Nouvelle méthode de localisation d'échantillons non Gaussiens : applications au signal sismique terrestre et martien

Arthur Cuvier

COLLEGE DOCTORAL PAYS DE LA LOIRE
ED MATIERE, MOLECULES MATERIAUX ET GEOSCIENCES

Nantes Université

THÈSE DE DOCTORAT

NANTES UNIVERSITE

ÉCOLE DOCTORALE N° 596
M*atière, Molécules, Matériaux et Géosciences*
Spécialité : Sciences de la Terre et des Planètes

Par

Arthur CUVIER

Nouvelle méthode de localisation d'échantillons non Gaussiens : applications au signal sismique terrestre et martien

Thèse présentée et soutenue à Nantes, le 18 décembre 2023
Unité de recherche : Laboratoire de Planétologie et Géosciences, UMR CNRS 6112

Rapporteurs avant soutenance :

Helle Anette Pedersen — Physicienne CNAP, Université de Grenoble-Alpes
Thomas Bodin — Directeur de recherche CNRS, Laboratoire de Géologie de Lyon : Terre, Planètes, Environnement

Composition du Jury :

Président :	Thomas Bodin	Directeur de recherche CNRS, Laboratoire de Géologie de Lyon : Terre, Planètes, Environnement
Examinateurs :	Philippe Lognonné	Professeur des universités, Université Paris Cité
	Anne Philippe	Professeur des universités, Nantes Université
Dir. de thèse :	Éric Beucler	Professeur des universités, Nantes Université
Co-enc. de thèse :	Mickaël Bonnin	Physicien-adjoint CNAP, Nantes Université

Invité
Raphaël Garcia — Professeur des universités, Université de Toulouse

Remerciements

Le travail présenté dans ce manuscrit représente l'aboutissement de trois années de thèse effectuées dans le contexte bienveillant et chaleureux du Laboratoire de Planétologie et Géosciences de Nantes. Tout ce travail n'aurait évidemment jamais été possible sans les différentes personnes m'ayant entouré et guidé, de près ou de loin, au cours de ces dernières années. Dans les quelques paragraphes ci-dessous, j'adresse mes sincères remerciements et ma reconnaissance à tous ceux qui m'ont accompagnés au cours de cette thèse, mais aussi plus généralement lors de mon parcours universitaire aboutissant finalement à la rédaction de ce manuscrit. Comme souvent lors de remerciements, les quelques noms cités ci-dessous ne constituent bien évidemment pas une liste exhaustive.

Tout d'abord, je tiens évidemment à remercier en premier lieu mes encadrants de thèse. **Éric**, **Mickaël**, **Raphaël**, je suis très heureux et fier d'avoir eu la chance de faire une thèse sous votre direction. L'encadrement que vous m'avez apporté au cours de ces dernières années était, bien évidemment excellent d'un point de vue professionnel, mais aussi profondément humain et bienveillant. Je tiens également à vous remercier sincèrement de m'avoir donné l'opportunité de pouvoir travailler sur un sujet aussi passionnant que la sismologie martienne en intégrant la mission spatiale InSight, chose que je n'aurais jamais imaginée après un simple parcours en mathématique.

Éric, tu as toujours fait preuve d'une grande disponibilité, et ce même malgré tes (très) nombreuses responsabilités au sein de l'OSUNA ces dernières années. Avec **Mickaël**, vous avez su me transmettre votre passion pour la sismologie, et avez répondu à mes innombrables questions naïves dans ce domaine. Je tiens également à vous remercier sincèrement pour avoir toujours su trouver le juste équilibre en m'accordant une pleine liberté dans mon travail de recherche, tout en me guidant perpétuellement dans la bonne direction. Comme je n'ai eu de cesse de le répéter dans mon entourage ces dernières années, ce fut un réel plaisir de travailler au quotidien avec vous, qui m'avez fourni un encadrement aussi complémentaire, pédagogique et bienveillant, m'ayant permis de m'épanouir pleinement dans mon travail. D'autre part, je tiens également à vous remercier de m'avoir offert l'opportunité de voyager pendant ma thèse, malgré un contexte sanitaire difficile, pour présenter mon travail lors de différents congrès et réunions d'équipes à Lyon, Nice, Toulouse, Londres, Grenade et Chicago. Enfin, je souhaite te remercier, **Raphaël**. Ton encadrement lors de cette thèse, malgré

la distance physique, m'a poussé à régulièrement te présenter mon travail, me permettant ainsi de toujours le remettre en perspective. Le regard nouveau que tu as su apporter durant dernières années sur mon travail interne au LPG et les idées originales que tu as proposées m'ont permis de constamment prendre du recul sur mes différents résultats. D'autre part, je tiens également à adresser mes remerciement à l'ensemble de mon jury de thèse : **Helle Pedersen**, **Thomas Bodin**, **Philippe Lognonné** et **Anne Philippe**, m'ayant fait l'honneur de bien vouloir juger de la qualité de mon travail.

Durant de ces dernières années, j'ai eu la chance d'intégrer l'équipe InSight, lors d'une période particulièrement captivante de la mission correspondant à la collecte des données sismiques martiennes, et ce, jusqu'à la fin de la mission le 21 décembre 2022. Je tiens donc à remercier tous ceux avec lesquels j'ai eu l'occasion de discuter de mon travail au sein de cette équipe, au cours par exemple de sessions posters, et qui malgré leur immense expérience, ont toujours été disponibles et bienveillants, sans jamais aucune arrogance de leur part. Parmi les membres de l'équipe InSight, je tiens particulièrement à adresser mes remerciements à **Ludovic Margerin** et **Martin Schimmel**, ayant tous deux accepté de faire partie de mon Comité de Suivi Individuel. Lors de l'exercice de présentation annuel de mon travail, vous avez su apporter un regard neuf sur mes différents résultats, ce qui a mené à de très intéressantes discussions sur certains détails techniques en traitement du signal. Je vous remercie également pour m'avoir conseillé de suivre la formation Géostructure Interne aux Houches en 2021, qui m'a été très bénéfique pour la suite de ma thèse.

Lors de mon arrivé au laboratoire en 2019, j'ai été accueilli spontanément par de nombreux doctorants m'ayant intégré rapidement au groupe déjà en place, parmi eux, je souhaite donc remercier tout particulièrement **Ludivine**, **Kevin**, **Maï**, **Maxime** et **Giovanni**. Par ailleurs, je tiens spécialement à remercier **Anthony**, auteur original de multiples blagues hilarantes du jeudi, ainsi que **Mathilde**, dont la capacité surhumaine à engloutir autant de coquillettes au beurre en un seul repas restera, je pense, l'un des plus grand mystère de ce siècle. Cette intégration rapide au cours notamment de (trop ?) nombreuses pauses cafés, soirées jeux de sociétés, ou même autour de bières au Berlin a contribué à un environnement bienveillant en leur compagnie, à l'intérieur et à l'extérieur du laboratoire, sans lequel je n'aurai sûrement pas pris la décision de candidater à cette thèse en 2020.

Par la suite, j'ai passé ces trois dernières années en compagnie de nombreux collègues, doctorants ou non, ayant tous contribué à l'ambiance chaleureuse quotidienne du laboratoire. Parmi eux, je tiens à remercier **Sami**, **Simon**, **Victorine**, **Alessandra**, **Guillaume** et **Justine**, pour avoir tous contribué à la belle ambiance régnant au LPG. Sans oublier **Anna**, dont le talent inné pour dessiner des œuvres d'arts sur les tableaux du bureau 110 n'est plus à démontrer, **Axel**, pourvu d'une admiration aussi débordante qu'incompréhensible pour Arnold Schwarzenegger, **Meven**, passé maître dans l'art des gâteaux au chocolat, et **Pauline**, sûrement dotée du plus grand ratio vitesse/taille connu à ce jour.
Au cours de ce doctorat, j'ai eu le plaisir de partager mon bureau avec deux autres doctorants, avec qui j'ai eu la chance de partager les différentes étapes de la thèse. Je remercie donc **Céline** et **Benoît**, qui ont toujours su entretenir une très belle am-

biance dans notre petit bureau 110, sans lesquels toutes ces journées passées au LPG n'auraient pas eu la même saveur, et avec qui j'espère sincèrement garder contact après mon départ du laboratoire. Par ailleurs, je ne saurais conclure ces remerciements associés au bureau 110 sans y inclure notre célèbre peluche de bureau, aka « **Charlotte la marmotte** », qui malgré son évidente inefficacité à coder en Python, s'est révélée être une merveilleuse mascotte.

Le parcours universitaire qui a abouti à la rédaction de ce manuscrit a débuté par une licence de mathématiques à Tours, il y a maintenant 10 ans, au cours de laquelle j'ai eu la chance de rencontrer de nombreux amis avec qui j'ai partagé de très beaux moments. Parmi eux, je voudrais donc remercier **Mika**, **Alex**, **Yoji**, **Cassy**, **Leuleu**, **Léo**, **Charles**, sans oublier **Marie Cochin**, que je ne saurai citer sans inclure son nom de famille. Enfin, je remercie tout particulièrement **Maedan** et **Paul**, pour les merveilleux conseils culinaires qu'ils m'ont apportés ces dernières années.

Pour conclure ces remerciements, je souhaite enfin remercier ma famille, et plus particulièrement mes parents et ma sœur, qui m'ont toujours soutenu et encouragé au cours de ces longues années d'études. Finalement, je tiens à remercier sincèrement toutes les personnes m'ayant apporté, de près ou de loin, un soutien précieux ces derniers mois, dans un contexte parfois difficile, et qui m'ont permis de mener à terme ce travail de rédaction.

Table des matières

Introduction générale

La sismologie constitue une discipline fascinante au croisement de la géologie, de la physique et de l'ingénierie, jouant un rôle essentiel dans notre compréhension de la Terre. En analysant les ondes sismiques qui parcourent notre planète, elle nous permet de sonder les profondeurs de la Terre et de décrypter les mystères de sa structure interne. De l'étude des tremblements de Terre, parfois dévastateurs, à l'exploration des processus tectoniques et volcaniques, la sismologie offre des informations cruciales pour la sécurité de nos sociétés et l'anticipation des événements naturels.

Afin d'étudier les propriétés géophysiques de notre planète, des progrès considérables ont eu lieu au cours du XXe siècle, dans le cadre de la détection et l'interprétation des signaux sismiques impulsifs associés aux tremblements de Terre. Cet effort s'est traduit par le développement de nombreuses méthodes de détection automatiques performantes (Allen, 1978; Berger et Sax, 1980; Allen, 1982), ainsi que des techniques d'inversions, visant à caractériser le milieu de propagation des ondes sismiques (Jeffreys et Bullen, 1940; Dziewonski et Anderson, 1981; Romanowicz, 1991). Cette avancée notable de la sismologie au cours de cette période s'explique notamment par le développement d'outils numériques de traitement du signal de plus en plus performants, mais aussi par un contexte géopolitique particulièrement tendu au cours de la guerre froide (1947-1991). En effet, la détection et la localisation d'événements sismiques se sont révélés être des outils particulièrement efficaces dans le cadre de la surveillance d'essais nucléaires, faisant de la sismologie un élément clé de la stratégie militaire (*e.g.* Barth, 2003). Par conséquent, le XXe siècle fut marqué par des études du signal sis-

mique se concentrant principalement sur la détection d'événements impulsifs, d'origines naturelles ou anthropiques.

Au cours des dernières décennies, une attention particulière a été apportée à l'exploration du signal sismique en dehors de ces arrivées d'ondes impulsives. Une telle étude a été rendue possible par l'acquisition désormais continue du signal sismique, l'augmentation exponentielle des capacités de stockage numérique, ainsi que l'importante extension du réseau de stations (Agnew et coll., 1976; Romanowicz et coll., 1984; Smith, 1987; Romanowicz et Dziewonski, 1987; Tytell et coll., 2016). Bien que ce signal sismique ne contienne pas de séisme notable, celui-ci se révèle toutefois riche en informations, étant composé de nombreux phénomènes d'origines naturelles ou anthropiques, provoquant des vibrations du sol (voir Díaz 2016 pour une revue complète). Parmi ces derniers, on peut citer la circulation routière (voiture, trains, etc...), l'influence des machines industrielles, les phénomènes de marées, ou encore la large influence de l'océan sur le signal sismique, connu sous le nom de « pics microsismiques » (*e.g.* Pedersen et Krüger, 2007; Ebeling, 2012; Stutzmann et coll., 2012; Beucler et coll., 2015; Gualtieri et coll., 2015). De plus, à ces informations viennent également s'ajouter de fortes contributions atmosphériques, venant influencer le signal sismique lors de fluctuations notables de la pression de l'air (Sorrells et coll., 1971; Tanimoto et Wang, 2018), de la température (Doody et coll., 2018), ainsi que de la vitesse du vent (De Angelis et Bodin, 2012; Dybing et coll., 2019).

Au vu de la grande quantité d'informations composant ce signal sismique d'intérêt, de nombreuses études se sont naturellement concentrées sur son analyse détaillée au début du XXIe siècle. Parmi ces dernières, on peut citer les approches d'inter-corrélations visant à reconstruire les fonctions de Green relatant des caractéristiques du milieu de propagation des ondes sismiques (Shapiro et Campillo, 2004; Bensen et coll., 2007; Hanasoge et Branicki, 2013; Schimmel et coll., 2017; Pedersen et Colombi, 2018), ou encore des méthodes statistiques de détection de redondance de phases dans le signal continu (*e.g.* Gaudot et coll., 2016).

Dans ce contexte marqué par un intérêt accru de l'étude du signal sismique continu, il convient alors d'effectuer une analyse approfondie de ce signal afin d'obtenir une

meilleure compréhension de ses propriétés. Afin de répondre à ce besoin, il est alors intéressant de mener une analyse des caractéristiques statistiques d'un tel signal. De plus, le choix d'une étude statistique se trouve particulièrement adaptée à l'analyse des données sismiques, de par l'importante quantité d'informations fournies par les stations enregistrant le signal sismique continu à une fréquence d'échantillonnage élevée (régulièrement supérieure à 10 points par seconde).

Lorsque le signal sismique n'est pas affecté par une perturbation notable, celui-ci relate alors d'infimes oscillations du sol et est alors communément appelé « bruit sismique » (Scales et Snieder, 1998; Weaver, 2005). Une inspection détaillée de ce bruit sismique, nous permet d'observer une propriété remarquable : sa distribution correspond parfaitement à celle d'un loi normale (voir la figure 1). Cette hypothèse de la distribution normale des données sismiques fut notamment avancée par Groos et Ritter (2009); Zhong et coll. (2015a,b); Aggarwal et coll. (2020), et est également régulièrement exploitée lors de la simulation numérique de bruit sismique, généré par un tirage aléatoire Gaussien (Tang et Ma, 2010; Asgedom et coll., 2012). En pratique, la distribution normale des données sismiques se trouve régulièrement altérée par des signaux transitoires de différentes amplitudes (séismes, influence anthropique, phénomènes atmosphériques, *glitchs*, ...). Par conséquent, l'inspection détaillée de la distribution Gaussienne, ou non, du signal sismique nous renseigne alors sur la probable présence d'une perturbation dans les données.

Dans ce travail, nous nous intéressons à l'estimation de la normalité du signal sismique et exploitons cette propriété statistique particulière, en tant qu'indicateur témoignant d'une altération des données. Afin de répondre à ce besoin, nous développons une approche innovante permettant de localiser, dans une série d'échantillons à priori Gaussien, les points s'écartant de la distribution normale du signal. Cette nouvelle méthode, nommée NG-loc (*Non-Gaussian sample locator*), permet de localiser, sans jamais imposer un quelconque choix de seuil, les échantillons ayant une amplitude anormalement élevée, ne respectant plus la distribution normale du bruit sismique. NG-loc se distingue particulièrement des tests d'adéquations classiques, concluant simplement quant à la normalité ou non d'une série de données (Pearson, 1900; Kolmogorov, 1933;

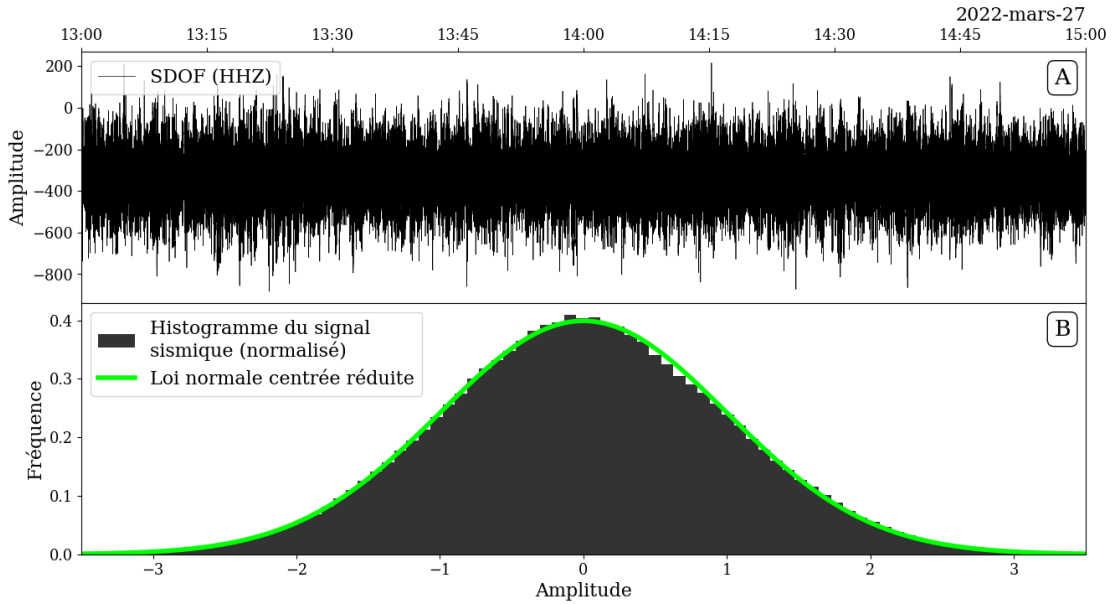

FIGURE 1 – Illustration de la distribution Gaussienne du bruit sismique. (A) : Signal sismique brut enregistré lors de la journée du 27 mars 2022. (B) : Histogramme normalisé du signal sismique (en noir), montrant une forte correspondance avec la densité de probabilité de la loi normale centrée réduite (courbe verte).

Shapiro et Wilk, 1965; Plackett, 1983), et apporte ici une information supplémentaire sur la localisation des éléments non Gaussiens.

Notre méthode NG-loc ouvre la voie à de nombreuses applications quant à l'analyse du signal sismique. En effet, celle-ci s'avère être un outil particulièrement efficace dans le cadre de la surveillance des stations sismiques, permettant d'estimer la qualité du signal enregistré, et de pointer dans certains cas, d'éventuelles dégradations. Cette propriété est d'autant plus intéressante car celle-ci présente la possibilité de localiser les périodes temporelles, gammes de fréquences et composantes associées à ces dégradations.

Par ailleurs, la très grande sensibilité de NG-loc, à l'échantillon près, permet de localiser un éventail extrêmement large de perturbations, dès lors que leurs amplitudes dans le signal se démarquent du niveau du signal de fond sismique. Par conséquent NG-loc se trouve aussi bien capable de détecter le passage de véhicule agricole que des phénomènes atmosphériques (tornades de poussière) ou encore certains types spécifiques de perturbations dans le signal sismique (*glitchs*).

À noter que notre définition de « perturbation » est ici à interpréter au sens sta-

tistique, et s'applique donc également aux tremblements de Terre, venant altérer la distribution en amplitude du signal sismique. Par conséquent, NG-loc peut apporter d'intéressantes contributions en termes de détections de séismes et trouve également une application pertinente dans l'estimation la durée de la coda des événements sismiques de faibles magnitudes. Bien que ce manuscrit se concentre sur une application de NG-loc au signal sismique, son utilisation dépasse largement le cadre de la sismologie et pourrait être appliquée à n'importe quelle série de données suivant une distribution Gaussienne.

Le travail présenté dans ce manuscrit est découpé en quatre chapitres. Nous proposons dans le chapitre 1 une introduction à la théorie des probabilités, aboutissant à la définition de la « fonction Probit » (Bliss, 1934). La fonction Probit caractérise la distribution attendue d'une série de données lorsque celle-ci suit une loi normale centrée réduite. En exploitant cette base mathématique solide, nous effectuons dans le chapitre 2 une présentation détaillée de la méthode NG-loc, permettant de retrouver les échantillons d'un signal appartenant à la distribution Gaussienne. L'étendue des capacités de détection de NG-loc est ensuite illustrée via l'analyse de nombreux signaux synthétiques, justifiant la sensibilité de notre approche à un large panel de perturbations.

Nous présentons dans le chapitre 3 l'application de NG-loc sur le signal sismique terrestre. Ce chapitre se concentre principalement sur la présentation d'un article intitulé *Seismic Station Quality Control using Deviation from the Gaussianity*, accepté pour publication (avec révisions mineures) au journal SRL (Seismological Research Letters). Quelques applications supplémentaires de NG-loc sont également proposées, incluant une étude globale de la qualité des stations du massif Armoricain ainsi que plusieurs contributions dans le cadre de la détection d'événements sismiques. Nous procédons dans le chapitre 4 à l'analyse d'un signal sismique particulier, enregistré au cours de la mission spatiale **InSight** (2018-2022), par le sismomètre SEIS (Lognonné et coll., 2019), déposé sur le sol de la planète Mars. Après une présentation du contexte de cette mission, de ses objectifs scientifiques, et du signal sismique atypique enregistré

sur Mars, nous présentons deux applications majeures de NG-loc sur ces données sismiques. La première application constitue en une analyse complète de la qualité du signal sismique acquis lors de l'entièreté de la mission mettant un certain type de perturbations appelées « *glitchs* » (Scholz et coll., 2020). La seconde application tire profit d'un important panel de perturbations détectées par NG-loc, afin de procéder à une discrimination permettant d'extraire celles d'entre elles correspondant à des tornades de poussière.

Chapitre 1

Introduction à la théorie des probabilités

Sommaire

1.1 Préambule

L'aléatoire est un élément incontournable reflétant la complexité et l'incertitude du monde qui nous entoure. Celui-ci est omniprésent dans notre quotidien et se retrouve par exemple dans les phénomènes météorologiques, les jeux de hasard, ou même les aléas liés aux accidents. De nombreux domaines sont également influencés par des phénomènes aléatoires, comme par exemple en économie, où les fluctuations du marché semblent parfois imprévisibles, et en biologie, où l'évolution des espèces est soumise à des mutations génétiques régies par le hasard. Par conséquent, l'étude de l'aléatoire se révèle donc nécessaire afin de mieux appréhender ces phénomènes physiques, difficilement prévisibles. En effet, une meilleure compréhension du hasard permet par exemple, de mieux prédire, anticiper, voire contrôler ces événements aléatoires (prévention des risques).

