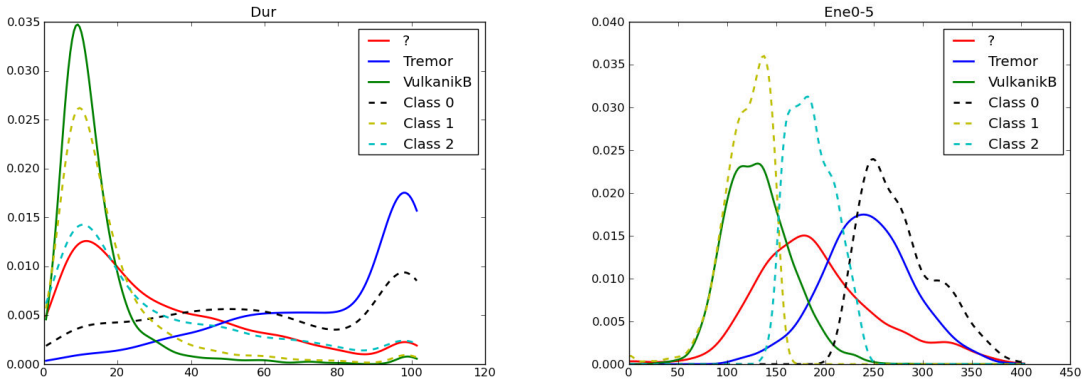


(a) Nombre d'attributs > 50



(b) Nombre d'attributs $= 11$

FIG. III.2.30: Résultats de la classification non supervisée appliquée au catalogue reclassfié. Pour chaque figure, la sous-figure (a) donne le diagramme de répartition du catalogue reclassé ; la (b) donne le diagramme de répartition issu des K-moyennes ; et la (c) présente les diagrammes de répartition de chacune des classes automatiques en termes de classe manuelle.



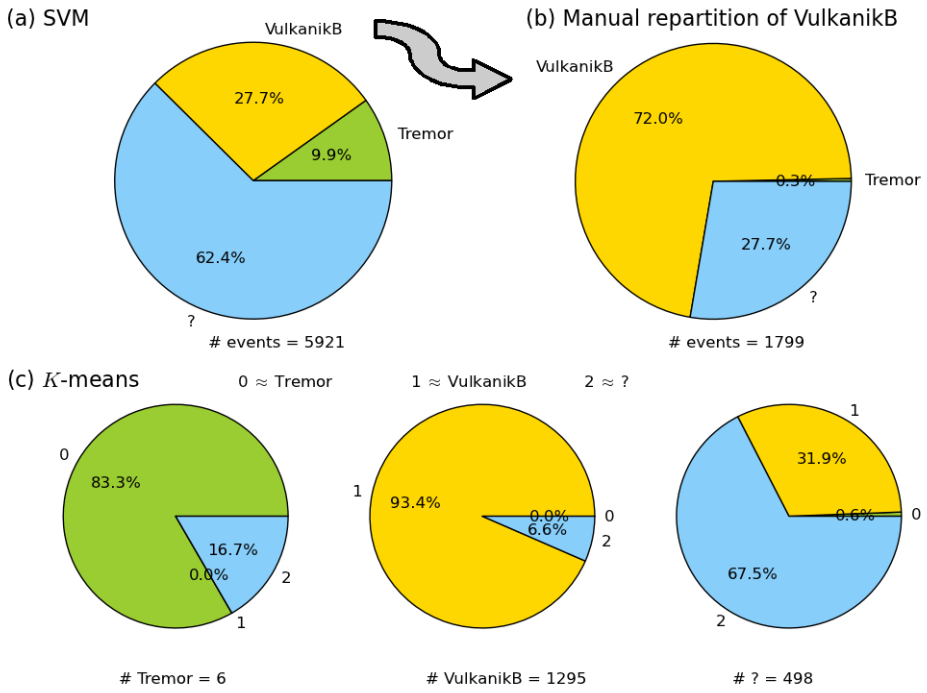
(a) Durée

(b) Energie entre 0 et 5 Hz

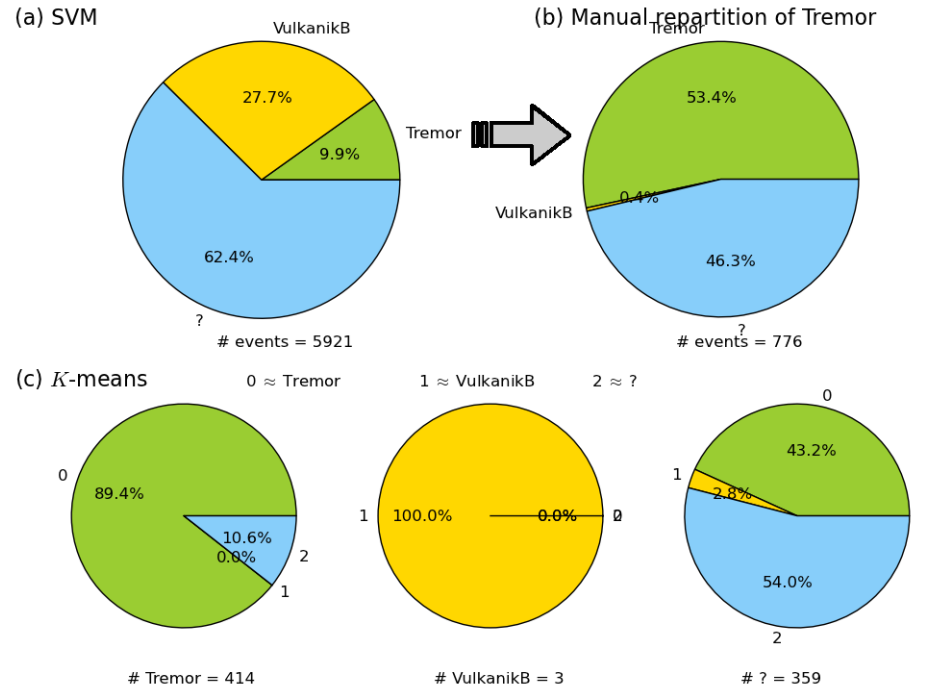
FIG. III.2.31: Densités de probabilités pour deux attributs et comparaison entre les classes manuelles (traits pleins) et les classes obtenues grâce à la classification non supervisée (traits tiretés).

Dans la section III.2.3.2, on avait vu qu'une proportion non négligeable d'événements indéterminés étaient classés en VB ou tremors par la SVM et la LR (FIG. III.2.22). On cherche maintenant à savoir s'il existe une éventuelle correspondance entre ces VB et tremors et les classes déterminées par les K-moyennes. Les résultats sont présentés sur la figure III.2.32. Par exemple, la SVM a classé 27.7% de VB (figure (a)). Parmi ces VB, 27.7% avaient été classés dans les indéterminés manuellement, et parmi ces indéterminés, 31.9% ont été classés dans la classe 1 par les K-moyennes, classe qui peut être assimilée à la classe des VB. Pour les tremors (figure (b)), 43.2% des 46.3% d'indéterminés manuels sont classés dans la classe 0, qui peut être assimilée à celle des tremors. Conformément à ce qu'on avait déjà pu entrevoir dans la section III.2.3.2, les événements indéterminés manuels qui sont classés par la SVM en VB ou tremors partagent bien des caractéristiques communes à ces deux classes mais ne peuvent y être classés assurément. De plus, on remarque que les pourcentages de classe 0 (≈tremor) pour les VB et de classe 1 (≈VB) pour les tremors sont faibles, ce qui montre que ces événements indéterminés avec un penchant vers l'une ou l'autre des deux classes ne peuvent être confondus avec la classe opposée.

Ces diagrammes nous apprennent également que l'immense majorité des événements qui sont bien classés par la SVM sont aussi classés dans les classes des K-moyennes correspondantes. Ainsi, 93.4% des 72% de VB bien classés par la SVM sont classés en 1 (soit la classe assimilée aux VB) ; et 89.4% des 53.4% de tremors bien classés pas la SVM sont classés en 0 (soit la classe assimilée aux tremors).



(a) Répartition des VB de la SVM



(b) Répartition des tremors de la SVM

FIG. III.2.32: Diagrammes de répartition montrant la correspondance entre les classes déterminées par l'algorithme des K-moyennes et par la SVM. Pour chaque figure, les sous-figures donnent : (a) le diagramme de répartition issu de la SVM appliquée au catalogue reclassé ; (b) le diagramme de répartition d'une classe (VB ou tremor) de la SVM en termes de classes manuelles ; (c) les diagrammes de répartition de ces classes manuelles en termes de classes déterminées par les K-moyennes.

III.2.6 Conclusion

Le but de cette partie était de réussir à classer de manière automatique les huit différents types d'événements sismiques principaux enregistrés sur le volcan du Kawah Ijen. La complexité du problème et les difficultés rencontrées dès le début nous ont amenés à développer un certain nombre de stratégies pour tirer le meilleur parti du jeu de données dont on disposait. Finalement, l'étude a permis de mettre en évidence les points génaraux suivants :

- l'importance de la classification manuelle. On s'en doutait déjà, vu que le système d'apprentissage supervisé "apprend" sur des données déjà classées. De manière plus générale, on voit surtout qu'avec une classification trop compliquée ou pas suffisamment claire, on ne peut espérer obtenir de bons résultats.

- lorsque le jeu de données comprend des classes de taille très hétérogènes, il devient difficile de classer correctement les petites classes. Un système d'apprentissage automatique peut et doit prendre en compte les déséquilibres qui peuvent exister, mais lorsque ceux-ci sont trop conséquents, on ne peut espérer bien classer les petites classes (sauf, bien sûr, si celles-ci se séparent vraiment bien du reste du jeu de données). De plus, le *training set* doit être construit de manière à respecter les proportions du *test set*, pour éviter les problèmes de généralisation et la favorisation de certaines classes peu importantes en termes de taille (par exemple, si une classe ne représente que 10% du jeu de données, il n'y a pas de raison que cette classe soit sur-représentée dans le *training set*, au risque de fausser l'apprentissage avec des probabilités d'apparition plus élevés qu'il ne faudrait).

Plus spécifiquement pour le volcan du Kawah Ijen, on peut dire que :

- la classification en huit classes semble présomptueuse : les résultats de la classification non-supervisée (Fig. III.2.29) ont montré qu'avec les données et attributs disponibles, on ne pouvait espérer distinguer au maximum que 5 classes d'événements.

- deux classes, celles des VT (VA+VB) et des tremors, sont relativement faciles à séparer.

- le reste des événements nécessiterait des investigations plus poussées en termes de recherche de caractéristiques discriminantes. Globalement, ces événements sont en effet "intermédiaires" entre VT et tremors et partagent des caractéristiques communes aux deux classes. L'intégration d'observations autres que sismologiques au problème pourrait aider à améliorer le système de classification automatique. Par exemple, un certain nombre de mesures liées aux phénomènes hydrothermaux du volcan sont effectuées (température, acidité du lac...) et sont peut-être corrélées d'une certaine manière à la sismicité.

CHAPITRE III.3

Localisation des séismes sur le Kawah Ijen

III.3.1 Objectifs

L'idée de cette partie était de localiser la sismicité enregistrée sur le Kawah Ijen, d'abord à partir des événements du catalogue fourni (III.1.2.2), donc sans passer par l'étape de détection. La première étape avant de commencer toute étude détaillée a été de faire des tests de résolution en fonction de la géométrie du réseau de stations.

III.3.2 Tests de résolution

Le réseau de stations installé autour du cratère du Kawah Ijen est beaucoup plus épars que celui du Piton de la Fournaise (Fig. III.1.1). L'extension est de l'ordre de la vingtaine de km dans les directions Est-Ouest et Nord-Sud, avec une répartition entre 600 et 2300 m d'altitude. La couverture spatiale n'est pas homogène : contrairement au Piton de la Fournaise où le barycentre du réseau correspondait approximativement au cratère du volcan, on dispose ici d'environ 5 stations à proximité immédiate du cratère, 3 stations au Sud et quelques stations à l'Ouest. Une seule station est située à l'Est, et très peu au Nord.

De plus, il faut rappeler que le catalogue qui nous a été fourni pour la station IJEN couvre une période allant de février 2011 à juin 2012. Une majorité de stations a été installée de manière temporaire fin 2011 (IBLW, IGEN, IMLB...- voir Fig. III.1.2). On a choisi de ne pas les intégrer pour la localisation avec Waveloc. On a finalement retenu 9 stations au total (en noir sur la figure III.1.1) qui ont fonctionné au moins la moitié de la période considérée. Certaines de ces stations sont localisées au même emplacement (MLLR et RAUN ; IJEN et TRWI), ce qui revient, pour le problème de localisation, à avoir 6 stations. Cinq stations sont situées sur le volcan même (POS, KWUI, PSG, IJEN, TRWI) ; 2 stations sont situées à une quinzaine de km au Sud-Ouest (MLLR, RAUN) et une station est isolée à environ 5 km au Sud du volcan (POSI).

La grille d'étude choisie comprend 41 points dans les directions x et y et 11 dans la direction z, ce qui correspond à des pas de 500 m. Les mailles sont donc relativement espacées, mais les resserrer aurait considérablement augmenté les temps de calcul (on a déjà près de 18500 points). L'origine de la grille en coordonnées UTM est X=-10 km, Y=-10 km et Z=-2 km

(positif vers le haut).

Les tests de résolution vont permettre d'avoir un aperçu des localisations que l'on peut espérer obtenir avec une telle répartition de stations (voir FIG. III.3.1). Le modèle de vitesse utilisé est un modèle 1D assez simple composé de 5 couches de vitesse homogène développé par le CVGHM (*Center for Volcanology and Geohazards Mitigation*, Indonésie - TAB. III.3.1).

Profondeur (km)	V_P (km/s)	V_P/V_S	V_S (km/s)
0	2.20	1.33	1.65
3	2.47	1.39	1.78
8	2.70	1.37	1.97
13	3.10	1.30	2.38
15	3.45	1.46	2.36

TAB. III.3.1: Modèle de vitesse 1D utilisé pour la localisation des séismes sur le Kawah Ijen. Le niveau de référence (profondeur zéro) est pris au sommet du volcan.

Pour ces tests, on a considéré des signaux de 40 s de long enregistrés avec une fréquence d'échantillonnage de 100 Hz (conformément aux stations). Le pulse triangulaire correspondant à l'événement se produit à 10 s (temps absolu) et a une amplitude égale à 5. Aucun bruit n'a été ajouté aux données synthétiques. Le niveau de détection a été fixé à 20 (soit la moitié des stations du réseau multipliée par l'amplitude).

Les tests de résolution effectués à différentes profondeurs et visibles sur la figure III.3.1 et en annexe A3 montrent que :

- la résolution en x et y est plutôt bonne pour des valeurs de z comprises entre -500 et 1000 m. Lorsque z augmente ou diminue, la résolution se dégrade rapidement, surtout pour les points les plus éloignés du réseau de stations. Il existe également une petite zone située au Nord-Est du cratère qui est systématiquement moins bien résolue que le reste de l'espace.

- le risque de localisations multiples augmente sensiblement lorsque l'altitude augmente, notamment dans la zone située autour du réseau principal de stations. Ceci peut poser problème puisqu'il s'agit de la zone qui nous intéresse (située sur le volcan lui-même). On rappelle que les localisations multiples se produisent lorsque plusieurs maxima locaux sont détectés (voir FIG. III.3.2).

- la précision espérée sur les temps origine est globalement comprise entre 0 et 0.1 s, avec toujours cette zone Nord-Est qui est moins bien résolue que le reste de l'espace (avec des temps origine plus précoces que les vrais).

Ces tests de résolution montrent qu'il sera difficile d'obtenir des localisations précises des événements sur le Kawah Ijen sans une amélioration du réseau de surveillance sismique.

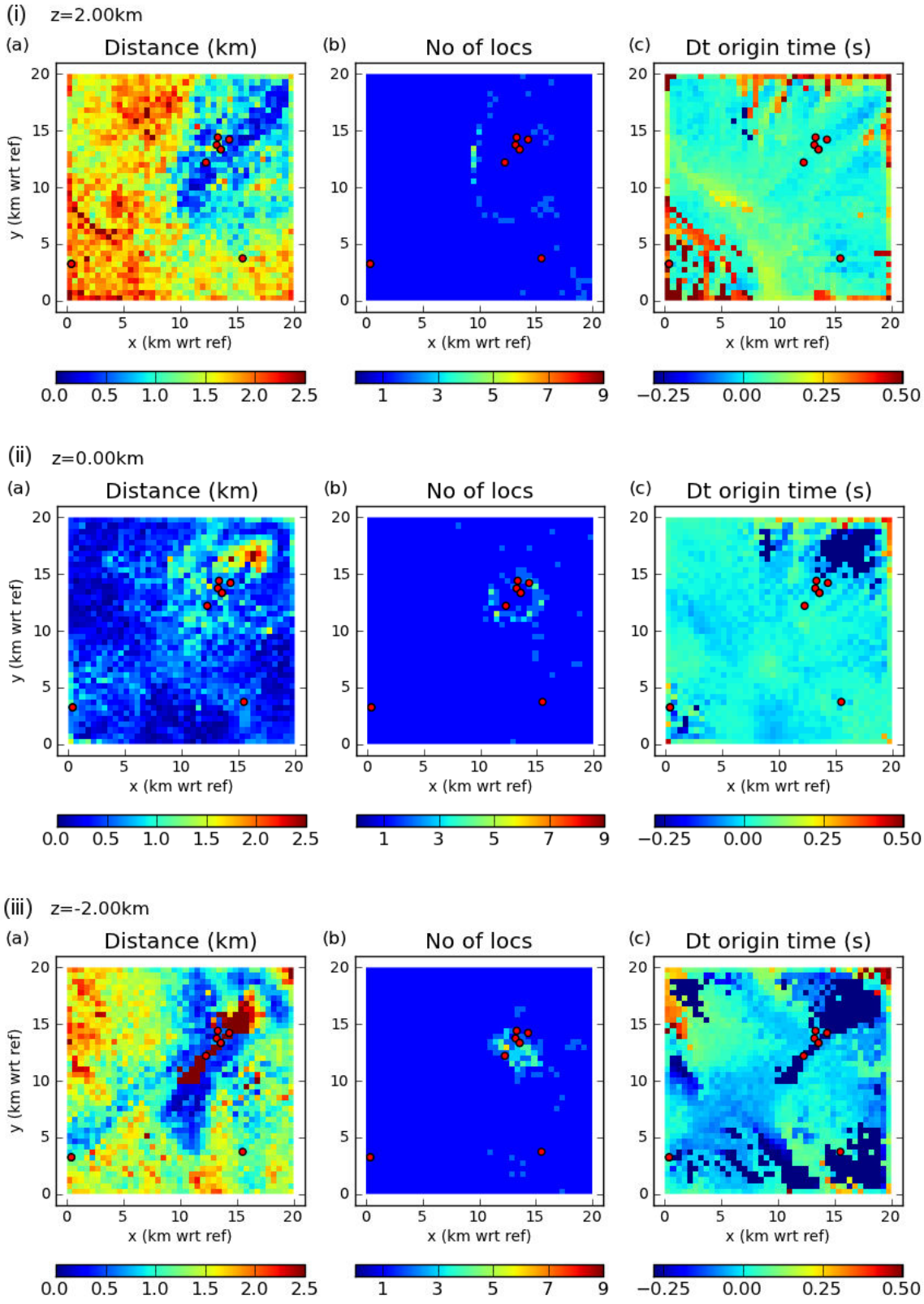


FIG. III.3.1: Tests de résolution effectués sur le réseau de 8 stations du Kawah Ijen pour différentes profondeurs : (i) 2 km, (ii) 0 km (niveau de la mer), (iii) -2 km. L'axe des z est positif vers le bas. La grille utilisée est de taille 20x20 km et a pour origine -10 km UTM en x et en y.

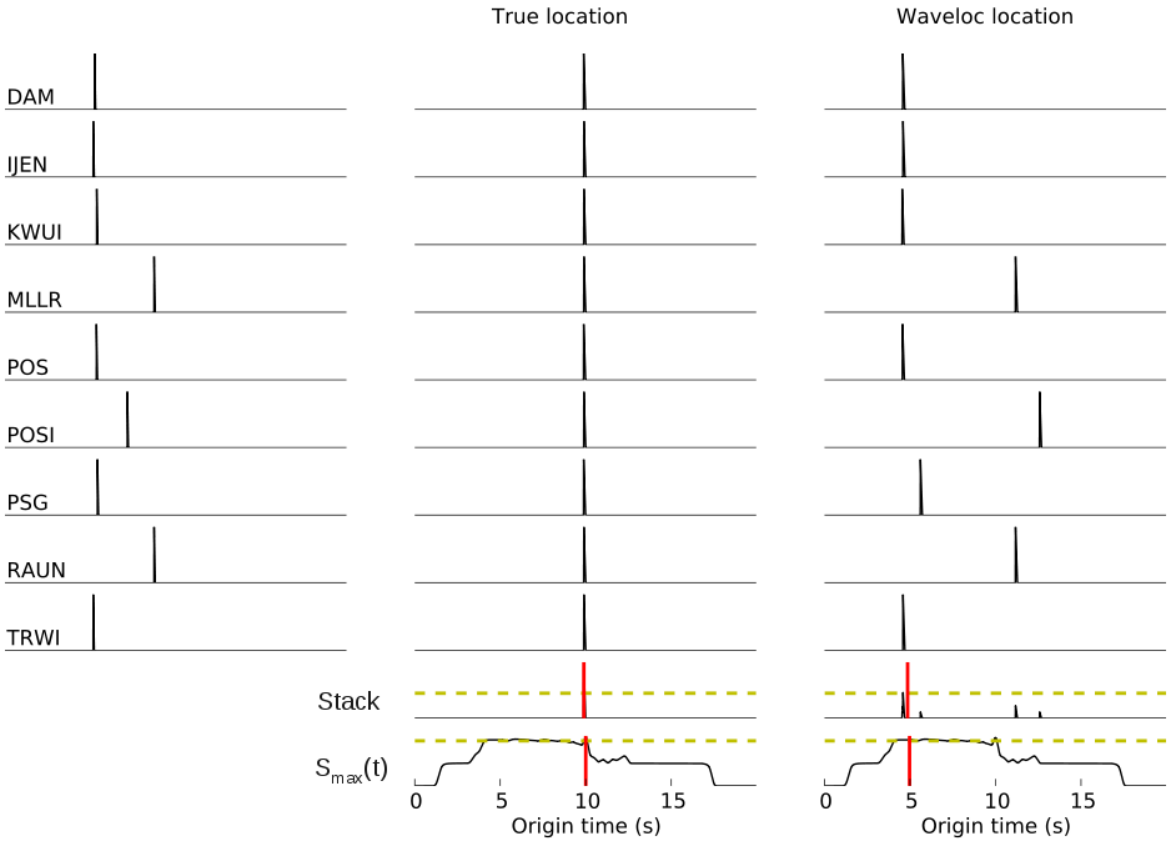


FIG. III.3.2: Un exemple d'illustration synthétique expliquant pourquoi il est possible d'avoir plusieurs localisations correspondant à une seule source. Plusieurs maxima locaux supérieurs au seuil de détection fixé sont repérés. Une localisation correspondant à chacun de ses maxima sera effectuée, bien qu'il n'existe qu'un seul maximum absolu. Le trait horizontal jaune correspond au seuil de détection (moitié du nombre de stations multipliée par amplitude du pic synthétique). Le trait vertical rouge correspond au maximum de $S_{max}(t)$ et au temps origine de l'événement détecté.

III.3.3 Conclusion

Dans un problème de localisation, un nombre suffisant de stations est requis : au moins 4 pour une estimation grossière des 4 paramètres hypocentraux. Une bonne distribution géographique est aussi essentielle. Ce besoin est encore accentué pour les méthodes automatiques pour lesquelles il faut pouvoir neutraliser l'effet d'un phénomène local autre qu'un séisme mais enregistré sur quelques stations.

Sur le Kawah Ijen, le réseau de stations est limité à 9, ce qui est faible. De plus, la répartition est loin d'être optimale et plusieurs stations sont situées au même endroit. Enfin, les données ne sont pas complètes pour toutes les stations et il n'est pas rare que le nombre d'enregistrements disponibles pour une période de temps donnée soit inférieur à 6 stations. Tout ceci complique fortement le problème de localisation automatique, voire le rend impossible.

Comme ce qui nous intéresse est une localisation de la microsismicité sur et/ou autour

du volcan, il est primordial de pouvoir obtenir une bonne précision. Une résolution de l'ordre du km n'est donc pas satisfaisante pour un problème de ce type, quand on sait, ou du moins on suppose, que la sismicité est localisée dans une région de quelques km sur quelques km seulement.

Une localisation des événements enregistrés sur le volcan du Kawah Ijen entre juin 2011 et juin 2012 a déjà été effectuée par un étudiant de master de l'Observatoire Royal de Belgique [Van Kriekinge, 2013] en utilisant un algorithme de STA/LTA pour la détection des événements et le logiciel HYPOELLIPSE [Lahr, 1999 - revised in 2012] pour la localisation des événements. Le modèle de vitesse 1D initial a ensuite été amélioré via ces localisations et l'utilisation du programme VELEST [Kissling et al., 1995]. La correction du modèle de vitesse se fait grâce au calcul du résidu (RMS) de localisation d'une cinquantaine d'événements pointés manuellement parmi le jeu de données, de manière à minimiser l'erreur systématique observée sur chacune des stations. La majorité des localisations obtenues avec ce modèle de vitesse se situe entre 0 et 1000 m d'altitude à l'aplomb du cratère du volcan.

Finalement, l'étude préliminaire effectuée avec Waveloc montre qu'une localisation automatique et précise des microséismes sur le volcan du Kawah Ijen sera possible uniquement si :

- la géométrie du réseau de stations sur et autour du volcan est nettement améliorée ;
- le modèle de vitesse 1D est aussi sensiblement amélioré ; même si une première amélioration du modèle a été effectuée par l'intermédiaire du programme VELEST, il demeure encore relativement simple pour décrire la structure du milieu profond.

Conclusion générale et perspectives

Rappels des principaux résultats

Pour la détection et la localisation
Le travail a consisté en la validation et l'amélioration de l'algorithme de Waveloc, qui avait été initialement développé par Maggi and Michelini [2010]. Cet algorithme présente l'avantage de détecter et localiser automatiquement et simultanément les événements sismiques par simple migration des formes d'ondes continues, et ne nécessite pas le pointé précis des différentes phases. La recherche d'une fonction caractéristique qui permet d'atteindre la meilleure précision possible sur les paramètres hypocentraux et de paramètres de calcul (e.g. taille des fenêtres, bande de filtrage...) et de décision (e.g. seuil de détection) adaptés ont constitué une part importante de ce travail, ceci afin d'optimiser le système en tenant compte du *trade-off* existant entre efficacité et fiabilité. L'ajout de fonctionnalités telles que le calcul des magnitudes locales ou la relocalisation précise de la sismicité par la méthode des double-différences permet une analyse plus détaillée de la sismicité. L'application aux données du Piton de la Fournaise a montré que Waveloc pouvait s'affirmer comme un outil fiable et efficace lorsque certaines conditions sont réunies : si la géométrie du réseau de stations est bonne, dans un espace géographique de taille réduite, l'information sur la première arrivée du signal est largement suffisante pour obtenir de bons résultats de localisation. Cependant, si la géométrie du réseau de stations ne s'y prête pas, on ne pourra pas espérer pouvoir localiser correctement les événements (ceci étant vrai dès qu'on essaie de rendre le processus automatique). Comme dans toute étude, un modèle de vitesse adapté est aussi un pré-requis essentiel pour obtenir de bonnes localisations.

Pour la classification
On a choisi de travailler avec deux méthodes de classification supervisée relativement peu utilisées en sismologie jusqu'à présent : la régression logistique et la SVM. Celles-ci sont relativement simples à mettre en œuvre et « minimisent » l'intervention de l'opérateur par rapport à d'autres méthodes telles que la logique floue (choix de l'intervalle et du type de fonction de possibilités), les réseaux de neurones (choix du nombre de couches). Par rapport à la méthode des modèles de Markov cachés (calcul des probabilités d'appartenance à une classe en chaque échantillon), la régression logistique et la SVM permettent une meilleure compréhension et une meilleure maîtrise de la manière dont la classification est effectuée. La SVM, surtout, est une méthode particulièrement puissante qui permet la résolution de problèmes de classification non linéaires. Dans ce travail, après d'un certain nombre d'attributs susceptibles de décrire au mieux les signaux sismologiques, on a montré que c'est surtout le choix, le nombre et le type de combinaison d'attributs qui priment et influencent les résultats de la classification.

Globalement, ce n'est donc pas tant la méthode choisie qui importe, mais ce qu'on lui donne en entrée, c'est-à-dire la manière dont on exploite l'information contenue dans les données. Ceci a été mis en évidence sur le jeu de données du Piton de la Fournaise qui avait déjà fait l'objet d'une classification automatique grâce à la méthode de logique floue. On a finalement trouvé qu'il était difficile de parvenir à obtenir une classification automatique au-delà des 93% du jeu de données, ce qui constitue déjà un excellent résultat.
L'application au jeu de données du Kawah Ijen s'est avérée plus complexe et fastidieuse mais a permis la mise en évidence de quelques autres points souvent déjà connus ou évidents mais qu'il paraît primordial de répéter, comme par exemple l'importance du contenu et de la représentativité du *training set* par rapport au jeu de données complet, ou encore la nécessité d'avoir un catalogue manuel initial fiable pour faire de la classification supervisée. Lorsque ce n'est pas le cas, une stratégie est d'utiliser des méthodes de classification non-supervisée comme les K-moyennes. Celles-ci s'avèrent être de bons outils pour tenter de mieux connaître ce que contiennent réellement les données, et donc ce qu'on peut espérer en obtenir. Cependant, le choix des attributs demeure toujours essentiel. Enfin, l'exemple du Kawah Ijen amène aussi à se poser la question sur la manière de définir un problème de classification. Dans le cas du Piton de la Fournaise, cela ne posait pas de problème car les deux classes des VT et des éboulements se séparaient bien et facilement. Pour le Kawah Ijen, en revanche, on a constaté que les deux classes des VB et des tremors se séparaient bien, mais que la présence d'une (ou plusieurs) classe(s) intermédiaire(s) venait « gêner » la classification. C'est la continuité qui existe entre les deux types d'événements qui complexifie le problème, et on peut clairement se demander si une classification nette comme dans le cas du Piton de la Fournaise est adaptée au cas du Kawah Ijen. Rappelons également que l'échelle à laquelle on souhaite travailler est aussi importante dès le départ : pour le Piton de la Fournaise, par exemple, il aurait aussi été possible d'essayer de classer les événements volcano-tectoniques uniquement dans différentes familles, de manière analogue à ce qui est fait lors de la recherche de multiplets par corrélation des formes d'ondes.

Perspectives et développements futurs envisageables

Le travail initié dans cette thèse a montré qu'il était possible d'automatiser de manière efficace et fiable les processus de dépouillement des données sismologiques, incluant la détection, la localisation et la classification des événements, à condition que le travail effectué en amont soit le plus complet possible. De nombreux développements visant à améliorer l'automatisation sont bien sûr possibles et souhaitables.

Pour la localisation
Sur le volcan du Piton de la Fournaise, on s'est contenté de localiser les événements de type volcano-tectoniques. Cependant, l'algorithme de Waveloc devrait pouvoir s'adapter à d'autres types d'événements, comme les éboulements, via l'utilisation d'une fonction caractéristique et d'un modèle de vitesse adaptés (voir Hibert [2012]). Dans l'exemple du Kawah Ijen, l'agencement peu avantageux du réseau de stations a malheureusement rendu impossible la localisation automatique. Une amélioration serait donc la bienvenue dans le futur, d'autant plus que la connaissance des localisations pourrait aussi permettre une meilleure appréhension du problème de classification (des événements se produisant au même endroit ont plus de chance d'appartenir à la même classe).

Pour la classification
On a vu que lorsque que le jeu de données est bon et bien défini, comme dans le cas du Piton de la Fournaise, les méthodes de régression logistique et de SVM donnaient d'excellents résultats. En ce qui concerne le volcan du Kawah Ijen, les résultats sont encourageants dans l'ensemble, mais il semble qu'une redéfinition claire et une reclassification manuelle des différents types d'événements soient nécessaires : en repartant de la base, en ne définissant qu'un nombre limité de classes, quitte à les complexifier au fur et à mesure (un peu à la manière d'un arbre de décision). Tant qu'une classification manuelle fiable n'est pas disponible, l'utilisation de plusieurs méthodes de classification non-supervisée peut s'avérer intéressante en fournissant des informations légèrement différentes. On peut globalement en distinguer deux grands types : les méthodes hiérarchiques, qui réalisent des mesures de similarité (type arbres de décision - pas mises en place dans ce travail) ; ou partitives, avec ré-allocation dynamique de chaque élément à chaque itération (type K-moyennes). De plus, on a vu sur ce jeu de données que la classification des classes de très petite taille était impossible et que l'application de stratégies « classiques » telles que les extracteurs ne permettait pas de l'améliorer. La détection de ces événements rares (potentiellement tous différents les uns des autres) reste donc un point crucial à approfondir, via l'utilisation de méthodes de détection d'anomalies déjà existantes dans d'autres domaines.
Concernant le choix (crucial) des combinaisons d'attributs, il paraît aussi possible de mettre en place diverses techniques : soit, à l'instar de Curilem et al. [2009], en mettant en œuvre un algorithme qui explore toutes les combinaisons possibles pour ne retenir que la meilleure (procédure relativement lourde à mettre en place et coûteuse en temps de calcul) ; soit en réduisant les dimensions de l'espace des caractéristiques en effectuant une analyse en composantes principales (PCA - *Principal Component Analysis*) ou une analyse en composantes indépendantes (ICA - *Independent Component Analysis*). Si la PCA permet de trouver les directions de l'espace des caractéristiques qui correspondent aux directions de variance maximale (on peut donc fixer un seuil de variance minimum), l'ICA trouve les directions de plus grande « non-gaussianité ». Ces deux techniques permettent de s'assurer qu'on utilise l'information la moins redondante possible et permettent normalement une meilleure optimisation de l'utilisation des informations contenues dans les données. Cependant, l'un des désavantages de ces méthodes est l'absence d'un sens physique facilement reliable aux observations effectuées.

Bilan

Les travaux présentés dans cette thèse ont permis de mettre en évidence qu'il est possible d'automatiser les premières étapes de traitement des données sismologiques, mais qu'un certain nombre de conditions doivent pour cela être réunies au préalable (Fig. 2). Cependant, les étapes préliminaires de détection, localisation et classification ne sauraient se substituer à des études plus détaillées de la sismicité. L'automatisation permet tout simplement de faciliter et favoriser le travail en aval en fournissant des pistes de recherche. Dans le cas du Piton de la Fournaise, on a par exemple mis en évidence une répartition intéressante de la microsismicité durant les périodes de crises, ainsi qu'une distribution très spécifique des magnitudes au sein des essaims. Ces observations devraient permettre aux spécialistes d'investiguer davantage les processus impliqués dans les éruptions du volcan. Pour le Kawah Ijen, il reste encore beaucoup

de travail à fournir. Néanmoins, les données sismologiques sont riches et contiennent des informations intéressantes. Comme l'importance des interactions avec les phénomènes hydrothermaux est connue pour ce volcan, il serait opportun de combiner les informations sismologiques avec d'autres types de sources d'informations dans le problème de classification : par exemple, prendre en compte la température, le pH ou les changements de couleur du lac acide... Ceci devrait permettre notamment de discriminer les événements qui sont directement liés à cette activité hydrothermale.
En outre, l'union des méthodes de localisation et de classification serait essentielle pour connaître la répartition spatio-temporelle des divers types d'événements et renseigner sur les mécanismes potentiels qui les génèrent. A terme, l'idéal serait de réussir à construire un outil qui combine détection, localisation et classification des événements pour un espace d'étude donné. Waveloc détecte et localise simultanément les événements sismiques. Les travaux de Beyreuther [2011], basés sur les méthodes utilisées en reconnaissance de la parole, permettent leur détection et classification simultanées. Une méthode qui allierait les 3 étapes permettrait un suivi de l'évolution spatio-temporelle des différents types d'événements.

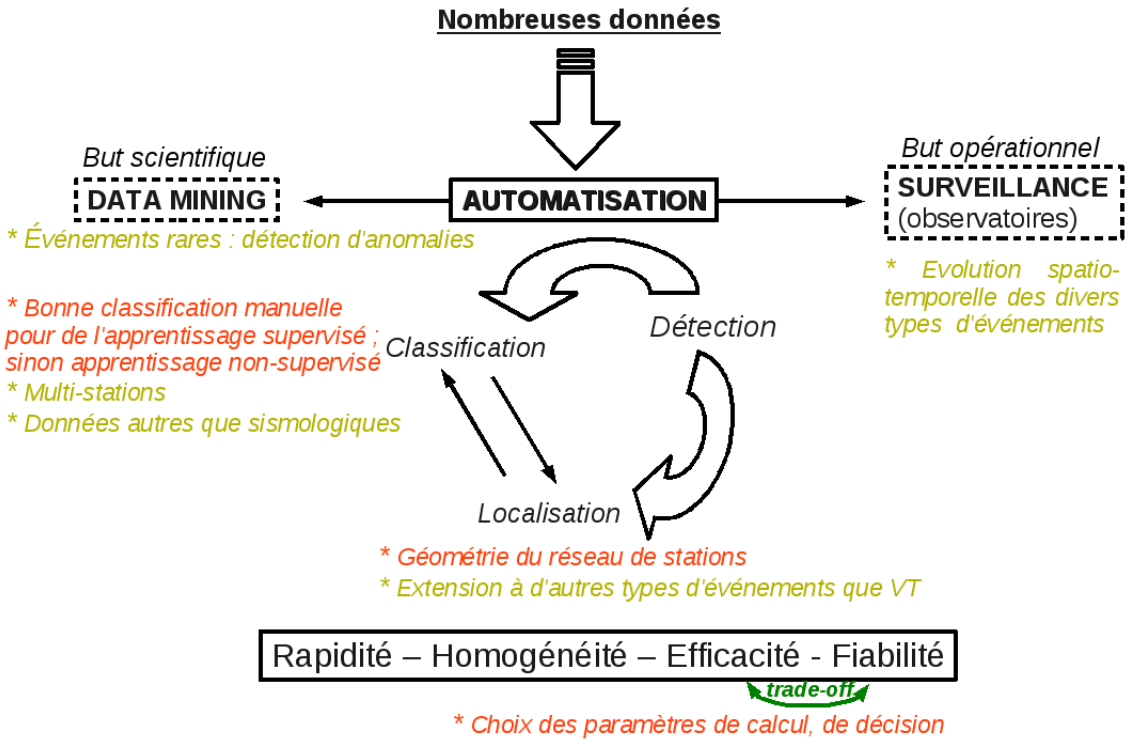


FIG. 2: Schéma bilan et perspectives.

Enfin, l'intérêt des méthodes automatiques réside aussi dans leur capacité à s'adapter à tous types de données et à divers problèmes. Dans cette thèse, on s'est plus particulièrement concentré sur des données sismologiques enregistrées en milieu volcanique, mais les algorithmes développés devraient pouvoir fonctionner pour d'autres domaines, comme par exemple la sismicité induite. Dans ce but, et dans un souci d'amélioration continue des algorithmes (plus on diversifie les exemples d'application, plus il est possible de mettre en place

un système général), ceux-ci ont été rendus publics. Le code de Waveloc est téléchargeable via http://github.com/amaggi/waveloc et les codes de discrimination sont accessibles via http://github.com/nlanget/discrimination.

Bibliographie

R. Allen. Automatic phase pickers : Their present use and future prospects. *Bull. Seismol. Soc. Am.*, 72(6B) :pp.S225–242, Dec. 1982. *Cité p.10*

J. Antoni. Fast computation of the kurtogram for the detection of transient faults. *Mechanical Systems and Signal Processing*, 21(1) :pp.108–124, 2007. *Cité p.10 et 55*

C. Baillard, W.C. Crawford, V. Ballu, C. Hibert, and A. Mangeney. An automatic kurtosis-based P- and S-phase picker designed for local seismic networks. *Bull. Seismol. Soc. Am.*, 104(1) :pp.394–409, Feb. 2014. *Cité p.51*

T. Baker, R. Granat, and R.W. Clayton. Real-time earthquake location using Kirchhoff reconstruction. *Bulletin of the Seismological Society of America*, 95(2) :pp.699–707, 2005. *Cité p.7*

W.H. Bakun and W.B. Joyner. The Ml scale in central California. *Bull. Seismol. Soc. Am.*, 74 :pp.1827–1843, 1984. *Cité p.19*

S. Baluja and M. Covell. Content fingerprinting using wavelets. *Proceedings of the 3rd European Conference on Visual Media Production (CVMP 2006)*, pages pp.198–207, Jan. 2006. *Cité p.61*

A.E. Barnes. Instantaneous spectral bandwidth and dominant frequency with applications to seismic reflection data. *Geophysics*, 58(3) :pp.419–428, Mar. 1993. *Cité p.67 et 68*

J. Battaglia. Etude détaillé des crises sismiques intrusives et pré-éruptives, 2012. *Cité p.95*

J. Battaglia and F. Brenguier. Seismogenic structures activated during the pre-eruptive and intrusive swarms of Piton de la Fournaise volcano (La Réunion island) between 2008 and 2011. AGU Fall meeting 2011, San Francisco, USA, 2011. *Cité p.95 et 96*

M.C. Benítez, J. Ramírez, J.C. Segura, J.M. Ibáñez, J. Almendros, A. García-Yeguas, and G. Cortés. Continuous HMM-based seismic-event classification at Deception Island, Antarctica. *IEEE Transactions on geoscience and remote sensing*, 45(1) :pp.138–146, Jan. 2007. *Cité p.30*

M. Beyreuther. *Speech recognition based automatic earthquake detection and classification.* PhD thesis, Fakultät für Geowissenschaften, Ludwig-Maximilians-Universität München, Feb. 2011. *Cité p.192*

M. Beyreuther and J. Wassermann. Continuous earthquake detection and classification using discrete Hidden Markov Models. *Geophys. J. Int.*, 175 :pp.1055–1066, 2008. *Cité p.30, 57, 67, 69 et 74*

M. Beyreuther, C. Hammer, J. Wassermann, M. Ohrnberger, and T. Megies. Constructing a Hidden Markov Model based earthquake detector : application to induced seismicity. *Geophys. J. Int.*, 189 :pp.602–610, 2012. *Cité p.30*

F. Brenguier, P. Kowalski, T. Staudacher, V. Ferrazzini, F. Lauret, P. Boissier, P. Catherine, A. Lemarchand, C. Pequegnat, O. Meric, C. Pardo, A. Peltier, S. Tait, N. M. Shapiro, M. Campillo, and A. Di Muro. First results from the UnderVolc high resolution seismic and GPS network deployed on Piton de la Fournaise Volcano. *Seismological Research Letters*, 83(1) :pp.97–102, Jan. 2012. *Cité p.79*

C. Caudron. *Multi-disciplinary continuous monitoring of Kawah Ijen volcano, East Java, Indonesia.* PhD thesis, Université Libre de Bruxelles, Sept. 2013. *Cité p.140 et 141*

E. Chassande-Mottin. Testing normality of gravitational wave data with a low cost recursive estimate of the kurtosis. In *Proc. of PSIP'2003*, pages 157–160, Grenoble (France), 2003. *Cité p.12*

B.A. Chouet and R.S. Matoza. A multi-decadal view of seismic methods for detecting precursors of magma movement and eruption. *Journal of Volcanology and Geothermal Research*, 252 :pp.108–175, 2013. *Cité p.139*

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Stein Clifford. *Introduction to Algorithms, second edition.* MIT Press and McGraw-Hill, 2001. *Cité p.21*

G. Curilem, J. Vergara, G. Fuentealba, G. Acuña, and M. Chacón. Classification of seismic signals at Villarrica volcano (Chile) using neural networks and genetic algorithms. *Journal of Volcanology and Geothermal Research*, 180 :pp.1–8, 2009. *Cité p.29, 30, 74 et 191*

F.U. Dowla, S.R. Taylor, and R.W. Anderson. Seismic discrimination with artificial neural networks : preliminary results with regional spectral data. *Bull. Seismol. Soc. Am.*, 80(5) : pp.1346–1373, Oct. 1990. *Cité p.29 et 49*

G. Ekström. Global detection and location of seismic sources by using surface waves. *Bulletin of the Seismological Society of America*, 96(4A) :pp.1201–1212, 2006. *Cité p.7*

G. Ekström, M. Nettles, and G.A. Abers. Glacial earthquakes. *Science*, 302(5645) :pp.622–624, Oct. 2003. *Cité p.7*

A.M. Esposito, F. Giudicepietro, L. D'Auria, S. Scarpetta, M.G. Martini, M. Coltelli, and M. Marinaro. Unsupervised neural analysis of Very-Long-Period events at Stromboli volcano using the self-organizing maps. *Bull. Seismol. Soc. Am.*, 98(5) :pp.2449–2459, Oct. 2008. *Cité p.30*

S. Falsaperla, S. Graziani, G. Nunnari, and Spampinato. Automatic classification of volcanic earthquakes by using multi-layered neural networks. *Natural Hazards*, 13 :pp.205–228, 1996. *Cité p.29, 49 et 61*

Y.V. Fedorenko, E.S. Husebye, B. Heincke, and B.O. Ruud. Recognizing explosion sites without seismogram readings : neural network analysis of envelope-transformes multistation sp recordings 3-6 hz. *Geophys, J. Int.*, 133 :F1–F6, 1998. *Cité p.29*

D. Gajewski and E. Tessmer. Reverse modelling for seismic event characterization. *Geophysical Journal International*, 163(1) :pp.276–284, 2005. *Cité p.8*

S. Gentili and A. Michelini. Automatic picking of P and S phases using a neural tree. *Journal of Seismology*, 10(1) :pp.39–63, 2006. *Cité p.10*

H.N. Gharti, V. Oye, M. Roth, and D. Kühn. Automated microearthquake location using envelope stacking and robust global optimization. *Geophysics*, 75(4) :MA27–MA46, 2010. *Cité p.8*

C. Hammer, M. Beyreuther, and M. Ohrnberger. A seismic-event spotting system for volcano fast-response systems. *Bull. Seismol. Soc. Am.*, 102(3) :pp.948–960, Jun. 2012. *Cité p.30, 57, 67, 69 et 74*

C. Hibert. *L'apport de l'écoute sismique pour l'étude des éboulements du cratère Dolomieu, Piton de la Fournaise, La Réunion.* PhD thesis, Institut de Physique du Globe de Paris, Sept. 2012. *Cité p.80, 107, 129, 130 et 190*

C. Hibert, A. Mangeney, G. Grandjean, C. Baillard, D. Rivet, N.M. Shapiro, C. Satriano, A. Maggi, V. Ferrazzini, and W. Crawford. Automated identification, location, and volume estimation of rockfalls at piton de la fournaise volcano. *J. Geophys. Res. : Earth Surf.*, 119 : pp.1082–1105, May 2014. *Cité p.30, 51, 52, 53, 74, 81, 108, 118, 120, 121, 122, 123 et 130*

J.M. Ibáñez, C. Benítez, L.A. Gutiérrez, G. Cortés, A. García-Yeguas, and G. Alguacil. The classification of seismo-volcanic signals using Hidden Markov Models as applied to the Stromboli and Etna volcanoes. *J. Volcanol. Geotherm. Res.*, 187 :pp.218–226, 2009. *Cité p.30*

M. Ibs-von Seht. Detection and identification of seismic signals recorded at Krakatau volcano (Indonesia) using artificial neural networks. *Journal of Volcanology and Geothermal Research*, 176 :pp.448–456, 2008. *Cité p.29 et 74*

C. Johnson, A. Bittenbinder, B. Bogaert, L. Dietz, and W. Kohler. Earthworm : a flexible approach to seismic network monitoring. *IRIS Newsletter*, 14 :pp.1–4, 1995. *Cité p.92*

C.E. Johnson, A.G. Lindh, and B. Hirshorn. Robust regional phase association. Open File Report 94-621, USGS, 1994. *Cité p.7*

A. Jurkevics. Polarization analysis of three-component array data. *Bull. Seismol. Soc. Am.*, 78(5) :pp.1725–1743, Oct. 1988. *Cité p.70*

H. Kao and S-J. Shan. The Source-Scanning Algorithm : mapping the distribution of seismic sources in time and space. *Geophysical Journal International*, 157(2) :pp.589–594, 2004. *Cité p.7 et 8*

E. Kissling, U. Kradolfer, and H. Maurer. Program VELEST user's guide - short introduction. *ETH Zürich open-file report*, Oct. 1995. *Cité p.187*

O. Kulhanek. Seminar on b-value. *University of Prague*, Dec. 2005. *Cité p.20*

L. Küperkoch, T. Meier, J. Lee, W. Friederich, and EGELADOS Working Group. Automated determination of P-phase arrival times at regional and local distances using higher order statistics. *Geophysical Journal International*, Jul. 2010. *Cité p.10*

J.C. Lahr. HYPOELLIPSE : a computer program for determining local earthquake hypocentral parameters, magnitude and first-motion pattern. *U.S. Geol. Surv. open-file report 99-23*, 1999 - revised in 2012. *Cité p.187*

J.C. Lahr, B.A. Chouet, C.D. Stephens, J.A. Power, and R.A. Page. Earthquake classification, location, and error analysis in a volcanic environment : implications for the magmatic system of the 1989-1990 eruptions at Redoubt Volcano, Alaska. *J. Volcanol. Geotherm. Res.*, 62 : pp.137–151, 1994. *Cité p.139*

H. Langer and S. Falsaperla. Seismic monitoring at Stromboli volcano Italy : a case study for data reduction and parameter extraction. *Journal of Volcanology and Geothermal Research*, 128 :pp.233–245, 2003. *Cité p.29*

H. Langer, S. Falsaperla, and G. Thompson. Application of Artificial Neural Networks for the classification of the seismic transients at Soufrière Hills volcano, Montserrat. *Geophysical Research Letters*, 30(21), 2003. *Cité p.29*

H. Langer, S. Falsaperla, T. Powell, and G. Thompson. Automatic classification and a-posteriori analysis of seismic event identification at Soufrière Hills volcano, Montserrat. *J. Volcanol. Geotherm. Res.*, 153 :pp.1–10, 2006. *Cité p.29*

N. Langet, A. Maggi, A. Michelini, and F. Brenguier. Continuous kurtosis-bases migration for seismic event detection and location, with application to Piton de la Fournaise volcano, La Réunion. *Bull. Seismol. Soc. Am.*, 104(1) :pp.229–246, Feb. 2014. *Cité p.7*

C. Larmat, J-P. Montagner, M. Fink, Y. Capdeville, A. Tourin, and E. Clévédé. Time-reversal imaging of seismic sources and application to the great Sumatra earthquake. *Geophys. Res. Lett.*, 33 :–, Oct. 2006. *Cité p.8*

W. H. K. Lee and S. W. Stewart. *Principles and Applications of microearthquake networks.* Academic Press, 1981. *Cité p.7*

W.H.K. Lee and J.C. Lahr. HYPO71 : a computer program for determining hypocenter, magnitude and first motion pattern of local earthquakes. *U.S. Geol. Surv. open report*, 1975. *Cité p.92*

W.H.K. Lee, R.E. Bennett, and K.L. Meagher. A method of estimating magnitude of local earthquakes from signal duration. *USGS open file report*, 1972. *Cité p.93*

P. Lesage. Interactive Matlab software for the analysis of seismic volcanic signals. *Computers & Geosciences*, 35 :pp.2137–2144, 2009. *Cité p.141*

Y.C. Liao, H. Kao, A. Rosenberger, S.K. Hsu, and B.S. Huang. Delineating complex spatio-temporal distribution of earthquake aftershocks : an improved Souce-Scanning Algorithm. *Geophys. J. Int.*, 189 :pp.1753–1770, 2012. *Cité p.7 et 8*

A. Lomax. The NonLinLoc software guide, 2011. URL http://alomax.free.fr/nlloc/. *Cité p.14 et 85*

A. Maggi and A. Michelini. Rapid waveform earthquake location in Italy. *Geophysical Research Abstracts*, 11, 2009a. EGU2009-4977. *Cité p.9*

A. Maggi and A. Michelini. Continuous waveform data stream analysis : Detection and location of the L'Aquila earthquake sequence. *Eos Trans. AGU*, 90(52), 2009b. Fall Meet. Suppl., Abstract U12A-01. *Cité p.9*

A. Maggi and A. Michelini. Waveloc - an algorithm for the detection and location of seismic sources within large, continuous waveform data volumes : the case of the l'Aquila earthquake sequence. *Geophysical Research Abstracts*, May 2010. EGU, Vienna, Austria. *Cité p.9 et 189*

M. Masotti, S. Falsaperla, H. Langer, S. Spampinato, and R. Campanini. Application of Support Vector Machine to the classification of volcanic tremor at Etna, Italy. *Geophysical Research Letters*, 33, 2006. *Cité p.30*

G.A. McMechan, J.H. Luetgert, and W.D. Mooney. Imaging of earthquake sources in Long Valley Caldera, California, 1983. *Bulletin of the Seismological Society of America*, 75(4) : pp.1005–1020, Aug. 1985. *Cité p.8*

S.R. McNutt. Observations and analysis of B-type earthquakes, explosions, and volcanic tremor at Pavlof volcano, Alaska. *Bull. Seismol. Soc. Am.*, 76(1) :pp.153–175, Feb. 1986. *Cité p.139*

S.R. McNutt. *International handbook of Earthquake & Engineering Seismology*, chapter 25 - Volcano Seismology and Monitoring for Eruptions, pages pp.383–406. Academic Press, 2002. *Cité p.93 et 139*

T. Minakami. Fundamental research for predicting volcanic eruptions (part 1) : Earthquakes and crustal deformations originating from volcanic activities. *Bulletin of the Earthquake Research Institute, Tokyo University*, 38 :pp.497–544, 1960. *Cité p.139*

T. Minakami. *Developments in Solid Earth Geophysics*, chapter Seismology of volcanoes in Japan, pages 1–6. Elsevier, Amsterdam, 1974. *Cité p.139 et 140*

M. Musil and A. Plešinger. Discrimination between local microearthquakes and quarry blasts by multi-layer perceptrons and Kohonen maps. *Bull. Seismol. Soc. Am.*, 86(4) :pp.1077–1090, Aug. 1996. *Cité p.29*

A. Ng. Machine learning - Coursera online courses, 2012. URL http://www.coursera.org. *Cité p.31*

M. Ohrnberger. *Continuous automatic classification of seismic signals of volcanic origin at Mt. Merapi, Java, Indonesia.* PhD thesis, Institut für Geowissenschaften, University of Potsdam, Apr. 2001. *Cité p.30, 70, 74 et 135*

O.J. O'Reilly, C.E. Yoon, and G.C. Beroza. Similarity search for continuous seismic data. *Poster, AGU Fall meeting 2013, San Francisco*, Dec. 2013. *Cité p.61*

E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3) :pp.1065–1076, 1962. *Cité p.27*

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :pp.2825–2830, 2011. *Cité p.73*

P. Podvin and I. Lecomte. Finite difference computation of traveltimes in very contrasted velocity models : a massively parallel approach and its associated tools. *Geophysical Journal International*, 105 :pp.271–284, Apr. 1991. *Cité p.14*

G. Poupinet, W.L. Ellsworth, and J. Fréchet. Monitoring velocity variations in the crust using earthquake doublets : An application to the Calaveras Fault, California. *J. Geophys. Res*, 89 :pp.5719–5731, 1984. *Cité p.20 et 21*

E. Prono, J. Battaglia, V. Monteiller, J. L. Got, and V. Ferrazzini. P-wave velocity structure of Piton de la Fournaise volcano deduced from seismic data recorded between 1996 and 1999. *Journal of Volcanology and Geothermal Research*, 184(1) :pp.49–62, 2009. *Cité p.85*

J.J. Pulli. An experiment in the use of trained neural networks for regional seismic event classification. *Geophysical Research Letters*, 17(7) :pp.977–980, Jun. 1990. *Cité p.29 et 49*

M. Rierola. Temporal and spatial transients in b-values beneath volcanoes. Master's thesis, ETH Zurich, Institute of Geophysics, 2000. *Cité p.20 et 93*

A. Rietbrock and F. Scherbaum. Acoustic imaging of earthquake sources from the Chalfant Valley, 1986, aftershock series. *Geophysical Journal International*, 119(1) :pp.260–268, 1994. *Cité p.8*

C.D Saragiotis, L.J. Hadjileontiadis, and S.M. Panas. PAI–S/K : A robust automatic seismic P phase arrival identification scheme. *IEEE Transactions on Geoscience and Remote Sensing*, 40(6) :pp.1395–1404, Jul. 2002. *Cité p.10*