Afin de répondre à ce besoin, de nombreux modèles mathématiques ont été développés aux cours de ces derniers siècles, menant alors à la création de la « Théorie des probabilités », branche fondamentale des mathématiques étudiant les phénomènes aléatoires et les incertitudes. Parmi les grands noms ayant apporté des contributions majeures au développement de cette théorie, on peut par exemple citer Blaise Pascal (1623-1662), Pierre de Fermat (1607-1665), Jacob Bernoulli (1654-1705), Carl Friedrich Gauss (1777-1855), ou encore Andrey Kolmogorov (1903-1987). Il est toutefois important de noter que, en pratique, un phénomène aléatoire est défini comme le résultat d'une expérience dont l'issue est jugée incertaine aux yeux d'un certain observateur. Or, cette définition peut sembler ambiguë car l'aléatoire est bien souvent purement déterministe. En effet, le résultat d'un lancer de dé pourrait en théorie être prédit, avec une connaissance parfaite des conditions initiales (style de lancer, poids du dé,...). Cependant, cette connaissance parfaite des paramètres initiaux étant en pratique impossible à réaliser, on considère alors que le résultat de l'expérience est donc purement aléatoire.

Nous présentons ici la théorie mathématique, servant de pilier fondamental à la méthode NG-loc, présentée dans le chapitre 2. Dans la première section 1.2, nous défi-

nissons les outils mathématiques basiques de la théorie des probabilités, en prenant soin de les illustrer par quelques exemples. Cette approche théorique nous permet notamment d'aboutir à l'énoncé du Théorème Limite Central (de Laplace, 1820), assurant, sous certaines conditions, la distribution normale d'une somme de variables aléatoires. Dans le but d'étudier par la suite les caractéristiques de loi normale, nous définissons dans la deuxième section 1.3, la « fonction quantile », caractérisant la distribution des données d'une loi de probabilité quelconque. Finalement, la troisième section 1.4 introduit la « fonction Probit », primordiale dans notre étude, définie par le cas particulier de la fonction quantile d'une loi normale centrée réduite. Une des propriétés remarquables réside dans le fait que tout ensemble d'échantillons dont les éléments ont été générés par une loi normale centrée réduite converge vers la fonction Probit, une fois ceux-ci triés par ordre croissant. Ce chapitre introduisant des définitions et propriétés relativement basiques de la théorie des probabilités, certains passages pourront alors, en fonction des connaissances du lecteurs, êtres sautés.

1.2 Bases de la théorie des probabilités

Définition 1.2.1 ► Univers

L'ensemble Ω des réalisations possibles d'une expérience aléatoire est appelé univers.

Exemple 1 : Lancer d'une pièce avec $\Omega = \{\text{Pile, Face}\}$, ou bien celui d'un dé à 6 faces avec $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Afin de décrire une expérience aléatoire, il est naturel d'introduire la notion d'événement, définie comme un sous-ensemble des résultats possibles pour cette expérience.

Définition 1.2.2 ▶ Tribu/Événement

Soit Ω un ensemble non vide et $\mathcal{P}(\Omega)$ l'ensemble de ses parties. Une tribu sur Ω est une partie $\mathcal{F} \in \mathcal{P}(\Omega)$ telle que :

1. $\mathcal{F}$ est non vide
2. $A \in \mathcal{F} \Rightarrow A^C \in \mathcal{F}$ (A est stable par complémentaire)
3. $\{A_n\}_{n\in\mathbb{N}} \in \mathcal{F}^n \Rightarrow \bigcup_{n\in\mathbb{N}} A_i \in \mathcal{F}$ (A est stable par union)

Les éléments A de $\mathcal{F}$ sont appelés événements.

Exemple 2 : Si on s'intéresse par exemple à la somme obtenu après deux lancers de dés à 6 faces, l'univers associé est donc défini par l'ensemble discret $\Omega = \{2, .., 12\}$. Un événement possible de cette expérience est par exemple celui d'obtenir un résultat supérieur ou égal à 10. Cet événement est donc ici noté $A = \{10, 11, 12\}$.

Exemple 3 : Dans le cadre d'un tirage à l'aveugle d'une carte dans un paquet, on peut également s'intéresser à l'événement $A = \{$"La carte obtenue est un carreau"$\}$.

Parmi l'ensemble des événements d'une expérience aléatoire, la réalisation simultanée de certains d'entre eux est parfois incompatible. On dira alors que ces événements sont « disjoints ».

Définition 1.2.3 ▶ Événements disjoints

Soit Ω un ensemble non vide et $\mathcal{F}$ une tribu sur Ω. Deux événements A et B appartenant à $\mathcal{F}$ sont définis comment disjoints si et seulement si $A \cap B = \emptyset$.

Exemple 4 : Lors d'un lancer de dés à 6 faces, l'intersection des deux événements $A = \{$"Le résultat est supérieur à 5"$\}$ et $B = \{$"Le résultat est inférieur à 2"$\}$ est nulle. La réalisation simultanée de ces deux événements est en effet incompatible. Par conséquent, ces deux événements sont donc disjoints.

Pour un événement donné, il est naturel de s'intéresser aux « chances » que celui-ci a de se produire. Cette propriété est décrite par la notion de probabilité.

Définition 1.2.4 ► Mesure de probabilité

Soit Ω un ensemble non vide et $\mathcal{F}$ une tribu de Ω. L'application $\mathbb{P} : \mathcal{F} \to [0,1]$ est une mesure de probabilité si elle vérifie les deux propriétés :

1. $\mathbb{P}(\Omega) = 1$
2. $\mathbb{P}\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} \mathbb{P}(A_i)$, où les $A_i \in \mathcal{F}$ sont deux à deux disjoints.

Le triplet $(\Omega, \mathcal{F}, \mathbb{P})$ est alors appelé espace de probabilité.

Exemple 5 : Dans le cas d'un lancer de dé équilibré à 6 faces, chaque probabilité p_i d'obtenir i comme résultat vaut donc $\frac{1}{6}$. Correspondant au ratio du nombre d'événements permettant cette réalisation sur le nombre d'événements total. Si on note A l'événement indiquant que le résultat est supérieur ou égal à 5,

$$\mathbb{P}(A) = \mathbb{P}(\{5\} \cup \{6\}) = \mathbb{P}(\{5\}) + \mathbb{P}(\{6\}) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

De manière générale, on peut noter X le résultat d'une expérience aléatoire. X définit donc ici une fonction, appelée « variable aléatoire ». Dans le cadre de ce travail, nous nous limitons ici au cas des variables aléatoires prenant des valeurs réelles.

Définition 1.2.5 ► Variable aléatoire réelle

Une variable aléatoire réelle (v.a.r) sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ est une application $X : \Omega \to \mathbb{R}$ mesurable (*i.e.* $X^{-1}(\mathbb{R}) \subset \mathcal{F}$). Si l'ensemble $X(\Omega)$ est fini ou dénombrable, X sera alors appelé variable aléatoire discrète. *A contrario*, une variable aléatoire X est dite continue, si celle-ci peut prendre un nombre infini de valeurs.

Exemple 6 : La variable aléatoire X représentant le résultat d'un lancer de dé à 6 faces est une v.a.r discrète car $|\Omega| = 6 < \infty$.

Exemple 7 : La variable X représentant la taille d'un individu est une v.a.r continue, car pouvant prendre un nombre infini de valeurs.

Si Ω est un ensemble fini, toute application $X : \Omega \to \mathbb{R}$ est mesurable.

Définition 1.2.6 ▶ Variable aléatoire à densité

Soit X une v.a.r continue, X est dite *à densité* s'il existe une fonction $f : \mathbb{R} \to \mathbb{R}$, positive, telle que

$$\forall (a,b) \in \mathbb{R}^2,\ \mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx.$$

On dit alors que f est la fonction de densité (ou également densité de probabilité).

Une propriété évidente, engendrée directement par la définition de la mesure de probabilité (1.2.4) est donc que $\int_{-\infty}^{+\infty} f(x)dx = 1$. Par la suite, nous nous concentrons, lors de l'étude d'une v.a.r continue, uniquement sur le cas particulier où celle-ci possède une fonction de densité.

Définition 1.2.7 ▶ Indépendance

Soient X et Y, deux v.a.r définies dans un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. On dit que X et Y sont indépendant si

$$\forall A, B \in \mathcal{F}^2, \mathbb{P}\big((X \in A) \cap (Y \in B)\big) = \mathbb{P}(X \in A)\mathbb{P}(X \in B).$$

Deux caractéristiques statistiques fondamentales décrivant le comportement d'une variable aléatoire sont les notions d'espérance et de variance. Celles-ci représentent respectivement le résultat moyen d'une variable aléatoire ainsi que la dispersion de ses valeurs.

Définition 1.2.8 ► Espérance

Soit X une v.a.r discrète. On dit que X admet un moment d'ordre 1 si $\sum_{x\in X(\Omega)} |x|\mathbb{P}(X=x) < \infty$. Dans ce cas, on définit l'espérance de X par

$$\mathbb{E}(X) = \sum_{x\in X(\Omega)} x\mathbb{P}(X=x).$$

Dans le cas d'une v.a.r continue X admettant une fonction de densité f, on dit que X admet un moment d'ordre 1 si $\int_{\mathbb{R}} |x|f(x) < \infty$. Dans ce cas, son espérance est définie par

$$\mathbb{E}(X) = \int_{\mathbb{R}} xf(x)dx.$$

Définition 1.2.9 ► Variance/Écart type

Soit X une v.a.r vérifiant $\sum_{x\in X(\Omega)} |x|^2\mathbb{P}(X=x) < \infty$ (on dira que X admet un moment d'ordre 2). Dans ce cas, on définit la variance de X par

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$

De plus, on appelle écart type de X la valeur $\sigma(X) = \sqrt{\mathbb{V}(X)}$.

Pour une expérience aléatoire donnée, il est possible de modéliser le comportement d'une v.a.r X, discrète ou continue, par une « loi de probabilité ». Parmi les lois classiques utilisées en théorie des probabilités, on retrouve :

A. Loi de Bernoulli : $X \sim \mathcal{B}(p)$

La loi de Bernoulli modélise un phénomène aléatoire comportant uniquement deux issues ($\Omega = \{0,1\}$), appelées souvent « succès » ou « échec », de probabilités p et $1-p$, respectivement. Parmi les exemples classiques de phénomènes aléatoires modélisés par une loi de Bernoulli, on note par exemple le lancer d'une pièce ($p = \frac{1}{2}$ d'obtenir pile ou face), ou bien la réalisation de l'événement "obtenir un 6" à l'issue d'un lancer de dé ($p = \frac{1}{6}$ pour un dé à 6 faces équilibré). On

parle alors souvent dans ce cas d'expérience de Bernoulli. On notera $X \sim \mathcal{B}(p)$ indiquant que la v.a.r X suit une loi de Bernoulli de paramètre p (une notation similaire sera également adoptée par la suite pour les autres lois de probabilités).

B. Loi binomiale : $X \sim \mathcal{B}(n, p)$

La loi binomiale décrit le nombre de succès lors de n répétitions indépendantes d'une même expérience de Bernoulli. La v.a.r X suivant une telle loi possède donc des valeurs discrètes de 0 à n. La probabilité que celle-ci soit égale à un certain entier $k \leq n$ vaut

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Par exemple, le nombre de « piles » obtenus après n lancers de pièce peut être modélisé par une v.a.r X suivant une loi binomiale de paramètres n et $p = \frac{1}{2}$.

C. Loi uniforme : $X \sim \mathcal{U}(x_1, ..., x_n)$

La loi uniforme discrète décrit un tirage aléatoire dans un ensemble fini (x_1,...,x_n) où chacun des résultat a la même probabilité d'être obtenu. On a donc dans ce cas

$$\mathbb{P}(X = x_i) = \frac{1}{n} \ \forall i \in [\![1, n]\!].$$

Cette loi permet, par exemple, de modéliser le résultat d'un lancer de dé équilibré, ou le tirage à l'aveugle d'une carte dans un paquet.

D. Loi géométrique : $X \sim \mathcal{G}(p)$

La loi géométrique modélise la répétition d'une expérience de Bernoulli (de paramètre p) jusqu'au 1er succès. La probabilité qu'une v.a.r $X \sim \mathcal{G}(p)$ soit égale à un certain entier positif k vaut

$$\mathbb{P}(X = k) = (1-p)^{k-1} p.$$

Elle peut donc décrire par exemple une répétition de lancers d'une pièce, jusqu'à l'obtention du premier « pile ».

E. Loi de Poisson : $X \sim \mathcal{P}(\lambda)$:

La loi de Poisson modélise les « événement rares », c'est-à-dire ceux ayant une petite probabilité de se produire. Cette loi se définit avec un certain paramètre $\lambda > 0$ fixé, et on a

$$\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

La loi de Poisson permet par exemple de modéliser la probabilité de gagner à une loterie.

Les 5 lois citées ci-dessus étant toutes modélisées par des v.a.r discrètes, on se propose maintenant de donner également quelques exemples de lois permettant de modéliser des v.a.r continues.

F. Loi uniforme continue : $X \sim \mathcal{U}([a, b])$

La loi uniforme continue est la généralisation de la loi uniforme discrète dans le cas particulier où l'ensemble des résultats n'est plus un ensemble fini mais un intervalle $[a, b] \subset \mathbb{R}$. Sa fonction de densité f est donnée par

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{si} \quad x \in [a, b] \\ 0 & \text{sinon.} \end{cases}$$

G. Loi normale : $X \sim \mathcal{N}(\mu, \sigma^2)$

La loi normale est une loi de probabilité continue, de paramètres μ et σ^2 représentant son espérance et sa variance, respectivement (avec $\sigma > 0$, son écart type). Sa fonction de densité est

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{1.1}$$

Par exemple, la distribution des tailles/poids d'individus au sein d'une popula-

tion suivent tous les deux une loi normale. Dans le cadre particulier $\mu = 0$ et $\sigma = 1$, on parle alors de « loi normale centrée réduite », et on note $X \sim \mathcal{N}(0, 1)$.

H. Loi exponentielle : $X \sim \mathcal{E}(\lambda)$

La loi exponentielle, de paramètre $\lambda > 0$ correspond à la version continue de la loi géométrique. Sa densité de probabilité est définie, par

$$f(x) = \begin{cases} \lambda e^{-\lambda} x \quad \text{si} \quad x > 0 \\ 0 \quad \text{si} \quad x \leq 0 \end{cases}$$

La loi exponentielle peut servir par exemple à modéliser la durée de vie d'une ampoule ou d'un atome radioactif.

Les lois classiques présentées ci-dessus admettent toutes des espérances/variances, dont les valeurs sont présentées ci-dessous.

TABLEAU 1.1 – Espérances et variances des lois de probabilités usuelles

Loi de probabilité	**Espérance**	**Variance**
Loi de Bernoulli	p	$p(1-p)$
Loi binomiale	np	$np(1-p)$
Loi uniforme discrète	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
Loi géométrique	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Loi de Poisson	λ	λ
Loi uniforme continue	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Loi normale	μ	σ^2
Loi exponentielle	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Dans le cadre de l'étude d'une suite de v.a.r, on peut alors introduire la notion de convergence en loi.

Définition 1.2.10 ▶ Convergence en loi

Soient X une v.a.r et $(X_n)_{n\in\mathbb{N}}$ une suite de v.a.r. On dit que X_n converge en loi vers X si, pour toute fonction $g : \mathbb{R} \to \mathbb{R}$ continue et borné,

$$\lim_{n\to\infty} \mathbb{E}\Big(g(X_n)\Big) = \mathbb{E}\Big(g(X)\Big).$$

Finalement, cette définition nous permet alors d'énoncer un résultat fondamental de la théorie des probabilités : le Théorème Limite Central.

Théorème 1.2.1 ▶ Théorème Limite Central

Soit $(X_i)_{1\leq i\leq n}$ une suite de variables indépendantes, identiquement distribuées admettant un moment d'ordre 2. Si on pose

$$\mu = \mathbb{E}(X_1)\ ,\ \sigma^2 = \mathbb{V}(X_1)\ ,\ S_n = \sum_{i=1}^{n} X_i \ \text{ et } \ Y_n = \frac{S_n - n\mu}{\sigma\sqrt{n}},$$

alors $(Y_n)_{1\leq i\leq n}$ converge en loi vers une variable aléatoire de loi $\mathcal{N}(0,1)$.

La démonstration de ce théorème utilisant des outils mathématiques n'étant pas définie dans le cadre ce chapitre d'introduction à la théorie des probabilités, est présentée dans l'annexe A.1.

La force du théorème limite central réside dans le fait qu'aucune condition sur la loi des X_i n'est imposée, signifiant que la somme de n'importe quelle somme de suite de variable aléatoire entraînera un comportement Gaussien, si celles-ci sont indépendantes et identiquement distribuées. Dans ce contexte, l'étude de la loi normale est donc essentielle en statistique et en sciences en général, car elle joue un rôle crucial dans la modélisation et l'analyse de nombreuses variables aléatoires. En pratique, des processus complexes résultant de l'interaction de multiples facteurs indépendants impliquent que de nombreux phénomènes naturels suivent une distribution normale. Parmi eux, on peut par exemple noter la distribution des tailles dans une population, les erreurs de mesures dans certains protocoles scientifiques, ou encore les notes obtenues à un examen.

La figure 1.1 propose une illustration simple de ce théorème dans le cas de n lancers de dés équilibrés à six faces (le nombre de dés variant ici de un à six). Dans chacun des cas, les points bleus représentent la fréquence d'apparition associée au résultat de la somme S des n dés. Lorsque le nombre de dés augmente, cette fréquence d'apparition converge vers une loi normale, comme énoncé dans le théorème limite central 1.2.1. Dans le cas particulier $n = 1$, un seul dé est lancé, chaque résultat a la même probabilité d'être obtenu et on a donc une fréquence constante égale à $\frac{1}{6} \approx 0{,}167$. Lorsque n augmente, on distingue alors une convergence très rapide de cette fréquence

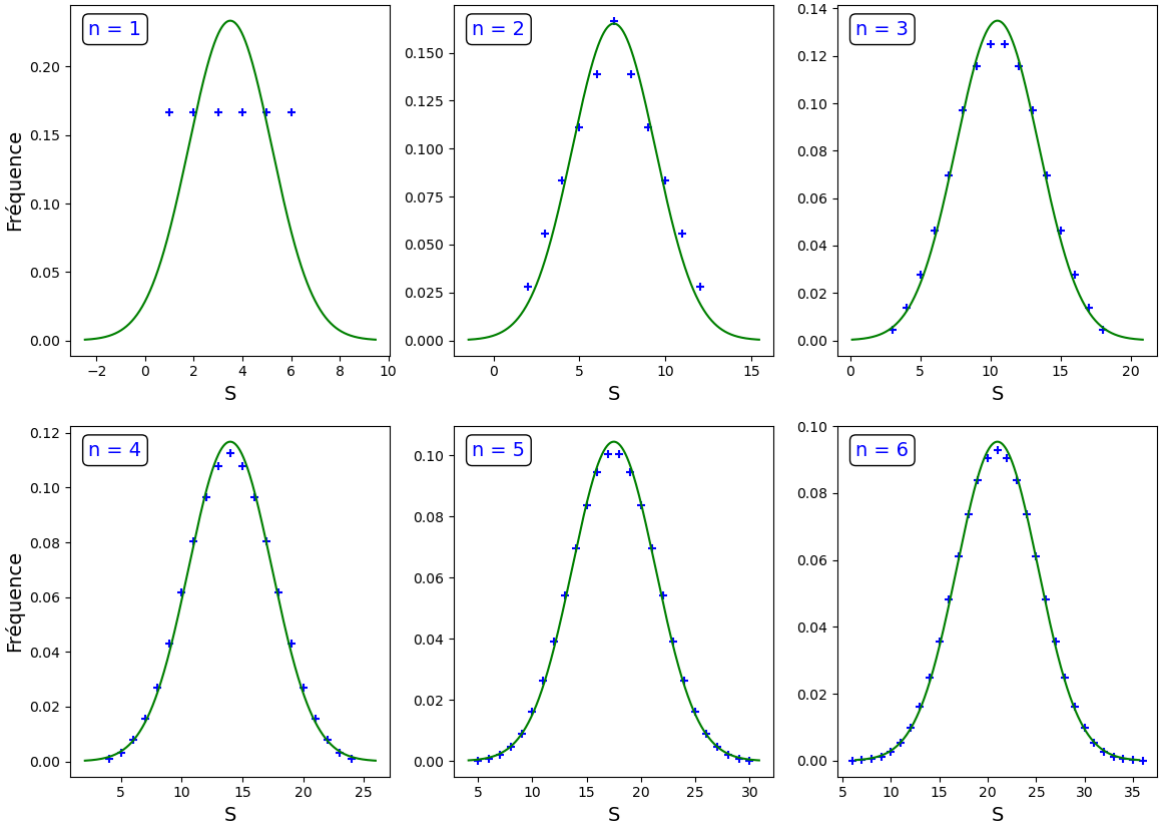

FIGURE 1.1 – Illustration du théorème limite central à travers la fréquence d'apparition des valeurs pour la somme S de n dés à six faces (bleus). Dans chacun des cas, la fréquence d'apparition est comparée à la densité de probabilité (courbes vertes) de la loi normale vers laquelle celle-ci converge d'après le théorème 1.2.1. La somme des dés n'étant ici pas normalisée comme dans l'expression de (Y_n) du théorème 1.2.1, la moyenne et l'écart type de ces densités de probabilités sont donc de $n\sigma$ et $\sigma\sqrt{n}$, respectivement.

vers la densité de probabilité de la loi normale associée (courbes verte). Comme la fréquence d'apparition correspond ici simplement à la somme des dés (*i.e.* $(S_n)_{1\leq i\leq n}$ dans l'énoncé du théorème 1.2.1), les courbes vertes ne correspondent pas à une densité de probabilité de loi normale centrée réduite. En accord avec le théorème limite central 1.2.1, chaque Gaussienne présentée ici a donc une moyenne de $n\sigma$ et un écart type de $\sigma\sqrt{n}$. Finalement, on notera que cette convergence est ici obtenue pour un nombre de lancers de dés relativement faible ($n \leq 6$). Ceci implique qu'il est possible d'obtenir une distribution normale, même pour un petit nombre de variables aléatoires.