S. Scarpetta, F. Giudicepietro, E.C. Ezin, S. Petrosino, E. Del Pezzo, M. Martini, and M. Marinaro. Automatic classification of seismic signals at Mt. Vesuvius volcano, Italy, using neural networks. *Bull. Seismol. Soc. Am.*, 95(1) :pp. 185–196, Feb. 2005. *Cité p.29*

D.P. Schaff, G.H.R. Bokelmann, W.L. Ellsworth, E. Zanzerkia, F. Waldhauser, and G.C. Beroza. Optimizing correlation techniques for improved earthquake location. *Bull. Seismol. Soc. of Am.*, 94(2) :pp.705–721, 2004. *Cité p.21*

A. Schmid. *Quelle prédictibilité pour les éruptions volcaniques ? De l'échelle mondiale au Piton de la Fournaise*. PhD thesis, Grenoble, Oct. 2011. *Cité p.80, 85 et 97*

P. Shearer. Global seismic event detection using a matched filter on long-period seismograms. *J. Geophys. Res.*, 99(B7) :pp.13713–13725, 1994. *Cité p.7*

E.J. Stollnitz, T.D. DeRose, and D.H. Salesin. Wavelets for computer graphics : A primer, part 1. *IEEE Computer Graphics and Applications*, 15(3) :pp.76–84, May 2005. *Cité p.61*

B. Taisne, F. Brenguier, N.M. Shapiro, and V. Ferrazzini. Imaging the dynamics of magma propagation using radiated seismic intensity. *Geophysical Research Letters*, 38 :L04304, 2011. *Cité p.83*

M.T. Taner, F. Koehler, and R.E. Sheriff. Complex seismic trace analysis. *Geophysics*, 44(6) : pp.1041–1063, Jun. 1979. *Cité p.67*

A. Ursino, H. Langer, L. Scarfi, G. Di Grazia, and S. Gresta. Discrimination of quarry blasts from tectonic microearthquakes in the Hyblean Plateau (Southern Sicily). *Annali di Geofisica*, 44(4), Aug. 2011. *Cité p.29*

G. Van Kriekinge. Localisation et évolution spatio-temporelle des événements sismiques de juin 2011 à juin 2012 du volcan kawah ijen, indonésie. Master's thesis, Observatoire Royal de Belgique, Université Libre de Bruxelles, 2013. *Cité p.187*

F. Waldhauser and W.L. Ellsworth. A double-difference earthquake location algorithm : method and application to the northern Hayward fault, California. *Bull. Seismol. Soc. Am.*, 90(6) :pp.1353–1368, Dec. 2000. *Cité p.20 et 21*

S. Wiemer. A software package to analyze seismicity : ZMAP. *Seismological Research Letters*, 72(3) :pp.373–382, May 2001. *Cité p.20*

M. Withers, R. Aster, C. Young, J. Beiriger, M. Harris, S. Moore, and J. Trujillo. A comparison of select trigger algorithms for automated global seismic phase and event detection. *Bulletin of the Seismological Society of America*, 88(1) :pp.95–106, 1998. *Cité p.7*

M. Withers, R. Aster, and C. Young. An automated local and regional seismic event detection and location system using waveform correlation. *Bulletin of the Seismological Society of America*, 89(3) :pp.657–669, 1999. *Cité p.7*

J. Woessner and S. Wiemer. Assessing the quality of earthquake catalogues : estimating the magnitude of completeness and its uncertainty. *Bull. Seismol. Soc. Am.*, 95(2) :pp.674–698, Apr. 2005. *Cité p.20*

C. Young, M. Harris, J. Beiriger, S. Moore, J. Trujillo, M. Withers, and Aster R. The waveform correlation event detection system project phase 1 : issues in prototype development and testing. Sandia Report SAND96-1916, Sandia National Laboratories, August 1996. *Cité p.7*

A1 Article publié dans BSSA - février 2014

Bulletin of the Seismological Society of America, Vol. 104, No. 1, pp. 229–246, February 2014, doi: 10.1785/0120130107

Ⓔ

Continuous Kurtosis-Based Migration for Seismic Event Detection and Location, with Application to Piton de la Fournaise Volcano, La Réunion

by Nadège Langet, Alessia Maggi, Alberto Michelini, and Florent Brenguier

Abstract We present an automatic earthquake detection and location technique based on migration of continuous waveform data. Data are preprocessed using a kurtosis estimator in order to enhance the first arrival information, then migrated onto a predefined search grid using precalculated *P*-wave travel times, and finally stacked. Local maxima in the resulting 4D space–time grid indicate the locations and origin times of seismic events. We applied our technique to earthquake swarms occurring on Piton de la Fournaise volcano, La Réunion, France. We located 5000 events from 12 different swarms that occurred between 2009 and 2011. Our automated locations are consistent with those performed using manual picks and indicate that the seismicity concentrates around sea level. Multiplet analysis of the detected events and subsequent double-difference relocation produce sharper images of the earthquake swarms. Our code, Waveloc, is released in open source.

Online Material: Figures of seismicity distributions from Waveloc, synthetic test, and stack amplitude values versus magnitudes.

Introduction

Traditional earthquake location is performed using phase arrivals and event association (e.g., Lee and Stewart, 1981). Extraction of phase arrivals from waveform data greatly reduces the volume of information processed during the location process (e.g., Withers *et al.*, 1998) but requires the introduction of complex logic in order to associate each arrival to a single event. After a large mainshock (e.g., $M \geq 6$) or during a volcanic swarm, a cascade of earthquakes inundates the acquisition/detection system with multiple, near-simultaneous events in different parts of the affected area. In such conditions, procedures based on phase picks and event association are often known to fail (e.g., Johnson *et al.*, 1994). As a result, many of these events may be either falsely associated and mislocated or missed altogether. To reduce this problem, we need to exploit fully and automatically more of the information contained in the recorded waveforms (e.g., Young *et al.*, 1996). The automation of procedures becomes especially important when a large number of data streams are available.

During the past two decades, a number of studies have been published on the use of full waveforms to locate earthquakes both at global and regional/local scales. These techniques rely on information coherence through back projection and reverse time migration. Early work at the global scale was carried out by Shearer (1994) through waveform match filtering, using a grid representing potential event locations. Young *et al.* (1996) further developed the grid technique and proposed the waveform correlation event detection system. More recently, a detection approach based on waveform deconvolution was developed by Ekstrom (2006) to identify sources lacking in body waves, an approach that led Ekstrom *et al.* (2003) to detect anomalous seismic sources such as glacial earthquakes. On a regional/local scale, a similar methodology was first proposed by Withers *et al.* (1999), who pursued the correlation technique of Young *et al.* (1996) and developed the local waveform correlation event detection system. Along the same lines and more recently, Kao and Shan (2004) proposed the source-scanning algorithm (SSA) recently improved by Liao *et al.* (2012), and Baker *et al.* (2005) proposed the real-time Kirchhoff location method. Time reverse migration is another approach using back projection, and initial developments were made by McMechan *et al.* (1985). Rietbrock and Scherbaum (1994) and, more recently, Gajewski and Tessmer (2005) and Larmat *et al.* (2006) also adopted the same basic idea.

Implementation of back projection or shift-and-stack approaches requires a constant volume of computation, regardless of the number of event detections, and contains no complex logic. The inherent simplicity of such systems, and their apparent lack of optimization, make them robust and stable, even during an energetic aftershock sequence or earthquake swarm, a feature that is highly desirable for seismic or volcanic monitoring centers.

The method we describe in this paper, Waveloc, is a variant of the shift-and-stack methodologies such as SSA (Kao and Shan 2004; Liao *et al.*, 2012) and the envelope stacking method of Gharti *et al.* (2010), with an important difference: instead of stacking the energy of the signal envelopes within predetermined windows, we directly and continuously stack the *P*-wave arrival information as highlighted by kurtosis waveforms. The kurtosis is the fourth statistical moment of a distribution and has recently started to be applied in the computation of characteristic functions for automated *P*-wave arrival time picking (Saragiotis *et al.*, 2002; Gentili and Michelini, 2006; Kuperkoch *et al.*, 2010). We continuously migrate the positive gradient of the kurtosis time series to obtain detections and hypocentral locations of seismic events.

The Waveloc method for detecting and locating seismic sources within large, continuous-waveform data volumes was first developed in a 2D context (epicentral location only) and applied to the aftershock sequence of the M_w 6.3 L'Aquila, Italy, event that occurred on 6 April 2009 (Maggi and Michelini, 2009, 2010). Since then, optimization and simplification of the method have permitted us to extend it to fully 3D contexts. In this paper we describe the Waveloc method in detail and apply it in the context of volcanic seismicity swarms, which, like aftershock sequences, produce large numbers of events closely grouped in both space and time. In this paper, we present results for 12 seismic swarms recorded at Piton de la Fournaise volcano by the UnderVolc project (Brenguier *et al.*, 2012) between 2009 and 2011. For one of these swarms, that of 14 October 2010, we had access to manual picks and locations (Schmid, 2011) that we used to fine tune our method and evaluate its robustness and accuracy. As automated single-event locations generally have low location precision compared to multiple-event relocation methods such as double-difference algorithms (e.g. Waldhauser and Ellsworth, 2000; Michelini and Lomax, 2004), especially when the latter are combined with waveform correlation differential travel-time measurements (e.g. Schaff and Waldhauser, 2005), we added an option in Waveloc to correlate all detected and located events and perform double-difference locations on clusters of highly similar events.

Method

The Waveloc algorithm is a three-step process, illustrated schematically in Figure 1. In the first step, we process the raw waveforms in order to highlight the onset of seismic events using an appropriate characteristic function (kurtosis). In the second step, we migrate and stack the characteristic functions themselves, according to an *a priori P*-wave velocity model. In the third and final step, we detect and simultaneously locate the seismic events by analyzing the local maxima of the 3D time-dependent migration stacks. In the following, we shall explain each step in detail.

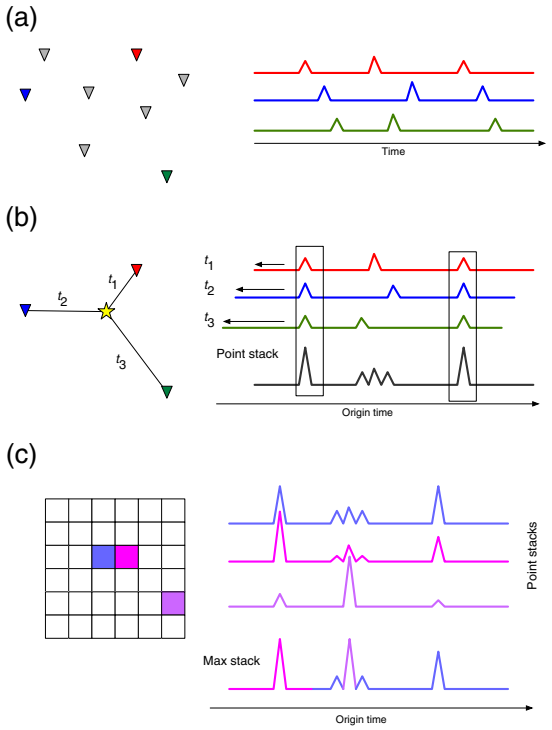


Figure 1. Schematic diagram of the Waveloc method. (a) Data processing: calculation of an appropriate characteristic function to maximize sensitivity to the onset of seismic events. (b) Migration: the processed waveforms are shifted back in time according to the predicted travel times from a potential hypocentral location then stacked. This procedure is repeated for a 3D grid of potential locations. (c) At each time step we store the maximum of the point stacks and the location it corresponds to, thereby allowing simple detection and event location. The color version of this figure is available only in the electronic edition.

First Arrival Enhancement: The Kurtosis Waveform

The purpose of this first step, and the purpose that Waveloc has in common with most of the more traditional phase-picking algorithms, is to highlight the first arrivals in the raw seismograms through an appropriate characteristic function. As we shall later use this characteristic function for migration, we require it to be sharply peaked at the phase arrival time. In automated phase picking, the most commonly used characteristic functions are variations on the short-term/long-term average (STA/LTA) method first introduced by Allen (1982) in which changes in the seismogram caused by an arriving phase are highlighted by taking the ratio of two sliding averages of the seismogram with two different window lengths. Such techniques form the basis of the automated pickers in both the Earthworm (Johnson *et al.*, 1995) and SeiscomP (e.g. Hanka *et al.*, 2010) software suites and have already been exploited by one of the authors for the automated selection of data windows for tomographic inversion

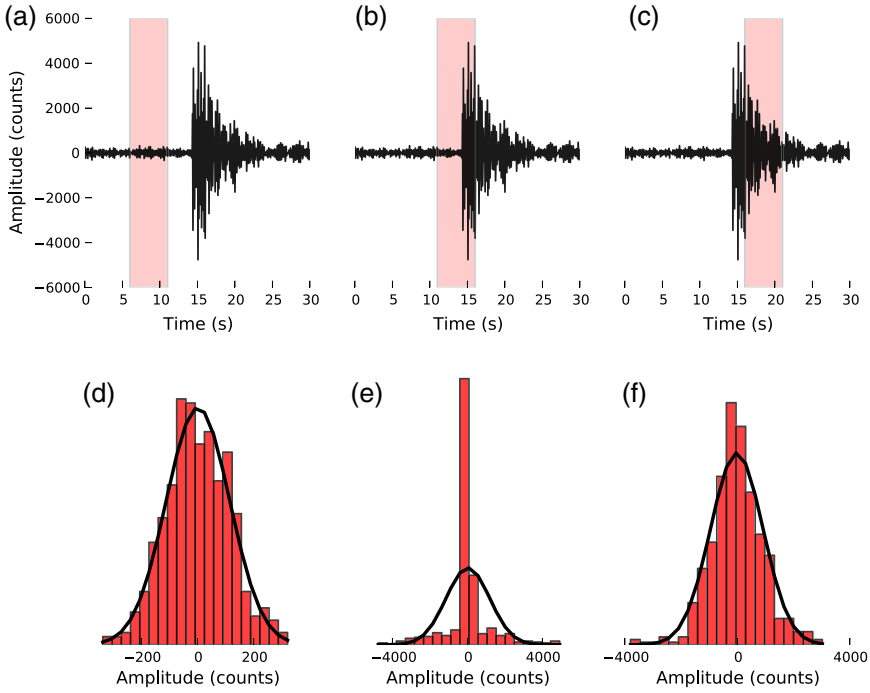


Figure 2. Amplitude distributions within different time windows along a seismic signal. (a–c) Raw seismic data from station UV15, starting at time 14 October 2010, 00:17 UTC. Time windows of 5 s duration are highlighted. (d–f) Normalized histograms of the amplitude distribution for each of the three highlighted time windows, overlain by the best-fit Gaussian (solid line). The amplitude distributions of both the seismic noise and the bulk of the seismic signal are approximately Gaussian. That of the transition between noise and signal is strongly peaked and will therefore have a correspondingly high value of kurtosis. The color version of this figure is available only in the electronic edition.

(Maggi *et al.*, 2009). Of the many alternative characteristic functions developed since STA/LTA was introduced, the ones that combine the two necessary conditions of simplicity of computation and strength of maximum at arrival times are those based on higher-order statistics, in particular the kurtosis and its gradient (Saragiotis *et al.*, 2002; Gentili and Michelini, 2006; Kuperkoch *et al.*, 2010).

The kurtosis is the fourth statistical moment of a distribution (the first and second statistical moments being the mean and variance, respectively, and the third its skewness). It is a nondimensional quantity that measures the peakedness (positive kurtosis values) or flatness (negative kurtosis values) of a distribution relative to a normal distribution. The standard definition of the kurtosis K is

$$K(x_1 \ldots x_n) = \left\{ \frac{1}{n} \sum_{j=1}^{n} \left[\frac{x_j - \bar{x}}{\sigma} \right]^4 \right\}, \qquad (1)$$

in which the distribution $(x_1 \ldots x_n)$ has mean $\bar{x}$ and variance σ^2. The kurtosis value of a Gaussian distribution computed using equation (1) is $+3$; it is common to subtract this value in order for the kurtosis of a Gaussian distribution to equal zero. As discussed by Kuperkoch *et al.* (2010) and illustrated in Figure 2, the amplitude distribution of a time window of seismic noise is close to being Gaussian, as is that of a time window within a seismic signal. However, in the transition between the two signals, the distribution becomes more highly peaked than the best-fit Gaussian distribution (its tail extends greatly) and will therefore give a strong-positive kurtosis value. We obtain a kurtosis waveform $K(t)$ by calculating the kurtosis within a sliding window and assigning it to the time of the last point in the window.

As the kurtosis measures a statistical property of the amplitude distribution of a signal, it will be sensitive to any preprocessing, such as filtering, that modifies this distribution. In order to enhance the values of kurtosis waveforms, we choose a frequency band that maximizes the kurtosis using the kurtogram method of Antoni (2007). A nonstationary signal $s(t)$ can be decomposed as follows (Wold–Cramér decomposition):

$$s(t) = \int_{-\infty}^{+\infty} e^{j2\pi f t} H(t, f) dX(f), \qquad (2)$$

in which $x(t)$ is the nonstationary signal, $H(t, f)$ is the complex envelope of $x(t)$, and $dX(f)$ is the spectral increment. The spectral kurtosis (SK) is then defined as

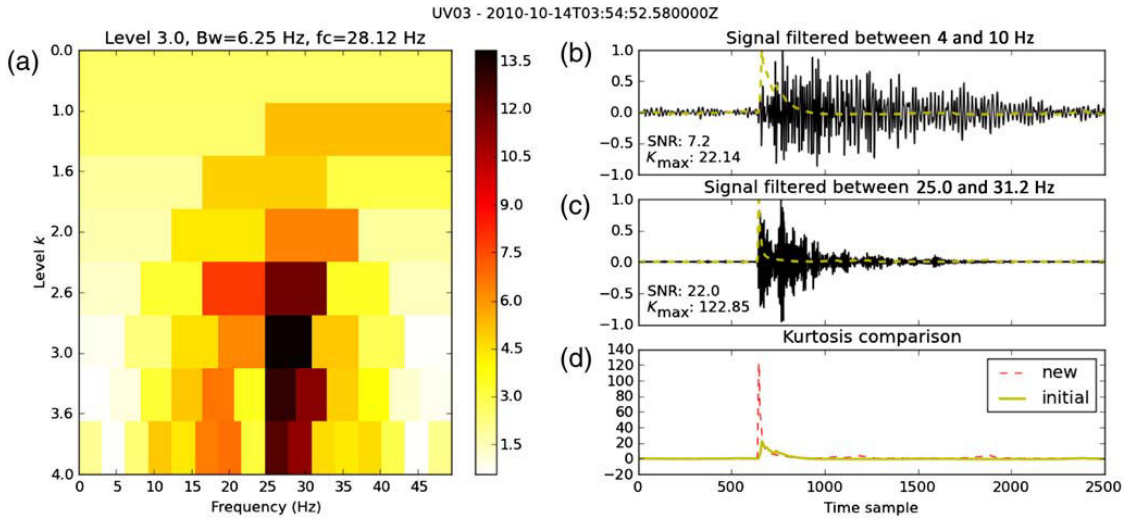


Figure 3. Kurtogram and choice of preprocessing filter. (a) Kurtogram of vertical-component seismic data from station UV03 starting at time 14 October 2010, 03:54:52.8 UTC and of duration 25 s. The kurtogram was obtained using a fast decimated filter bank tree algorithm and filter decomposition in thirds of power 2 (Antoni, 2007). The maximum value is reached at level 3.0 (bandwidth of 6.25 Hz) and centered on frequency 28.12 Hz. (b) Normalized data filtered between 4 and 10 Hz (standard filter parameters for routine analysis at Piton de la Fournaise observatory; F. Brenguier, personal comm., 2011) and the corresponding kurtosis waveform (dashed line). (c) Normalized data and corresponding kurtosis after filtering between 25.0 and 31.2 Hz, as indicated by the kurtogram maximum. (d) Direct comparison of the kurtosis waveforms for the two choices of filter. The color version of this figure is available only in the electronic edition.

$$SK_s(f) = \frac{|H(t,f)|^4}{\langle |H(t,f)|^2 \rangle^2} \tag{3}$$

and constitutes a representation of the transient features of the signal as a function of frequency. If we plot SK values as a function of frequency f and bandwidth Δf, we obtain a 2D plot, the kurtogram, an example of which is shown in Figure 3a. The $(f, \Delta f)$ pair corresponding to the maximum value of SK gives an indication of the best filtering parameters to use in preprocessing the data before calculating $K(t)$ (Fig. 3b–d).

The form of the kurtosis waveforms $K(t)$ also depends on the length of the sliding window used. Figure 4 shows an example of kurtosis calculated with three different window lengths. The maximum value of the kurtosis and its tail both increase with increasing window length. The maximum of the kurtosis is delayed with respect to the true first arrival, and this delay would induce a significant bias in the migrated origin times if the kurtosis waveforms themselves were used for the migration. In order to reduce this bias, we take the positive time derivative of K:

$$\dot{K}_+ = \begin{cases} \dot{K} & \text{if } \dot{K} \geq 0 \\ 0 & \text{otherwise} \end{cases}. \tag{4}$$

The resulting waveform is shown in Figure 4d and is a much more impulsive candidate for migration. A small bias in detection time still remains and shall be discussed in more detail later.

The kurtosis waveform $K(t)$ can be more efficiently calculated using a recursive method, such as that of Chassande-Mottin (2003). However, this method can become unstable for strongly nonstationary signals such as seismic events with strong onsets. Here we shall use a less formal recursive kurtosis estimator, which is stable in the presence of strong onsets. Let x be a signal whose standard deviation is σ_x. We estimate the recursive mean and standard deviation at the time of the ith sample of x, respectively, by

$$\bar{x}_i = C\bar{x}_{i-1} + (1 - C)x_i \tag{5}$$

and

$$\sigma_i = C\sigma_{i-1} + (1 - C)(x_i - \bar{x}_i)^2. \tag{6}$$

We then compute the kurtosis as follows:

$$K_i = \begin{cases} CK_{i-1} + (1 - C)\frac{(x_i - \bar{x}_i)^4}{\sigma_i^2} & \text{if } \sigma_i > \sigma_x \\ CK_{i-1} + (1 - C)\frac{(x_i - \bar{x}_i)^4}{\sigma_x^2} & \text{otherwise} \end{cases}, \tag{7}$$

in which the constant $C = 1 - (dt/w)$, dt is the sampling interval, and w an appropriately chosen time scale that serves a similar function to the window length for the more standard calculation in equation (1). The use of σ_x in equation (7) stabilizes the calculation when the instantaneous variances σ_i are small. When using the same window length in the two methods for calculating kurtosis waveforms, the sensitivity remains the same. However, the amplitudes are much larger

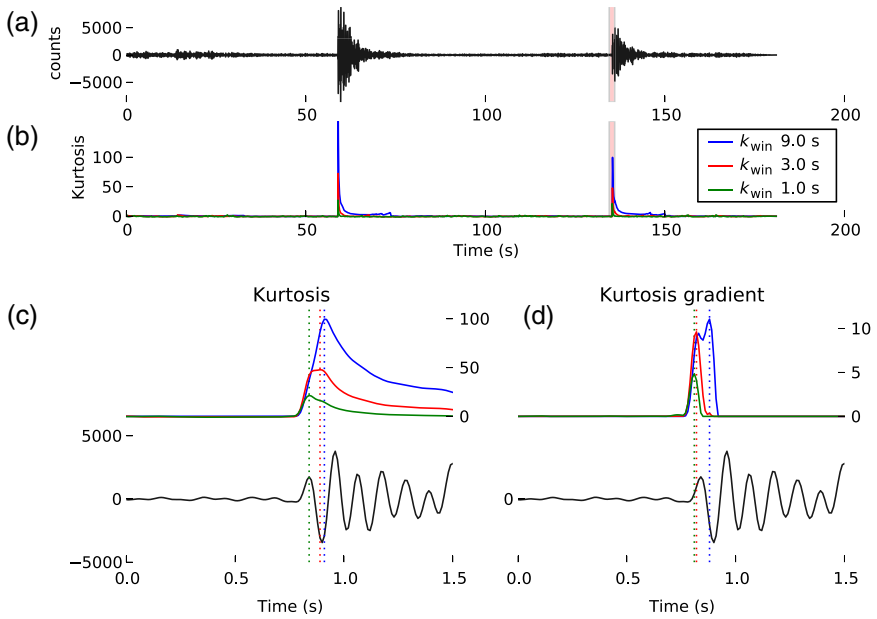


Figure 4. The sliding-window kurtosis $K(t)$ and its positive-valued gradient $\dot{K}_+(t)$. (a) Raw seismic data from station UV15 of the UnderVolc experiment starting at time 14 October 2010, 00:14:50 UTC. (b) The corresponding K waveform calculated using three different sliding window lengths k_{win}. The K waveforms are sharply peaked at the start of the seismic events, and low elsewhere. (c) Zoomed-in view around the portion of signal and $K(t)$ highlighted in (a) and (b). The maximum of $K(t)$ is systematically late with respect to the onset time of the event. (d) The same plot as in (c) showing the gradient $\dot{K}_+(t)$. For sufficiently small sliding windows (here 1–3 s) the maximum of $\dot{K}_+(t)$ is much closer to the onset time. The color version of this figure is available only in the electronic edition.

for recursive kurtosis and decrease more slowly. Kurtosis waveforms have similar amplitudes when w is approximately three times smaller than the standard kurtosis window length. We compare the results obtained using the two methods in the Application to Piton de la Fournaise Volcano section.

Migration

During the second step of the Waveloc process, we migrate the $\dot{K}_+(t)$ waveforms obtained in the first step onto a target grid of potential hypocentral locations, using an *a priori* P-wave velocity model for the region (see Fig. 1). We are exploiting here the coherence of first-arrival information across a network of stations, subject to sufficient knowledge of the regional velocity structure. We calculate and store travel times from each point of the target grid to each station through a 1D or 3D P-wave velocity model, using the Podvin and Lecomte (1991) eikonal solver as implemented by Lomax (2011) in the NonLinLoc package. For each point of the target grid, we shift each $\dot{K}_+(t)$ waveform back by the corresponding travel time, then stack the shifted waveforms to create a point stack (Fig. 1b):

$$\tilde{\dot{K}}_+^{ij}(t) = \dot{K}_+^j(t + \tau^{ij}) \tag{8}$$

and

$$S^i(t) = \sum_j \tilde{\dot{K}}_+^{ij}(t), \tag{9}$$

in which $\dot{K}_+^j(t)$ is the positive kurtosis time-derivative waveform for station j, τ^{ij} is the travel time from target point i to station j, $\tilde{\dot{K}}_+^{ij}(t)$ is the shifted waveform, and the point stack $S^i(t)$ is obtained by summing over all stations. We thereby obtain a 3D volume of point stacks $S(\mathbf{x}_i; t)$, in which $\mathbf{x}_i$ represents the geographical coordinates of the target-point i in some coordinate system, and t is absolute time. A strong local space–time maximum in $S(\mathbf{x}; t)$ indicates the occurrence of a seismic event at origin time t and position $\mathbf{x}$.