Au vu de l'intérêt majeur que représente l'étude de la loi normale comme justifié par le théorème limite central, l'objectif dans la suite de ce chapitre consiste alors à concentrer nos efforts sur le cas Gaussien. Afin de répondre à ce besoin, nous introduisons tout d'abord dans la section 1.3 la définition générale de la fonction quantile et ses propriétés, qui seront appliquées par la suite à la loi normale dans la section 1.4.

1.3 Fonction quantile, définition et propriétés

Bien que cette section commence par quelques définitions générales, l'objectif est ici d'introduire la « fonction Probit », définissant la distribution théorique des valeurs d'une suite de v.a.r, dont les éléments suivent une loi normale centrée réduite. Ainsi, étant donné un échantillon $(X_1, ..., X_n)$ fixé, il est donc possible de comparer la distribution de ses valeurs à la fonction Probit pour attester, ou non, de la gaussianité de ce dernier.

Commençons par définir la fonction de répartition d'une variable aléatoire.

Définition 1.3.1 ▶ Fonction de répartition

La fonction de répartition d'une v.a.r X est définie par

$$F(t) = \mathbb{P}(X \leq t) \quad \forall t \in \mathbb{R}.$$

Il est important de noter que la fonction de répartition d'une v.a.r n'est pas nécessairement continue. Si l'on possède désormais un échantillon $(X_1, ..., X_n)$ fixé, il est possible de définir sa fonction de répartition discrète comme ci-dessous.

Définition 1.3.2 ▶ Fonction de répartition empirique

Soit $(X_1, ..., X_n)$, un échantillon fixé, on définit sa fonction de répartition empirique,

$$F_n(t) = \frac{\text{Card}\big(\{X_i\,;\, X_i \leq t\}\big)}{n},$$

où Card représente la fonction cardinale.

La fonction de répartition empirique représente donc simplement la proportion de points plus petits qu'un certain réel t dans un ensemble donné $(X_1, ..., X_n)$. Elle est donc

égale à 0 lorsque t est plus petit que le minimum de cet ensemble, et vaut 1 lorsque t est plus grand que son maximum. L'intérêt majeur de la fonction de répartition empirique réside dans le fait qu'elle permet, dans certain cas, d'obtenir une très bonne approximation de la fonction de répartition. En effet, si les échantillons $(X_1, ..., X_n)$ sont indépendants, identiquement distribués, avec une fonction de répartition commune F, alors la fonction de répartition empirique F_n converge vers F. Cette convergence est énoncé dans le théorème de Glivenko-Cantelli (Glivenko, 1933), dont la démonstration est présentée dans l'annexe A.2.

Théorème 1.3.1 ▶ Glivenko-Cantelli

Soit $(X_1, ..., X_n)$, un échantillon fixé de variables aléatoires réelles indépendantes et identiquement distribuées ayant la même fonction de répartition F. Alors, F_n converge uniformément, presque sûrement, vers F, *i.e.*

$$\|F_n - F\|_\infty = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n \to +\infty]{} 0 \ p.s.$$

La figure 1.2 propose une illustration de cette convergence, dans le cas de la loi normale centrée réduite. Les $(X_1, ..., X_n)$ sont donc ici obtenus suite à une simulation numérique aléatoire de n variables X_i suivant toutes une loi normale centrée réduite. Les simulations numériques sont effectuées pour différentes tailles d'échantillons $n = 50$ (A), $n = 200$ (B) et $n = 1000$ (C). En accord avec le théorème de Glivenko-Cantelli, on observe une convergence de la fonction de répartition empirique vers la fonction de répartition lorsque n augmente.

Tout comme il est possible d'étudier une fonction donnée par sa fonction inverse, il est possible d'étudier le comportement d'une variable aléatoire par l'étude de sa « fonction quantile », notée Q. Cette fonction est intrinsèquement liée à la fonction de répartition F. En effet, si F est continue et strictement croissante, la fonction quantile est définie par l'inverse de F. Comme son nom l'indique, Q définit, pour une v.a.r X donnée, la position de chacun de ses quantiles. L'étude de la fonction quantile nous permet d'obtenir une information sur la distribution des valeurs d'une variable aléatoire donnée. En général (F quelconque), la fonction quantile est définie comme ci-dessous.

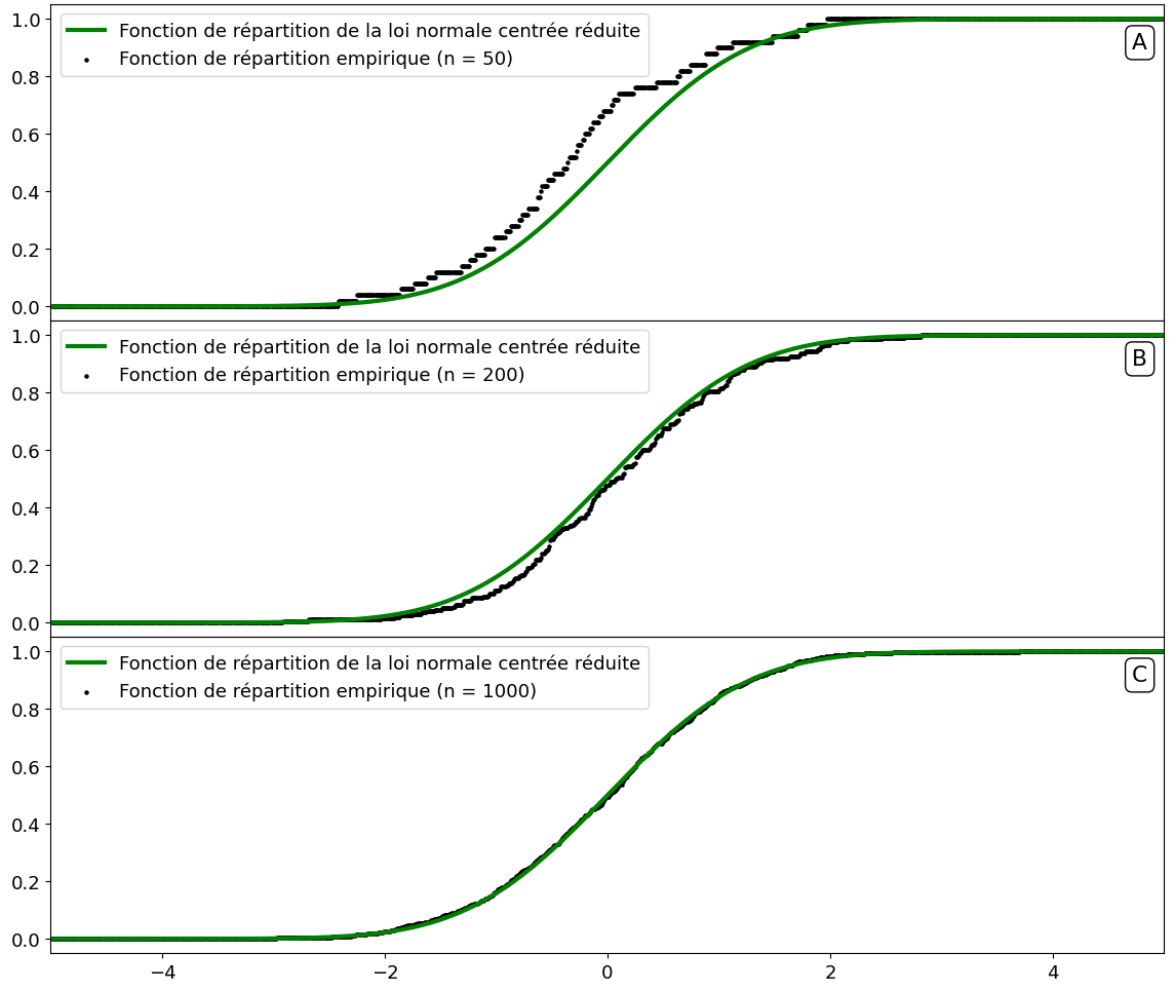

FIGURE 1.2 – Illustration de la convergence énoncée dans le théorème de Glivenko-Cantelli (1.3.1) dans le cas de la loi normale centrée réduite.

Définition 1.3.3 ▶ Fonction Quantile

Pour toute fonction de répartition F, on définit la fonction quantile Q associée par

$$Q(u) = \inf\{x \in \mathbb{R} \ ; \ F(x) \geq u\}, \quad \forall u \in [0,1].$$

Par conséquent, Q est donc l'inverse à gauche de F.

De manière similaire à la fonction de répartition empirique, il est également possible de définir une fonction quantile empirique.

Définition 1.3.4 ► Fonction quantile empirique

Soient $(X_1, ..., X_n)$ un échantillon de variables aléatoires réelles, en notant F_n sa fonction de répartition empirique, on définit sa fonction quantile empirique

$$Q_n(u) = \inf\{x \in (X_1, .., X_n) \ ; \ F_n(x) \geq u\}, \quad \forall u \in [0, 1].$$

La fonction quantile empirique représente, pour un échantillon $(X_1, ..., X_n)$ donné, ses valeurs triées par ordre croissant d'amplitude. En effet, on étudie ici le terme $F_n(x)$, c'est-à-dire la proportion de points dont l'amplitude est inférieure à x. On se restreint ensuite au cas où ce pourcentage est plus grand qu'une certaine valeur $u \in [0, 1]$. Parmi les éléments de $(X_1, ..X_n)$ vérifiant cette condition, on conserve enfin celui avec la plus petite valeur d'amplitude. Ceci correspond donc exactement à un simple tri des valeurs de $(X_1, ..., X_n)$ par ordre croissant.

Afin de mieux appréhender la notion de fonction quantile empirique, nous proposons de l'illustrer par un exemple simple dans la figure 1.3, où l'échantillon est réduit à 3 éléments avec $X = (3, 1, 2)$. Celle-ci est donc une fonction en escalier, continue par morceaux, composée dans cet exemple de $n = 3$ plateaux, dont les amplitudes correspondent à celle des éléments de X, une fois triés. Cette fonction quantile possède $n - 1 = 2$ points de discontinuités, en $x = \frac{1}{3}$ et $x = \frac{2}{3}$, représentés par des cercles creux/pleins à ces positions, indiquant que $Q_n(\frac{1}{3}) = 1$ et $Q_n(\frac{2}{3}) = 2$.

Dans le cas général, le nombre d'éléments n que l'on étudiera sera toujours très grand, ce qui augmentera alors considérablement le nombre de plateaux de notre fonction quantile empirique et la longueur de chacun d'entre eux, égale à $\frac{1}{n}$, convergera alors vers zéro. Par conséquent, chaque plateau pourra alors être approximé par un unique point (le plateau numéro k sera alors approché par le k-ème point plus petit de X).

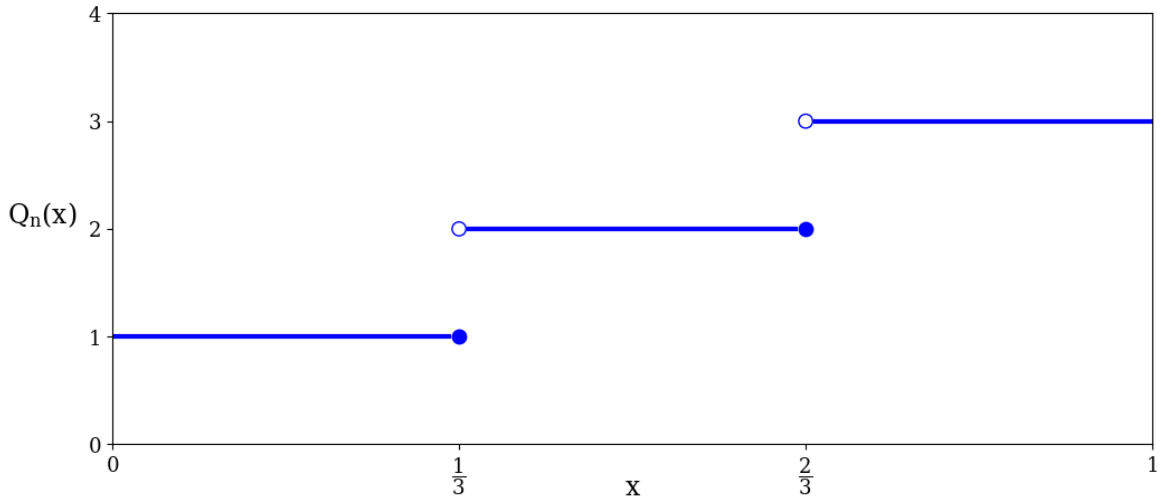

FIGURE 1.3 – Fonction quantile empirique associée à l'échantillon $X = (3, 1, 2)$.

De manière similaire au théorème de Glivenko-Cantelli (1.3.1), il existe, sous les mêmes conditions, une propriété de convergence de la fonction quantile empirique Q_n vers la fonction quantile Q. Cette convergence est illustrée par le théorème ci-dessous. Voir Van der Vaart (2000) pour la démonstration.

Théorème 1.3.2

Soit $(X_1, ..., X_n)$, un échantillon de variables aléatoires réelles indépendantes et identiquement distribuées ayant la même fonction de répartition F et fonction quantile Q. On note F_n la fonction de répartition empirique des $(X_i)_{1\leq i\leq n}$, et Q_n sa fonction quantile empirique. Alors,

$$F_n \xrightarrow[n\to+\infty]{} F \iff Q_n \xrightarrow[n\to+\infty]{} Q.$$

La convergence de la fonction quantile empirique vers la fonction quantile représente une étape cruciale dans notre étude. L'objectif par la suite consiste donc à appliquer cette propriété intéressante de convergence au cas particulier d'une variable aléatoire X suivant la loi normale centrée réduite (dont l'étude a été motivée par le théorème limite centrale 1.2.1).

1.4 Étude de la loi normale : Vers la fonction Probit

1.4.1 Loi normale centrée réduite

Définition 1.4.1 ▶ Fonction Probit

Soit $X \sim \mathcal{N}(0,1)$, la fonction quantile de X s'appelle la « fonction Probit », notée ϕ^{-1}.

La fonction Probit, fut introduite pour la première fois par Bliss en 1934, l'ayant nommée par la contraction du terme *probability unit*. Cette fonction probabiliste a été développée à l'origine pour mesurer l'efficacité d'un insecticide utilisé dans la lutte contre les nuisibles. Cependant, il s'est avéré que son application dépasse le cadre de la biologie et concerne de nombreux domaines (*e.g.* Hoffman et Low, 1981; Kockelman et Kweon, 2002; Pourhoseingholi et coll., 2008). De plus, les progrès mathématiques réalisés au cours des dernières décennies permettent une meilleure compréhension de la fonction de Probit et de ses caractéristiques (Finney, 1971; Alu, 2011). Dans notre étude, nous choisissons la notation ϕ^{-1}, découlant directement du passage à l'inverse de la fonction de répartition associée ϕ, dont une expression analytique est donnée par la proposition ci-dessous.

Proposition 1.4.1

Soit $X \sim \mathcal{N}(0,1)$, sa fonction de répartition ϕ peut s'écrire sous la forme

$$\phi(t) = \frac{1}{2}\Big(\text{erf}(\frac{t}{\sqrt{2}}) + 1\Big) \tag{1.2}$$

où erf désigne la fonction erreur.

Démonstration:

L'expression analytique de la fonction erreur (Silverman et coll., 1972) donne

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

En effectuant le changement de variable $t = \frac{z}{\sqrt{2}}$, on obtient alors,

$$\begin{aligned}
\operatorname{erf}(x) &= \frac{2}{\sqrt{2\pi}} \int_0^{x\sqrt{2}} e^{-\frac{z^2}{2}} dz \\
&= 2\Big(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x\sqrt{2}} e^{-\frac{z^2}{2}} dz - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-\frac{z^2}{2}} dz\Big) \\
&= 2\Big(\phi\big(x\sqrt{2}\big) - \phi(0)\Big) \\
&= 2\phi\big(x\sqrt{2}\big) - 1.
\end{aligned}$$

En posant maintenant $t = x\sqrt{2}$, on a finalement

$$\operatorname{erf}(\frac{t}{\sqrt{2}}) = 2\phi(t) - 1$$

et ainsi,

$$\phi(t) = \frac{1}{2}\Big(\operatorname{erf}(\frac{t}{\sqrt{2}}) + 1\Big).$$

□

Par conséquent, ce résultat nous permet d'obtenir une expression analytique simple de la fonction Probit. En effet, par passage à l'inverse de l'équation 1.2,

$$\phi^{-1}(u) = \sqrt{2}\,\operatorname{erf}^{-1}(2u - 1) \quad \forall u \in]0, 1[. \tag{1.3}$$

Une illustration de la fonction Probit est proposée dans la figure 1.4. En pratique, cette expression analytique peut être obtenue par une approximation de la fonction erf^{-1} via un développement de Mac Laurin (voir Blair et coll., 1976; Carlitz, 1963). La fonction erf^{-1} étant définie sur]-1,1[, une conséquence directe est donc (d'après l'équation 1.3) la restriction de la définition de ϕ^{-1} sur l'intervalle ouvert $I =]0,1[$.

En s'appuyant sur le théorème 1.3.2, l'objectif en pratique consiste enfin, à partir d'un échantillon donné $X = (X_i)_{1 \leq i \leq n}$, de trier ses éléments par ordre croissant puis de les comparer avec la fonction Probit. Une telle comparaison permet d'obtenir une information sur la gaussianité de l'échantillon X. En effet, en accord avec le théorème 1.3.2, une bonne correspondance entre les données triées et la fonction Probit

atteste d'un échantillon X dont les éléments suivent une loi normale centrée réduite. *A contrario*, une mauvaise correspondance entre ces derniers indique que les éléments de l'échantillon X ne suivent pas une telle loi. Le dernier détail technique nécessaire à cette comparaison est la génération d'une fonction Probit discrète de taille n. Pour répondre à ce besoin, la stratégie adoptée consiste ici à 1) effectuer une discrétisation linéaire de l'ensemble I, puis 2) appliquer l'expression analytique de l'équation 1.3 à cet intervalle fini.

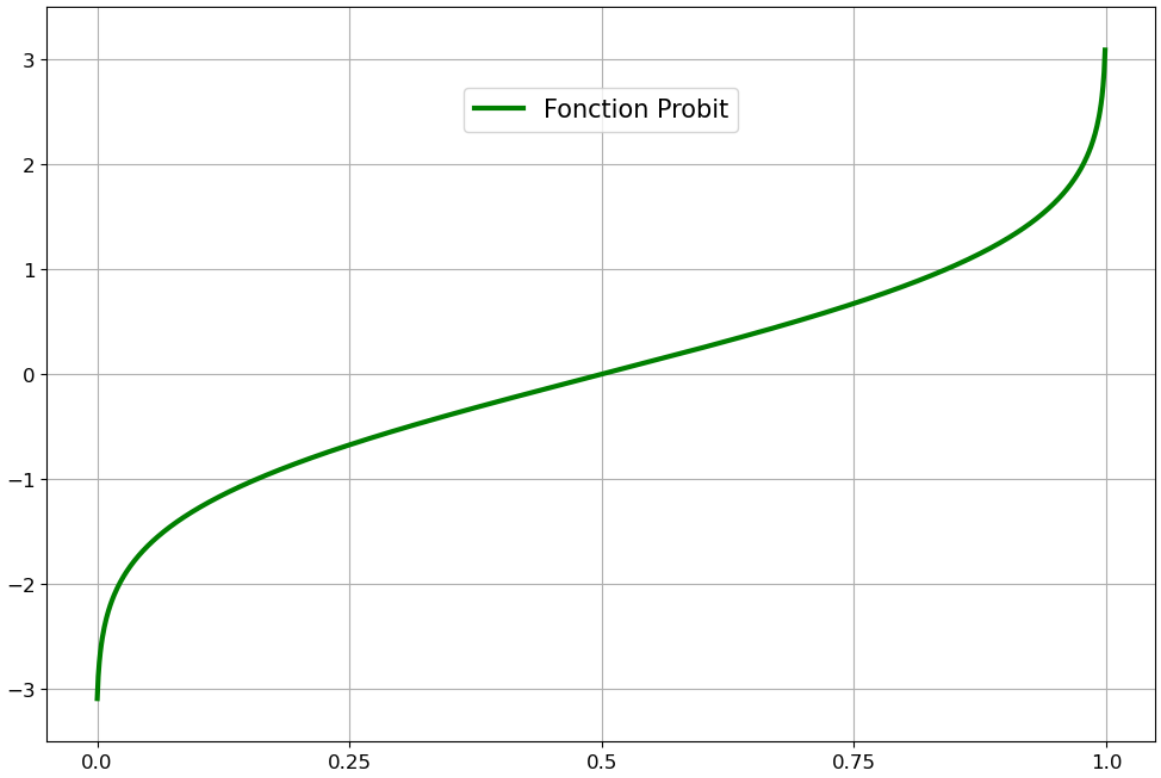

FIGURE 1.4 – La fonction Probit, générée par l'expression analytique (eq. 1.3).

Dans la figure 1.5, on compare deux échantillons de taille $n = 1000$ points (A et B) à la fonction Probit. L'échantillon (A) est généré via une loi normale centrée réduite alors que (B) est le résultat d'un tirage aléatoire uniforme. Une fois les deux signaux triés par ordre croissant, il est donc possible de les comparer avec la fonction Probit (C et D). On observe une très bonne correspondance dans le cas centré réduit (C), alors que celui du tirage uniforme présente de très grand écarts entre les deux courbes (D). Cette observation est confirmée par les valeurs de l'estimateur M, représentant pour chaque cas la différence moyenne, en valeur absolue, entre chacun des points composant les deux

courbes. En effet, on constate une très nette différence entre $M = 0{,}05$ dans l'exemple (C) alors que celui correspondant à la loi uniforme correspond à une valeur bien plus élevée avec $M = 0{,}71$ (D). Par conséquent, ceci confirme donc que l'échantillon (A) suit bien un loi normale centrée réduite, ce qui n'est pas le cas dans l'exemple (B). Bien que l'exemple proposé dans cette figure soit relativement trivial, il permet cependant de mettre en lumière les prémices d'un critère d'adéquation à la loi normale centrée réduite.