In order to illustrate the focussing achieved by $S(\mathbf{x}; t)$, we show in Figure 5 the result of a synthetic point-source retrieval test in a volcanic setting. It was set up using stations placed at the coordinates of the UnderVolc network on Piton de la Fournaise volcano, Réunion Island, France (Brenguier *et al.*, 2012), and a uniform target grid with 250 m node spacing. We used time grids calculated through the Prono *et al.* (2009) 3D velocity model to compute travel times to all stations from a grid node chosen as the hypocentral location, then constructed waveforms with simple triangular pulses at the corresponding times in order to simulate the $\dot{K}_+(t)$ waveforms. The width of these pulses was set to match that of the peaks in the observed $\dot{K}_+(t)$ waveforms (0.1 s). The

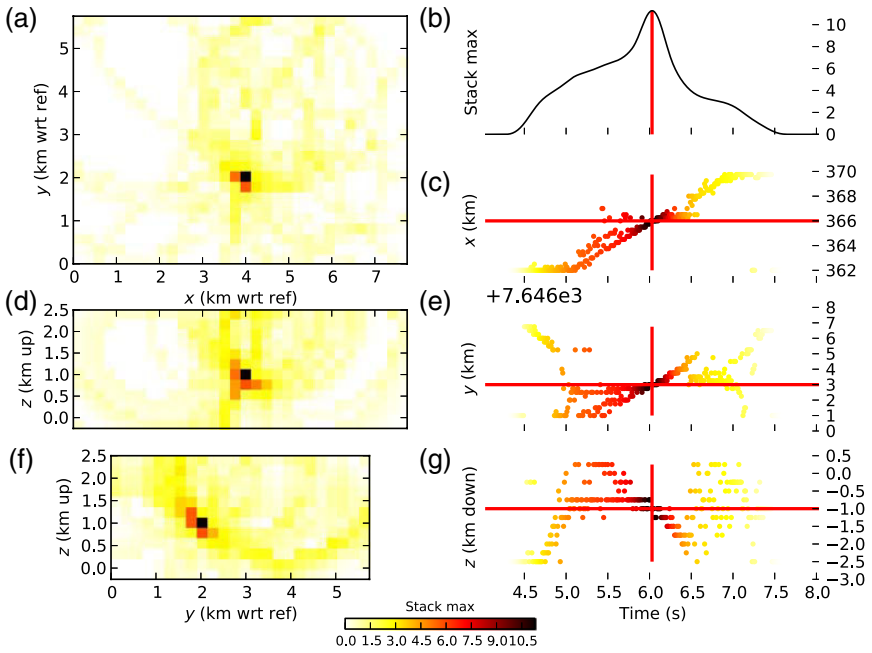


Figure 5. Synthetic test for the Piton de la Fournaise geometry. (a), (d), and (f) Cuts through the stack volume $S(\mathbf{x};t)$ at the synthetic hypocenter, showing focusing of the stack: (a) x–y plane, (d) x–z plane, and (f) y–z plane. In (d) and (f), z indicates elevation (positive up) in order to facilitate spatial orientation of the reader. (b) The summary stack $S_{\max}(t)$ for a 4 s time window centered around the synthetic origin time (vertical bar). (c), (e), and (g) The three components of $\mathbf{x}(t)$ within the same time window as (b), displayed using one point per time step. Shades represent the corresponding values of $S_{\max}(t)$ and use the same shades as for the stack cuts in (a), (d), and (f). The synthetic hypocentral coordinates are indicated by horizontal bars, and the origin time by a vertical bar. The bars intersect $\mathbf{x}(t)$ at the maximum stack values. In (g), z indicates event depth (positive down) in order to be consistent with most earthquake catalogs. The color version of this figure is available only in the electronic edition.

synthetic waveforms were then migrated back through the grid using the same time grids as for the forward problem.

We plotted orthogonal cut planes through the stack volumes at the coordinates of the true hypocenter in Figure 5a,d, and f. These cuts show that the migrated waveforms form spherical shells that intersect at the hypocenter. Depending on the source receiver geometry, these shells overlap over more or less of their surface and may smear the point of maximum focus. In Figure 5f, for example, there is some smearing in the y and z directions. Similar smearing will be expected for a real event at the same location, giving a first indication of the maximum location precision obtainable using this method.

For a typical problem with 20 stations, a search grid of $30 \times 30 \times 10 = 9000$ target points, and a sample rate of 100 Hz, the stack volume $S(\mathbf{x}, t)$ will contain 3.24×10^9 data points per hour of recorded data. In order to simplify the task of sifting through such a large volume of data, and to reduce storage needs, we create and store a summary stack $S_{\max}(t)$ containing for each time t the maximum value of $S(\mathbf{x}, t)$ (see Fig. 1c) and also store the coordinates of the corresponding target points $\mathbf{x}(t)$. The summary stack for the point retrieval test (Fig. 5b,c,e, and g) shows the peak in $S_{\max}(t)$ occurs at the correct origin time and that the corresponding $\mathbf{x}(t)$ time series cross the correct coordinates at the same instant in time. This correspondence will be exploited in the detection and location process described below. Note that values of $\mathbf{x}(t)$ outside the times at which an event occurs are meaningless, as the coordinates of the local maximum move all over the grid as a function of time. The large dispersion of z-values in Figure 5g is due to a combination of the uncertainty of event depth due to the geometry of recording stations and that due to a trade-off with origin time.

Detection and Location

We intend detection to refer to the determination of the occurrence of a seismic event and location to the determination of its origin time, its hypocentral coordinates, and their corresponding uncertainties.

As illustrated by the synthetic test in Figure 5, strong local maxima in $S_{\max}(t)$ occur at the origin times of seismic events. Therefore, our detection algorithm is a simple triggering algorithm applied directly on the $S_{\max}(t)$ time series. The choice of the triggering threshold depends on the impulsivity of the events and the number of stations and may be adjusted by visually comparing the $S_{\max}(t)$ time series to the raw waveforms on a representative portion of the data.

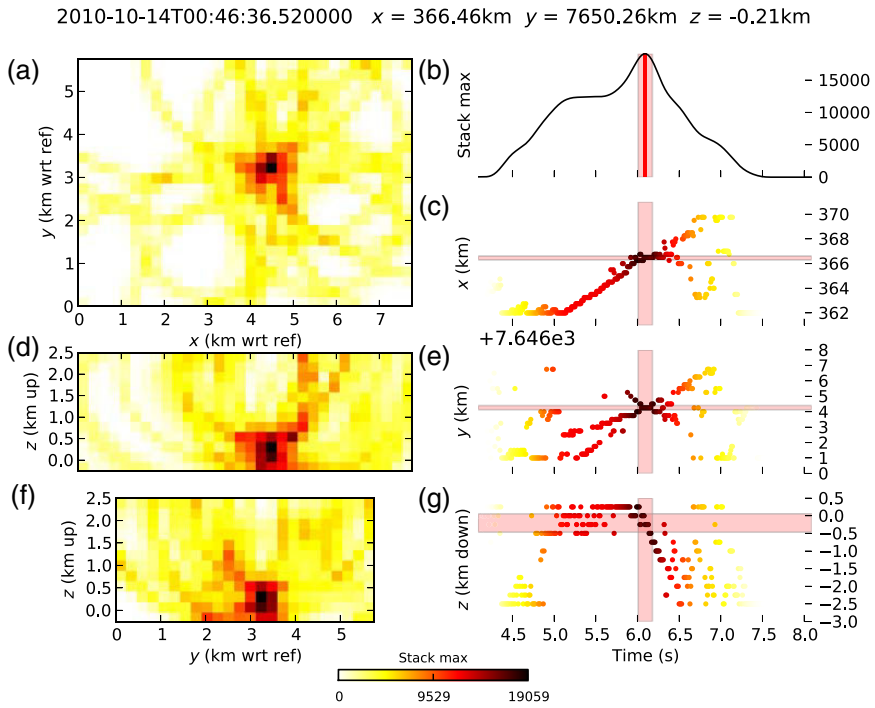


Figure 6. Migration and location using summary information of a volcano-tectonic event on Piton de la Fourn aise (14 October 2010, 00:46:36 UTC). Layout of figure as in Figure 5. Uncertainty bounds are indicated by shading in (b), (c), (e), and (g). The color version of this figure is available only in the electronic edition.

Location uses the summary information in $S_{\max}(t)$ and $\mathbf{x}(t)$ and is performed as follows. Once the triggering algorithm detects the occurrence of an event on $S_{\max}(t)$, we assign the time at which the corresponding local maximum occurs as the origin time t_0 and define the left and right uncertainty bounds t_1 and t_2 around this origin time by taking the times at which $S_{\max}(t)$ descends to 95% of its value at the local maximum. The mean and standard deviation of the corresponding components of $\mathbf{x}(t)$, between the times t_1 and t_2, are taken as hypocentral coordinates and their uncertainties, respectively.

Figure 6 shows an example of the results of this simple location algorithm for an event recorded during the pre-eruptive seismic crisis of 14 October 2010 on Piton de la Fournaise. This event, the waveforms of which are shown in Figure 7, is typical of the Piton de la Fournaise dataset: although it is clear that an event of some kind has occurred, data quality is variable across the network, and first arrivals are emergent at many stations. The cuts through the stack volume (Fig. 6a,d, and f) show more complexity than those for the synthetic test (Fig. 5a,d, and f), complexity that is due to the noisiness of the data and the spread in values for the maxima of the kurtosis gradients $\dot{K}_+(t)$ (Fig. 7b). The peak in $S_{\max}(t)$ is broader than for the synthetic test, and the location has an origin time uncertainty of ± 0.1 s (indicated by shading in the figure). Uncertainties in x, y, and z evaluated as described above (standard deviations of the $\mathbf{x}(t)$ values within the shaded time span in the figure) are respectively ± 0.2, ± 0.1, and ± 0.2 km.

We have carried out synthetic tests in order to evaluate Waveloc's ability to separate events occurring close in time (Fig. 8). The effective time resolution actually depends on how the kurtosis preprocessing was performed and on the aspect of the waveforms. To a lesser extent, it also depends on the threshold chosen by the user as well as on the depth of the trough between the two successive peaks in $S_{\max}(t)$. This test shows that triggering on a single summary waveform is too simple a method to detect events occurring close in time and that the development of a more complex detection algorithm that takes into account the full space–time stack should be considered in the future. Other synthetic tests show that Waveloc is able to locate an event properly with a signal-to-noise ratio (SNR) as low as 1.5 (Ⓔ see Fig. S12 in the electronic supplement).

Application to Piton de la Fournaise Volcano

Piton de la Fournaise volcano, a basaltic shield volcano located on La Réunion island, is characterized by a high level of activity (one eruption per year on average since 1998) and ranks amongst the most active volcanoes in the world. In

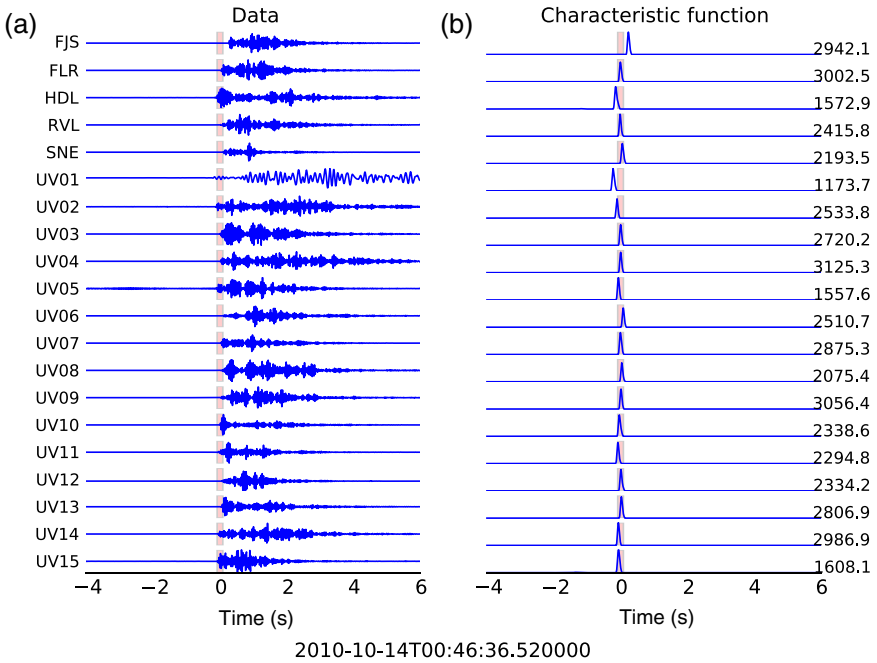


Figure 7. Migrated waveforms for the event in Figure 6: (a) raw waveforms and (b) kurtosis gradient waveforms $\dot{K}_+(t)$ underlain by the summary stack $S_{\max}(t)$. Peak values of $\dot{K}_+(t)$ are given beside the corresponding traces. All waveforms are time shifted by the expected travel times from the hypocenter, and the origin time uncertainty is indicated by shading. The color version of this figure is available only in the electronic edition.

order to monitor the volcanic activity and better understand the processes that occur in the edifice, 15 broadband seismic stations were installed in the framework of the UnderVolc project (2009–2012, Brenguier *et al.*, 2012) in addition to the 6 already set up by the Volcanological Observatory of Piton de la Fournaise (OVPF). Given the relatively steep topography of the volcano, the seismic network is effectively 3D, which partially compensates the disadvantages of using only *P*-wave information to locate seismic events.

Data from the 12 seismic swarms recorded between 2009 and 2011 are available from the portal of the French national seismological network RESIF (see Data and Resources). Not all swarms on Piton de la Fournaise are followed by eruptions, and the swarms are cataloged after the fact as either pre-eruptive or intrusive (this term is not the most appropriate, as eruptions are also preceded by intrusive phases, but we shall use it in the following to indicate the swarms that are not followed by eruptions).

This paper discusses three seismic swarms that occurred on 23 September 2010 (intrusive), 02 January 2010 (pre-eruptive), and 14 October 2010 (pre-eruptive). Ⓔ Results for the nine other swarms in the 2009–2011 period are shown in the electronic supplement to this article (see Figs. S2–S11). The 14 October 2010 swarm is of particular interest as it was picked manually and located by Agathe Schmid as part of her doctoral thesis (Schmid, 2011). She located 447 events from this swarm using the NonLinLoc software (Lomax, 2011) and the 3D *P*-wave velocity model for the Piton de la Fournaise volcano of Prono *et al.* (2009). In the following, we shall use the same travel-time grids as those used by Schmid (2011) in order to (1) facilitate direct comparison between our automated locations and her manual ones and (2) to fine-tune our algorithm and estimate its efficiency and reliability.

We considered that automated and manual events correspond to the same event when the difference between their origin times is less than 1 s. Then, for common events, we built histograms of the differences between automatic and manual hypocentral parameters (Figs. 9 and 10). These histograms give an idea of the accuracy we can expect to reach with Waveloc and illustrate the influence of various Waveloc parameters (band-pass filter, characteristic function used for migration, etc.) on the accuracy of the locations.

Figure 9 illustrates the importance of choosing an appropriate band-pass filter before computing the kurtosis waveform. In our discussion of the preprocessing stage of the Waveloc process, we showed how the amplitude of the kurtosis peak at event onset could be maximized by using the kurtogram technique (Antoni, 2007) to select the best pair $(f, \Delta f)$ (Fig. 3). Here we show the influence of band-pass selection on location accuracy: the dotted and dashed curves in Figure 9 were obtained respectively using a

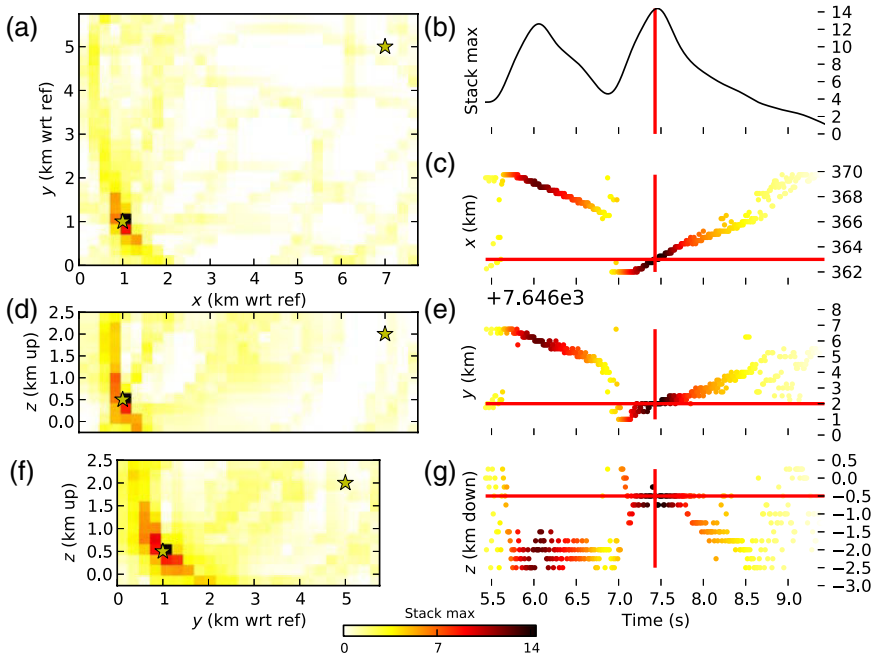


Figure 8. Synthetic test for event separation. We consider two events of the same amplitude occurring in a short time period (one at 6.0 s, the other at 7.4 s) and at two different places (stars). The figure was made for the event occurring at 7.4 s. Waveloc is able to distinguish both events: (b) two well-separated peaks are visible on the stacking trace; on (c), (e), and (g), coordinates are clearly identifiable and are not mixed with those of the other event. The color version of this figure is available only in the electronic edition.

4–10 Hz filter, used in standard processing at OVPF, and a 20–35 Hz filter selected after applying the kurtogram technique to all detected events. Both the number of detections and their accuracy are improved by the higher-frequency filter. The total number of events is increased from 150 to 500 events, and the number of common events is increased from 150 to 350.

Also shown in Figure 9 (solid curve) is the effect of migrating the positive kurtosis derivative $\dot{K}_+(t)$ instead of the kurtosis $K(t)$ itself. The improvement in the correspondence between the position of the maximum and the onset time illustrated in Figure 4d leads to a reduction in the origin-time bias from approximately 0.3 to 0.1 s. The remaining bias is consistent with the lag remaining between maxima in $\dot{K}_+(t)$ and the actual onset times. The accuracy of the three other hypocentral parameters is not affected significantly.

Figure 10a illustrates the effect of using the recursive calculation in equation (7) instead of the sliding window method in equation (1) for computing the kurtosis waveforms. The dotted and dashed curves were obtained using respectively the standard formulation with a 3 s sliding window and the recursive formulation with a 1 s time scale. Location accuracy is not affected by the method chosen; however, both the speed of computation and the number of detections are significantly increased with the recursive method, the first by a factor of up to 10, the second by 40%. The 200 extra events were individually verified, and all correspond to coherent signals. Their increased number may be due to the increased sensitivity the shorter window gives to the recursive method. We conclude that the recursive kurtosis calculation is to be preferred, as it leads to greater computational efficiency without loss of accuracy.

The kurtosis gradient waveforms are asymmetric, they ramp up quickly and decay over a longer time (Fig. 4d), which may cause skewness in the stacking process and dispersion in the coordinates at times close to the origin time of the event. We tested replacing the $\dot{K}_+(t)$ waveforms by a series of Gaussian functions centered on the local maxima, with half-width $\sigma = 0.1$ s (the approximate width of the peaks in $\dot{K}_+$). This procedure can be thought of as replacing discrete *P*-wave arrival picks by Gaussian distributions in time that have widths encapsulating the uncertainty in the pick times. The solid curve in Figure 10a illustrates the effect of using these Gaussian waveforms instead of $\dot{K}_+(t)$ in the migration process. First, the efficiency of Waveloc is improved as the number of events increases (approximately 50 additional events); second, the accuracy of the hypocenter coordinates is improved, especially in z (from 0.26 km to 0.09 km). The increase of the number of events could be explained by the fact that the Gaussian is two times larger than the usual characteristic function: the traces are then more likely to stack constructively.

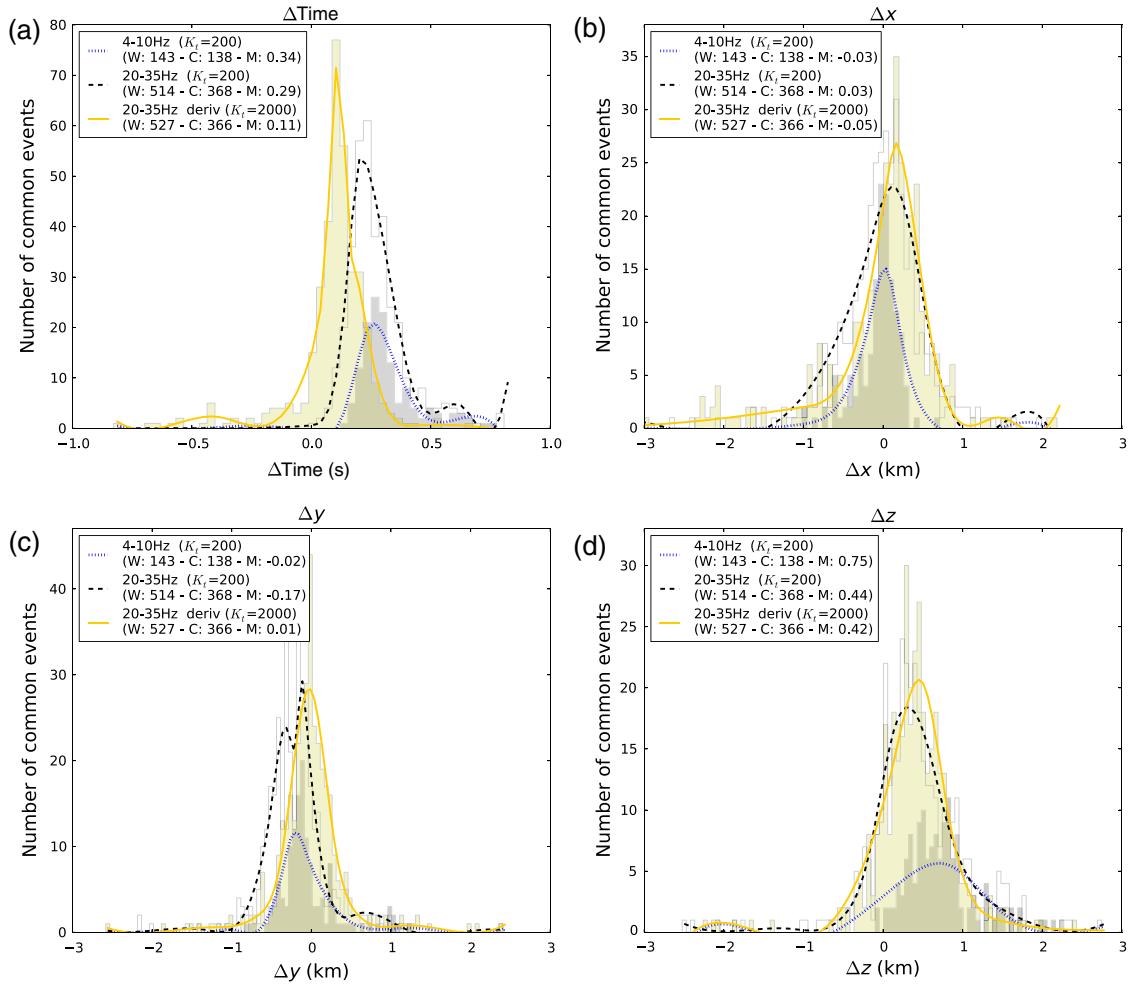


Figure 9. Histograms of common events showing the accuracy on hypocentral parameters and the influence of filtering and use of positive derivatives. Kurtosis computations use a 3 s window. Dark bars (dotted line) correspond to data filtered between 4 and 10 Hz; white bars (dashed line) to data filtered between 20 and 35 Hz; and light bars (solid line) to data filtered between 20 and 35 Hz and migrated with the positive derivatives of the kurtosis. For comparison purposes, the detection threshold ($K_t = 200$) is the same for the two first cases. W gives the total number of events located by Waveloc, C is the number of common events with the manual locations, and M is the mean error on the considered hypocentral parameter. The color version of this figure is available only in the electronic edition.

Finally, for the crisis of 14 October 2010, Waveloc recovered more than 85% of the manual events. The 15% of missed events can be classed into two categories: events that are well detected but are associated to a large number of low SNR traces (so the criterion on the minimum number of stations recording the event is not fulfilled) and, more rarely, events which are not detected because of an insufficient maximum stacking value (i.e., below the trigger threshold).

We collected the seismic catalogs provided by the OVPF (see Data and Resources) and compared them with our automatic results. The OVPF catalog is constructed using the automatic detection and picking tools from the Earthworm software suite (Johnson *et al.*, 1995), and the locations are carried out using hypo71 (Lee and Lahr, 1975). In most cases, Waveloc detects more events than the OVPF; however, the OVPF catalog for the 2 January 2010 swarm is missing, and some catalogs are incomplete. Similarly to our previous statistical analysis, we considered that detections from the two catalogs correspond to a single event when their origin time difference was smaller than 1 s. The detection rate varies between swarms: the number of detections for Waveloc is close to that for OVPF for some swarms and 3–4 times greater for others. In general, we observed that half to two-thirds of the detections for a given swarm are common between the

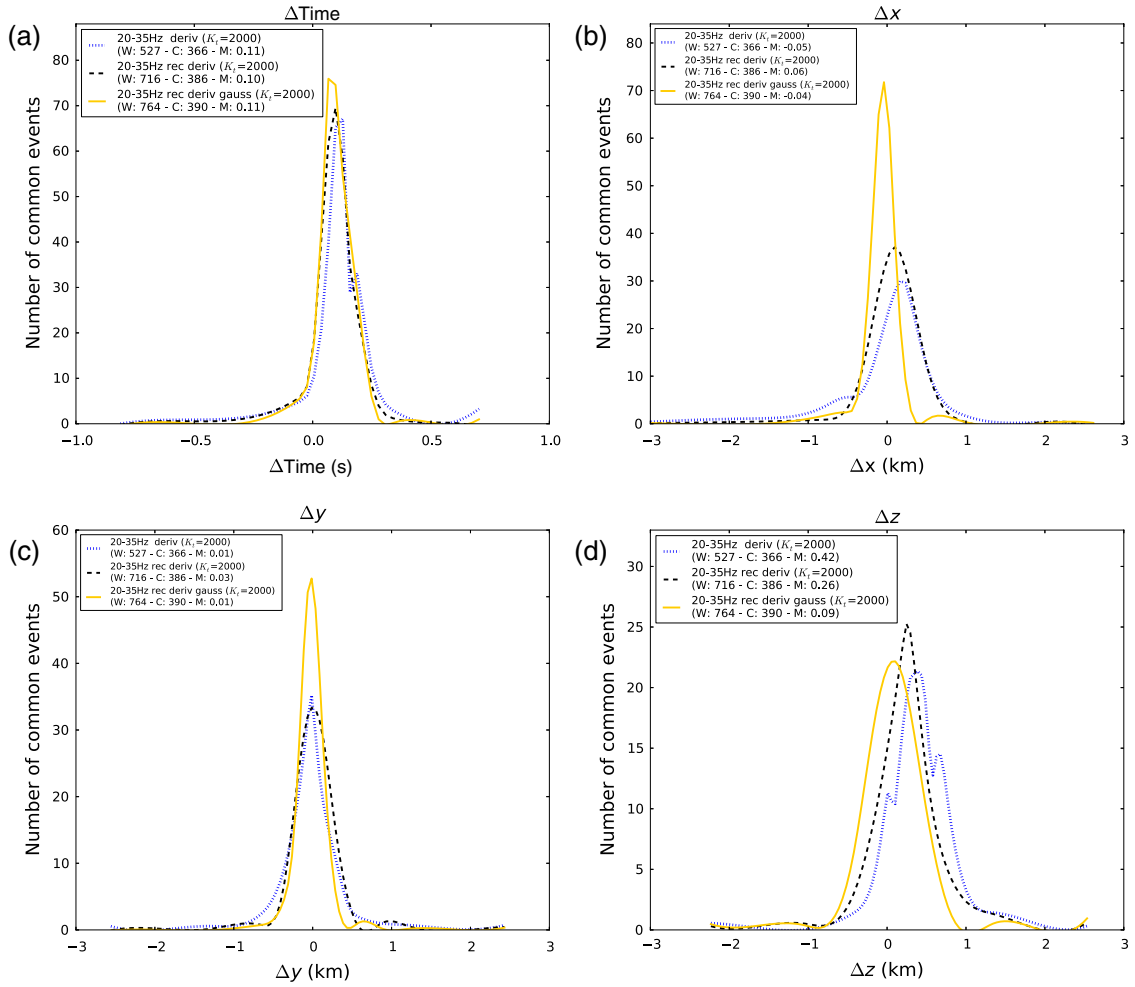


Figure 10. Histograms of common events showing the accuracy on hypocentral parameters and the influence of the recursive kurtosis and Gaussian convolution. All data are filtered between 20 and 35 Hz. Positive derivatives of kurtosis are used in all cases. The dotted line corresponds to nonrecursive kurtosis computed on a 3 s window; the dashed line to recursive kurtosis is computed with a 1 s time scale; and the solid line to recursive kurtosis is computed with a 1 s window and replaced by Gaussian distributions with an aperture of 0.1 s. The detection threshold is indicated by K_1. W gives the total number of events located by Waveloc, C is the number of common events with the manual locations, and M is the mean error on the considered hypocentral parameter. The color version of this figure is available only in the electronic edition.

Waveloc and OVPF catalogs, independently of the total number of detections. Of the 81 available locations in the OVPF catalog, 72 events were also located by Waveloc (taking a origin time difference of 2 s only added two common events). The comparison of their hypocentral parameters (see Fig. 11) shows that x and y coordinates are similar (differences do not exceed 200 m on average). The fits in t and z are not as good: the OVPF origin times tend to be later than ours, and the depths tend to be shallower. However, one should note that the origin times published by the OVPF catalogs are less precise than ours (to the nearest second only) which may partly explain our observation.

The OVPF seismic catalog also contains information on the event magnitudes. These duration magnitudes were computed on four summit stations (SNE, UV05, UV11, and UV15) using the Lee *et al.* (1972) formula, $M_D = -0.87 + 2\log\tau + 0.0035\Delta$, in which τ is the duration of the earthquake in seconds and Δ is the hypocentral distance in kilometers. We added a magnitude module in Waveloc that computes the local magnitude of the seismic events based on the Bakun and Joyner (1984) definition, given by $M_l = \log A + \log(\Delta/100) + 0.00301(\Delta - 100) + 3$, in which A is the zero-to-peak displacement amplitude as measured on a Wood–Anderson seismometer and Δ is the

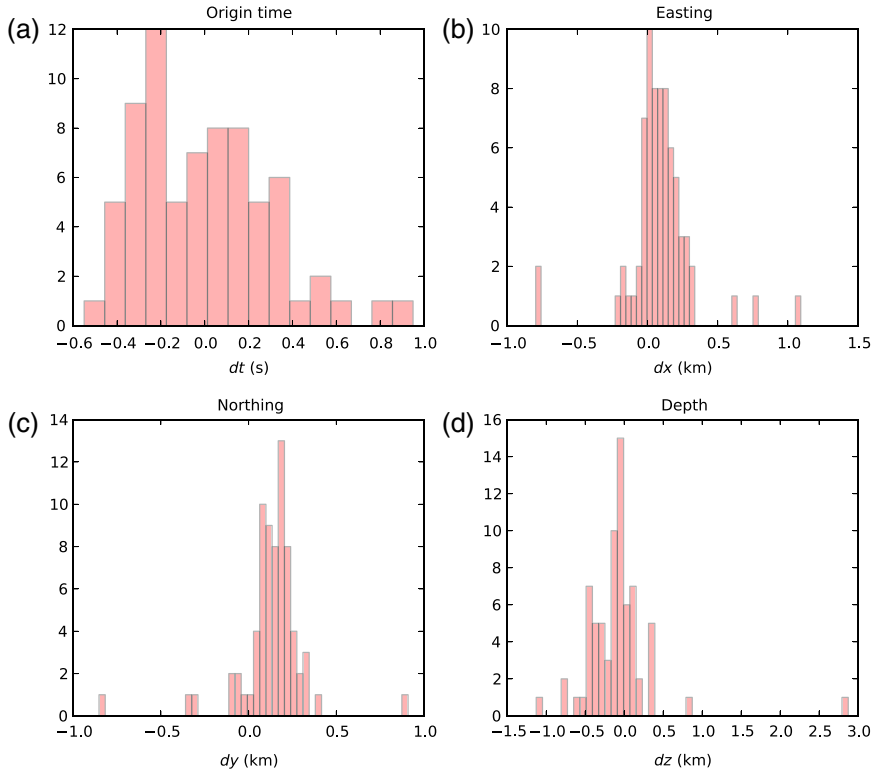


Figure 11. Comparison of Waveloc and the OVPF hypocentral parameters for the 72 locations provided by OVPF. The color version of this figure is available only in the electronic edition.

hypocentral distance. We use three-component data and work with peak-to-peak amplitudes measured within a 5.5 s window after the origin time. In order to compare our magnitudes with the OVPF ones, we also limit our computation to the four summit stations. Magnitudes for all swarms in both catalogs range from −1 to 2 (see Fig. 12). The numbers of events are comparable (only 500 additional events for Waveloc). The shapes of the two histograms are relatively similar, once the irregular binning of the OVPF magnitudes is taken into account, however Waveloc computes more negative magnitudes and proportionally lacks positive magnitudes close to 0. This explains why Waveloc's Gutenberg–Richter law is slightly curved whereas the one for OVPF forms a straight line. The *b*-value, obtained by simple linear fitting, equals 1 for Waveloc and 1.1 for the OVPF. There is no correlation between the magnitude and the stack amplitude, but there is a clear lower limit to the stack value possible for an event of a given magnitude (Ⓔ see Fig. S13 in the electronic supplement).

The statistical analysis developed in the previous paragraphs has allowed us to choose suitable parameters for locating earthquakes at Piton de la Fournaise: all results and seismicity plots presented in the next paragraphs are derived from the migration of the Gaussian waveforms obtained using the gradient of the recursive kurtosis computed with a 1 s time scale. Other parameters such as the location threshold were adjusted with respect to the amplitude of the stack, which may vary from one swarm to another, especially if the number of recording stations is not the same. We found that a threshold set to 100 times the number of stations was suitable, and we considered that each event should be detected at least on half of the stations to be correctly located.

Figure 13a represents the seismicity of the 14 October 2010 pre-eruptive crisis and shows a main seismic swarm located right under the volcano crater. Seismic events are mostly situated above sea level and do not reach the surface (they stop at about 1500 m), which is consistent with other studies (Battaglia and Brenguier, 2011). The orientation of the swarm at depth seems to be slightly north–south. By making a comparison with manual locations (Fig. 13b), we notice that Waveloc seismic events are much more scattered, which is very likely due to the higher number of events (about 300 additional events). This is confirmed when we strictly compare the manual and Waveloc locations of the common events (i.e., 390 events): both seismic swarms are similar (Ⓔ see Fig. S1 in the electronic supplement).

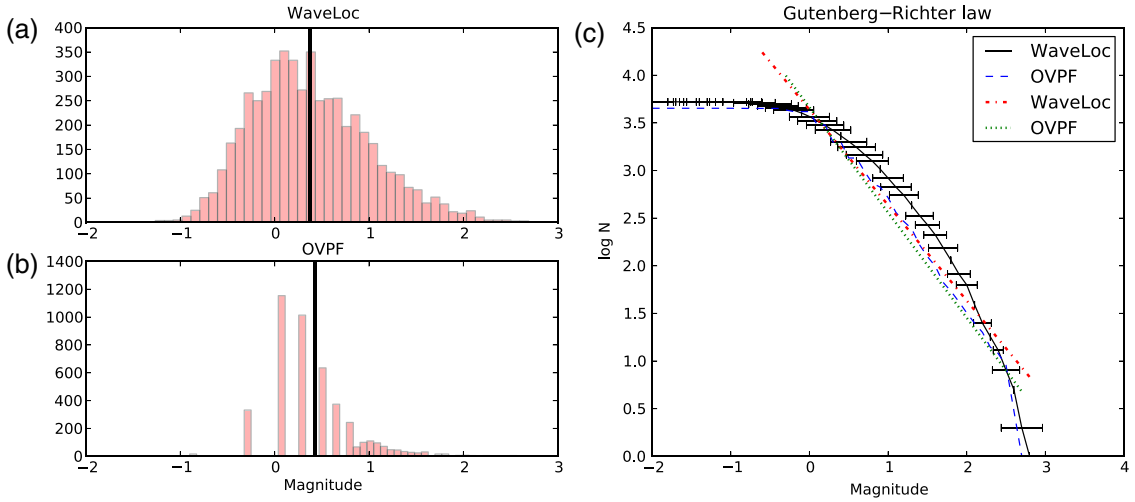


Figure 12. Magnitude distribution for all crises: (a) local magnitude computed from Waveloc located events; (b) duration magnitude computed by the OVPF (vertical bar represents the mean magnitude); and (c) Gutenberg–Richter laws. Horizontal bars represent the mean uncertainties we measured for each magnitude range. Linear regressions are also plotted (straight dotted lines). The color version of this figure is available only in the electronic edition.

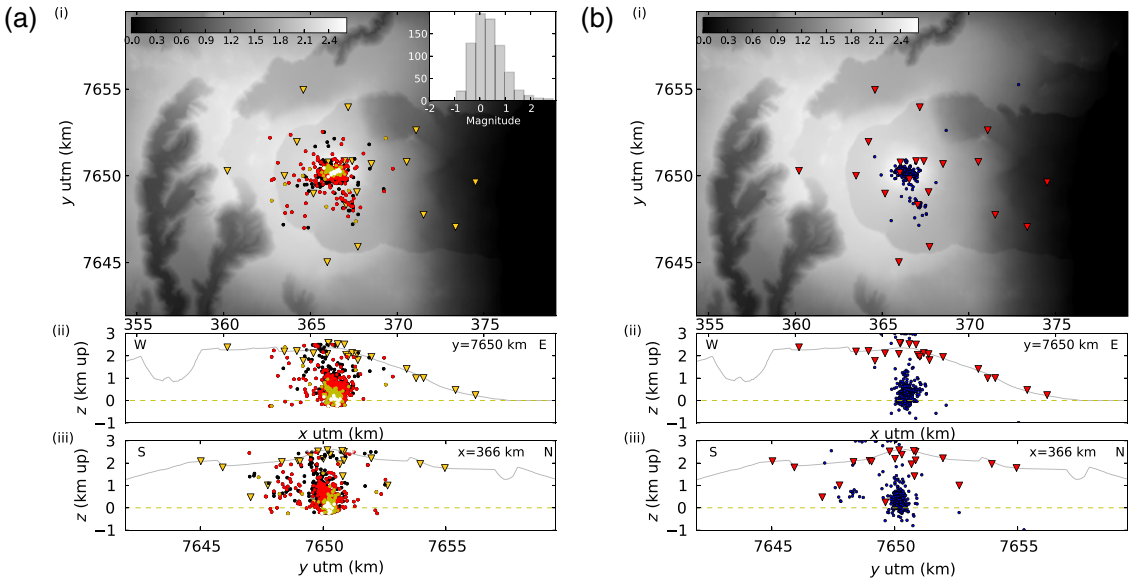


Figure 13. Location of the 14 October 2010 seismic events in the three Cartesian planes (i)–(iii) and comparison between (a) Waveloc locations and (b) manual locations. Stations are plotted as reversed triangles. Sea level is symbolized by the dashed line. Shades of gray on (i) indicate the altitude in kilometers. The shade of the events in (a) depends on their magnitude. Four ranges were defined, respectively from darker to lighter shades: $M_L < 0$, $0 \leq M_L < 1$, $1 \leq M_L < 2$, and $M_L \geq 2$. The magnitude distribution is shown by the inset plot on (i). The color version of this figure is available only in the electronic edition.

Figure 13a shows a repartition of the seismicity according to magnitude, with the larger events occurring at greater depth. Large events seem to be in the middle of the deeper part of the swarm, surrounded by events of lower magnitude.

Figure 14a is the seismicity plot of the pre-eruptive swarm of January 2010. Similarly to our previous observations, the seismic event depths range from 0 to 1 km above sea level and do not reach the volcano summit where the

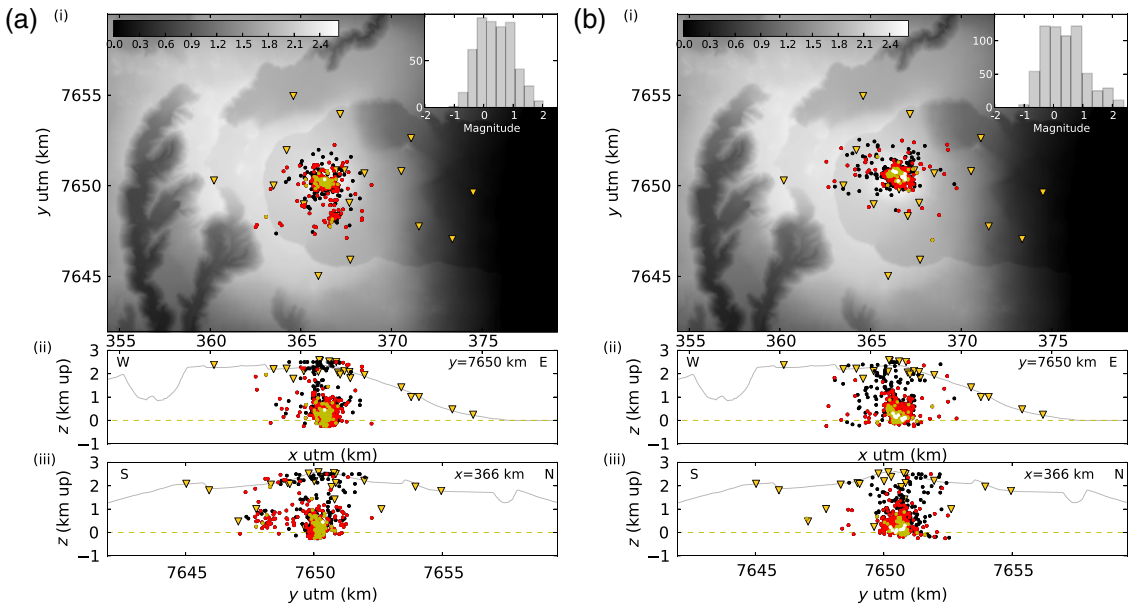


Figure 14. Waveloc locations for two swarms: (a) 02 January 2010 pre-eruptive swarm and (b) 23 September 2010 intrusive swarm. Shades and symbols are as in Figure 13a. The color version of this figure is available only in the electronic edition.

eruption took place. This would mean the dyke propagation just before an eruption is not (or only weakly) coupled with the seismicity. Figure 14b is the seismicity plot of the intrusive swarm of 23 September 2010. The plot does not show any clear difference with what is observed for pre-eruptive case, that is, the swarm is still situated just under the volcano, between 0 and 1 km above sea level. This suggests we cannot distinguish the two types of swarms so easily: a more detailed study of the seismicity must be performed for that purpose (Battaglia and Brenguier, 2011). The magnitude distribution for these two swarms shows the same pattern as before (i.e., a decreasing size when going up). Seismic events seem to be spatially distributed within the swarm: for instance, events with a magnitude greater than 2 are almost always located around $x = 366$ km, $y = 7650$ km, and $z = 0.1$ km.

Improving Waveloc Locations Using Correlation and Double-Difference Relocation

Waveloc gives us a first overview of the seismicity location during the numerous swarms that occurred at Piton de la Fournaise. However, the scattering of some events around the main seismic swarm indicates that better accuracy on locations could be obtained by relocating the events relative to each other. This is all the more important as it might resolve dike propagation or the seismogenic structures inside the volcano. The relocation procedure within Waveloc contains two steps: we cross correlate all events in order to create clusters of similar events and then relocate events within each cluster using a double-difference algorithm.

Two earthquakes are assumed to be similar when they occur at the same place with a similar focal mechanism. In practice, this implies they have similar waveforms. We look for events with similar waveforms by cross correlating seismic signals of 4.5 s length (0.5 s before and 4.0 s after the origin time, see Fig. 15a). Correlation values and time delays are first calculated in the time domain. When the correlation value is greater than a chosen threshold (taken to be 0.7 here), the cross correlation is also performed in the Fourier domain, which enables us to get a subsample precision on the time delay, the latter being deduced from the slope of the phase spectrum (Poupinet *et al.*, 1984).

The next step consists of creating clusters of events (Fig. 15b). We set two parameters: the minimum correlation value allowing a pair of events to be part of the same cluster (0.7) and the minimum number of stations for which this threshold has to be reached for a given event doublet (8). Clustering is performed using the depth-first search algorithm (Cormen *et al.*, 2001), which consists of exploring each possible path until it ends. This algorithm presents the advantage of being much faster than the more traditional breadth-first search algorithm. During this process, we brought to light some events with very similar waveforms that Waveloc initially located far from each other (Fig. 15c, dark symbols).

We relocate the events belonging to each cluster relative to each other by using a reimplementation of the double-difference algorithm (Waldhauser and Ellsworth, 2000), which hypothesizes the travel-time difference between two events recorded at a given station is due to their

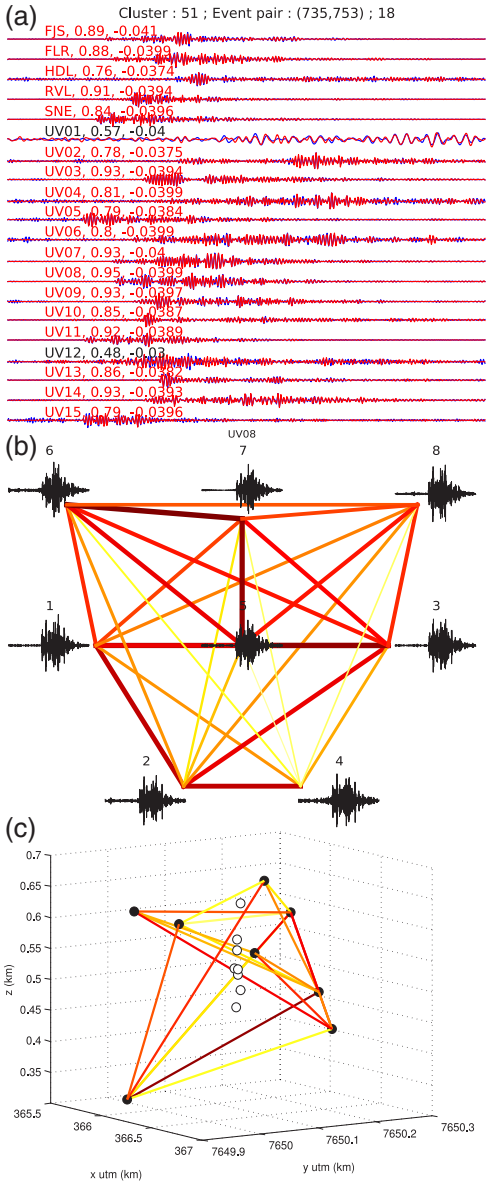


Figure 15. (a) Fine correlation between two events plotted on an 8 s window for all stations. Both events are superimposed. The correlation and delay values are indicated above each trace. The correlation value is greater than 0.7 at 18 stations. (b) Connectivity plot within a cluster of eight similar events at station UV08. Waveforms are 7 s long. The shade and thickness of the link depend on the correlation degree between two events: the darker and thicker it is, the better the correlation. Shade scale begins at 0.7. (c) The 3D plot represents the initial locations (dark points) and the new locations obtained by double difference (white points). The shade of the link between two events is a function of the number of stations where the correlation value is greater than 0.7. Shade scale begins at 8 and darkens when the correlation is strong. The color version of this figure is available only in the electronic edition.

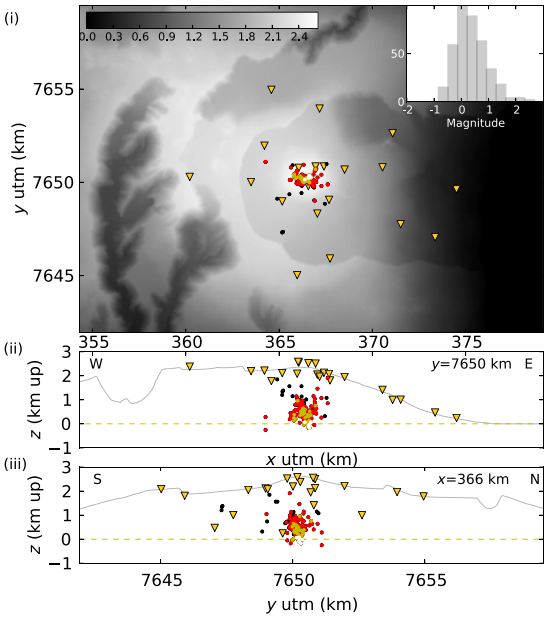


Figure 16. Relocated events for the 14 October 2010 crisis. Shades and symbols are as in Figure 13a. The color version of this figure is available only in the electronic edition.

spatial offset, provided the latter is small enough in comparison with the distance to the station.

Figure 15c (white symbols) shows an example of relocation for a small cluster: the events become closer to each other. Figure 16 shows the seismicity map after relocation for the 14 October 2010 swarm. Approximately 400 events were relocated in this manner. We note that the seismic swarm is much thinner and less scattered than previously (Fig. 13a), which highlights the usefulness of implementing relocation. The event depths are not modified and still range between 0 and 1 km.

Concerning the other swarms (Ⓔ see Figs. S2–S11 in the electronic supplement), we still observe that the seismicity concentrates between 0 and 1 km above sea level and under the volcano crater, but with no preferred orientation. In some cases, the number of events after relocation is not sufficient to detect any particular trend (in particular for the 14 October 2009 swarm). However, other swarms give interesting results, such as that of 09 December 2010, which is actually divided into one intrusive and one eruptive swarm. Although the main seismic swarm during the eruptive phase is at depth, the seismicity seems to go up and reach the surface (east–west section). Another intrusive swarm of interest is that of 29 October 2009, during which a small seismic swarm situated around $z = 1.5$ km and $y = 7650$ km is clearly identifiable.

Computational Requirements

Waveloc runs on a desktop computer with a good-quality Python distribution. It can straightforwardly be par-

allelized using embarrassingly parallel techniques and ported to computational clusters for dealing with large volumes of data. One of the main advantages of developing an automatic method for detecting and locating earthquakes is the time saving it permits. We have already seen that Waveloc is superior to the Earthworm + hypo71 implementation at OVPF in terms of number of events located. In the following, we give the computation time necessary to process the entire 14 October 2010 swarm (16 hr records from 20 stations with 100 Hz sampling and a grid of 9216 trial hypocenters) on a desktop computer with a 2.67 GHz CPU and 8 GB of RAM, using a serial (nonparallel) implementation of Waveloc. For the first step, data processing, the computation of nonrecursive kurtosis using a 3 s window takes 10 min/station, whereas the computation time of recursive kurtosis with a 1 s time scale is reduced to only 1 min/station. The migration step lasts approximately 3 hr. By comparison with other swarms, we estimate the migration of 5 hr of data with 100 Hz sampling takes about 1 hr. The duration of the location step varies with the chosen trigger threshold (smaller thresholds require longer location times). For the 14 October 2010 crisis with a location threshold set to 2000, which corresponds to almost 1200 detected events, the computation takes 8 min. The time required for correlation depends on the number of located events and on the number of stations: we estimate a cost of 2 min/station. Only a few minutes are necessary for clustering and double-difference relocation. In summary, Waveloc took 7 hr to process the whole of the 14 October 2010 swarm; this time was reduced to only 4.5 hr when using the recursive kurtosis estimation. In comparison, the time necessary for manual processing was estimated to 2–3 weeks (800 picked and 450 located events, whereas Waveloc detected around 1200 events and located 750 of them), which illustrates that Waveloc allies both robustness and speed.

Conclusion

We present Waveloc, an algorithm for automated detection and location of earthquakes based on the continuous migration of kurtosis gradient waveforms. Filtering parameters maximizing the kurtosis are determined by the kurtogram analysis and ensure the best enhancement of the beginning of an event, thereby increasing the chances to get a coherent stack after the migration step. We demonstrate Waveloc's reliability by applying it to real data recorded at Piton de la Fournaise volcano and by comparing the results with manual ones when possible. This comparison shows that Waveloc recovers most events (more than 85%) with a good accuracy on origin times (0.1 s) and other hypocentral parameters. This accuracy is due in part to the 3D geometry of the seismic network. The few events that are missed are attributed to the choice of the automatic parameters and are counterbalanced by the higher total number of locations. Local magnitudes are also computed and are coherent with duration magnitudes provided by the OVPF catalogs. Waveloc is also efficient with regards to the time it takes to process data: only 4–5 hr were necessary to process 16 hr of data recorded at 20 stations with a 100 Hz sampling, which obviously is much faster than manual processing, while maintaining a good accuracy.

Volcano seismicity is generally complex, with a large number of events occurring in a short period of time. Here, we prove Waveloc is well suited for such swarm analysis, mainly because the method is less affected by the density of events in time than automatic phase-picking methods, and in particular avoids the phase association step. However, we find that the separation of events close in time is still limited and may be improved with a more complex detection algorithm. Addition of a systematic cross correlation and double-difference relocation step further improves the location accuracy. Finally, our results clearly highlight a main seismic zone on Piton de la Fournaise that is systematically activated during swarms whether they are followed by eruptions or not. We also find that event magnitudes are linked to their spatial distribution, with size increasing with depth. These preliminary locations and magnitudes, provided by Waveloc, could be used for further, more detailed analysis.

First results obtained by Waveloc are encouraging. We are currently applying it to data coming from other study areas. We have demonstrated here the strong potential of the method for analyzing seismic swarms, which occur not only in volcanic contexts, but also in the fields of geothermal and hydrothermal exploitation and in aftershock sequences. Future developments of Waveloc include an application for real-time monitoring.

Data and Resources

The dataset used for the analysis was collected by the Institut de Physique du Globe de Paris, Observatoire Volcanologique du Piton de la Fournaise (IPGP/OVPF) and the Institut des Sciences de la Terre (ISTerre) within the framework of ANR-08-RISK-011/UnderVolc project. The sensors are properties of the réseau sismologique mobile français, Sismob (Institut National des Sciences de l'Univers–Centre National de la Recherche Scientifique [INSU-CNRS]). This work has been supported by Agence Nationale de la Recherche (ANR) (France) under contract ANR-08-RISK-011 (UnderVolc). More detailed information on data can be found in Brenguier *et al.* (2012). Data from the UnderVolc experiment will be freely available from www.resif.fr/portal (last accessed December 2013). The OVPF catalogs are available from http://volobsis.ipgp.fr (last accessed February 2013). The Waveloc code is open source, released under the CeCILL license, and can be downloaded from http://github.com/amaggi/waveloc (last accessed December 2013). Waveloc is written in Python, and was developed using the Enthought Python distribution under an academic license.

Acknowledgments

The development of the Waveloc software is actively supported by Network of European Research Infrastructures for Earthquake Risk Assess-

ment and Mitigation (NERA, EC Grant Number 262330) and the PYROPE project (ANR-09-0229-000). The application of Waveloc to Piton de la Fournaise seismicity was partly funded by Institut National des Sciences de l'Univers–Centre National de la Recherche Scientifique (INSU-CNRS) through the 2011 Risk program. N. Langet and A. Maggi benefitted from the numerous workshops organized by the UnderVolc project (ANR08-RISK011), and from discussions with volcano seismologists with expertise on Piton de la Fournaise, most notably Jean Battaglia.

References

Allen, R. V. (1982). Automatic phase pickers: Their present use and future prospects, *Bull. Seismol. Soc. Am.* **72,** no. 6B, S225–242.

Antoni, J. (2007). Fast computation of the kurtogram for the detection of transient faults, *Mech. Syst. Signal Process.* **21,** no. 1, 108–124.

Baker, T., R. Granat, and R. W. Clayton (2005). Real-time earthquake location using Kirchhoff reconstruction, *Bull. Seismol. Soc. Am.* **95,** no. 2, 699–707.

Bakun, W., and W. Joyner (1984). The Ml scale in central California, *Bull. Seismol. Soc. Am.* **74,** 1827–1843.

Battaglia, J., and F. Brenguier (2011). Seismogenic structures activated during the pre-eruptive and intrusive swarms of Piton de la Fournaise volcano (La Réunion island) between 2008 and 2011, *AGU Fall Meeting 2011*, San Francisco, California, 5–9 December 2011.

Brenguier, F., P. Kowalski, T. Staudacher, V. Ferrazzini, F. Lauret, P. Boissier, P. Catherine, A. Lemarchand, C. Pequegnat, O. Meric, C. Pardo, A. Peltier, S. Tait, N. M. Shapiro, M. Campillo, and A. Di Muro (2012). First results from the UnderVolc high resolution seismic and GPS network deployed on Piton de la Fournaise Volcano, *Seismol. Res. Lett.* **83,** no. 1, 97–102.

Chassande-Mottin, E. (2003). Testing the normality of the gravitational wave data with a low cost recursive estimate of the kurtosis, in *Proc. 3rd workshop on Physics in Signal and Image Processing 2003*, Grenoble, France, 157–160.

Cormen, T. H., C. E. Leiserson, R. L. Rivest, and S. Clifford (2001). *Introduction to Algorithms*, Second Edition, MIT Press and McGraw-Hill, Cambridge, Massachusetts.

Ekstrom, G. (2006). Global detection and location of seismic sources by using surface waves, *Bull. Seismol. Soc. Am.* **96,** no. 4A, 1201–1212.

Ekstrom, G., M. Nettles, and G. A. M. Abers (2003). Glacial earthquakes, *Science* **302,** no. 5645, 622–624.

Gajewski, D., and E. Tessmer (2005). Reverse modelling for seismic event characterization, *Geophys. J. Int.* **163,** no. 1, 276–284.

Gentili, S., and A. Michelini (2006). Automatic picking of *P* and *S* phases using a neural tree, *J. Seismol.* **10,** no. 1, 39–63.

Gharti, H. N., V. Oye, M. Roth, and D. Kuhn (2010). Automated microearthquake location using envelope stacking and robust global optimization, *Geophysics* **75,** no. 4, MA27–MA46.