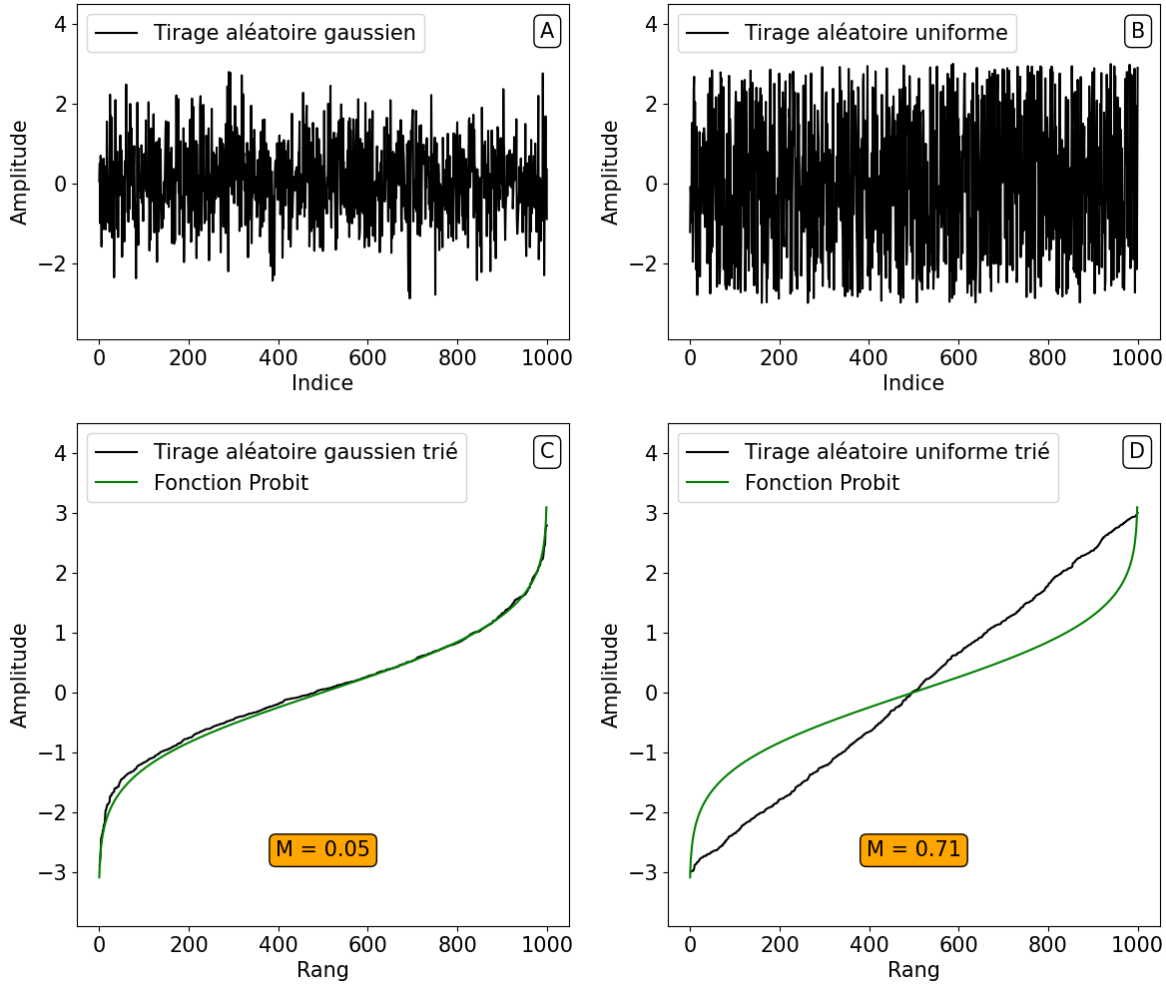

FIGURE 1.5 – Illustration de la convergence d'un tirage aléatoire trié (généré par une simulation normale centrée réduite) vers la fonction Probit. La valeur M représente dans chaque cas la correspondance entre les deux courbes, définie par la moyenne de la différence absolue entre chaque points.

Bien que le théorème 1.3.2 nous assure de la convergence simple de la fonction quantile empirique vers la fonction quantile, celle-ci peut être généralisée dans le cas

de la loi normale, à une convergence uniforme en vertu du second théorème de Dini.

Théorème 1.4.1 ▶ Second théorème de Dini

Soient $a, b \in \mathbb{R}^2$ tels que $a \leq b$. Si Q_n est une suite de fonctions croissantes de $[a, b]$ dans $\mathbb{R}$ qui converge simplement vers une fonction Q continue, alors Q_n converge uniformément vers Q.

Dans notre cas, il suffit alors de choisir $[a, b] = [0, 1]$ et de remarquer que la fonction Probit est continue afin de conclure de la convergence uniforme. La démonstration du second théorème de Dini est proposée en Annexe A.3.

1.4.2 Généralisation à une loi normale quelconque

Bien que le théorème 1.3.2 s'applique au cas particulier d'une loi normale centrée réduite, la généralisation à tout signal Gaussien peut être obtenue par des opérations de translation et d'homothétie.

Proposition 1.4.2

Soient $X \sim \mathcal{N}(0, 1)$ et $Y = aX + b$ avec $a, b \in \mathbb{R}^2$. Alors, Y suit une loi normale d'espérance b et d'écart type $|a|$.

La démonstration de cette proposition est directement déduite par la linéarité de la somme (ou de l'intégrale dans le cas d'une variable aléatoire continue), associée aux définitions de l'espérance et de l'écart-type (voir les définitions 1.2.8 et 1.2.9). Pour tout échantillon $(X_1, ..., X_n)$ de moyenne μ et d'écart type σ, il est donc possible de comparer ses valeurs triées avec une fonction probit dite « modifiée ». La fonction probit modifiée sera alors définie comme la fonction Probit auquel on impose moyenne μ et l'écart type σ, par translation et homothétie, respectivement. La figure 1.6 propose un exemple de cette généralisation, où l'échantillon (A), est le même que celui présenté dans la figure 1.5(A), auquel on a appliqué une translation/homothétie, modifiant alors les valeurs de sa moyenne et son écart type à $\mu = 200$ et $\sigma = 100$, respectivement. En imposant ces mêmes opérations de translation et d'homothétie à la fonction Probit, on obtient alors la fonction probit modifiée. Finalement, la correspondance entre les données triées et

la fonction probit modifiée (B) est donc exactement la même que dans le cas centré réduit avec la fonction Probit classique présentée dans la figure 1.5(C).

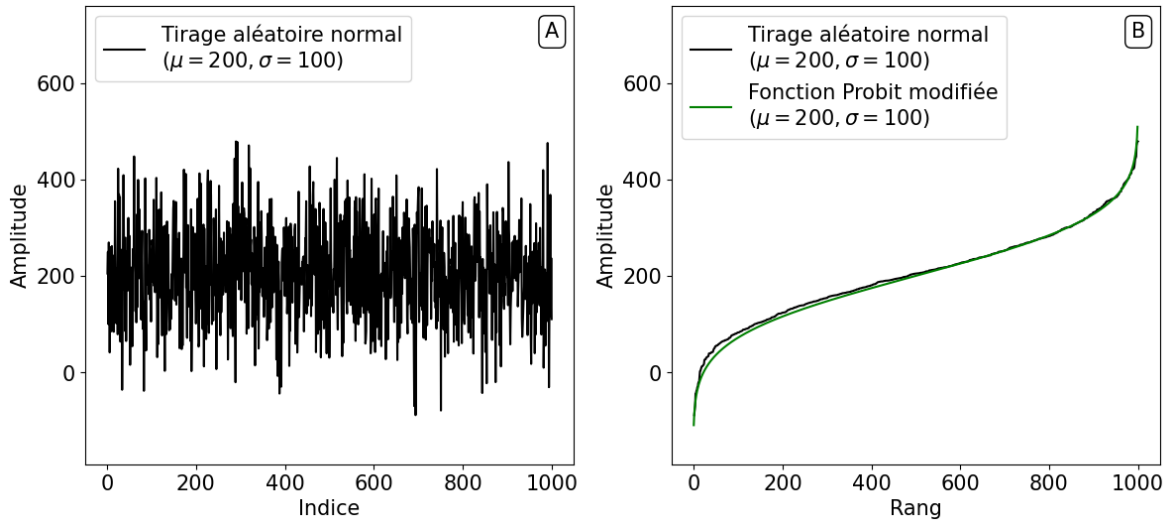

FIGURE 1.6 – Généralisation de la convergence énoncée dans le théorème 1.3.2 pour n'importe quel tirage aléatoire Gaussien (de moyenne/écart type quelconques).

En conclusion de ce chapitre, nous avons ici commencé à poser les bases d'une approche permettant, pour un échantillon $X = (X_1, ..., X_n)$ donné, d'estimer (en triant les échantillons et en les comparant à la fonction Probit modifiée) si celui-ci a été généré par une loi normale, ou non, comme illustré par les figures 1.5 et 1.6. Une application directe de ce résultat pourrait par exemple servir au développement d'un test d'adéquation à la loi normale (c'est-à-dire un test d'hypothèse servant à vérifier la gaussianité d'une série de donnée). Cependant, la stratégie que nous adoptons ici ne se limite pas simplement au test de normalité d'un échantillon statistique, mais aussi à la localisation probable d'éléments ne faisant pas partie de la distribution Gaussienne. Par conséquent, nous proposons par la suite de nous intéresser au cas où la convergence des données triées vers la fonction Probit modifiée peut être observée uniquement sur un sous-intervalle de quantiles. Le développement d'une telle méthode fournit alors plus d'informations qu'un simple test d'adéquation, puisqu'elle permet 1) de déterminer si un échantillon donnée suit une loi normale, mais aussi 2) de retrouver chacun des éléments de $(X_1, ..., X_n)$ s'écartant d'une distribution Gaussienne.

Chapitre 2

Discrimination des échantillons non Gaussiens dans un signal : la méthode NG-loc

Sommaire

L'objectif de ce chapitre est de tirer profit de la propriété de convergence énoncée dans le théorème 1.3.2 afin de développer une méthode de détection nommée « NG-loc » (*Non-Gaussian sample locator*) permettant, de localiser les points s'écartant de la distribution Gaussienne dans un échantillon statistique donné.

Dans la section 2.1, nous proposons une description détaillée de l'algorithme NG-loc, discriminant le « signal de fond Gaussien » des « éléments perturbés ». La section 2.2, se concentre quant à elle sur l'application de NG-loc à un large panel d'exemples synthétiques, permettant d'illustrer l'étendue des perturbations pouvant être détectées, donnant ensuite l'occasion de discuter des différents avantages et limites de cette méthode. Afin de faciliter son utilisation pratique, nous présentons ensuite dans la section 2.3 une approche par dichotomie du parcours de l'espace des paramètres associé à l'algorithme NG-loc, menant alors à une réduction drastique de son temps de calcul, divisé par un facteur 5000.

Plusieurs points clefs de la méthode NG-loc font alors l'objet de discussions détaillées dans la section 2.4. Nous abordons tout d'abord les différentes singularités de notre approche par comparaison aux autres méthodes de détections, justifiant de la contribution intéressante apportée par NG-loc dans ce domaine (2.4.1). Par la suite, nous discutons du nombre minimal d'éléments n de la série de données analysée afin d'assurer un fonctionnement optimal de NG-loc et analysons dans 2.4.3 la capacité de notre approche à détecter une proportion majeure d'éléments perturbés, pouvant représenter jusqu'à 90% de ses éléments. Quelques caractéristiques techniques de notre méthode sont alors abordées dans les sous-sections 2.4.4 et 2.4.5, justifiant la définition du « misfit » tel qu'introduit lors de l'algorithme de NG-loc, et discutant des conséquences de l'étude d'un échantillon statistique trié par ordre croissant. Finalement, bien que la méthode présentée dans ce chapitre se concentre sur l'analyse de séries de données comprenant un nombre de points inférieur à $n = 10\ 000$, une généralisation de NG-loc via une approche par fenêtre glissante peut également être effectuée afin de procéder à l'étude de n'importe quel signal continu ($n >> 10\ 000$).

2.1 Présentation de la méthode NG-loc

Présentons désormais la méthode NG-loc, permettant de localiser les éléments s'écartant de la distribution Gaussienne dans un échantillon statistique donné. Par la suite, l'échantillon statistique sera simplement appelé « signal », définissant une série de données fixée, un vecteur d'information, dont le nombre d'éléments n pourra varier, mais sera toujours fini. Dans le cadre général de ce chapitre, les signaux étudiés sont purement synthétiques. De la même manière qu'un vecteur, un signal synthétique $X = (X_i)_{1\leq i\leq n}$ est défini par l'ordre de ses indices i allant de 1 à n. Il est important de noter que l'arrangement de ses indices est cependant quelconque, et ne correspond donc pas nécessairement à un ordre temporel.

Dans la continuité des résultats obtenus à la fin du chapitre 1, notre étude se concentre sur les signaux suivant une loi normale. Comme nous l'avons vu dans la figure 1.6, un signal Gaussien $X = (X_i)_{1\leq i\leq n}$, de moyenne μ et d'écart type σ se définit par une forte correspondance entre ses éléments triés par ordre croissant, et la fonction probit modifiée :

$$\widehat{\phi}^{-1} = \mu + \sigma\phi^{-1} \ , \tag{2.1}$$

où ϕ^{-1} désigne la fonction Probit. Dans ce contexte, nous nous posons alors la question suivante : Que se passe t-il lorsque une « perturbation » vient modifier les extremums du signal Gaussien ?

Si X est un signal Gaussien, la notion de perturbation est ici définie par une modification de l'amplitude de un ou plusieurs de ses éléments. De plus, la nouvelle valeur prise par les éléments perturbés doit être de plus grande amplitude que le signal Gaussien d'origine. Une conséquence directe de la présence d'une perturbation implique que le signal considéré ne suit donc plus une loi normale, il est alors possible de séparer ses éléments en deux classes distinctes : d'une part le signal de fond Gaussien, dont les éléments triés par ordre croissant correspondent toujours à une distribution Probit, et d'autre part les éléments constituant la perturbation. L'altération d'un signal Gaussien par une perturbation entraîne alors une modification des caractéristiques statistiques de l'intégralité du signal, et donc également de l'ordre de ses quantiles.

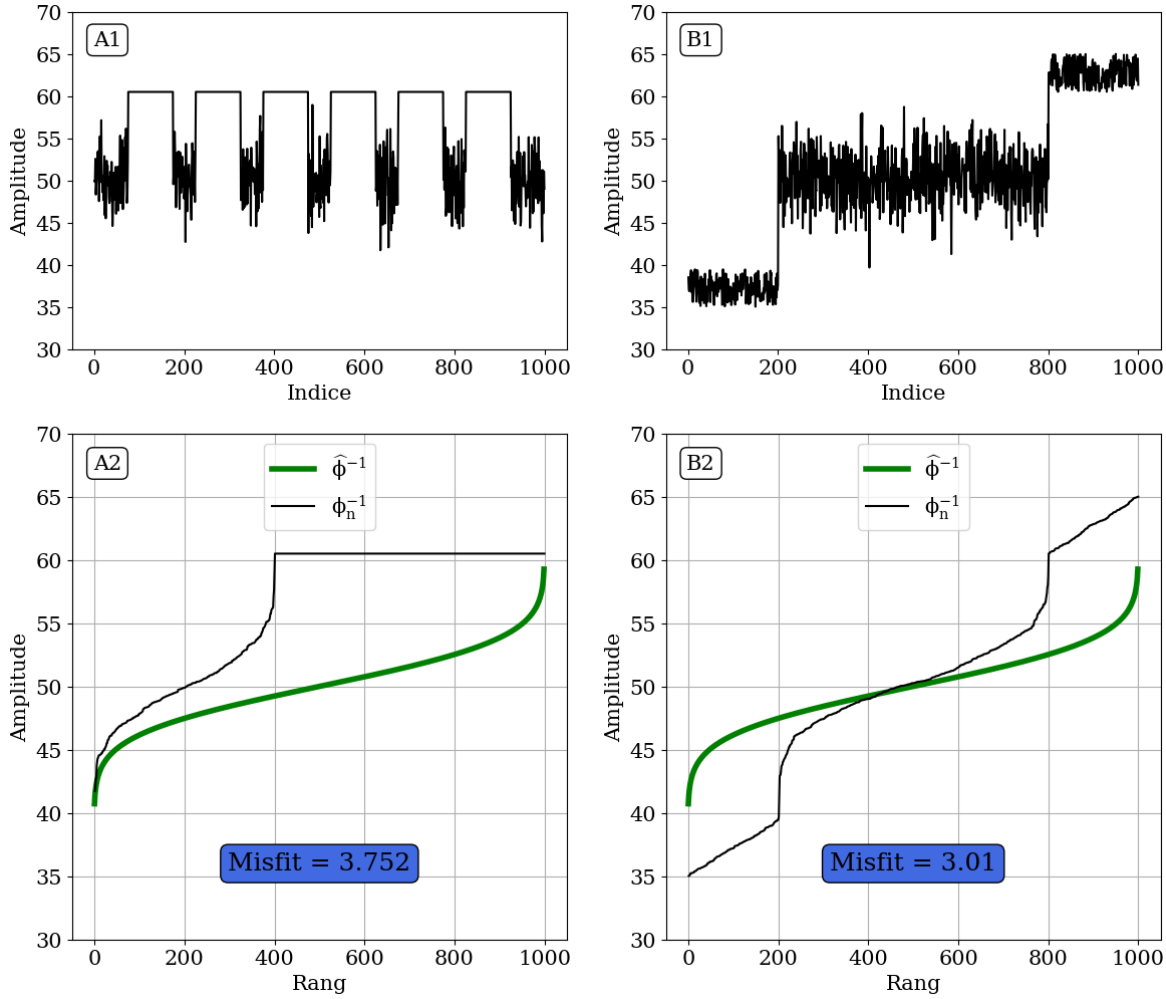

FIGURE 2.1 – Exemples de deux signaux Gaussiens ($\mu = 50$, $\sigma = 3$, $n = 1000$) altérés par des perturbations (A1 et B1, voir le texte pour une description complète). La présence de ces éléments perturbés entraîne une mauvaise correspondance entre ces signaux triés par ordre croissant ϕ_n^{-1} et la fonction probit modifiée $\widehat{\phi}^{-1}$ (A2 et B2), telle que définie dans l'équation 2.1.

Dans la figure 2.1, deux signaux Gaussiens ($\mu = 50$, $\sigma = 3$ et $n = 1000$) ayant subi des perturbations sont présentés. Dans le premier cas A1, on remarque 6 zones où l'amplitude du signal augmente fortement et atteint une valeur de 60 (autour des indices 125, 275, 425, 575, 725 et 875). L'amplitude du signal est, dans chaque cas, altérée sur 100 points autour des ces 6 indices, correspondant donc ici à un total de 600 éléments perturbés pour un signal de fond Gaussien de 400 points. Le signal B1 quant à lui est affecté sur ses 200 premiers et derniers éléments, où l'amplitude de ces indices ont été générés par une loi uniforme (entre $[35, 40]$ et $[60, 65]$, respectivement).

Par conséquent, 400 éléments sont ici considérés comme étant perturbés, et le signal de fond Gaussien est donc formé de 600 points. Dans les deux cas (A2 et B2), la présence de ces perturbations modifie la répartition des quantiles de chaque signal, entraînant alors de grands écarts entre les éléments triés ϕ_n^{-1} et la fonction probit modifiée $\widehat{\phi}^{-1}$.

Cependant, en considérant que ces signaux soient composés d'un unique signal de fond Gaussien, il est alors possible de retrouver les éléments le constituant par comparaison à la fonction probit modifiée. En effet, les éléments associés au signal de fond Gaussien sont présents dans la figure A2 sur les 400 premiers rangs, tandis que ceux de B2 se situent entre les rangs 200 à 799. On a ici exclu pour A2 les 600 éléments perturbés avec des amplitudes anormalement élevées, et pour B2 les 200 plus petits (et plus grands) points s'écartant de la distribution normale.

Sur ces intervalles de quantiles, nommés $[Q_A, Q_B]$, on observe sur la figure 2.2 une très bonne correspondance entre ϕ_n^{-1} et $\widehat{\phi}^{-1}$. Ceci signifie que, comme les perturbations affectent uniquement les extremums des signaux (et donc les bords de leurs données triées), il existe alors un intervalle continue de quantiles $[Q_A, Q_B]$, où l'on peut retrouver la distribution de la loi normale. Cette observation sert de base à la méthode NG-loc, qui repose sur une exploration du domaine des quantiles pour trouver cet intervalle $[Q_A, Q_B]$, sur lequel la meilleure correspondance entre ϕ_n^{-1} et $\widehat{\phi}^{-1}$ est observé. De plus, bien que les caractéristiques statistiques du signal Gaussien $\mu = 50$ et $\sigma = 3$ soient connues dans cet exemple, l'objectif de NG-loc consiste, en pratique, à retrouver ces paramètres.

Afin d'estimer la correspondance entre ϕ_n^{-1} et $\widehat{\phi}^{-1}$ sur un intervalle $[Q_A, Q_B]$ donné, nous introduisons un paramètre de « misfit », défini par

$$\|\phi_n^{-1}(Q_A, Q_B) - \hat{\phi}^{-1}(Q_A, Q_B)\|_{L^\infty}, \tag{2.2}$$

en choisissant ici les notations $\phi_n^{-1} = \phi_n^{-1}(Q_A, Q_B)$ et $\hat{\phi}^{-1} = \hat{\phi}^{-1}(Q_A, Q_B)$, étant tous deux des fonctions de l'intervalle quantile analysé. En effet, $\phi_n^{-1}(Q_A, Q_B)$ correspond au signal trié sur $[Q_A, Q_B]$. Les valeurs de moyenne μ et d'écart-type σ étant en pratiques inconnues, celles-ci sont alors calculées directement sur le signal trié $\phi_n^{-1}(Q_A, Q_B)$,

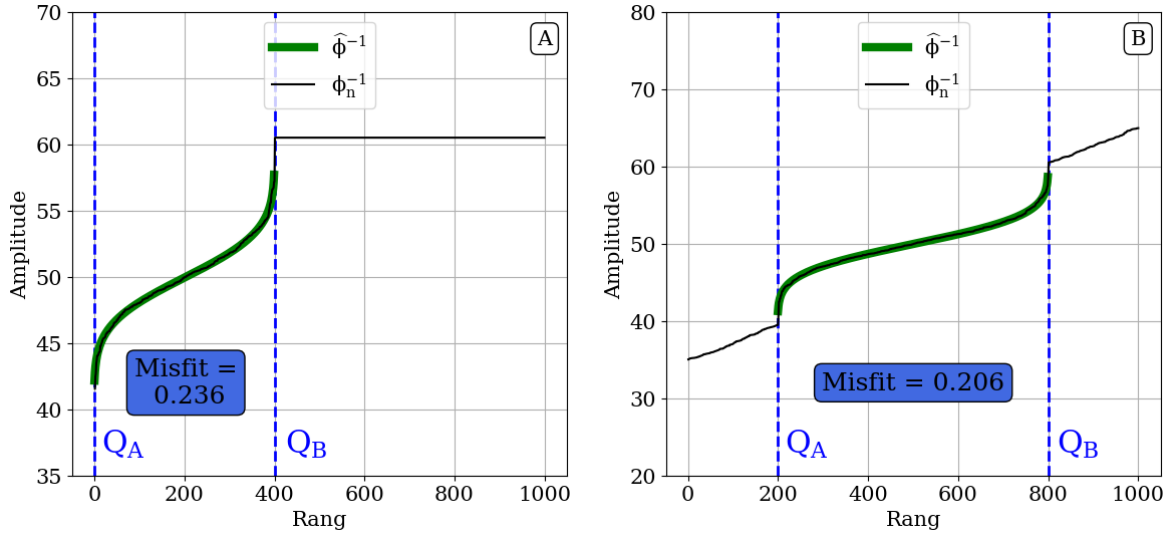

FIGURE 2.2 – Même signaux triés que ceux présentés dans la figure 2.2. La comparaison avec la fonction probit modifiée se fait cependant sur l'intervalle $[Q_A, Q_B] = [0, 399]$ pour A et $[200, 799]$ pour B.

permettant ainsi de définir $\hat{\phi}^{-1}(Q_A, Q_B)$ via l'équation 2.1. Ces deux courbes partageant désormais les mêmes moyenne/écart-types, il est alors possible de les comparer afin de vérifier si les échantillons triés sur $[Q_A, Q_B]$ suivent une loi normale. Par conséquent, si les éléments dans l'intervalle $[Q_A, Q_B]$ correspondent au signal de fond Gaussien, la différence entre ϕ_n^{-1} et $\hat{\phi}^{-1}$ sera alors très faible. Par ailleurs, en supposant l'existence et l'unicité du signal de fond Gaussien, ceci implique également l'existence d'un unique intervalle $[Q_A, Q_B]$ comportant ses éléments. De plus, cet intervalle $[Q_A, Q_B]$ est associé au plus petit misfit, en comparaison de ceux calculés sur les autre intervalles de rangs.

De manière cohérente avec la convergence uniforme énoncée dans la section 1.4, le misfit est défini par une norme L^∞ (voir la sous-section 2.4.4 pour des arguments supplémentaires justifiant ce choix). En pratique, le misfit est divisé par l'écart type σ (calculé sur chaque intervalle de rang), afin que cet estimateur soit indépendant de l'amplitude du signal analysé. Comme indiqué dans les figures 2.1 et 2.2, le misfit est un estimateur efficace témoignant de la correspondance entre les courbes ϕ_n^{-1} et $\hat{\phi}^{-1}$. En effet, on constate une énorme différence de misfit passant de 3,752 à 0,236 et de 3,010 à 0,206 pour les exemples (A) et (B), respectivement. Les valeurs de misfit sont donc ici environ 15 fois plus faible lorsque l'intervalle de rang utilisé pour la comparaison

correspond à celui du signal de fond Gaussien.

En résumé, l'algorithme de la méthode NG-loc est décrit par :

1. L'exploration de l'espace des paramètres des quantiles pour tout intervalle $[Q_A, Q_B]$.
2. Pour chaque intervalle $[Q_A, Q_B]$, les données triées ϕ_n^{-1} sont comparées à la fonction probit modifiée $\hat{\phi}^{-1}$, à l'aide du misfit.
3. Enfin, l'intervalle $[Q_A, Q_B]$ finalement sélectionné correspond à celui associé au plus petit misfit.

Il est important de noter que l'intervalle $[Q_A, Q_B]$ peut donc recouvrir l'intégralité du signal trié avec $Q_A = 0$ et $Q_B = n - 1$ si celui-ci est entièrement Gaussien. La seule condition imposée à cette exploration réside dans la taille minimale de la fenêtre $[Q_A, Q_B]$, devant être supérieure à 10% de la longueur totale du signal n (ce choix sera discuté dans la section 2.4.3). Cette propriété remarquable de NG-loc n'impose donc pas nécessairement que le signal de fond Gaussien représente plus de 50% du signal étudié. Une très grande liberté est alors laissée à la méthode NG-loc dans le choix de l'intervalle $[Q_A, Q_B]$.

La méthode NG-loc, telle que décrite par l'algorithme ci-dessus a été développée en Python 3 (Van Rossum et Drake, 2009) en utilisant notamment les bibliothèques Numpy (Harris et coll., 2020), Scipy (Virtanen et coll., 2020) et Numba (Lam et coll., 2015). Finalement, en appliquant NG-loc sur les deux signaux présentés dans la figure 2.1, on obtient alors $[Q_A, Q_B] = [0, 399]$ et $[200, 799]$ pour (A1) et (B1), respectivement. Ce résultat correspond exactement à la discrimination attendue, à l'échantillon près, entre le signal de fond Gaussien et les éléments perturbés, telle qu'annoncée dans la figure 2.2. Finalement, en revenant à une étude des signaux dans leur ordre d'origine (par indice), on peut alors retrouver la position de chacun des éléments perturbés, comme le montre la figure 2.3. Dans cette figure, les éléments faisant partie de la perturbation sont alors affichés en rouge tandis que le signal de fond Gaussien est représenté par des points gris.

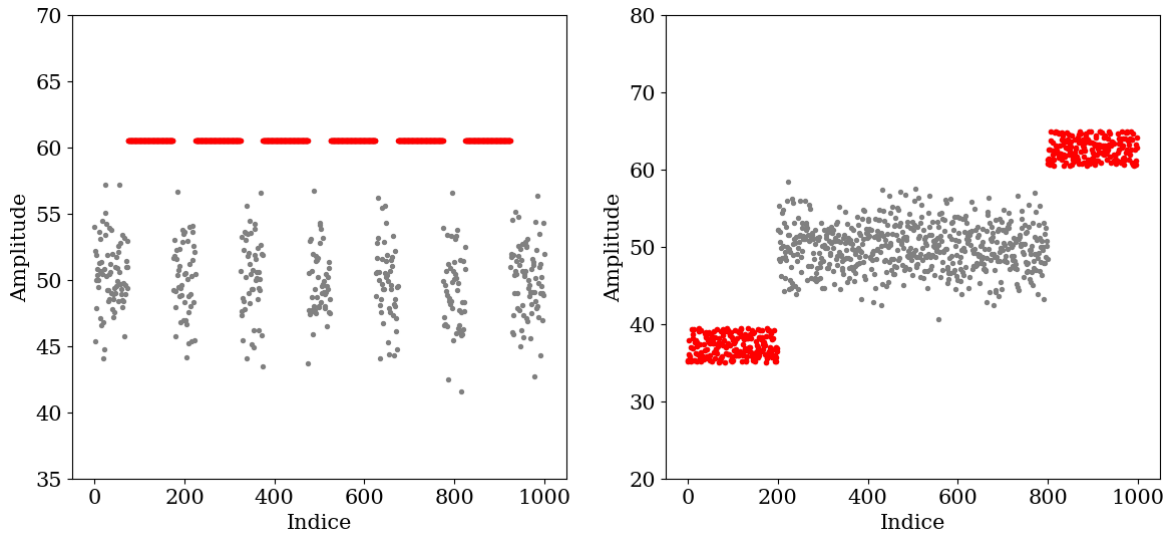

FIGURE 2.3 – Localisation des éléments perturbés (en rouge), visualisés dans leur ordre d'origine suite à l'application de la méthode NG-loc.

2.2 Applications de NG-loc à des signaux synthétiques

Suite à l'introduction de la méthode NG-loc, nous proposons désormais dans la section 2.2.1, d'illustrer par de nombreux exemples synthétiques variés, l'étendue des perturbations pouvant être détectées par cette approche. Il est toutefois important de noter que ces exemples ne constituent évidemment pas une liste exhaustive de tous les types de perturbations pouvant être détectés par NG-loc. Finalement, nous discutons dans la section 2.2.2 des avantages de la méthode NG-loc, et nous intéressons également aux limites de son application.

2.2.1 Aperçu des perturbations détectées

Afin de présenter le large éventail de perturbations pouvant être détectés par NG-loc, nous proposons ici d'illustrer son application sur de nombreux exemples synthétiques. Il est cependant important de souligner que ces signaux synthétiques sont obtenus via des simulations numériques, ne permettant pas exactement d'obtenir une véritable variable aléatoire au sens mathématique du terme. Par conséquent, on parlera plutôt de « pseudo-aléatoire », pour décrire le résultat d'une approximation du

hasard obtenu via une simulation numérique. Bien que cette notion sorte du cadre de cette section consacrée à l'application de la fonction NG-loc, on pourra cependant se référer à l'annexe B, pour une discussion permettant de 1) s'assurer de la bonne approximation du hasard par une variable pseudo-aléatoire et 2) proposer une méthode efficace pour simuler une telle variable, en utilisant la fonction quantile (voir la définition 1.3.3).

La figure 2.4 présente l'application de NG-loc sur trois exemples, numérotés individuellement par les lettres A (en haut), B (au milieu) et C (en bas). Chaque exemple est illustré par 1) le signal analysé (en haut à gauche), 2) la sélection de l'intervalle $[Q_A, Q_B]$ déterminée par NG-loc (à droite) et 3) le résultat de notre approche (en bas à gauche), permettant de discriminer les éléments faisant partie du signal de fond Gaussien (en gris), des éléments perturbés (en rouge). Dans les différents exemples analysés au cours cette section, le nombre d'échantillons de chaque signal est fixé à $n = 10\ 000$.

Dans l'exemple A, nous proposons tout d'abord l'étude d'un signal purement Gaussien, obtenu par la simulation numérique d'une loi normale de paramètres $\mu = 50$ et $\sigma = 10$. Le signal étudié dans ce cas ayant été généré via un loi normale, l'intégralité de sa distribution ordonnée correspond alors parfaitement à la fonction probit modifiée et l'application de NG-loc résulte donc à $[Q_A, Q_B] = [0, 9999]$, recouvrant donc l'intégralité des rangs (A2). Par conséquent, la sélection de cet intervalle nous indique que la totalité des 10 000 éléments de A1 font partie du signal de fond Gaussien, comme le montre A3, où tous les points sont gris (aucun point rouge, associés aux éléments non Gaussiens, n'est alors présent dans cette figure). Ceci nous permet de nous assurer que ce signal suit bien une distribution normale et qu'il n'est soumis à aucune altération.

En simulant des signaux Gaussiens avec les mêmes paramètres ($\mu = 50$ et $\sigma = 10$), nous modifions, dans les prochains exemples, l'amplitude d'une partie de leurs éléments afin d'introduire des perturbations artificielles. Le premier signal perturbé B1, subi une modification de seulement 6 éléments, aux indices 2000, 3000, 4000, 6000, 7000 et 8000 où ces derniers ont été translatés d'une amplitude de 300, 200, 100, -100, -200, et -300, respectivement. Le problème de minimisation de la méthode NG-loc sélectionne l'intervalle de quantiles $[Q_A, Q_B] = [3, 9996]$ (B2), afin d'exclure les 6 éléments altérés,

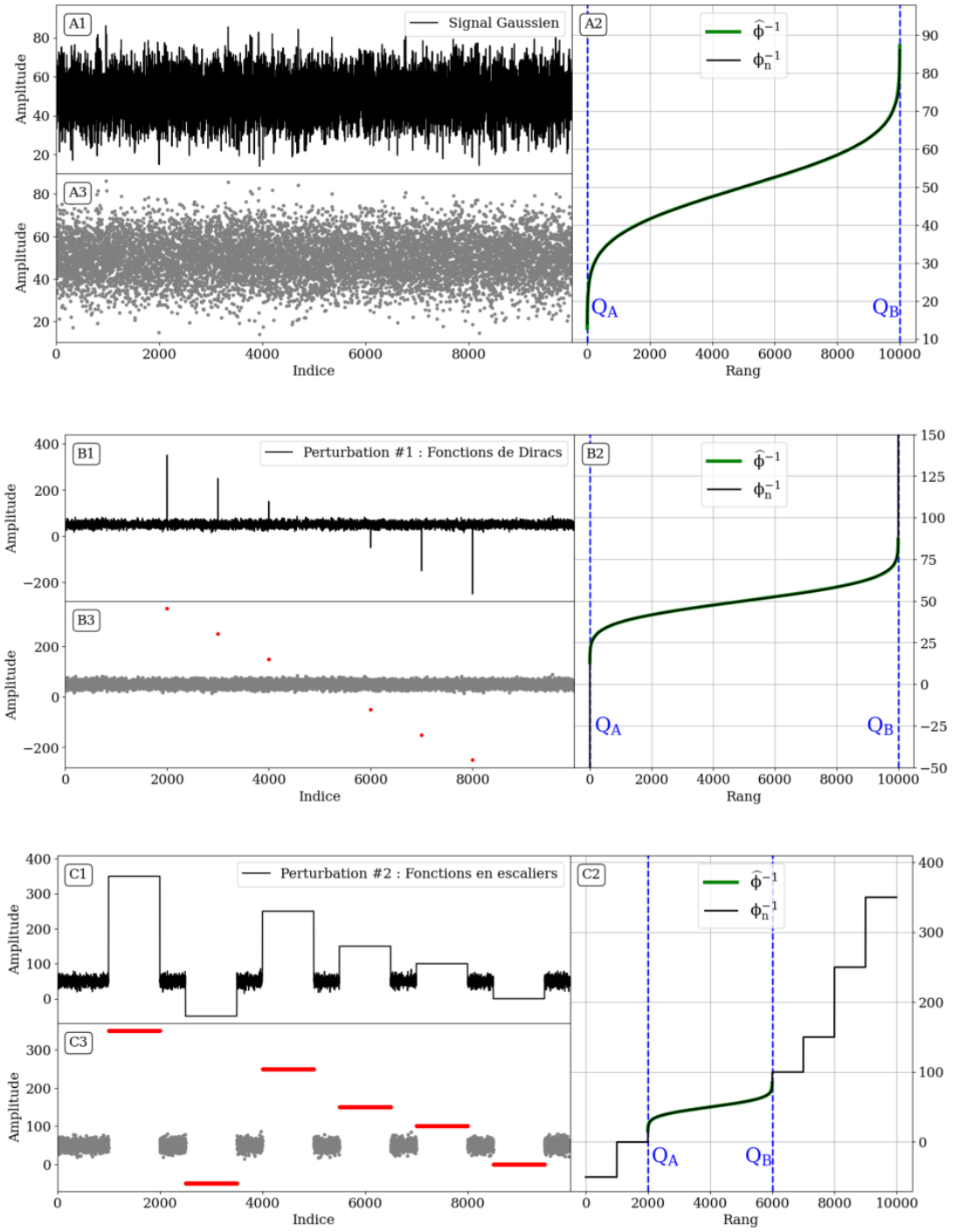

FIGURE 2.4 – Application de NG-loc sur 3 signaux synthétiques (numérotés par une lettre $X = A, B$ ou C). (X1) : Signal analysé par NG-loc. (X2) : Choix de l'intervalle $[Q_A, Q_B]$ tel que décrit dans la section 2.1. (X3) : Résultat de NG-loc, discriminant chaque point du signal analysé (gris pour le signal de fond Gaussien et rouge pour les éventuels points perturbés). Voir le texte pour la description des signaux.

représentés par des points rouges dans B3. Ce premier exemple, permet de mettre en avant la précision de la méthode NG-loc, capable de détecter des perturbations isolées dans le signal, même si celle-ci ne sont composées que d'un seul point.

L'exemple C, présente quant à lui un signal altéré par de nombreuses fonctions en escaliers de différentes amplitudes, entraînant des perturbations sur un total de 6000 éléments. Malgré une proportion d'éléments perturbés représentant plus de la moitié du signal (60% exactement), la méthode NG-loc est capable retrouver les 4000 échantillons non altérés, en sélectionnant ici $Q_A = 2000$ et $Q_B = 5999$. Finalement, C3 nous assure que les 4000 éléments de $[Q_A, Q_B]$ sont exactement les échantillons dont l'amplitude n'a pas été altérée par les fonctions en escaliers. La détection d'une proportion majeure d'éléments non Gaussien (60%) a ici été rendue possible par le choix de la fenêtre $[Q_A, Q_B]$ pouvant être réduite, lors de l'exploration des intervalles de quantiles, jusqu'à une longueur minimale de 10% de $n = 10\ 000$. Par conséquent, la méthode NG-loc permet alors de retrouver un signal de fond Gaussien, même si celui-ci est altéré par 90% d'éléments perturbés. Bien que les quelques exemples présentés dans la figure 2.4 permettent de mieux appréhender les atouts de la méthode NG-loc, les exemples de perturbations analysées sont cependant assez triviaux. En effet, les altérations dans les exemples B1 et C1 présentent tous deux une discontinuité conséquente entre l'amplitude du signal de fond Gaussien, et celle des leurs éléments perturbés, permettant alors une discrimination plus évidente entre ces deux classes (voir B2 et C2). Par la suite, nous nous intéressons alors à quelques exemples, moins triviaux, où les perturbations se caractérisent par une amplitude s'écartant progressivement du niveau de base.

La figure 2.5 présente trois autres exemples synthétiques perturbés, dont le signal de fond Gaussien est toujours généré par des simulations de $n = 10\ 000$ éléments, de moyenne $\mu = 50$ et d'écart type $\sigma = 10$. Les signaux D1 et E1 sont affectés, entre les indices 4000 et 6000, par des perturbations sinusoïdales (période complète pour E1 et demi période pour D1). Bien que l'application de la méthode NG-loc permette de détecter la majeure partie de ces deux altérations sinusoïdale (D3 et E3), on remarque cependant que l'intégralité des éléments entre ces indices ne sont pas tous considérés comme perturbés (et donc, de couleur grise). En effet, ceci est une conséquence di-

recte de notre définition d'une perturbation, imposant que l'amplitude de celle-ci soit supérieure à celle du signal de fond Gaussien, ce qui n'est pas le cas au début et à la fin de la perturbation sinusoïdale. Dans l'exemple E3, on observe de plus un « retour à l'équilibre » de l'amplitude, au centre de la perturbation, considéré pour les mêmes raisons comme étant une partie non perturbée du signal. Une conséquence directe de cet artefact est directement visible sur le misfit final de la méthode NG-loc, reflétant de la bonne correspondance entre ϕ_n^{-1} et $\hat{\phi}^{-1}$ sur $[Q_A, Q_B]$ lorsque celui-ci est faible. En effet, le misfit des perturbations sinusoïdales est égal à 0,473 (D2) et 0,394 (E2) alors qu'il vaut 0,210 dans l'exemple purement Gaussien A2 de la figure 2.4. Bien que la correspondance visuelle entre ϕ_n^{-1} et $\hat{\phi}^{-1}$ soit satisfaisante dans tous les cas, ceci atteste néanmoins d'une très légère différence dans les exemples D2 et E2, à cause de la présence de quelques éléments faisant partie de l'altération sinusoïdale dans l'intervalle $[Q_A, Q_B]$.

Le signal synthétique F1 présente une altération du signal entre les indices 4000 et 6000, où ses éléments ont été remplacés par le résultat d'une loi uniforme continue sur $[-130, 230]$ (correspondant à $\mu \pm 18\sigma$). Cette altération soudaine entraîne alors une modification de l'amplitude du signal, mais également de sa distribution. Cette fois encore, la méthode NG-loc permet de localiser de nombreux éléments perturbés et sélectionne ici $[Q_A, Q_B] = [782, 9195]$ (F2). Notre approche détecte donc un total de 1587 points perturbés, ne correspondant pas aux 2000 éléments uniformes, car seuls ceux ayant une variation d'amplitude plus grande que le signal de fond Gaussien ont été considérés comme altérés. De manière similaire aux exemples D et E, quelques éléments uniformes ont été intégrés à l'intervalle $[Q_A, Q_B]$, provoquant une légère augmentation du misfit, ayant ici pour valeur 0,339. Finalement, bien que la discrimination entre l'amplitude des perturbations et celle des signaux de fond Gaussiens dans la figure 2.5 soit moins évidente que dans la figure 2.4, les résultats de NG-loc sont tout de même très satisfaisants au vu de la localisation des points non Gaussiens (D3, E3 et F3). En effet, notre approche a permis de retrouver systématiquement le signal Gaussien original entre les rangs Q_A et Q_B et a aussi permis de détecter la présence de toutes altérations, dès que leur amplitude a excédé celle de la loi normale.

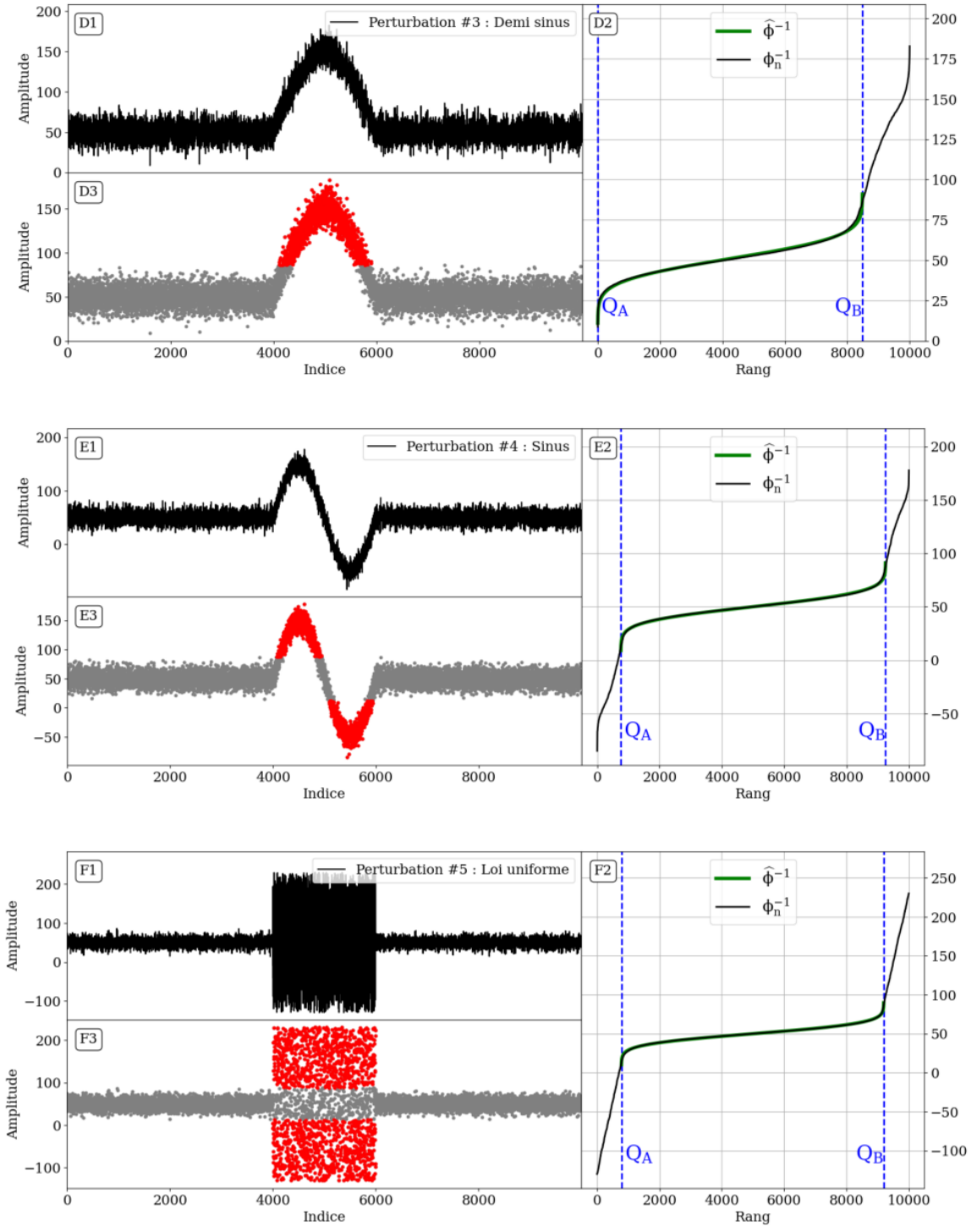

FIGURE 2.5 – Application de NG-loc sur 3 signaux synthétiques (même légende que pour la figure 2.4).