Hanka, W., J. Saul, B. Weber, J. Becker, P. Harjadi, Fauzi, and Gitews Seismology Group (2010). Real-time earthquake monitoring for tsunami warning in the Indian Ocean and beyond, *Nat. Hazards Earth Syst. Sci.* **10,** no. 12, 2611–2622.

Johnson, C., A. Bittenbinder, B. Bogaert, L. Dietz, and W. Kohler (1995). Earthworm: A flexible approach to seismic network monitoring, *IRIS Newsletter* **14,** 1–4.

Johnson, C., A. Lindh, and B. Hirshorn (1994). Robust regional phase association, *Tech. Rept.* 94–621.

Kao, H., and S.-J. Shan (2004). The source-scanning algorithm: Mapping the distribution of seismic sources in time and space, *Geophys. J. Int.* **157,** no. 2, 589–594.

Kuperkoch, L., T. Meier, J. Lee, W. Friederich, and EGELADOS Working Group (2010). Automated determination of p-phase arrival times at regional and local distances using higher order statistics, *Geophys. J. Int.* **181,** no. 2, 1159–1170.

Larmat, C., J.-P. Montagner, M. Fink, Y. Capdeville, A. Tourin, and E. Clévédé (2006). Time-reversal imaging of seismic sources and application to the great Sumatra earthquake, *Geophys. Res. Lett.* **33,** doi: 10.1029/2006GL026336.

Lee, W. H. K., and J. Lahr (1975). Hypo71: A computer program for determining hypocenter, magnitude and first motion pattern of local earthquakes, *U.S. Geol. Surv. Open-File Rept. 75-1311*, 64 pp.

Lee, W. H. K., and S. W. Stewart (1981). *Principles and Applications of Microearthquake Networks*, Advances in Geophysics, Supplement 2, Academic press, New York.

Lee, W. H. K., R. Bennett, and K. Meagher (1972). A method of estimating magnitude of local earthquakes from signal duration, *U.S. Geol. Surv. Open-File Rept.*, 29 pp.

Liao, Y., H. Kao, A. Rosenberger, S. Hsu, and B. Huang (2012). Delineating complex spatiotemporal distribution of earthquake aftershocks: An improved source-scanning algorithm, *Geophys. J. Int.* **189,** 1753–1770.

Lomax, A. (2011). The NonLinLoc software guide, http://alomax.free.fr/nlloc/ (last accessed December 2013).

Maggi, A., and A. Michelini (2009). Continuous waveform data stream analysis: Detection and location of the L'Aquila earthquake sequence (abstract U12A-01), *Eos Trans. AGU* **90,** no. 52 (Fall Meet. Suppl.), U12A–01.

Maggi, A., and A. Michelini (2010). Waveloc—An algorithm for the detection and location of seismic sources within large, continuous waveform data volumes: The case of the l'aquila earthquake sequence, *Geophys. Res. Abstr*, EGU, Vienna, Austria, 2–7 May 2010.

Maggi, A., C. Tape, M. Chen, D. Chao, and J. Tromp (2009). An automated time-window selection algorithm for seismic tomography, *Geophys. J. Int.* **178,** no. 1, 257–281.

McMechan, G. A., J. H. Luetgert, and H. M. Mooney (1985). Imaging of earthquake sources in Long Valley Caldera, California, 1983, *Bull. Seismol. Soc. Am.* **75,** no. 4, 1005–1020.

Michelini, A., and A. J. Lomax (2004). The effect of velocity structure errors on double-difference earthquake location, *Geophys. Res. Lett.* **31,** no. 9, 4.

Podvin, P., and I. Lecomte (1991). Finite difference computation of traveltimes in very contrasted velocity models: A massively parallel approach and its associated tools, *Geophys. J. Int.* **105,** 271–284.

Poupinet, G., W. Ellsworth, and J. Fréchet (1984). Monitoring velocity variations in the crust using earthquake doublets: An application to the Calaveras Fault, California, *J. Geophys. Res.* **89,** 5719–5731.

Prono, E., J. Battaglia, V. Monteiller, J. L. Got, and V. Ferrazzini (2009). *P*-wave velocity structure of Piton de la Fournaise volcano deduced from seismic data recorded between 1996 and 1999, *J. Volcanol. Geoth. Res.* **184,** no. 1, 49–62.

Rietbrock, A., and F. Scherbaum (1994). Acoustic imaging of earthquake sources from the Chalfont valley, 1986, aftershock series, *Geophys. J. Int.* **119,** no. 1, 260–268.

Saragiotis, C. D., L. J. Hadjileontiadis, and S. Panas (2002). Pai–s/k: A robust automatic seismic *P*-phase arrival identification scheme, *IEEE Trans. Geosci. Remote Sens.* **40,** no. 6, 1395–1404.

Schaff, D. P., and F. Waldhauser (2005). Waveform cross-correlation-based differential travel-time measurements at the Northern California seismic network, *Bull. Seismol. Soc. Am.* **95,** no. 6, 2446–2461.

Schmid, A. (2011). *Quelle prédictibilité pour les éruptions volcaniques? De l'échelle mondiale au Piton de la Fournaise*, *Ph.D. Thesis*, Université de Grenoble, Grenoble, France.

Shearer, P. M. (1994). Global seismic event detection using a matched filter on long-period seismograms, *J. Geophys. Res.* **99,** no. B7, 13713–13725.

Waldhauser, F., and W. Ellsworth (2000). A double-difference earthquake location algorithm: Method and application to the northern Hayward fault, California, *Bull. Seismol. Soc. Am.* **90,** no. 6, 1353–1368.

Withers, M., R. Aster, and C. Young (1999). An automated local and regional seismic event detection and location system using waveform correlation, *Bull. Seismol. Soc. Am.* **89,** no. 3, 657–669.

Withers, M., R. Aster, C. Young, J. Beiriger, M. Harris, S. Moore, and J. Trujillo (1998). A comparison of select trigger algorithms for

automated global seismic phase and event detection, *Bull. Seismol. Soc. Am.* **88,** no. 1, 95–106.

Young, C., M. Harris, J. Beiriger, S. Moore, J. Trujillo, M. Withers, and R. Aster (1996). The waveform correlation event detection system project phase 1: Issues in prototype development and testing, *Tech. Rept. SAND96-1916*.

Institut de Physique du Globe de Strasbourg
CNRS et Université de Strasbourg (EOST)
5, Rue René Descartes
67084 Strasbourg cedex
France
(N.L., A.M.)

Istituto Nazionale di Geofisica e Vulcanologia
Via di Vigna Murata, 605
00143 Roma, Italy
(A.M.)

Institut des Sciences de la Terre
CNRS et Université de Grenoble
ISTerre–BP 53
38041 Grenoble cedex 9
France
(F.B.)

Manuscript received 2 May 2013;
Published Online 21 January 2014

A2 Cartes de sismicité des crises du Piton de la Fournaise (2009-2011)

Dans le chapitre II.2 de ce manuscrit, on a détaillé l'ensemble de la procédure ayant permis d'établir les cartes de sismicité des crises du 14 octobre et du 23 septembre 2010. Les résultats des 10 autres crises sont présentés dans cette annexe. Chaque figure montre les localisations dans les 3 plans cartésiens ((i)-(iii)) : (i) est une projection de la sismicité sur le plan x-y. La coloration en niveaux de gris donne l'altitude en (km). (ii) est une coupe Ouest-Est. (iii) est une coupe Sud-Nord. Les triangles jaunes renversés symbolisent le réseau de stations. La ligne pointillée sur les coupes (ii) et (iii) correspond au niveau de la mer.

L'échelle de couleurs utilisée pour représenter les événements en (a) dépend de l'intervalle de magnitude auquel ils appartiennent : noir pour $M_L < 0$, rouge pour $0 \leq M_L < 1$, jaune pour $1 \leq M_L < 2$ et blanc pour $M_L \geq 2$. L'histogramme de répartition des magnitudes est également affiché dans l'encart en haut à droite.

Chaque paire de figures correspond aux localisations automatiques de Waveloc "brutes" (a), puis issues de la relocalisation par double-différence (b).

On rappelle également que le terme *pré-éruptif* est utilisé pour une crise sismique suivie d'une éruption, au contraire d'une crise dite *intrusive*.

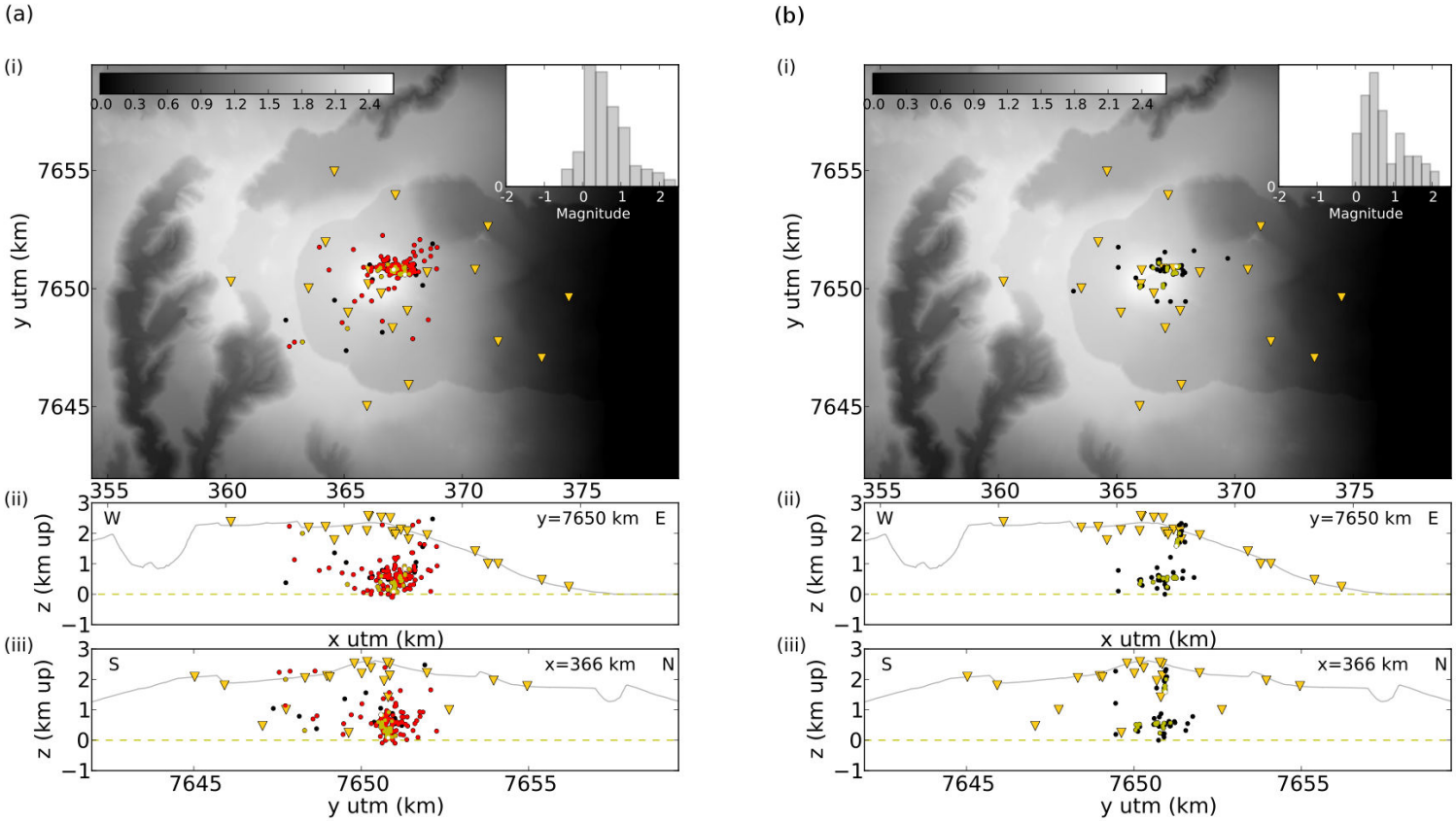


FIG. A2.1: Cartes de sismicité obtenues pour la crise intrusive du 14 octobre 2009.

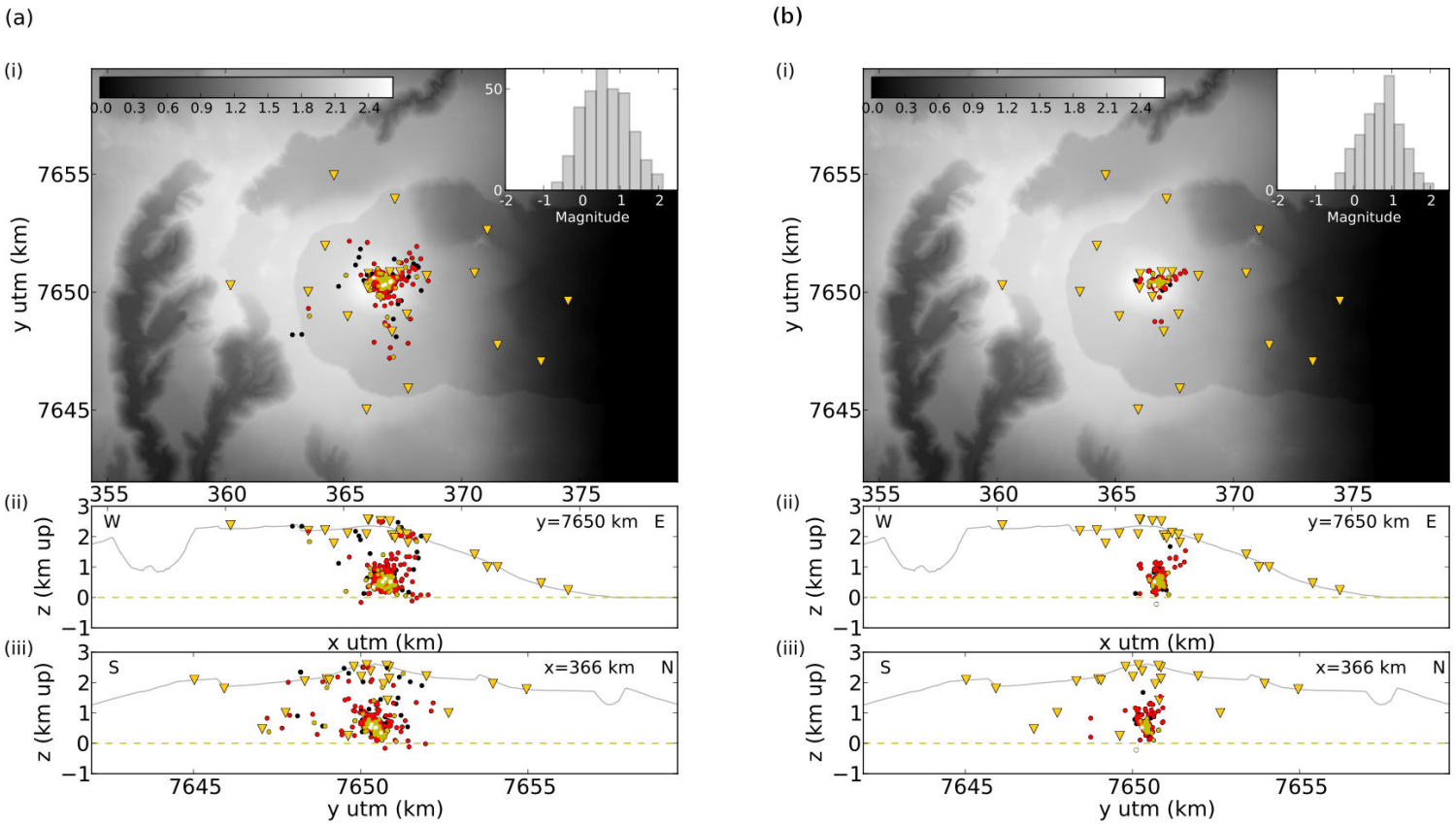


FIG. A2.2: Cartes de sismicité obtenues pour la crise intrusive du 18 octobre 2009.

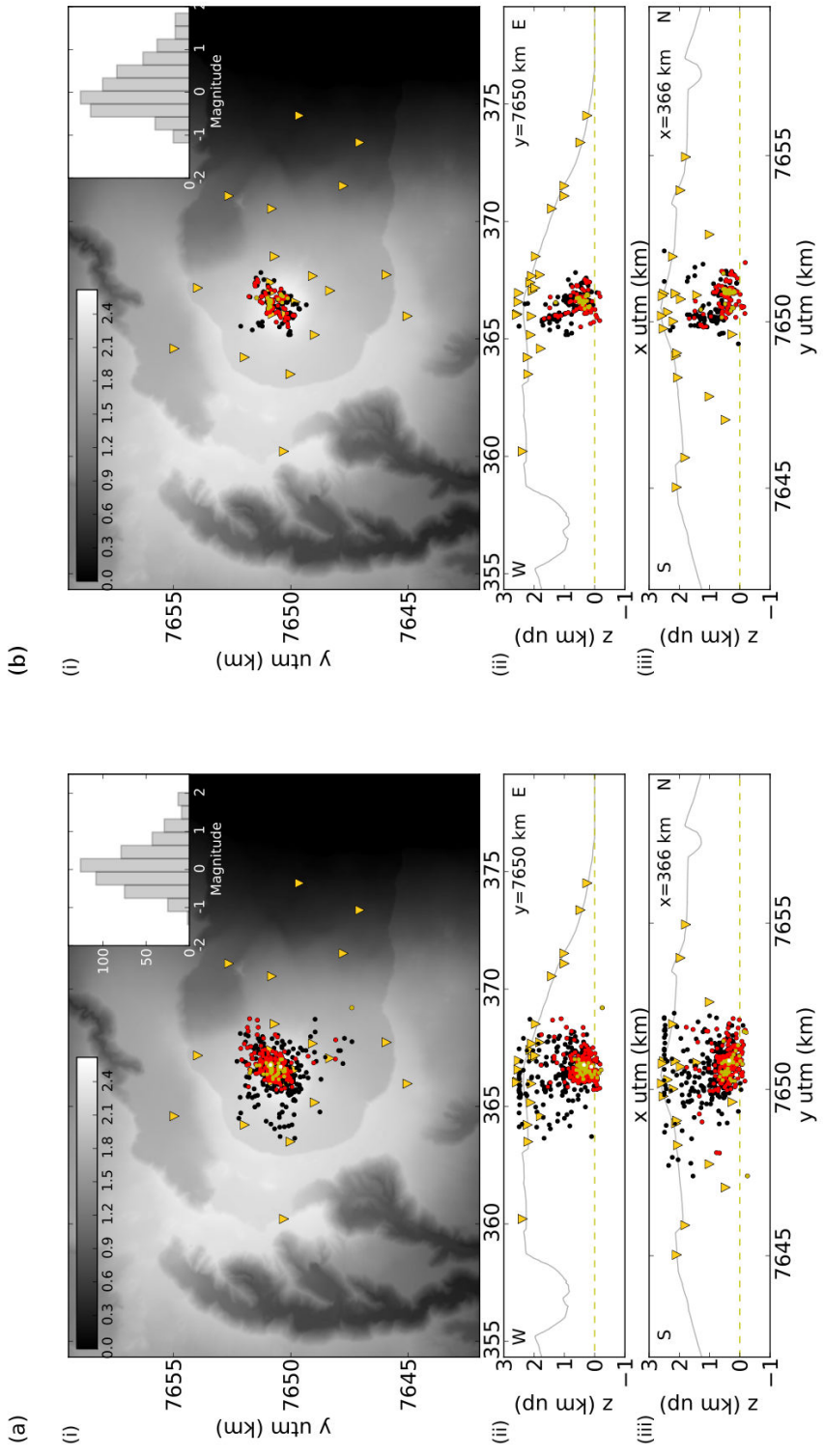


FIG. A2.3: Cartes de sismicité obtenues pour la crise intrusive du 29 octobre 2009.

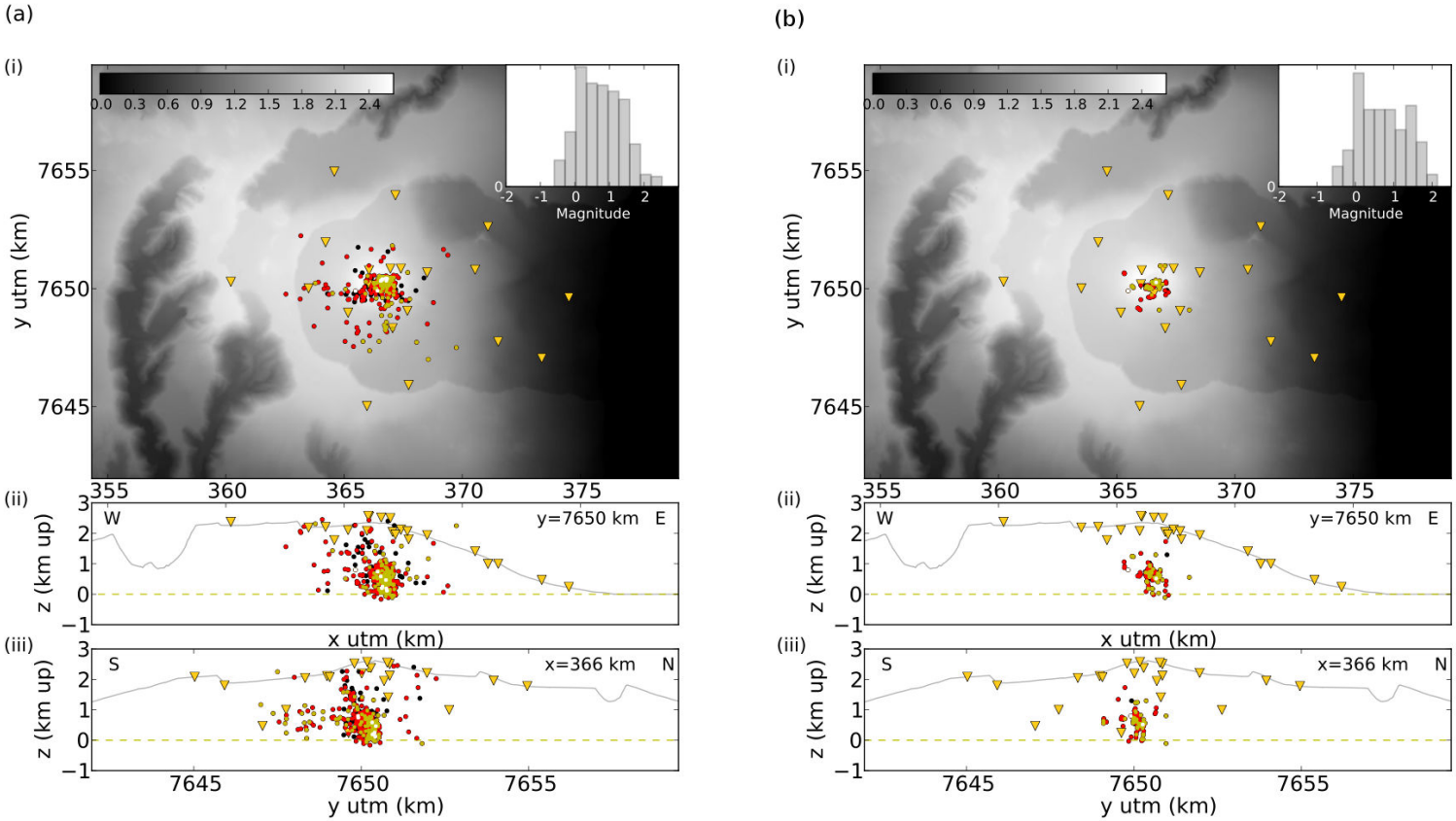


FIG. A2.4: Cartes de sismicité obtenues pour la crise pré-éruptive du 5 novembre 2009.

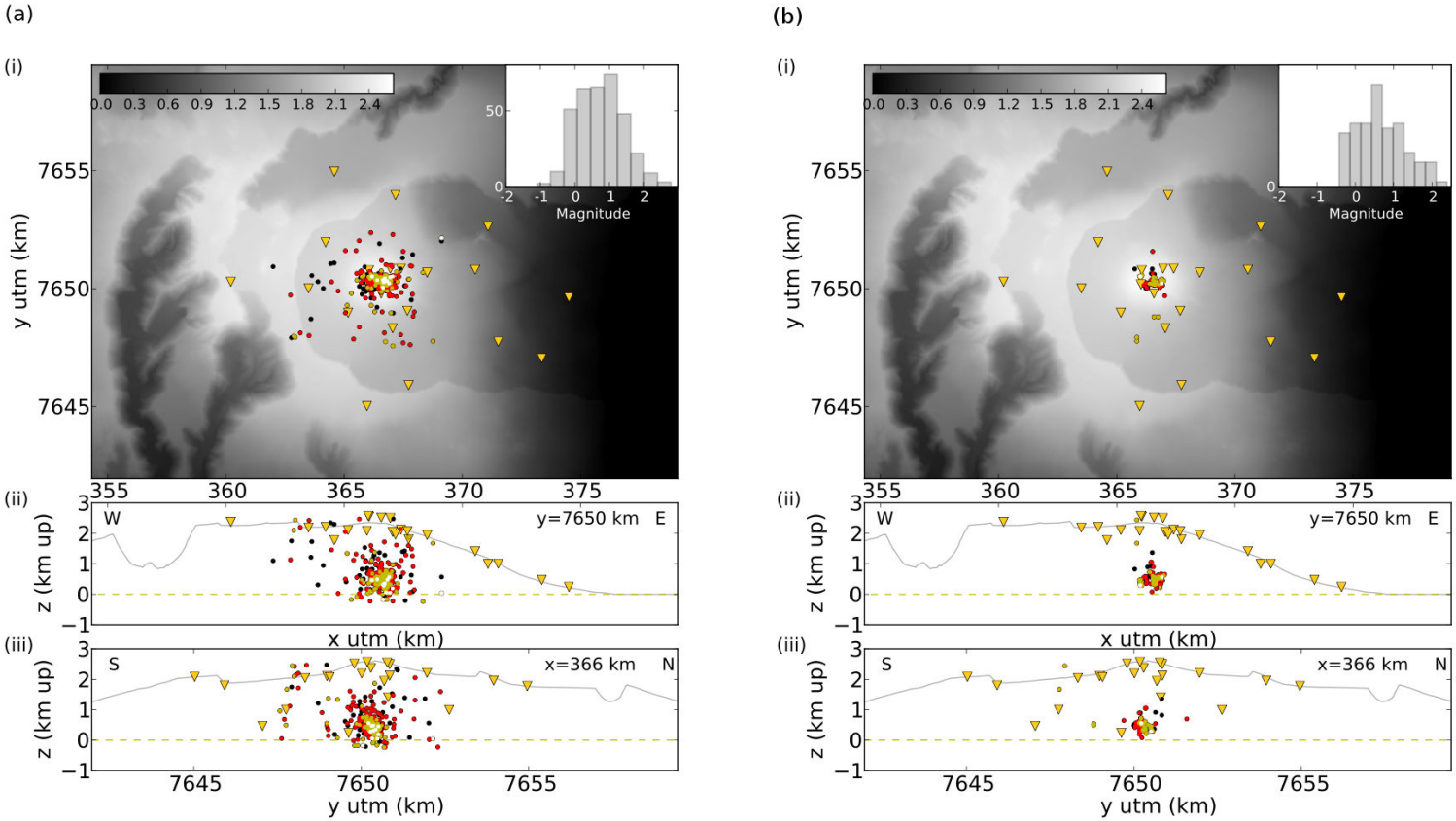


FIG. A2.5: Cartes de sismicité obtenues pour la crise pré-éruptive du 14 décembre 2009.