Dans la figure 2.6, nous proposons d'étudier trois dernières perturbations synthétiques, de natures différentes que les précédents exemples. Les signaux G1 et H1 présentent des altérations atypiques, modifiant uniquement l'amplitude du signal (entre les indices 4000 et 6000 pour G1 et les indices 3000 à 4000 et 6000 à 7000 pour H1), mais conservant leur distribution Gaussienne. Ces modifications se caractérisent par une augmentation (translation de +70), et une multiplication (homothétie de rapport 5) de leurs amplitudes pour G1 et H1, respectivement. Par conséquent, ceci revient donc simplement à modifier sur une portion du signal (2000 éléments), la moyenne μ pour G1 et l'écart type σ pour H1, de la distribution normale.

Il est toutefois important de noter que, en théorie, le cas d'une translation comme définie dans l'exemple G1, mais sur un nombre d'indices plus large, pourrait se révéler problématique concernant notre définition d'une perturbation. En effet, si l'on imagine une translation de l'intégralité du signal, celui-ci serait alors considéré comme non perturbé par NG-loc, alors que ses éléments ont bien subi une modification de leurs amplitudes. De manière rigoureuse, il est donc nécessaire de préciser que notre définition de perturbation correspond à une modification de l'amplitude du signal de fond Gaussien, en imposant que celle-ci ne génère pas une nouvelle distribution normale. Bien que l'étude de ce cas particulier soit intéressant, nous ne rencontrerons pas en pratique, dans les prochains chapitres d'applications de la méthode NG-loc, ce type de translations instantanées conservant la gaussianité du signal. Par ailleurs, le résultat de l'exemple G révèle une propriété intéressant de NG-loc, qui a favorisé la sélection de $[Q_A, Q_B] = [0, 7999]$, plutôt que $[Q_A, Q_B] = [8000, 9999]$, alors que la tendance autour des ces rangs correspond également à une distribution Gaussienne. La raison de cette sélection réside dans la définition du misfit, plus stable/proche de zéro lorsque celui-ci est calculé sur un plus grand nombre d'éléments vérifiant une distribution Gaussienne.

Le traitement du cas particulier de l'homothétie H1 est quant à lui moins problématique et donne lieu à des résultats différents. En effet, l'approche NG-loc repose sur la recherche d'un intervalle quantile continue $[Q_A, Q_B]$ permettant d'obtenir une bonne correspondance entre ϕ_n^{-1} et $\hat{\phi}^{-1}$. Cependant, dans le cas H1 de l'analyse de signaux suivants tous deux une loi normale, celui ayant le plus petit écart type sera systéma-

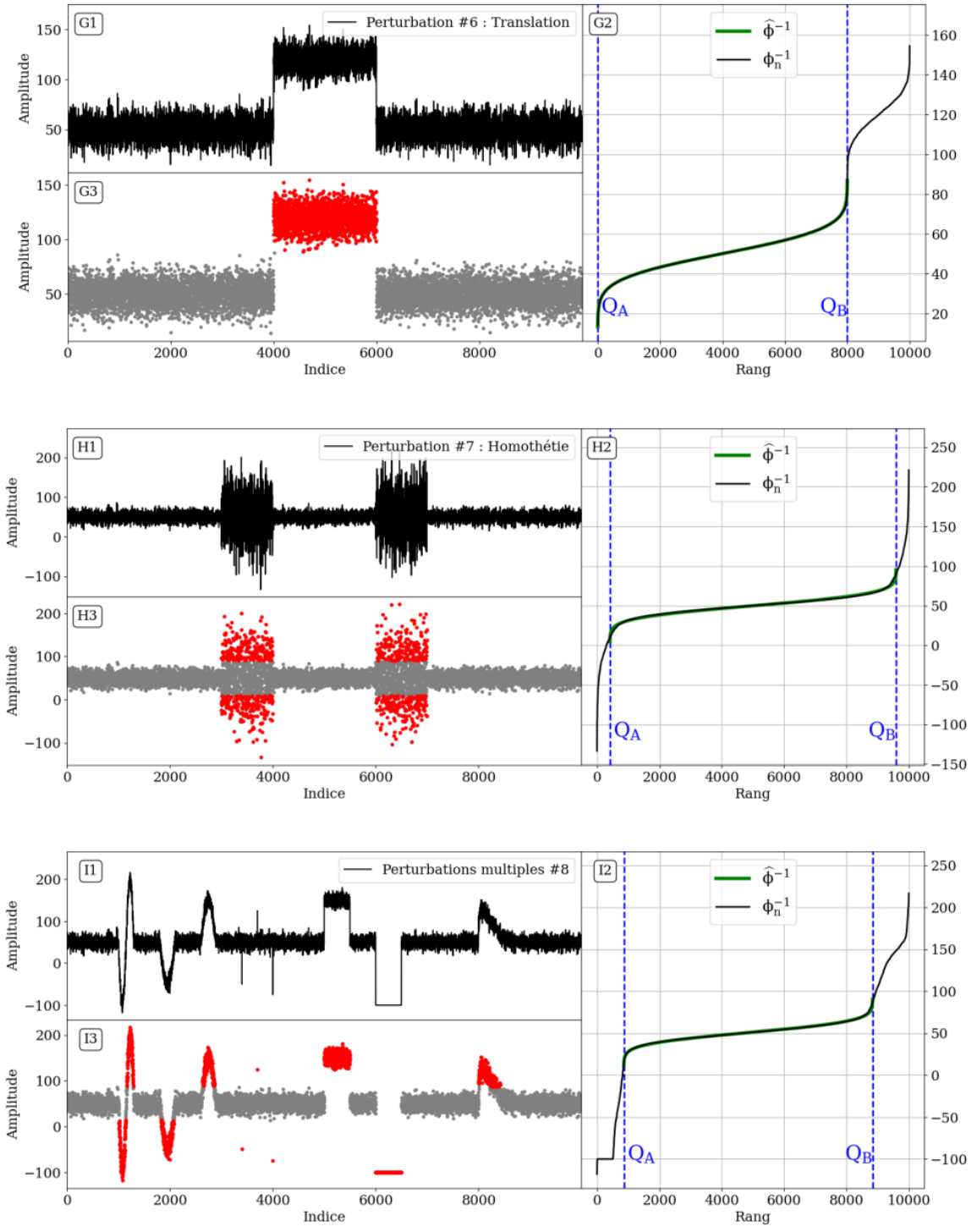

FIGURE 2.6 – Application de NG-loc sur 3 signaux synthétiques (même légende que pour la figure 2.4).

tiquement sélectionné comme étant le signal de fond Gaussien. Ceci est causé par le fait qu'il est possible de choisir un certain intervalle $[Q_A, Q_B]$ incluant l'intégralité des éléments de la loi normale de plus faible écart-type, tout en excluant une partie de ceux associés à la distribution Gaussienne de plus grande variance (voir H2). *À contrario*, l'inverse n'est cependant pas réalisable, par continuité de l'intervalle $[Q_A, Q_B]$.

Finalement, on analyse dans le cas I un exemple chaotique où le signal est altéré simultanément par de multiples perturbations. Parmi ces dernières, on retrouve (de gauche à droite) des fonctions sinusoïdales, de Dirac, une translation du signal, une fonction en escalier, ainsi qu'une perturbation asymétrique jamais analysée dans les exemples précédents. Même dans ce cas complexe altéré par de nombreuses perturbations de natures, périodes, et d'amplitudes différentes, la méthode NG-loc permet de retrouver la distribution normale du signal (voir I2). En observant I3, on remarque finalement que tous les points dont les rangs sont en dehors de l'intervalle $[Q_A, Q_B]$ (c'est-à-dire des points rouges) correspondent alors effectivement aux points dont nous avons altéré les amplitudes.

En complément de l'analyse des 9 signaux synthétiques présentés dans les figures 2.4, 2.5 et 2.6, nous proposons d'étudier dans le tableau 2.1 le misfit associé aux résultats de l'application de NG-loc sur chacun d'entre eux. En inspectant les valeurs de misfits, on remarque que le signal Gaussien et les perturbations 1,2 et 6 ont toutes des valeurs de misfits assez faibles, comprises entre 0,210 et 0,362, alors que ceux des autres altérations sont sensiblement plus élevés et varient entre 0,339 et 0,663. Ceci est causé par le fait que, les signaux associés à un misfit faible correspondent tous à une discrimination évidente en amplitude entre le signal de fond Gaussien et les éléments perturbés. Par conséquent, ces derniers ne possèdent aucun élément perturbé dans le signal de fond. *À contrario*, les signaux associés à des misfits légèrement plus élevés ne présentent pas de fortes discontinuité en amplitude entre le signal de fond Gaussien et les altérations, impliquant alors que quelques éléments perturbés sont alors compris dans l'intervalle $[Q_A, Q_B]$. Bien que la présence de quelques points perturbés dans l'intervalle quantile ne remette pas en cause l'efficacité de NG-loc dans ces exemples, le misfit est cependant sensible à ces légers changements de distribution. Ceci reflète alors

d'une correspondance légèrement moindre entre ϕ_n^{-1} et $\hat{\phi}^{-1}$, même si cet artefact est à peine visible sur les figures 2.4, 2.5 et 2.6.

TABLEAU 2.1 – Misfit obtenu suite à l'application de NG-loc, sur chacun des 9 signaux synthétiques présentés dans les figures 2.4, 2.5 et 2.6. La valeur du misfit affichée est celle obtenue en moyenne suite à l'analyse de chaque signal pour 100 tirages aléatoires Gaussien différents (avec une incertitude correspondant à deux fois son écart-type).

	Misfit
Signal Gaussien	$0{,}210 \pm 0{,}134$
Perturbation 1 : Fonctions de Diracs	$0{,}222 \pm 0{,}166$
Perturbation 2 : Fonctions en escaliers	$0{,}362 \pm 0{,}277$
Perturbation 3 : Demi sinus	$0{,}473 \pm 0{,}094$
Perturbation 4 : Sinus	$0{,}394 \pm 0{,}223$
Perturbation 5 : Loi uniforme	$0{,}339 \pm 0{,}046$
Perturbation 6 : Translation	$0{,}218 \pm 0{,}154$
Perturbation 7 : Homothétie	$0{,}445 \pm 0{,}034$
Perturbation 8 : Perturbations multiples	$0{,}663 \pm 0{,}432$

En conclusion, l'étude du misfit obtenu par la méthode NG-loc nous offre une information supplémentaire intéressante permettant de quantifier la gaussianité des données dans l'intervalle $[Q_A, Q_B]$. Nous verrons notamment dans les prochaine sections 2.2.2 et 2.4.3 qu'une valeur de misfit élevée (par exemple supérieure à 1) peut par exemple indiquer un problème, causé lors de l'analyse de cas très atypiques pouvant provoquer la non-gaussianité d'un grand nombre éléments dans l'intervalle $[Q_A, Q_B]$.

2.2.2 Avantages et limites de la méthode

La méthode NG-loc, présentée et illustrée dans les sections 2.1 et 2.2.1 se révèle être un outil simple et efficace permettant de discriminer le signal de fond Gaussien des éventuelles altérations.. Un des atouts de NG-loc réside dans la polyvalence des perturbations détectées, permettant à la fois de localiser des altérations isolées impactant seulement un ou plusieurs points, mais aussi celles affectant un très grand nombre d'éléments comme montrées par exemple dans les figures 2.4(C) et 2.5(D, E et F). De

plus, cette discrimination est extrêmement précise et permet, de détecter individuellement chaque échantillon ne faisant pas partie du signal de fond Gaussien, dès lors que son amplitude est assez élevée. Il est également important de noter que l'analyse effectuée par NG-loc ne porte que sur la distribution en amplitudes d'une série de données indépendamment de sa moyenne et/ou de son écart type, permettant alors une application de notre méthode à n'importe quel signal. Par ailleurs, un des avantages principaux de NG-loc réside dans le fait que celle-ci est capable de discriminer le signal de fond Gaussien des éléments perturbés uniquement via la résolution d'un problème de minimisation. Par conséquent, cela permet de s'affranchir du choix d'un seuil arbitraire, comme souvent utilisé dans d'autres méthodes de détections (Taylor, 2006; Hamamoto et coll., 2018; Lu et Ghorbani, 2008).

NG-loc est une approche permettant la détection de n'importe quelle artefact venant altérer la gaussianité du signal, dès lors que celui-ci correspond à notre définition de perturbation. Cette méthode diffère des approches de type cross-corrélation (Derrick et Thomas, 2004), permettant de détecter un type précis de perturbation dans un signal, dont la forme d'onde est déjà connue au préalable. Cette différence tire son origine de notre choix de la caractérisation, non pas des perturbations, mais du bruit de fond du signal, ici considéré comme Gaussien. Connaissant en amont la nature du signal de fond recherchée, ceci permet alors de localiser les éléments perturbés et ce, même s'ils représentent une très large proportion du signal pouvant altérer jusqu'à 90% de ses éléments (voir la sous-section 2.4.3). Finalement, le domaine des applications de NG-loc est extrêmement large, la seule contrainte nécessaire à son application étant la présence d'un signal de fond Gaussien.

Même si la méthode NG-loc semble être une méthode efficace permettant de retrouver le signal de fond Gaussien, il est toutefois intéressant de discuter également des limites de celle-ci, illustrées par trois signaux synthétiques dans la figure 2.7 ($n = 10\,000$). Les exemples J et K présentent des signaux Gaussien affectés par une croissance graduelle de leurs moyennes et écarts types. Le niveau de base de J1 est altéré par une forte tendance linéaire tandis que l'amplitude de K1 croît progressivement en fonction de son indice. L'analyse de ces deux exemples sort du cadre de notre théorie, reposant

sur la recherche d'un signal de fond Gaussien dont les propriétés statistiques (μ et σ) sont supposées constantes. L'approche NG-loc n'est donc dans ce cas pas adaptée, car il n'existe aucune partie de ces signaux (au moins 10%) où la moyenne et l'écart type est constant. Par conséquent, l'application de NG-loc sur ces signaux entraîne une légère différence entre ϕ_n^{-1} et $\hat{\phi}^{-1}$, confirmée par un misfit de 1,20 pour J2 et 0,39 pour K2, marquant dans les deux cas une nette différence avec le misfit de 0,210 de l'exemple purement Gaussien de la figure 2.4(A).

Dans l'exemple J, on obtient alors $[Q_A, Q_B] = [1, 9999]$, excluant seulement un seul échantillon de la distribution normale. Cependant, il convient d'être prudent avec l'interprétation de ce résultat, dépendant de la nature de l'augmentation de la moyenne ainsi que de son amplitude (avec dans notre cas une augmentation linéaire de la moyenne faisant varier l'amplitude de 0 à 100). Le choix de l'intervalle $[Q_A, Q_B]$, recouvrant presque la totalité des rangs, a ici été causé car la linéarité des données triées, coïncidant avec celle de la fonction probit (en dehors de ses extrémités).

L'application de NG-loc dans l'exemple K2 conduit à des résultats différents avec $[Q_A, Q_B] = [45, 9948]$, impliquant que la distribution des données ne correspond pas à une loi normale. Une augmentation de l'amplitude du signal n'est cependant pas, en théorie, problématique. La difficulté vient ici principalement du fait que cette croissance de l'amplitude est ici graduelle et est observée sur la totalité du signal analysé, compliquant alors la localisation d'un signal de fond Gaussien. Par conséquent, il est difficile pour la méthode de minimisation de NG-loc de sélectionner un intervalle fiable $[Q_A, Q_B]$, où celui-ci discrimine clairement le signal de fond Gaussien des éléments perturbés. Bien que les signaux synthétiques traités dans les exemples J et K puissent paraître, en théorie, problématiques, ceux-ci ne constituent cependant pas un obstacle majeur dans l'application de la méthode présenté dans les chapitres 3 et 4. En pratique, des opérations de filtres appliquées au signal ainsi qu'un choix de taille de fenêtre analysée adaptée aux perturbations nous permettront d'éviter ces deux cas pathogènes.

Le signal L1 est obtenu par un tirage aléatoire Gaussien ($\mu = 0$, $\sigma = 10$), de la même manière que le tout premier exemple synthétique étudié dans la figure 2.4(A1). Cependant, l'intervalle quantile sélectionné par NG-loc ne recouvre pas l'intégralité des

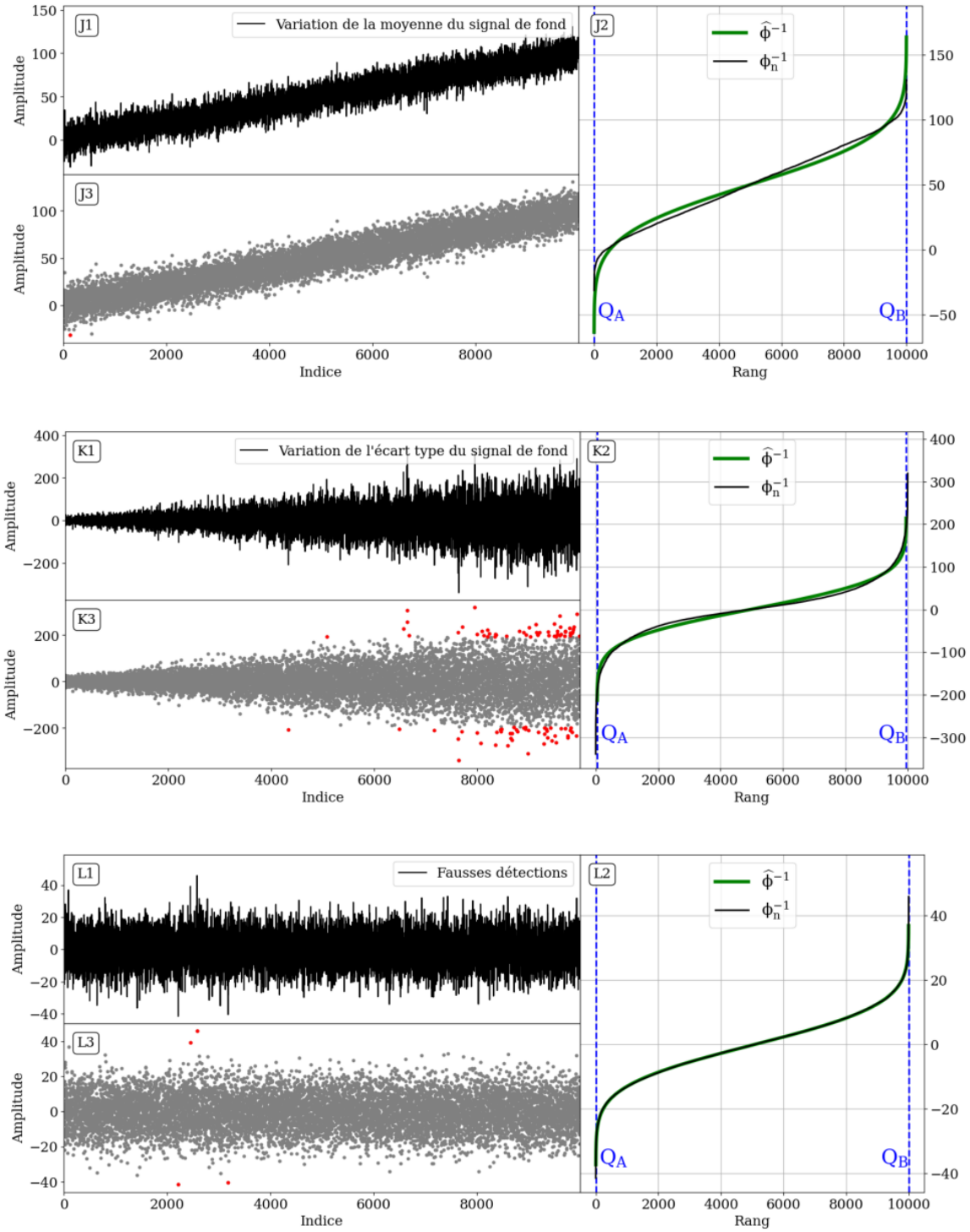

FIGURE 2.7 – Limites de la méthode NG-loc (même légende que pour la figure 2.4).

rangs, mais seulement $[Q_A, Q_B] = [2, 9997]$, excluant alors 4 éléments du signal de fond Gaussien (voir L3). Ce phénomène est la cause de fluctuations statistiques provenant du tirage aléatoire, pouvant se produire lorsque celui-ci est atypique et ne correspond pas exactement et intégralement à une distribution probit. Il est toutefois important de noter que le nombre de points considérés comme étant perturbés dans ce genre de cas représente une infime partie du signal, avec ici seulement 4 points perturbés sur 10 000.

2.3 Optimisation du temps de calcul : résolution du problème de minimisation par dichotomie

L'approche NG-loc, présentée dans ce chapitre, repose sur une exploration complète de l'espace des quantiles afin de trouver l'intervalle $[Q_A, Q_B]$, correspondant, au signal de fond Gaussien. Bien que cette exploration permette une discrimination à l'échantillon près, celle-ci nécessite toutefois de rechercher le plus petit misfit sur un très large panel d'intervalles de quantiles. En effet, la mise en pratique de NG-loc, telle que présentée dans la section 2.1, nécessite par exemple l'analyse de 40 504 500 d'intervalles quantiles différents pour $n = 10\ 000$ éléments. L'étude d'un tel nombre d'intervalles conduit alors, en pratique, à un temps de calcul d'environ 500 s. Bien que ce temps de calcul, trop important, puisse être contourné en choisissant par exemple une exploration moins fine de l'espace des quantiles, une approche de résolution de notre problème de minimisation par dichotomie a été développée, permettant de concilier vitesse et précision. De plus, une utilisation rapide et efficace de NG-loc se révèle être une condition incontournable dans le cadre des applications à grande échelles qui seront présentées dans les prochains chapitres 3 et 4.

De manière théorique, notre problème de minimisation consiste, pour un signal de taille n fixé, à trouver les entiers Q_A et Q_B minimisant le misfit défini par l'équation

2.2. De plus, on impose également les conditions suivantes :

$$\begin{aligned} Q_A &< Q_B \\ Q_A &\geq 0 \text{ et } Q_B \leq n-1 \\ Q_B - Q_A &\geq \frac{n}{10} \quad \text{(taille minimale de la fenêtre).} \end{aligned}$$

Par la suite, nous étudions ce problème de minimisation, non pas par la définition d'une fenêtre via ses extrémités Q_A et Q_B, mais plutôt, de manière équivalente, par son centre q et sa longueur dq. L'objectif de notre approche par dichotomie est d'analyser dans un premier temps, de manière grossière, l'espace des paramètres, puis d'affiner progressivement la recherche de la solution. En d'autres termes, l'idée consiste en premier lieu à trouver une solution approchée q_0 et dq_0 de notre problème avec une précision donnée (par exemple $p_0 = 100$). Sachant désormais que notre solution q, dq appartient respectivement aux intervalles $[q_0 - p_0, q_0 + p_0]$ et $[dq_0 - p_0, dq_0 + p_0]$, nous concentrons alors notre recherche sur cet intervalle, mais pour précision plus fine (par exemple $p_1 = 10$), nous permettant de réduire notre zone de recherche. Finalement, en répétant l'opération avec $p_2 = 1$, on obtient alors la solution q, dq, et donc l'intervalle $[Q_A, Q_B]$ correspondant. Les entiers $P_l = [p_0, p_1, p_2]$ définissent ici une « liste de précisions », que l'on peut généraliser au cas où celle-ci est composée de N_p éléments. P_l est donc toujours strictement décroissante, composée d'entiers, et dont le dernier élément est égal à 1, permettant ainsi à NG-loc de fournir la solution la plus précise possible.

La figure 2.8 illustre l'application de NG-loc, utilisant la recherche de minimum par dichotomie sur un tirage aléatoire Gaussien de $n = 10\ 000$ éléments, avec $P_l = [100, 10, 1]$. Le résultat attendu de la méthode est donc, dans ce cas Gaussien, l'obtention d'une fenêtre $[Q_A, Q_B]$ recouvrant l'intégralité des rangs, correspondant alors à $q = 5000$ et $dq = 10\ 000$. Les trois étapes d'explorations de l'espace des paramètres associées aux précisions $p = 100, 10$ et 1 sont représentées sur la figure 2.8 (en haut, à gauche et à droite, respectivement), où le code couleur de chaque couple (q,dq) analysé correspond à sa valeur de misfit associée. À noter que l'espace des paramètres se présente sous la forme d'un triangle, et non d'un simple carré, car il est impos-

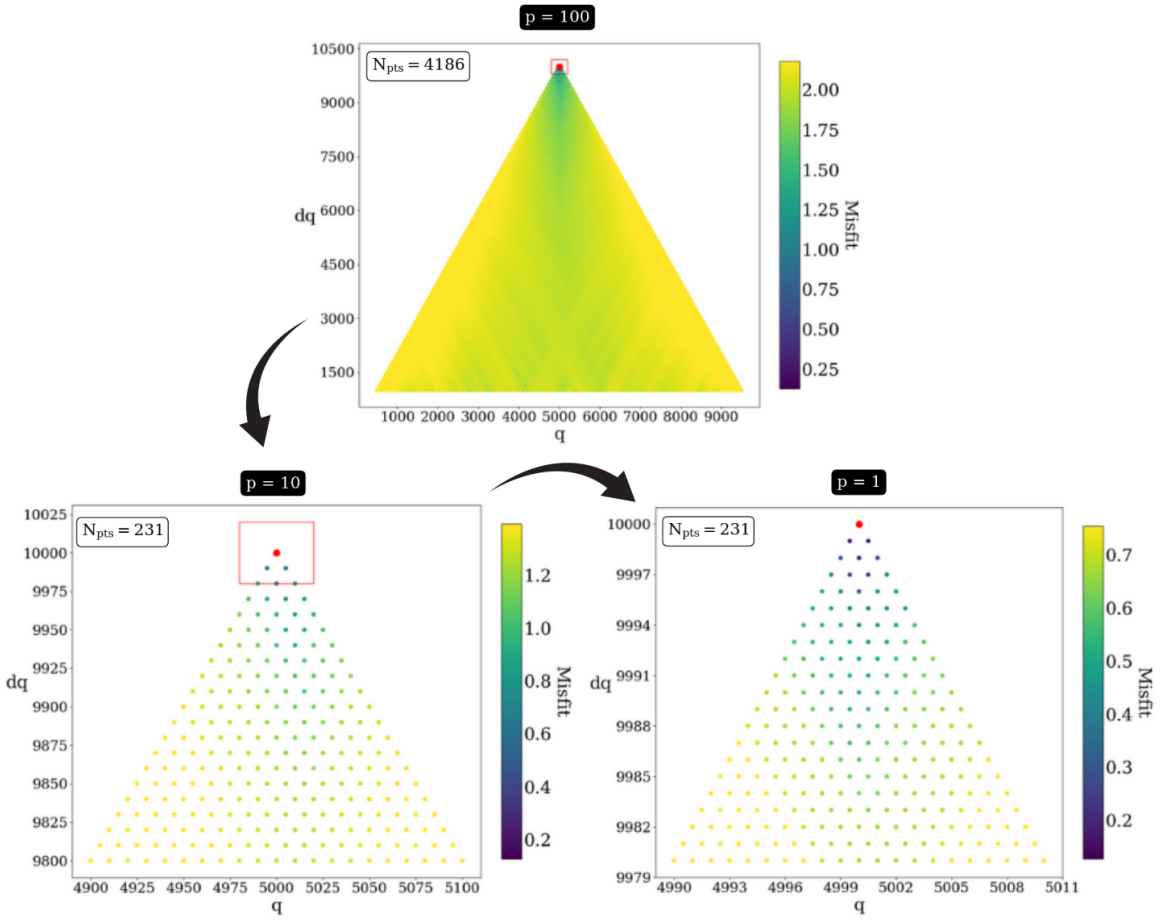

FIGURE 2.8 – Exemple d'application de NG-loc par résolution du problème de minimisation via la méthode de dichotomie. Approximations successives de la solution pour une précison $p = 100$ (en haut), $p = 10$ (en bas à gauche), puis $p = 1$ (en bas à droite).

sible d'analyser une taille de fenêtre trop grande, si son centre q est proche du bord (correspondant aux respects des conditions $Q_A \geq 0$ et $Q_B \leq n - 1$). On remarque en observant le triangle associé à $p = 100$, que la localisation d'un minimum situé sur son sommet semble être assez évidente, comme indiquée par les valeurs de misfits, nettement plus faibles dans cette zone d'étude (points bleu foncés). Par conséquent, cette étape conduit à une première approximation $q_0 = 5000$ et $dq_0 = 10\ 000$ de la solution, indiquée par le point rouge au sommet du triangle. La zone d'étude sur laquelle se concentrera la prochaine étape est alors représentée par le carré rouge, autour du couple (q_0, dq_0). À noter que, par mesure de précaution, nous choisissons d'effectuer une exploration sur $[q_0 - 2p_0, q_0 + 2p_0]$ et $[dq_0 - 2p_0, dq_0 + 2p_0]$ (incluant évidemment $[q_0 - p_0, q_0 + p_0]$ et $[dq_0 - p_0, dq_0 + p_0]$). La seconde étape $p = 10$ sélectionne une nou-

velle fois $q_1 = 5000$ et $dq_1 = 10\ 000$ comme solution approchée associée au plus faible des misfits analysés et affine donc encore l'étude sur la pointe du triangle. À noter ici que l'intensité du code couleur n'est cependant pas la même à chaque itération : tous les misfit dans cette zone d'étude sont désormais inférieurs à 1,4 (contre 2,2 lors de l'étape $p = 100$). Finalement, la dernière étape $p = 1$ sélectionne alors pour solution finale $(q, dq) = (5000, 10\ 000)$. Dans ce cas précis, l'étape $p = 100$ ayant déjà proposé l'analyse de $(q, dq) = (5000, 10\ 000)$ (car tous deux multiples de 100), la première approximation de la solution correspond alors également à la solution finale. En pratique, la valeur du centre de la fenêtre q pourra être un nombre décimal, permettant alors de tester n'importe quel couple d'entier (Q_A, Q_B).

Cette recherche de minimum par dichotomie est donc une manière performante de résoudre notre problème de minimisation, sans perte de précision. En effet, la figure 2.9 illustre l'application brute de NG-loc, sans recourir à notre optimisation, aboutissant au même résultat $(q, dq) = (5000, 10\ 000)$. Cependant, son utilisation est néanmoins laborieuse et requiert un très grand nombre d'itérations : $N_{iter} = 40\ 504\ 500$, contre un nombre total de $N_{iter} = 4186+231+231 = 4648$ avec la méthode de dichotomie (voir les figures 2.8 et 2.9). Par conséquent, cette méthode d'optimisation entraîne naturellement un énorme gain de temps de calcul, et permet d'effectuer l'analyse de cette fenêtre de $n = 10\ 000$ points en 0,087 secondes, contre 500 s auparavant, correspondant alors à une vitesse d'exécution 5000 fois plus rapide.

Finalement, il est également important de discuter de la pertinence du choix de la recherche d'un minimum global par dichotomie dans le cadre de notre étude. En effet, certains cas de problèmes d'optimisations complexes peuvent présenter de multiples minimums locaux, ou une forte instabilité de la solution. Cependant, notre étude suppose l'unicité du signal de fond Gaussien, impliquant alors également l'unicité de la solution recherchée. De plus, le misfit à l'intérieur de l'espace des paramètres (voir figure 2.9) ne subit pas d'énormes variations rapides au sein du triangle (grâce à la continuité de la norme L^∞), et une très nette différence est observée entre le misfit finalement obtenu ($M = 0{,}127$), et le reste du triangle (de couleur jaune, associé à $M \geq 2$). Finalement, de nombreuses observations des résultats fournis par NG-loc

avec, et sans dichotomie, nous ont permis de constater une correspondance presque parfaite des solutions obtenues dans les deux cas.

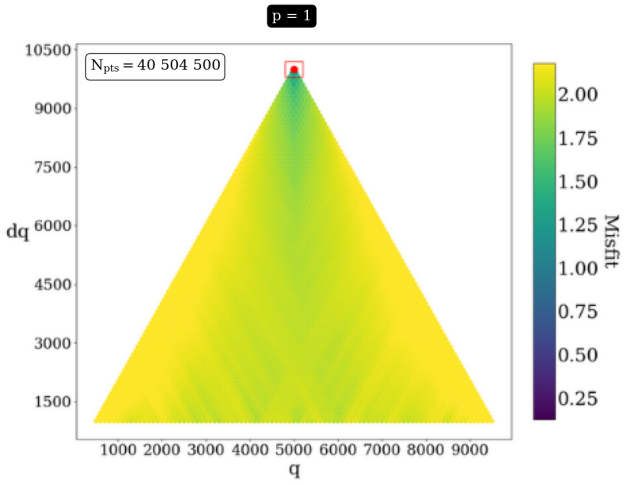

FIGURE 2.9 – Application de NG-loc, sans recourir à la recherche de minimum par dichotomie.

2.4 Discussions générales sur la méthode NG-loc

2.4.1 Singularités de NG-loc par rapport aux autres méthodes de détections

Afin de justifier des contributions apportées par NG-loc, il est essentiel de discuter des différentes méthodes de détections existantes permettant, pour une série de données, i) de localiser la déviation statistique de certains de ses éléments ou ii) d'attester de sa normalité. Concernant le premier point, plusieurs méthodes de *Change-Point Analysis* ont été développées (Taylor, 2006; Page, 1957), permettant la localisation des modifications soudaines de la statistique d'un signal. Bien que ces méthodes soient efficaces pour repérer des changements subtils modifiant la statistique d'une série de données, elles ne sont cependant pas capable de localiser un unique point atypique et se révèlent dépendantes du choix d'un seuil fixé par l'utilisateur.

L'intérêt porté à l'étude de la loi normale (justifié par le théorème limite central 1.2.1) a conduit à un effort scientifique conséquent, nous proposant alors de nombreux outils pour attester ou non de la normalité d'une série statistique. Commençons tout d'abord par la manière la plus triviale de vérification de la gaussianité d'une série de données, consistant en une simple comparaison entre son histogramme et la célèbre courbe en cloche définissant la fonction de densité d'une loi normale (voir figure 2.10). Bien que cette première approche permette une visualisation intéressante de la distribution des données aboutissant à une prise de décision sur sa normalité, ou non, dans des cas simples (A étant ici Gaussien, contrairement à B), la définition d'un histogramme est cependant problématique. En effet, la forme de l'histogramme repose notamment sur le choix d'une discrétisation arbitraire des amplitudes en un nombre fini de classes, ne permettant alors pas une analyse précise des données.

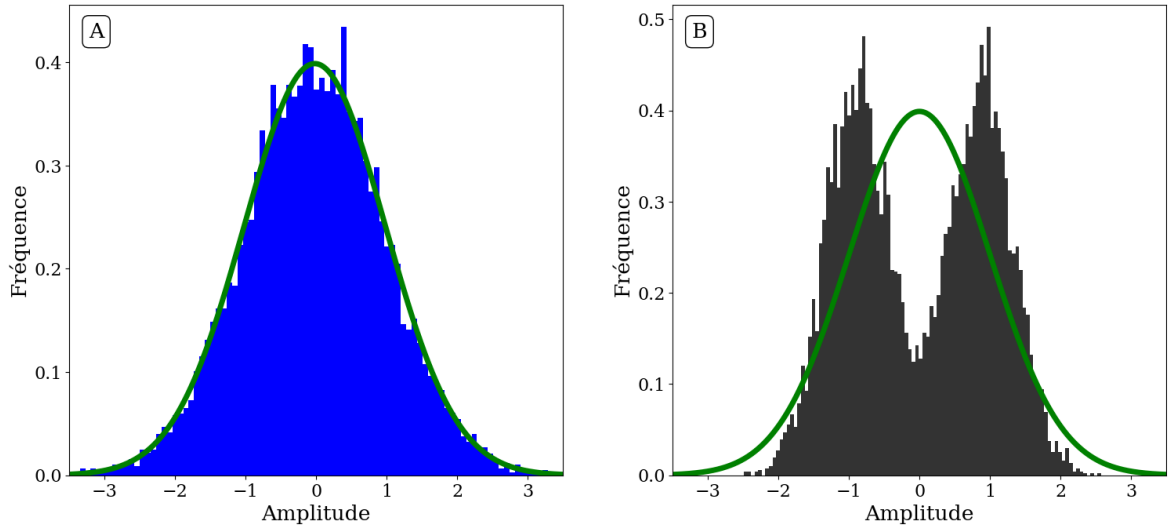

FIGURE 2.10 – Observation de l'histogramme normalisé de deux séries de données (A et B). Pour chaque exemple, la moyenne/variance du signal associée à l'histogramme est calculée, permettant d'afficher la courbe Gaussienne correspondante (en vert), définie par la fonction de densité de la loi normale (eq. 1.1)

Une méthode différente d'observation repose sur la visualisation de la droite de Henry (Filliben, 1975), en comparant les quantiles d'une série statistique avec sa distribution théorique. Cette approche utilise des outils similaires à ceux que nous exploitons dans la méthode NG-loc, en recourant notamment à la comparaison entre la fonction Probit et les données triées, mais se limite cependant à une simple visualisation de

la correspondance à la normalité et ne permet ni de la quantifier, ni de localiser les échantillons non Gaussiens.

Au 20ème siècle, les définitions de « coefficient d'asymétrie » (moment d'ordre 3) (Pearson, 1895) et « coefficient d'aplatissement » (moment d'ordre 4) ont alors ouvert de nouvelles portes permettant de tester la normalité d'une série statistique. En effet, il a été démontré que ces deux coefficients se révèlent être des outils efficaces permettant de tester l'hypothèse de gaussianité d'une série de données, comme démontré dans une série d'articles de Pearson (1930, 1931, 1963, 1965), mais aussi par Williams (1935) et Fisher (1930a,b). La nécessité d'attester de manière probabiliste de la normalité d'une série statistique a alors mené au développement de nombreux « tests d'adéquations » (Emmert-Streib et Dehmer, 2019). Parmi les plus célèbres, on pourra par exemple citer la méthode du χ^2 (Pearson, 1900; Plackett, 1983), celle de Shapiro-Wilk (Shapiro et Wilk, 1965, 1968; Drezner et coll., 2010) ou encore la méthode de Kolmogorov-Smirnov (Kolmogorov, 1933) (voir Yazici et Yolacan (2007) pour une comparaison des différents tests d'adéquations).

Cependant, ces tests d'adéquation, de par leur nature, permettent uniquement de conclure sur la normalité ou non des données, avec une certaine « probabilité de confiance » nommée p-valeur (fixée la plupart du temps à 95 ou 99%). En comparaison, notre approche NG-loc ne se limite pas seulement à répondre à la question de la normalité des données, mais permet également de localiser individuellement les éléments non Gaussiens. Cette nouvelle information fournie par notre approche, associée à sa précision « au point près » et sa vitesse d'exécution, font de NG-loc une méthode singulière et efficace, n'ayant à notre connaissance pas d'équivalent dans la littérature. Au vu des différents atouts de NG-loc, nous sommes alors persuadés que celle-ci peut apporter des contributions intéressantes dans le domaine de la détection de séries de données non Gaussiennes, en complément des informations fournies par les méthodes existantes.

2.4.2 Nombre minimal d'éléments analysés

La méthode NG-loc repose sur l'analyse d'un signal donné, de longueur n. Bien que la taille maximale d'un signal étudié pourrait être, en théorie, infiniment grande, il convient cependant de discuter du nombre minimal d'éléments pouvant être analysé par notre approche. Dans ce contexte, nous nous posons alors la question suivante :

Quel est le nombre minimal d'éléments analysés, conduisant à un résultat fiable de NG-loc ?

Bien que cette question soit évidemment sujette à interprétation, de par la subjectivité de la notion de « fiabilité », nous apportons ici des éléments de réponses, par application de NG-loc sur des signaux Gaussiens de tailles variées ($10 \leq n \leq 2000$). L'idée consiste ici à étudier le résultat de l'application de NG-loc sur chacun de ces signaux de taille n, sachant que le résultat attendu dans le cas Gaussien est $[Q_A, Q_B] = [0, n-1]$. Pour chaque résultat, l'analyse se porte sur le ratio $\mathcal{G}$, représentant le nombre de points faisant partie du signal de fond Gaussien divisé par le nombre total d'éléments du signal étudié. Son expression mathématique est donc donnée par

$$\mathcal{G} = \frac{Q_B - Q_A + 1}{n}.$$

Le résultat attendu pour un tel ratio est donc $\mathcal{G} = 1$, lorsque le signal analysé suit une distribution normale.

La figure 2.11 présente les résultats de ce ratio $\mathcal{G}$ pour chacun des tirages aléatoires Gaussien de taille $10 \leq n \leq 2000$. On remarque que $\mathcal{G}$ est très instable lorsque $n \leq 250$, et atteint régulièrement des valeurs comprises entre 0,1 et 0,9, causées par l'étude de signaux composés d'un nombre insuffisant d'éléments. Lorsque n est compris entre 250 et 600, on observe une convergence de $\mathcal{G}$ vers 1, même si celui-ci contient toujours quelques points s'éloignant fortement de cette valeur. Finalement, lorsque n est supérieur à 600, $\mathcal{G}$ devient alors remarquablement stable et reste très proche de 1 (supérieur à 0,984 lorsque $n \geq 600$, et à 0,992 lorsque $n \geq 1000$). Au vu de ce résultat, et par mesure de précaution, nous préconisons alors de privilégier l'étude d'un signal

composé au minimum de 1000 éléments (représenté par la ligne verticale rouge sur la figure 2.11). Ce nombre minimal d'éléments permet donc d'obtenir des résultats fiables de NG-loc, avec une proportion d'erreur inférieure à 1% (égale à 0,8% exactement).

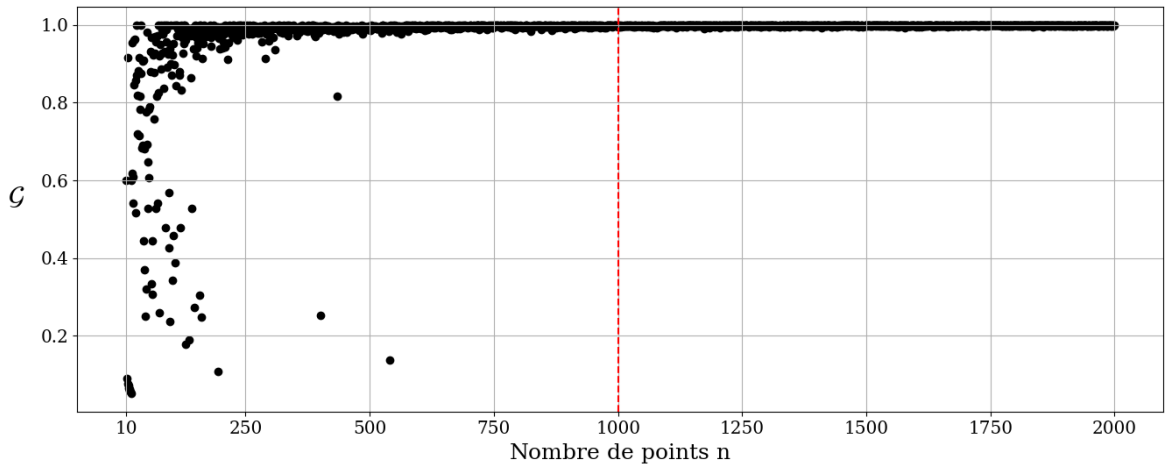

FIGURE 2.11 – Application de NG-loc sur des tirages aléatoires Gaussiens de tailles différentes ($10 \leq n \leq 2000$). Pour chaque application, le ratio $\mathcal{G}$, définit le nombre d'éléments appartenant à $[Q_A, Q_B]$ divisé par n.