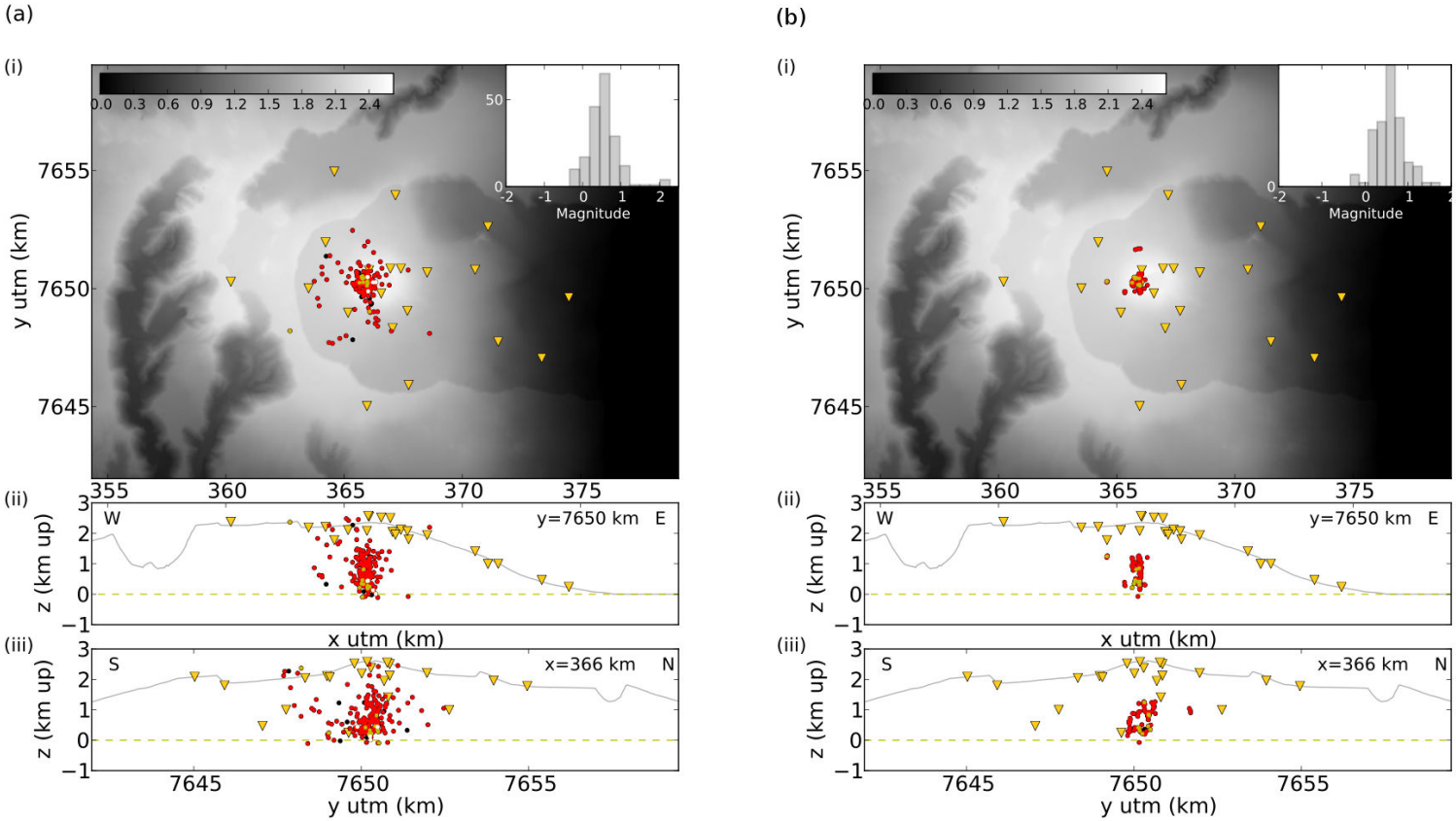


FIG. A2.6: Cartes de sismicité obtenues pour la crise intrusive du 29 décembre 2009.

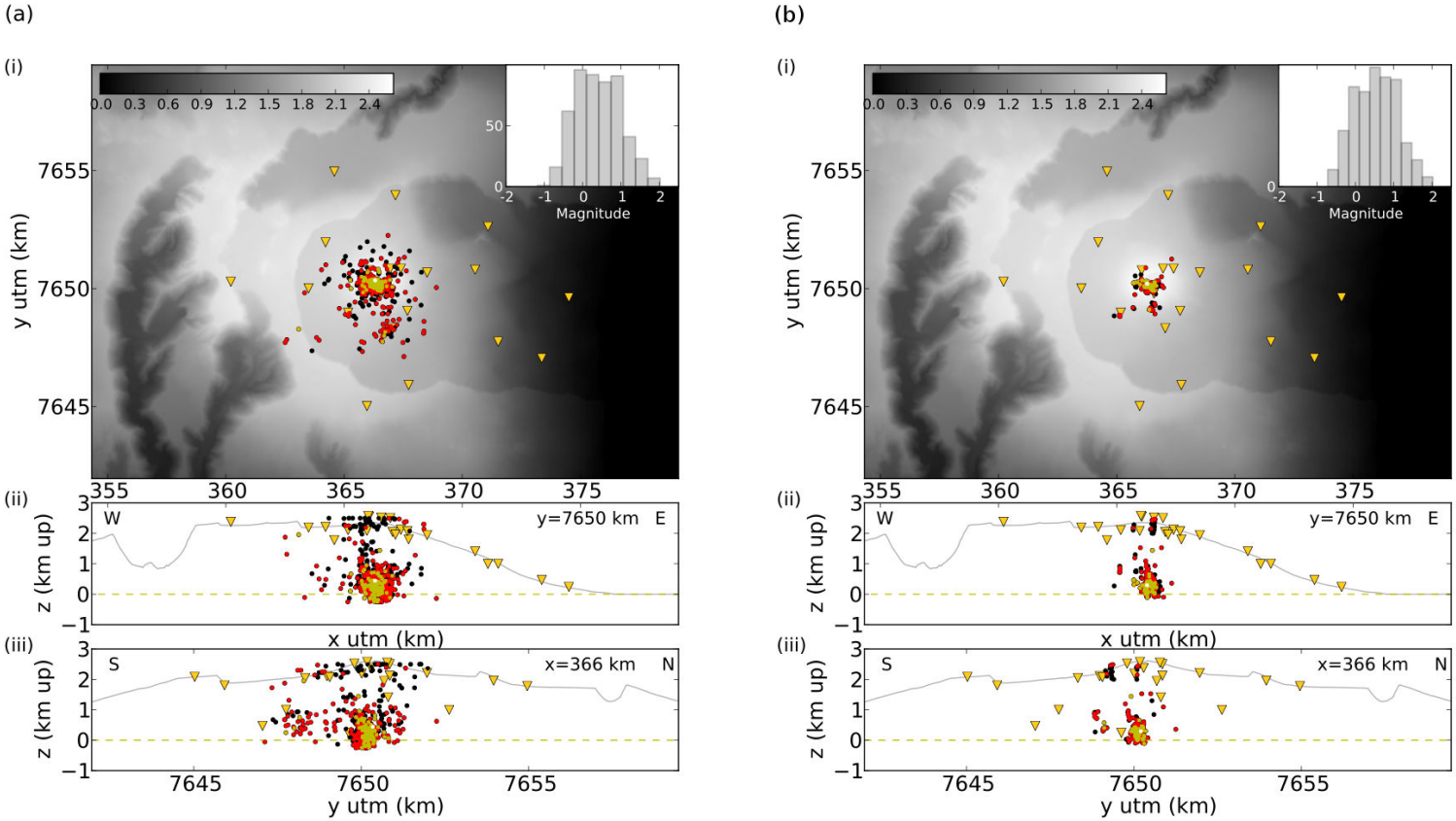


FIG. A2.7: Cartes de sismicité obtenues pour la crise pré-éruptive du 2 janvier 2010.

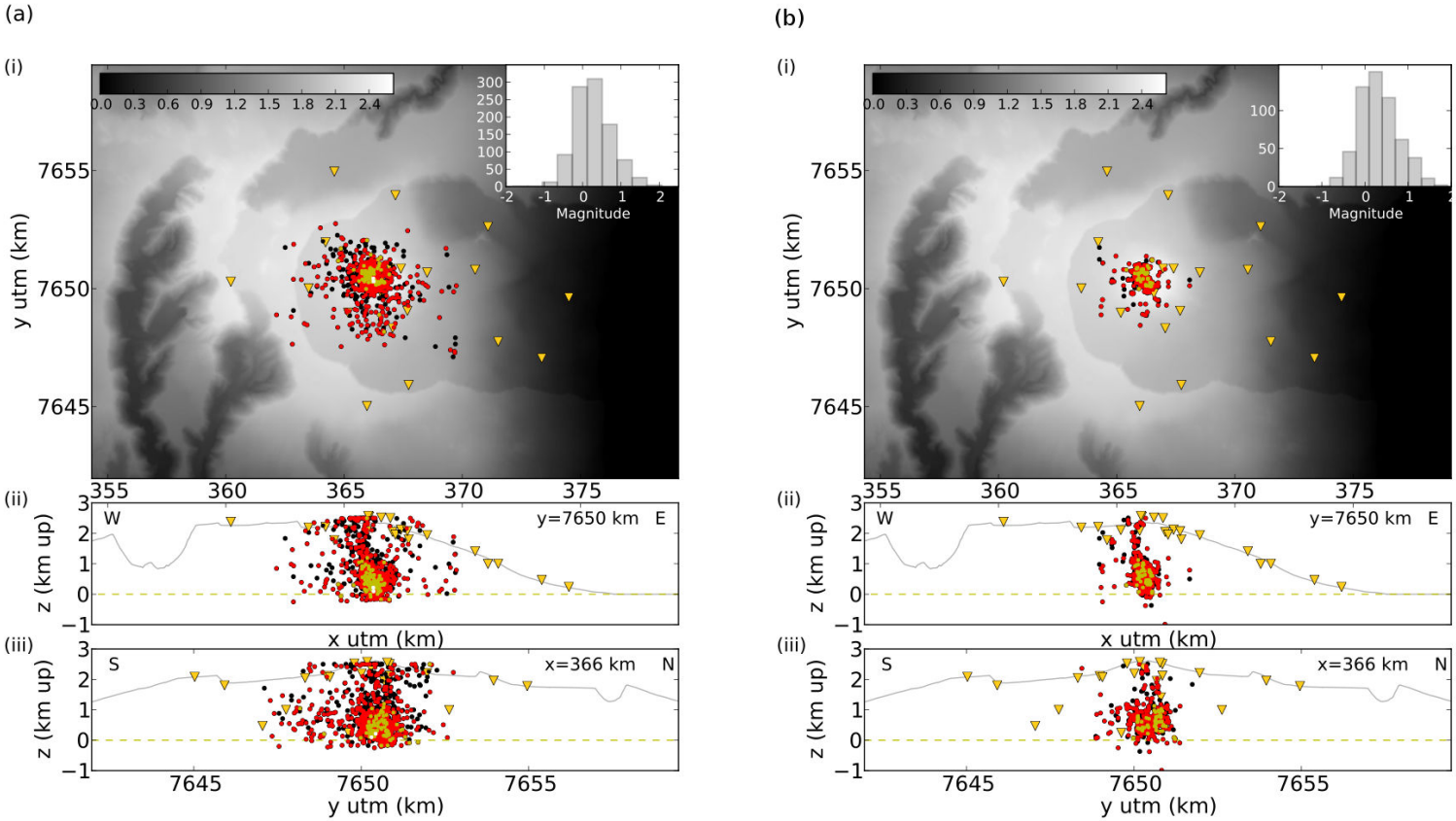


FIG. A2.8: Cartes de sismicité obtenues pour la crise du 9 décembre 2010. Cette crise se décompose en deux crises : une crise intrusive (bleu), suivie d'une crise pré-éruptive (jaune).

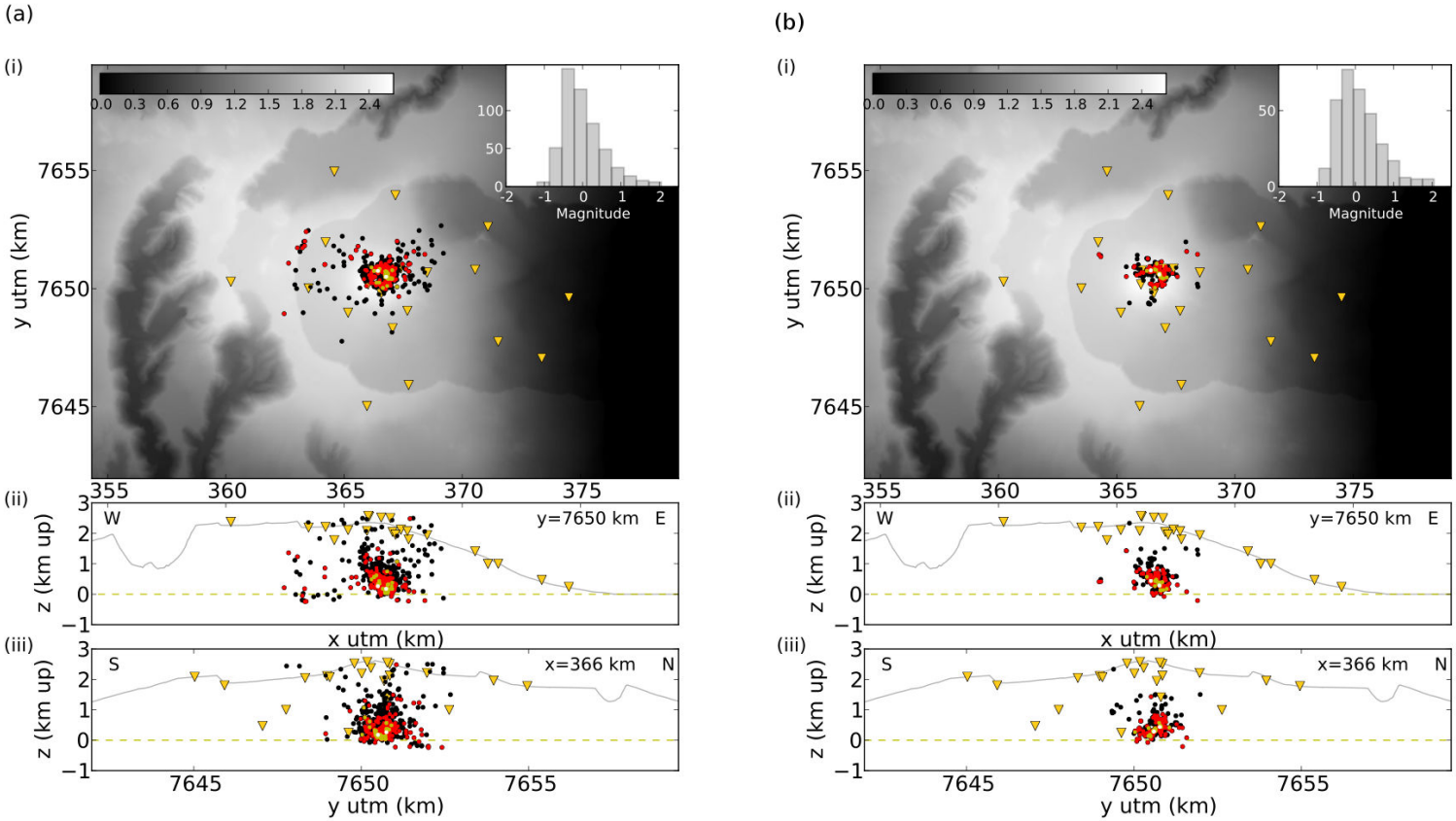


FIG. A2.9: Cartes de sismicité obtenues pour la crise intrusive du 2 février 2011.

A3 Tests de résolution avec Waveloc

Les tests de résolution sont menés de la manière suivante :

- pour chaque point de la grille définie dans l'espace d'étude, on crée des traces synthétiques non bruitées telles qu'elles seraient enregistrées à chacune des stations du réseau. L'événement est simulé par un pulse triangulaire de largeur 0.1 s ;

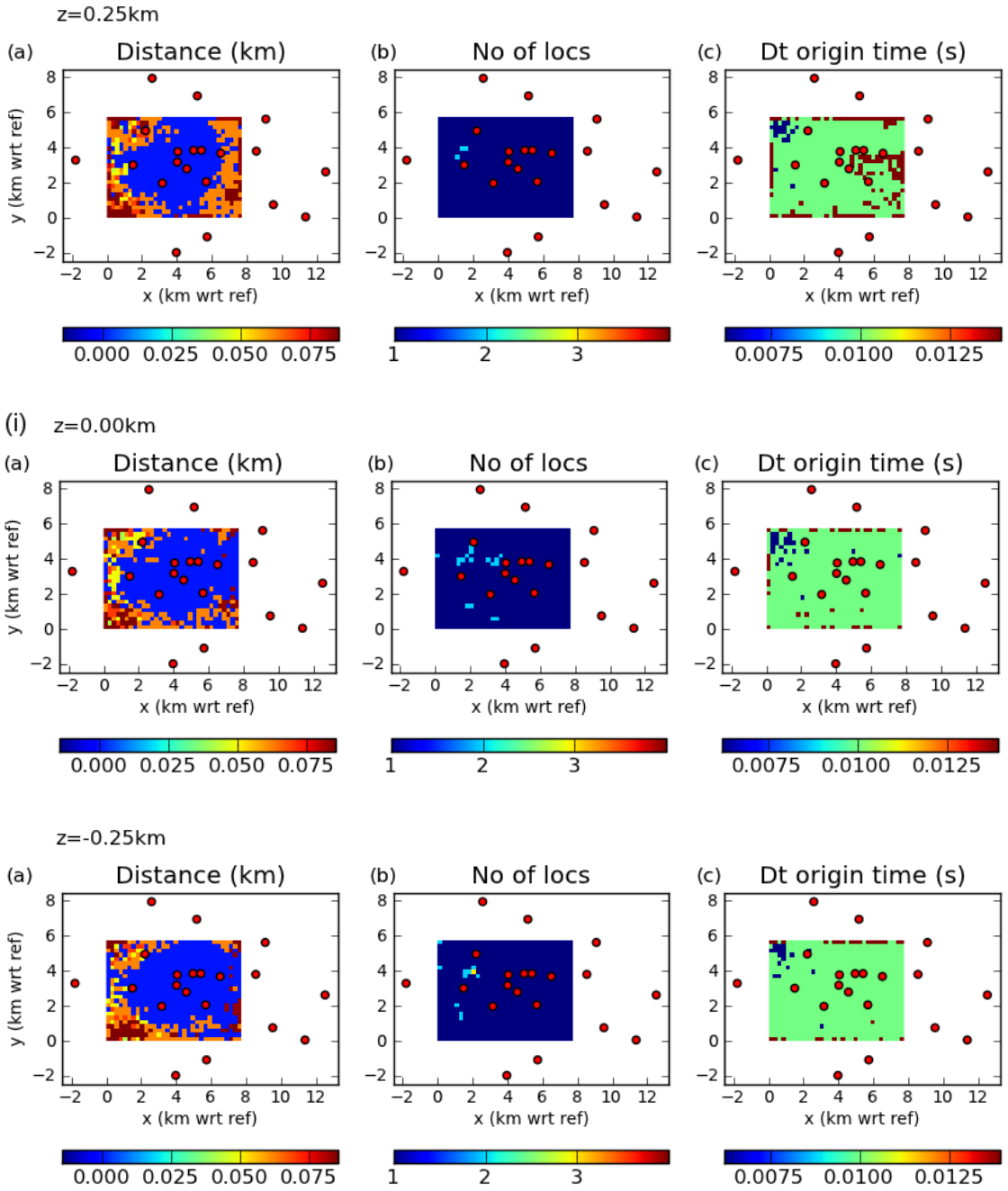
- on applique ensuite l'algorithme de Waveloc en utilisant comme fonction caractéristique les traces synthétiques créées ;

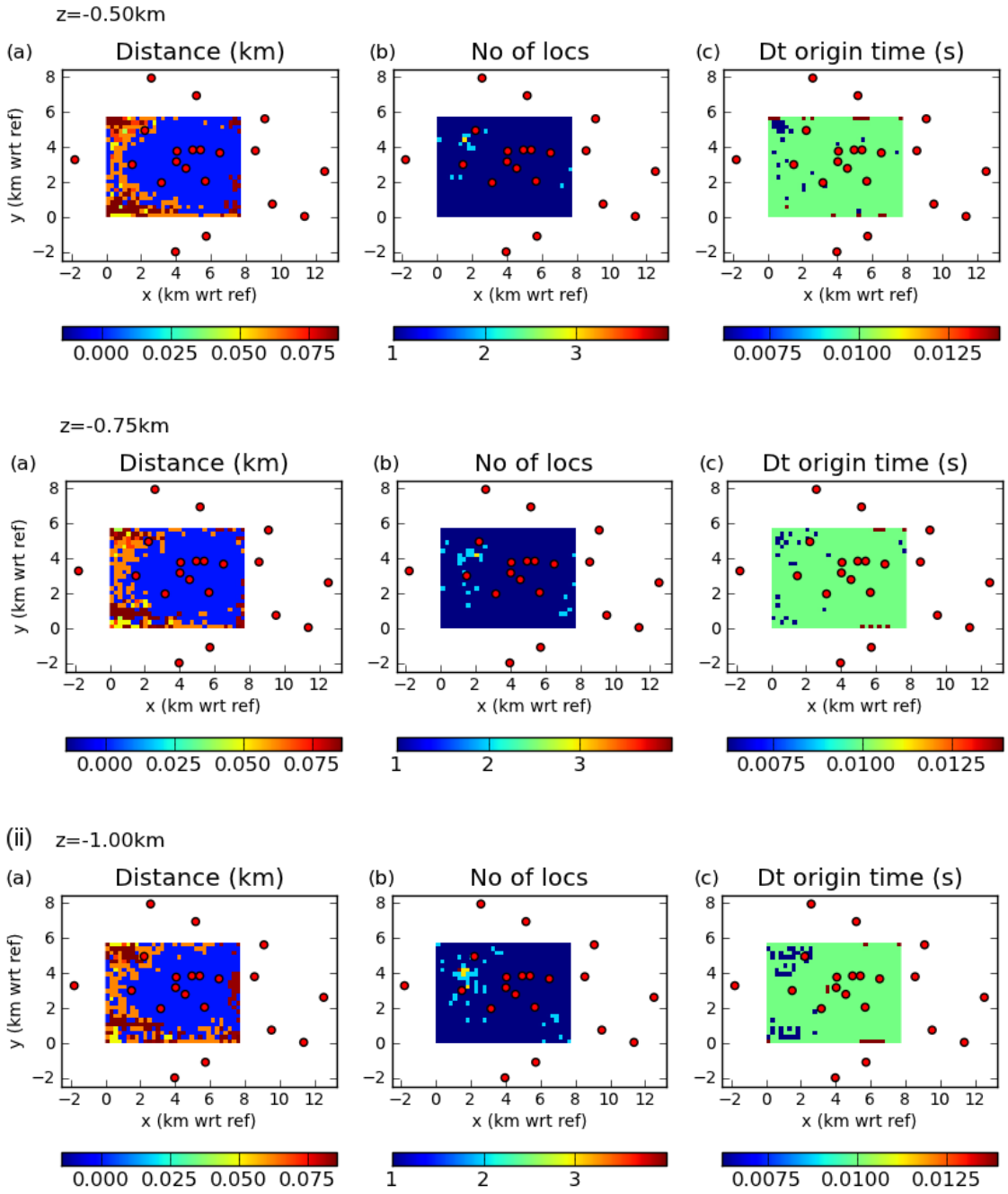
- puis on calcule la distance qui sépare la localisation trouvée par Waveloc de la localisation « vraie », le nombre de localisations faites par Waveloc, et l'écart sur les temps origine.

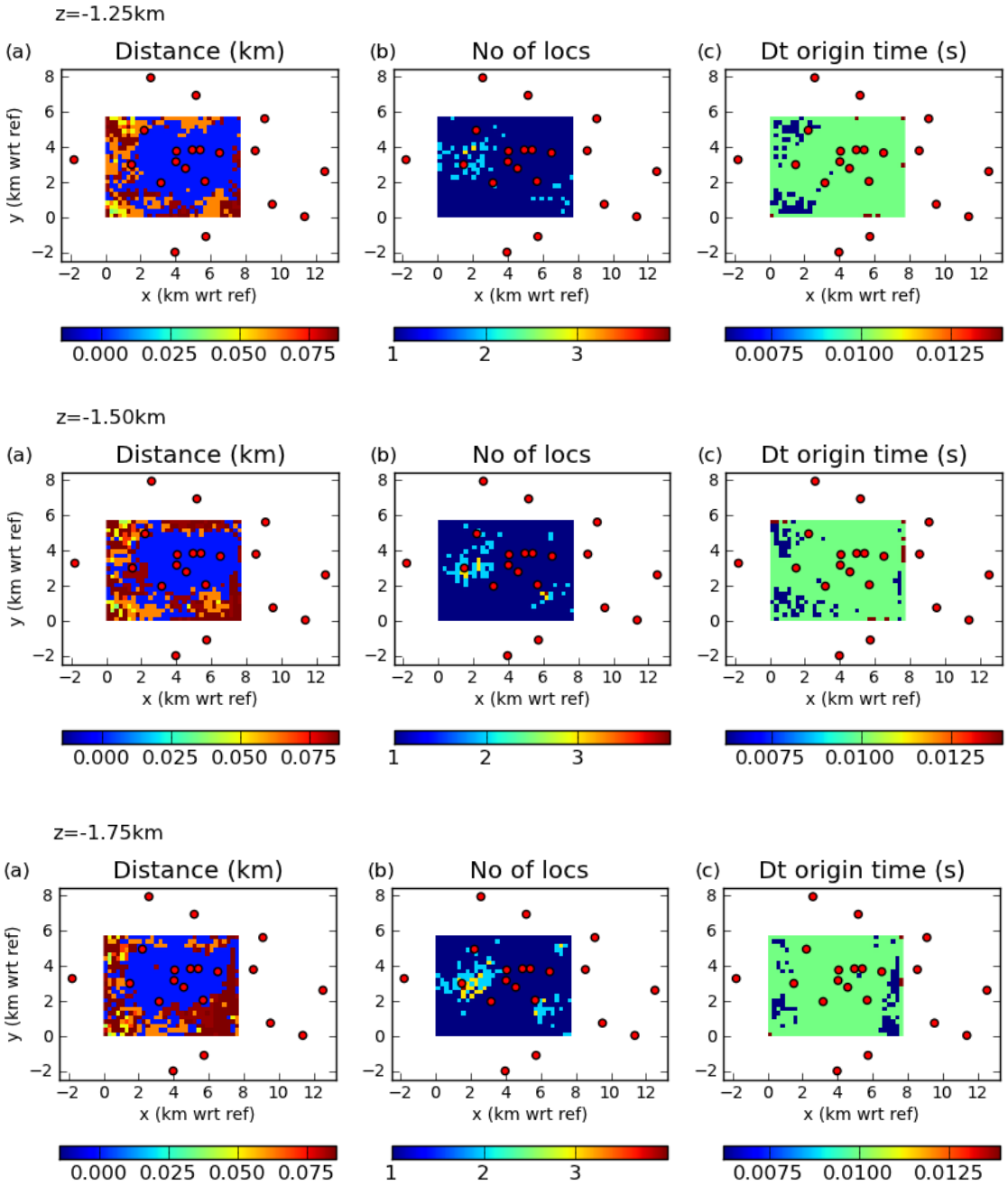
- les résultats sont finalement cartographiés pour différentes valeurs de profondeur z (axe positif vers le bas).

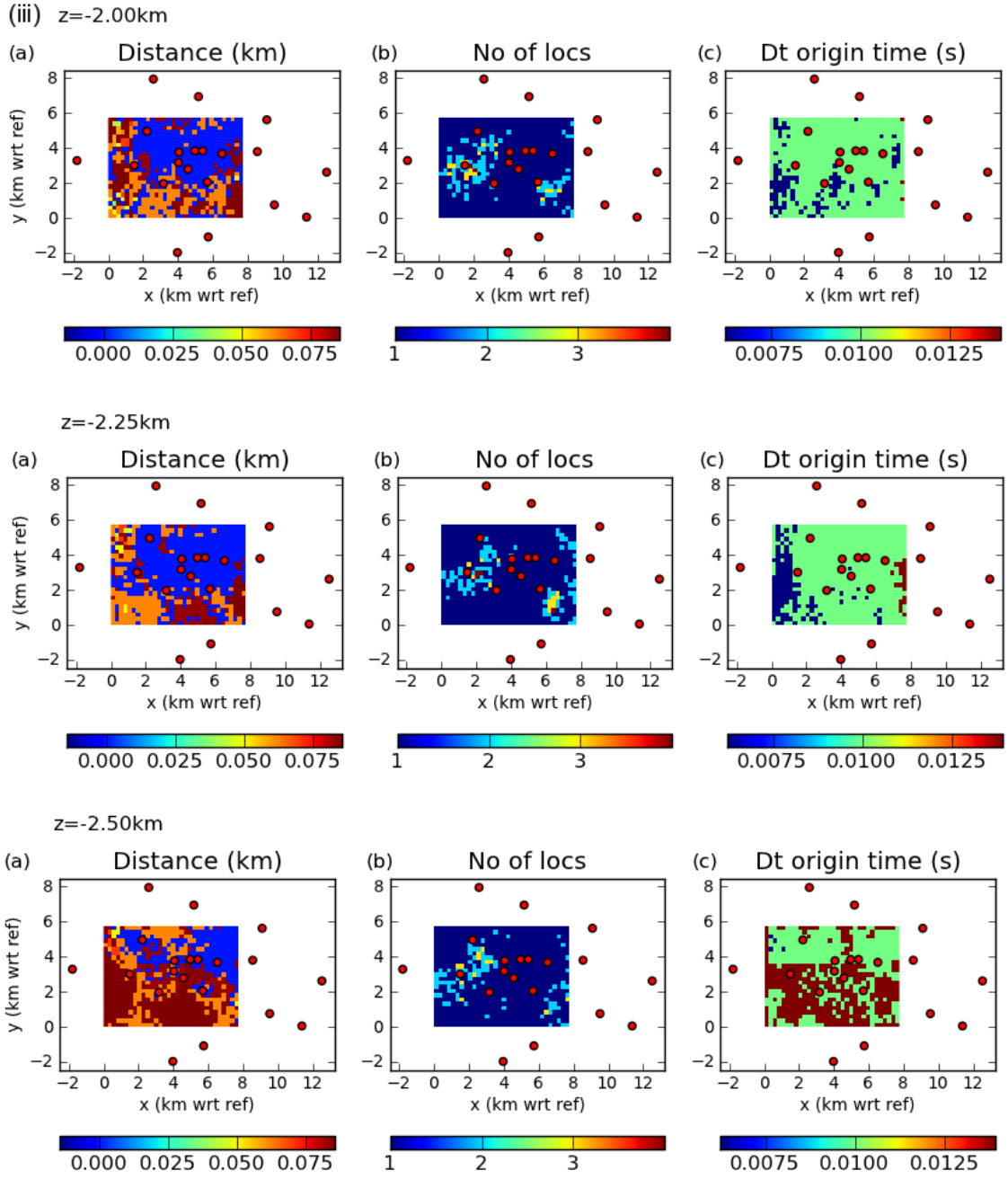
Sur chacune des figures qui suit, la sous-figure (a) cartographie les erreurs de localisation, la sous-figure (b) le nombre de localisations trouvées par Waveloc et la sous-figure (c) les erreurs sur les temps origine. Les ronds rouges correspondent aux stations.

A3.1 Configuration de stations sur le Piton de la Fournaise

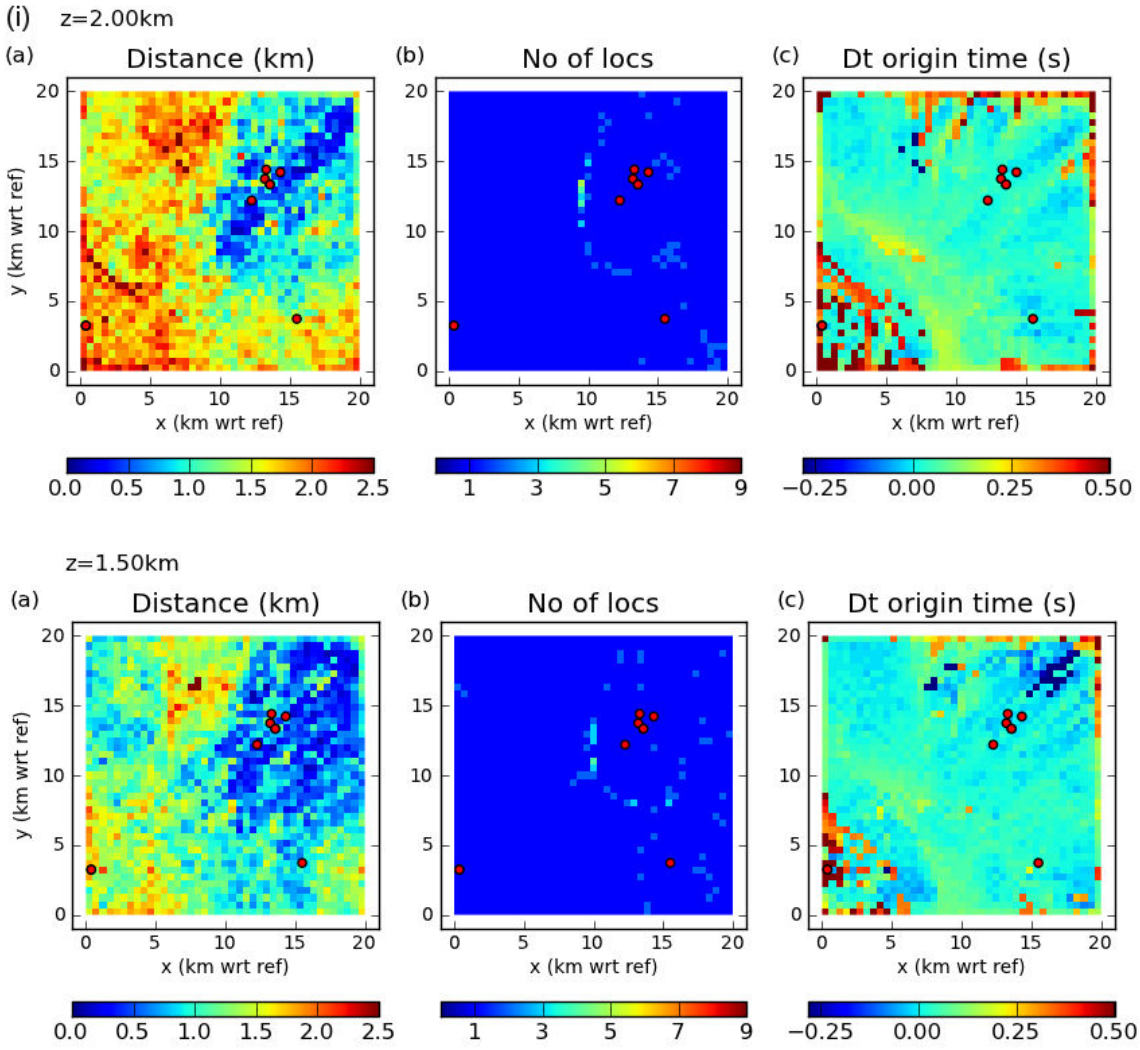


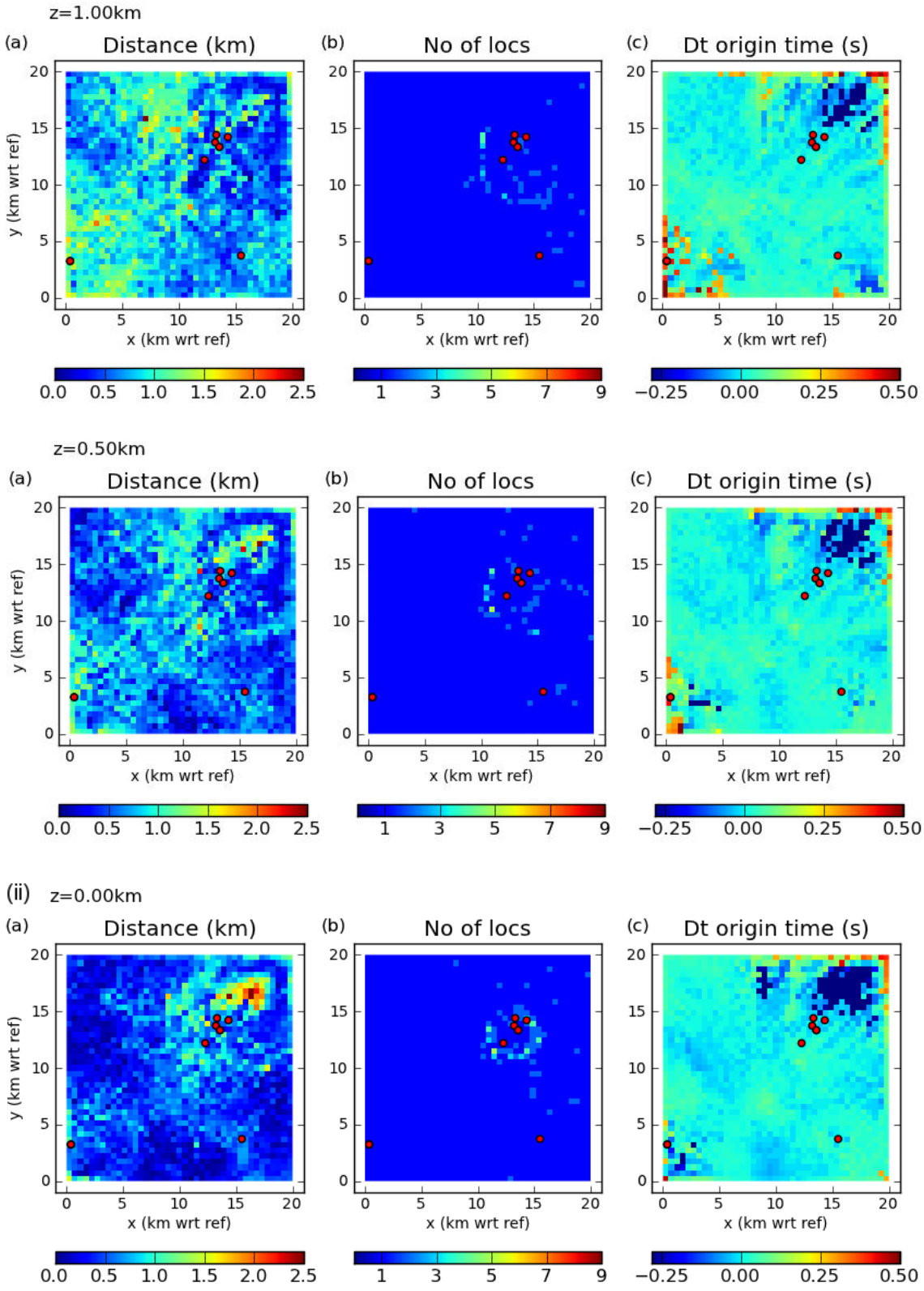

z=-0.50km
(a) Distance (km)
(b) No of locs
(c) Dt origin time (s)
y (km wrt ref)
x (km wrt ref)
0.000 0.025 0.050 0.075
1 2 3
0.0075 0.0100 0.0125
z=-0.75km
(a) Distance (km)
(b) No of locs
(c) Dt origin time (s)
y (km wrt ref)
x (km wrt ref)
0.000 0.025 0.050 0.075
1 2 3
0.0075 0.0100 0.0125
(ii) z=-1.00km
(a) Distance (km)
(b) No of locs
(c) Dt origin time (s)
y (km wrt ref)
x (km wrt ref)
0.000 0.025 0.050 0.075
1 2 3
0.0075 0.0100 0.0125

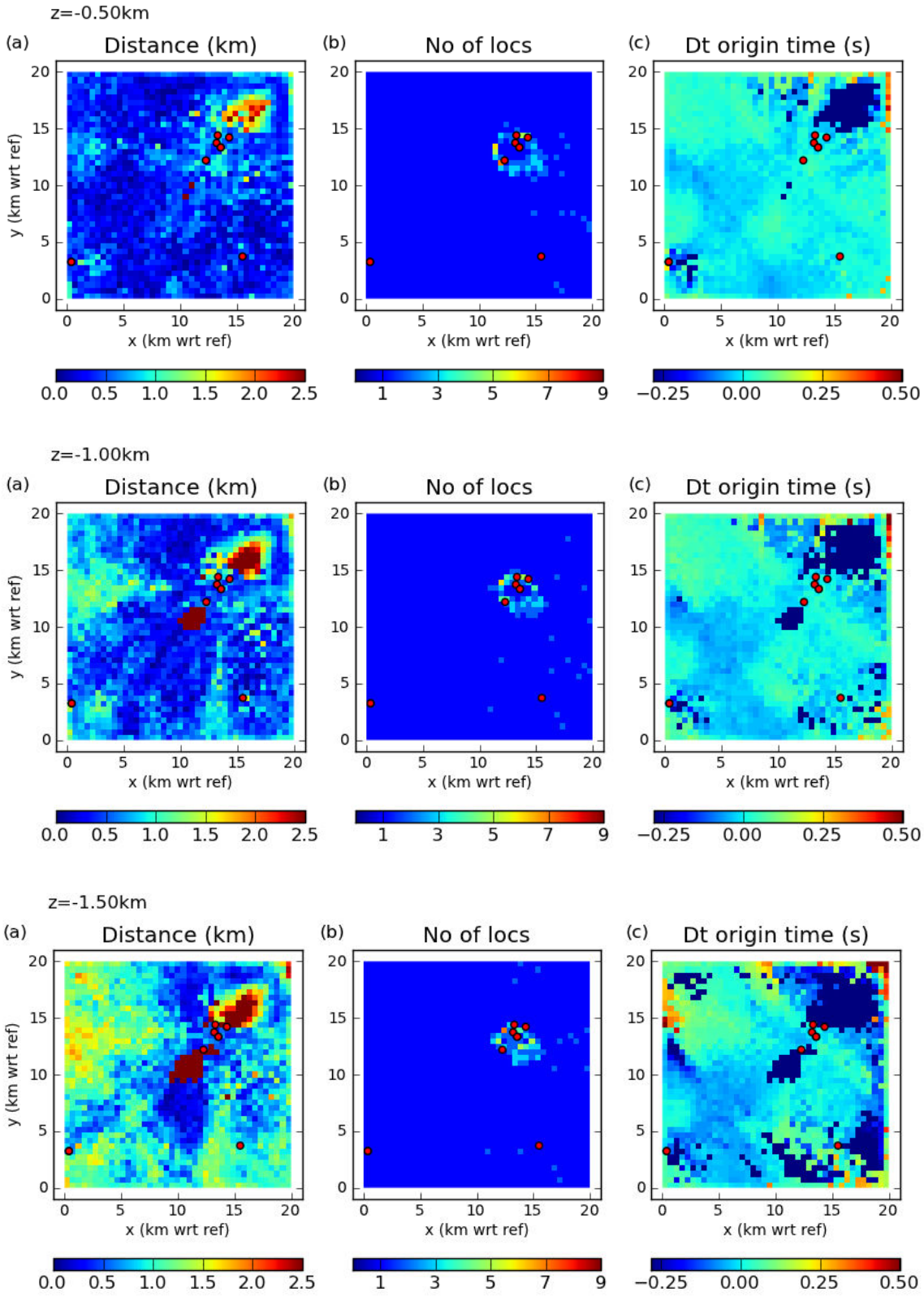

z=-1.25km
(a)
Distance (km)
(b)
No of locs
(c)
Dt origin time (s)
y (km wrt ref)
x (km wrt ref)
0.000 0.025 0.050 0.075
1 2 3
0.0075 0.0100 0.0125
z=-1.50km
(a)
Distance (km)
(b)
No of locs
(c)
Dt origin time (s)
y (km wrt ref)
x (km wrt ref)
0.000 0.025 0.050 0.075
1 2 3
0.0075 0.0100 0.0125
z=-1.75km
(a)
Distance (km)
(b)
No of locs
(c)
Dt origin time (s)
y (km wrt ref)
x (km wrt ref)
0.000 0.025 0.050 0.075
1 2 3
0.0075 0.0100 0.0125

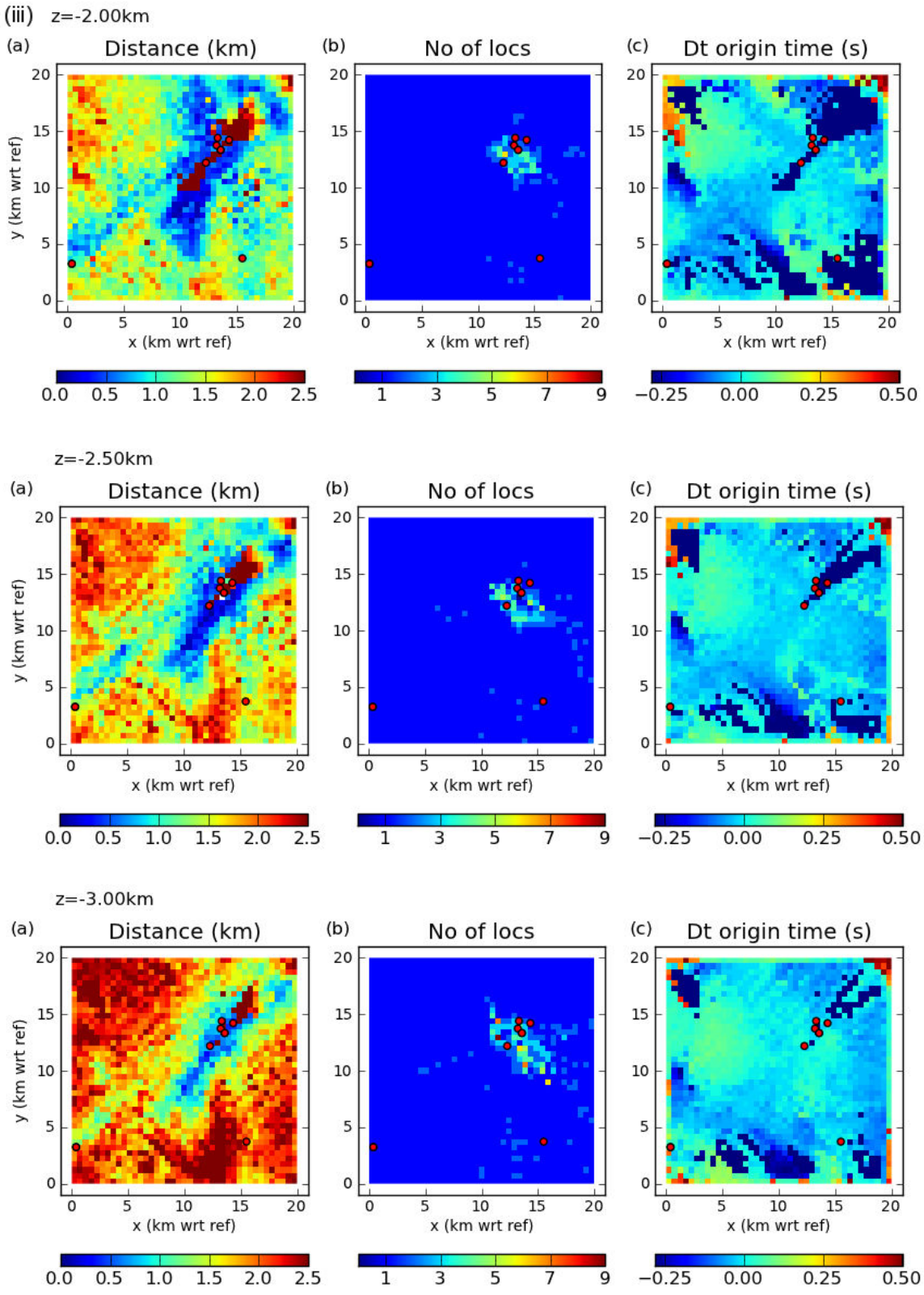

(iii) z=-2.00km
(a) Distance (km)
(b) No of locs
(c) Dt origin time (s)
y (km wrt ref)
x (km wrt ref)
0.000 0.025 0.050 0.075
1 2 3
0.0075 0.0100 0.0125
z=-2.25km
(a) Distance (km)
(b) No of locs
(c) Dt origin time (s)
y (km wrt ref)
x (km wrt ref)
0.000 0.025 0.050 0.075
1 2 3
0.0075 0.0100 0.0125
z=-2.50km
(a) Distance (km)
(b) No of locs
(c) Dt origin time (s)
y (km wrt ref)
x (km wrt ref)
0.000 0.025 0.050 0.075
1 2 3
0.0075 0.0100 0.0125

A3.2 Configuration de stations sur le Kawah Ijen




z=1.00km
(a)
Distance (km)
(b)
No of locs
(c)
Dt origin time (s)
y (km wrt ref)
x (km wrt ref)
0.0 0.5 1.0 1.5 2.0 2.5
1 3 5 7 9
−0.25 0.00 0.25 0.50
z=0.50km
(a)
Distance (km)
(b)
No of locs
(c)
Dt origin time (s)
y (km wrt ref)
x (km wrt ref)
0.0 0.5 1.0 1.5 2.0 2.5
1 3 5 7 9
−0.25 0.00 0.25 0.50
(ii)
z=0.00km
(a)
Distance (km)
(b)
No of locs
(c)
Dt origin time (s)
y (km wrt ref)
x (km wrt ref)
0.0 0.5 1.0 1.5 2.0 2.5
1 3 5 7 9
−0.25 0.00 0.25 0.50


z=-0.50km
(a)
Distance (km)
(b)
No of locs
(c)
Dt origin time (s)
y (km wrt ref)
x (km wrt ref)
0.0 0.5 1.0 1.5 2.0 2.5
1 3 5 7 9
−0.25 0.00 0.25 0.50
z=-1.00km
(a)
Distance (km)
(b)
No of locs
(c)
Dt origin time (s)
y (km wrt ref)
x (km wrt ref)
0.0 0.5 1.0 1.5 2.0 2.5
1 3 5 7 9
−0.25 0.00 0.25 0.50
z=-1.50km
(a)
Distance (km)
(b)
No of locs
(c)
Dt origin time (s)
y (km wrt ref)
x (km wrt ref)
0.0 0.5 1.0 1.5 2.0 2.5
1 3 5 7 9
−0.25 0.00 0.25 0.50


(iii) z=-2.00km
(a) Distance (km)
(b) No of locs
(c) Dt origin time (s)
y (km wrt ref)
x (km wrt ref)
0.0 0.5 1.0 1.5 2.0 2.5
1 3 5 7 9
−0.25 0.00 0.25 0.50
z=-2.50km
(a) Distance (km)
(b) No of locs
(c) Dt origin time (s)
y (km wrt ref)
x (km wrt ref)
0.0 0.5 1.0 1.5 2.0 2.5
1 3 5 7 9
−0.25 0.00 0.25 0.50
z=-3.00km
(a) Distance (km)
(b) No of locs
(c) Dt origin time (s)
y (km wrt ref)
x (km wrt ref)
0.0 0.5 1.0 1.5 2.0 2.5
1 3 5 7 9
−0.25 0.00 0.25 0.50

A4 Tests synthétiques de classification

Génération des synthétiques

Les tests synthétiques qu'on a réalisés concernent la discrimination de deux ou plus de deux classes caractérisées par deux attributs x_1 et x_2 :

- le *test set* comprend 1200 éléments, quel que soit le nombre de classes total ; le *training set* fait 40% du *test set*, soit 480 éléments.
- les proportions de chacune des classes au sein des deux *sets* peuvent varier indépendamment l'une de l'autre (c'est-à-dire qu'on peut avoir un *training set* déséquilibré lorsque le *test set* est équilibré, ou inversement ; ou les deux déséquilibrés, mais pas de la même manière...)
- les deux attributs sont issus du tirage aléatoire de deux distributions gaussiennes dont on choisi les paramètres de forme et de position. Elles peuvent être légèrement différentes entre le *training set* et le *test set* afin d'étudier l'effet de la représentativité du *training set*.

Pour chaque test, on calcule à la fois les séparateurs de la régression logistique et de la SVM afin de pouvoir comparer les deux méthodes. Lorsque ceux-ci sont réprésentés sur la même figure, on cartographie en fond la classification trouvée par la régression logistique (par défaut).

Les PDFs sont modélisées par de simples gaussiennes définies par l'écart-type et la moyenne des distributions x_1 et x_2.

Dans les cas où les classes ont des tailles différentes, la classe de plus grande taille est modélisée en noir ; celle de taille intermédiaire, en gris ; et la plus petite en blanc.

Objectifs des synthétiques

Les test synthétiques ont d'abord permis de montrer que les algorithmes fonctionnaient bien. Ensuite, ils permettent de mieux appréhender ce qu'il se passe sur des données réelles :

- comparaison des comportements de la régression logistique et de la SVM pour des jeux de données très facilement, correctement ou difficilement séparables ;
- effet de l'ajout d'une classe "intermédiaire" ;
- effet d'un déséquilibre dans le *test set*, dans le *training set*, ou dans les deux à la fois ;
- effet d'une variabilité entre le *training set* et le *test set* ;
- ...

Nadège LANGET

Détection et caractérisation massives de phénomènes sismologiques pour la surveillance d'événements traditionnels et la recherche systématique de phénomènes rares

Résumé

Du fait de la multiplication du nombre de réseaux sismiques temporaires, on assiste aujourd'hui à une explosion du nombre de données sismologiques disponibles. Ces données demandent ensuite à être traitées afin d'en tirer les informations voulues et accroître la connaissance que l'on a d'une zone d'étude donnée. Cette phase de traitement, si elle est effectuée manuellement, peut s'avérer d'autant plus longue et fastidieuse que les données sont nombreuses.
L'automatisation du traitement préliminaire des données est donc devenu une nécessité. Elle inclut la détection des événements, leur classification et leur localisation.
Le but d'une telle automatisation peut être double. D'abord, comme on l'a dit, en constituant une aide aux observatoires qui enregistrent et surveillent continuellement les signaux sismiques (en remplacement du traitement manuel). Ensuite, dans un intérêt plus scientifique, pour la recherche, l'identification et la reconnaissance de phénomènes plus rares et moins connus. Les domaines d'application sont divers (crises sismiques (précurseurs/répliques ; volcans...), microsismicité induite (géothermie, stockage du CO_2...),...). Ce travail de thèse est plus spécifiquement focalisé sur la sismicité volcanique, avec le traitement des données enregistrées sur 2 volcans : le Piton de la Fournaise, situé sur l'île de la Réunion et le Kawah Ijen, situé sur l'île de Java en Indonésie.
Cette thèse se décompose en 2 axes majeurs :
- l'un pour la détection/localisation des événements sismiques, avec le logiciel Waveloc dont les grandes lignes directrices ont été mises en place par *Maggi & Michelini* (2010). Le travail a consisté essentiellement en l'amélioration des outils déjà existants en terme de précision ou de temps de calcul, et en l'ajout de nouvelles fonctionnalités pour une analyse plus précise et/ou plus détaillée de la sismicité (recherche des multiplets, relocalisation après double-différence et calcul des magnitudes locales). L'application aux données du Piton de la Fournaise a permis de valider la fiabilité et la robustesse du code.
- l'autre pour la classification des événements sismiques, avec l'utilisation de deux méthodes d'apprentissage supervisé principales (la régression logistique et la SVM). Cette partie comprend la recherche des attributs permettant de caractériser au mieux les signaux sismiques, puis la classification proprement dite. L'application aux données du Piton de la Fournaise, pour lequel seuls deux types de signaux sont enregistrés majoritairement, permet de démontrer l'efficacité des méthodes mises en place. Le jeu de données du Kawah Ijen est beaucoup plus complexe (8 types définis *a priori*) et il a été nécessaire d'appliquer diverses stratégies de classification.

Mots-clé : sismologie, localisation, classification automatique, micro-sismicité, volcan

Résumé en anglais (Abstract)

For some time now the quantity of available seismological data has kept increasing. Processing such data may be time-consuming and quite tedious, especially when it is done manually, although it remains fundamental. Thus the automation of processing steps such as detection, location and classification of seismic data has become necessary and should help to provide a good overview of the seismicity in a given study area. Such an automation aims to (1) help the local observatories by replacing, or at least by completing manual processing ; (2) search, identify, recognize and characterize some rarer or not well-known phenomena.
Automation should also have the advantage of being potentially applicable to a large number of datasets: from seismic crises (foreshocks-aftershocks sequences, volcanoes...) to induced seismicity (geothermal, CO_2 storage...). In this work we only focus on volcanic seismicity, with one dataset recorded on the Piton de la Fournaise volcano (La Réunion island) and another on the Kawah Ijen volcano (Java, Indonesia).
The work is divided into two main research directions:
- firstly, the detection and location of seismic events with the Waveloc software whose main outline was set up by *Maggi & Michelini* (2010). The accuracy and computation time of this software have been improved during this thesis. The addition of new functions such as multiplet analysis, double-difference relocation and local magnitude computation also allows a more detailed analysis of the seismicity. The application to the Piton de la Fournaise volcano dataset has proved Waveloc's reliability and robustness while underlying the importance of choosing adapted parameters to exploit the full potential of the dataset.
- secondly, the seismic event classification using two main machine learning methods (logistic regression and SVM). This part also includes the search for some seismic attributes that best describe and characterize seismic signals. The application to the Piton de la Fournaise dataset proved the validity of these methods. This dataset is quite simple as only two main types of events occur on the volcano. The dataset from the Ijen volcano is much more complex, with 8 pre-defined classes, and has led us to apply some other strategies, such as extractors, in order to improve the classification. It has also highlighted how crucial the knowledge we have on data before starting the classification is.

Keywords : seismology, location, automated classification, microseismicity, volcano