2.4.3 Proportion maximale d'éléments perturbés dans le signal

Une des forces de NG-loc réside dans sa capacité à détecter une grande proportion de perturbations dans un signal Gaussien, pouvant altérer jusqu'à 90% de ses éléments. Cette affirmation n'étant jusqu'à présent pas démontrée, nous proposons d'illustrer cette propriété intéressante sur un cas simple dans la figure 2.12, et étudions également le comportement de NG-loc lorsque ce seuil de 90% est dépassé.

L'exemple A1 présente un signal de $n = 10\ 000$ éléments, composé de fonctions en escaliers, ayant cependant une distribution purement Gaussienne entre les indices 4500 et 5499. Par conséquent, ceci représente le cas d'un signal où 90% de ses éléments sont perturbés, et l'application de NG-loc est alors capable de retrouver son signal de fond Gaussien, et sélectionne exactement $[Q_A, Q_B] = [4500, 5499]$ (voir A2). Notre approche est donc ici efficace, et permet de discriminer dans A3 chacun des éléments (en rouge), ne faisant pas partie de la distribution Gaussienne.

Le signal B1 est semblable à A1, mais la proportion d'éléments Gaussien est désormais réduite entre les indices 4750 et 5249, représentant alors seulement 5% du signal total. Nous sortons ici du cadre de la méthode, telle que nous l'avons définie dans la section 2.1, et NG-loc n'est donc plus capable de retrouver le signal de fond Gaussien et propose un intervalle $[Q_A, Q_B]$ incluant des perturbations (voir B2 et B3). De plus, la mauvaise correspondance entre ϕ_n^{-1} et $\hat{\phi}^{-1}$ est aussi nettement visible en inspectant le misfit, étant ici égal à 1,24, ce qui est donc très nettement supérieur à tous ceux étudiés dans le tableau 2.1.

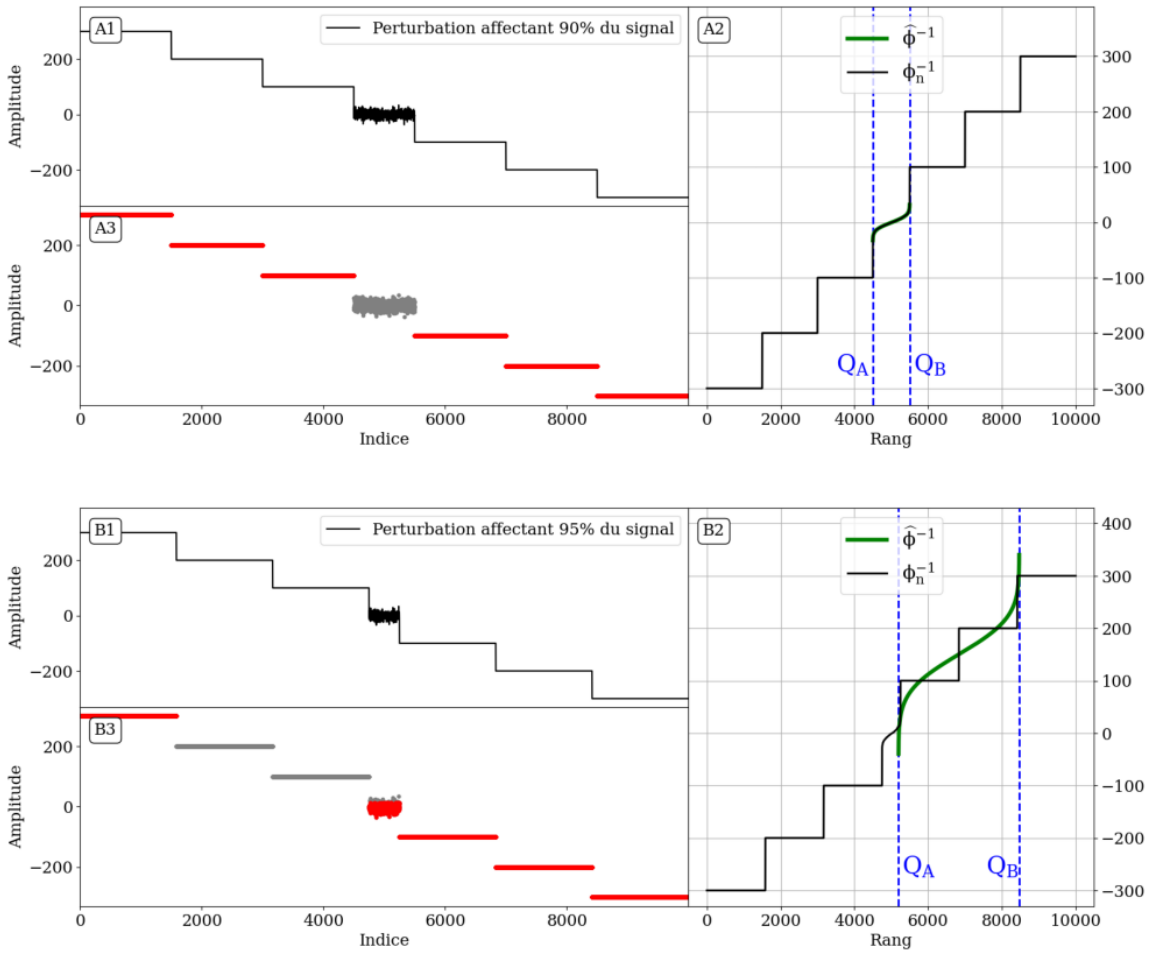

FIGURE 2.12 – Illustration des limites de NG-loc, lorsque la perturbation représente 90% et 95% du signal dans les exemples A (en haut) et B (en bas), respectivement (même légende que pour la figure 2.4).

En conclusion, nous avons justifié, par un exemple, la capacité de NG-loc à retrouver la distribution Gaussienne d'un cas synthétique extrême, lorsque celle-ci ne représente

que 10% du signal (A). De plus, l'exemple B nous indique que cette limite de 90% de perturbations ne doit en aucun cas être dépassée et caractérise alors un seuil brut au delà duquel la méthode NG-loc n'est plus adaptée.

2.4.4 Définition du misfit : le choix de la norme L^∞

Le misfit est un paramètre central de la méthode NG-loc, permettant la décision menant à la discrimination entre le signal de fond Gaussien et les éléments perturbés. Il est alors évident que le misfit joue un rôle déterminant dans notre approche et que sa définition se doit donc d'être justifiée avec soin. Dans l'algorithme présentant NG-loc (section 2.1), le misfit a été défini par la norme L^∞ de la différence entre ϕ_n^{-1} et $\hat{\phi}^{-1}$. Bien que le choix d'étudier la différence entre ces deux termes semble évident afin de retranscrire de la correspondance des courbes, la décision de la norme L^∞ mérite cependant d'être justifiée par des arguments solides.

Dans cette section, nous comparons les résultats de NG-loc, obtenus via des misfits calculés par des normes L^1, L^2 et L^∞. Rappelons tout d'abord la définition de ces normes pour des vecteurs $x = (x_1, ..., x_n) \in \mathbb{R}^n$:

$$\|x\|_{L^1} = \sum_{i=1}^{n} |x_i| \tag{2.3}$$

$$\|x\|_{L^2} = \Big(\sum_{i=1}^{n} |x_i|^2\Big)^{\frac{1}{2}} \tag{2.4}$$

$$\|x\|_{L^\infty} = \max_{1\leq i\leq n}\Big(|x_i|\Big). \tag{2.5}$$

La figure 2.13, compare les résultats de NG-loc, obtenus dans les cas de la définition du misfit par des normes L^1, L^2 et L^∞. Nous étudions dans cet exemple 1000 signaux Gaussiens synthétiques de mêmes tailles ($n = 10\ 000$), où chacun d'entre eux est alors analysé 3 fois par la méthode NG-loc, pour des normes différentes. Par conséquent, le résultat de chaque analyse nous permet d'obtenir un intervalle de quantiles $[Q_A, Q_B]$, dont la valeur attendue est alors, dans ces cas Gaussiens, de $[Q_A, Q_B] = [0, n-1 = 9999]$. En remarquant ceci, nous affichons dans la figure 2.13, la valeur $n - (Q_B -$

$Q_A + 1$), étant alors supposée être égale à zéro. Cependant, nous observons quelques différences entre les 3 choix de normes définissant le misfit. En effet, les résultats associés à la norme L^1 sont bien supérieurs à ceux des autres normes, avec une moyenne de 7,55 tandis qu'elles valent 2,13 et 1,77 pour les normes L^2 et L^∞, respectivement. Ce résultat indique alors que le choix de la norme L^∞ semble être le plus intéressant, car sa fenêtre $[Q_A, Q_B]$ est plus régulièrement susceptible de recouvrir l'intégralité des points, désignant alors l'entièreté du signal comme étant Gaussien.

Il est tout de même important de constater que, même si cette illustration favorise le choix de la norme L^∞ pour une utilisation plus précise de NG-loc, les résultats des 3 normes sont tout de même très satisfaisants. En effet, les résultats affichés dans la figure 2.13 sont toujours inférieurs à 60 points (pour toutes les normes), et cette quantité est pourtant presque négligeable au vu de l'analyse de la fenêtre totale représentant ici $n =$ 10 000 échantillons. Cette erreur maximale de 60 points représenterait alors seulement une erreur de discrimination de 0,6% des échantillons. En choisissant la norme L^∞, cette erreur est alors réduite à une proportion encore plus infime, représentant en moyenne 0,0177% du signal total.

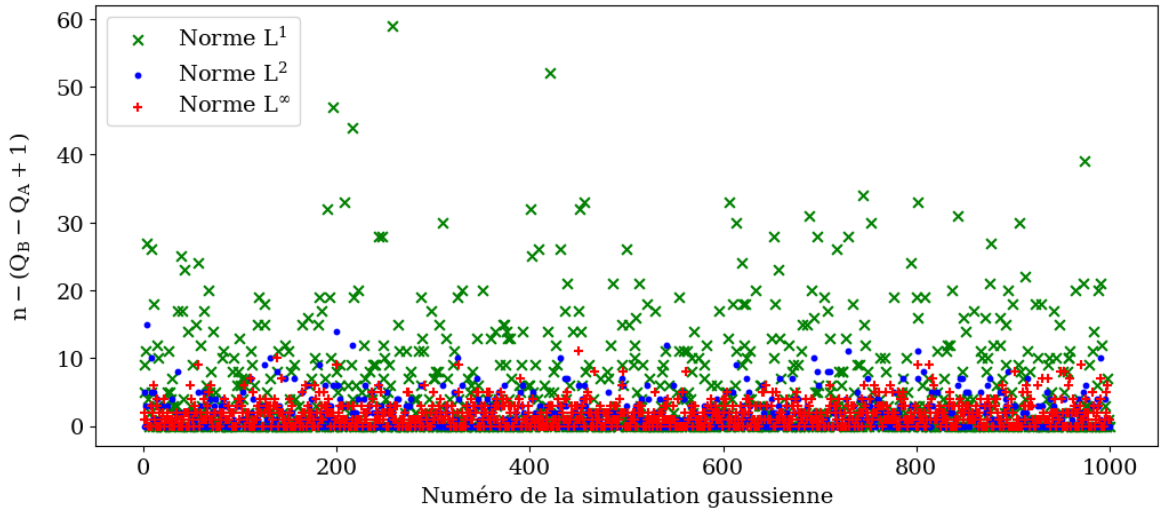

FIGURE 2.13 – Influence du choix de la norme du misfit sur le résultat de NG-loc, appliqué à 1000 signaux Gaussiens de tailles $n = 10\ 000$. Chaque signal est analysé 3 fois, pour des définitions de misfit utilisant des normes différentes (L^1, L^2 et L^∞). Le résultat $n - (Q_B - Q_A + 1)$ obtenu pour chaque norme est finalement affiché en remarquant que celui-ci est supposé être égal à zéro dans le cas purement Gaussien.

La définition du misfit par la norme L^∞ est en réalité un avantage majeur, car elle permet à NG-loc une sensibilité « au point près ». En effet, supposons qu'on ajoute à un signal Gaussien un unique élément perturbé, de très grande amplitude A, disons 10 fois plus grande que son maximum noté M. Dans ce cas, si l'intervalle de rang sélectionné lors de la recherche de $[Q_A, Q_B]$ inclut ce point, alors le misfit (en norme L^∞) sera donc égal à $A - M$. Finalement, les fenêtres de rangs contenant cet élément perturbé seront alors associées à un misfit très élevé, excluant donc ce point de grande amplitude dans la sélection finale de l'intervalle $[Q_A, Q_B]$. *À contrario*, si l'on choisit une norme L^1 ou L^2, leurs définitions (eqs. 2.3 et 2.4) impliquent que cette amplitude n'influencera alors qu'un seul terme de la somme. Par conséquent, le résultat sera donc évidemment affecté par ce point de grande amplitude, mais dans une moindre mesure qu'avec la norme L^∞.

En conclusion, au vu de ce raisonnement et des résultats de la figure 2.13, nous choisissons alors d'opter pour une définition du misfit via une norme L^∞. De plus, ce choix est également cohérent avec notre théorie mathématique énonçant (dans le cas Gaussien) la convergence uniforme de la fonction quantile empirique vers la fonction probit modifiée.

2.4.5 Commentaire sur l'étude de la statistique ordonnée d'un signal

Afin de retrouver une distribution Probit dans une série de donnée, la méthode NG-loc analyse un signal dont les éléments sont triés par ordre croissant d'amplitude. L'étude d'un signal trié étant toutefois une manière atypique d'analyser une série de donnée, nous discutons dans cette section des conséquences d'une tel choix. D'un point de vue mathématique, l'opération de tri consiste simplement en une bijection discrète, réorganisant les indices du signal $(X_i)_{1\leq i\leq n}$ en une nouvelle série de données $(X_{(i)})_{1\leq i\leq n}$, telles que $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$. Ceci permet alors de définir, ce qu'on appelle une « relation d'équivalence » entre deux séries de données $X = (X_i)_{1\leq i\leq n}$ et $Y = (Y_i)_{1\leq i\leq n}$ ayant exactement les mêmes signaux triés. On notera alors $X \sim Y$ (c'est-à-dire « X

est équivalent à Y ») si leurs signaux ordonnés sont égaux (voir Ramis et coll. (2013) pour la définition mathématique d'une relation d'équivalence).

La figure 2.14 présente 4 séries de données (A, B, C et D), dont la particularité réside dans le fait que chacune d'entre elles, une fois triée par ordre croissant, permet d'obtenir le même signal E. Ces 4 signaux, bien que très différents au regard de leurs ordres d'origines, sont donc équivalents, au sens de notre définition. Par conséquent, l'application de NG-loc sur ces 4 signaux aboutira alors exactement au même résultat. La méthode NG-loc « perd » donc l'information de l'ordre de base (indice) d'un signal et analyse uniquement la distribution triée des données (rang). Finalement, ceci illustre alors une propriété importante de NG-loc résidant dans le fait que celle-ci est alors totalement indépendante de l'ordre original du signal.

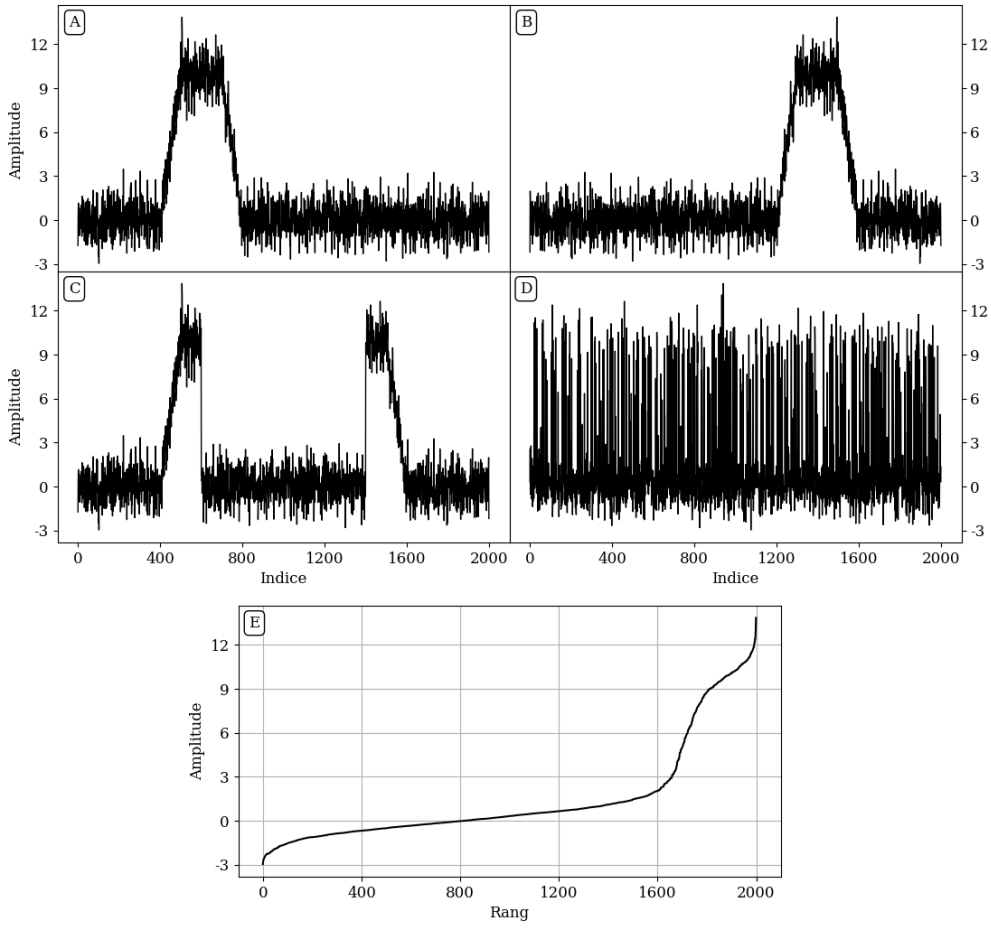

FIGURE 2.14 – Illustration de 4 signaux synthétiques différents A, B, C et D, correspondant tous à la même distribution E, une fois ceux-ci triés par ordre croissant.

Chapitre 3

Applications : Signal sismique terrestre

Sommaire

Nous proposons dans ce chapitre, une application de la méthode NG-loc à un type de série de données particulier : le signal sismique. La section 3.1 introduit tout d'abord la méthode d'acquisition du signal sismique ainsi que ses différentes caractéristiques, aboutissant à l'hypothèse de la distribution Gaussienne de son bruit de fond. La section 3.2 se concentre ensuite sur une application de notre approche, permettant d'estimer la qualité du signal sismique enregistré sur plusieurs stations. Cette application, centrale dans ce chapitre, fera alors l'objet de la présentation de l'article *Seismic Station Quality Control using Deviation from the Gaussianity*. Finalement, nous détaillerons dans la section 3.3 quelques applications supplémentaires de NG-loc au signal sismique. En pratique, les opérations de traitement du signal sont effectuées sur Python (Van Rossum et Drake, 2009), en utilisant la bibliothèque Obspy (Beyreuther et coll., 2010; Megies et coll., 2011; Krischer et coll., 2015), permettant une manipulation simple des données sismiques enregistrées au format standard MiniSEED.

3.1 Caractéristiques et traitement du signal sismique

3.1.1 Le signal sismique

L'analyse du signal sismique, enregistré par un sismomètre (figure 3.1), revêt une importance cruciale dans l'étude et la compréhension des mouvements de la Terre. Ces instruments sont conçus pour capter les ondes sismiques, qu'elles proviennent de tremblements de Terre, d'activités volcaniques, d'autres phénomènes géologiques ou environnementaux, ou encore de l'activité humaine. La mesure et l'interprétation de ces signaux jouent un rôle essentiel dans la surveillance de l'activité sismique et la recherche en géophysique. Un sismogramme est constitué de 3 composantes, représentant le mouvement du sol (position, vitesse ou accélération) dans les trois directions de l'espace : l'axe vertical Z, ainsi que les deux axes horizontaux Est et Nord.

En pratique, l'acquisition du signal sismique continu est effectuée à une certaine fréquence d'échantillonnage, représentant le nombre de points par seconde enregistrés. Le signal sismique se présente sous la forme d'une série temporelle, illustrée par la figure

3.2, représentant la vitesse du sol enregistrée par la station MTNF (Montenay, France) le 28 mai 2023. Ce signal se distingue sur les trois composantes par une oscillation régulière d'amplitude quasi constante au cours de la journée, ne comportant ici pas d'événement sismique évident.

FIGURE 3.1 – Installation d'un sismomètre en surface (modèle T120QA) provenant de la station BOUF, situé à Bouguenais, France.

À noter que chaque signal est associé à un canal : HHE (A), HHN (B) et HHZ (C), identifié par trois lettres, selon la norme SEED (Seed Reference Manual, 2012). La première lettre caractérise la fréquence d'échantillonnage (H si celle-ci est comprise entre 80 et 250 Hz, B lorsqu'elle se situe entre 10 et 80 Hz), tandis que la deuxième relate le type d'instrument utilisé : H pour un vélocimètre et N s'il s'agit d'un accéléromètre. La troisième lettre correspond quant à elle à la composante étudiée, Z, N ou E. Dans notre cas, la figure 3.2 correspond donc à des enregistrements de la vitesse du sol, acquise à une fréquence d'échantillonnage de 100 points par secondes.

Certaines arrivées d'ondes contenues dans le signal sismique étant parfois difficilement observables sur le signal sismique brut, il est parfois plus intéressant d'analyser son

contenu fréquentiel, par exemple via l'étude de son spectrogramme. La création d'un spectrogramme repose sur la transformée de Fourier, qui permet d'extraire le contenu fréquentiel d'un signal. L'application de transformées de Fourier sur des fenêtres glissantes permet alors d'obtenir le spectrogramme d'un signal sismique, caractérisant son énergie au cours du temps.

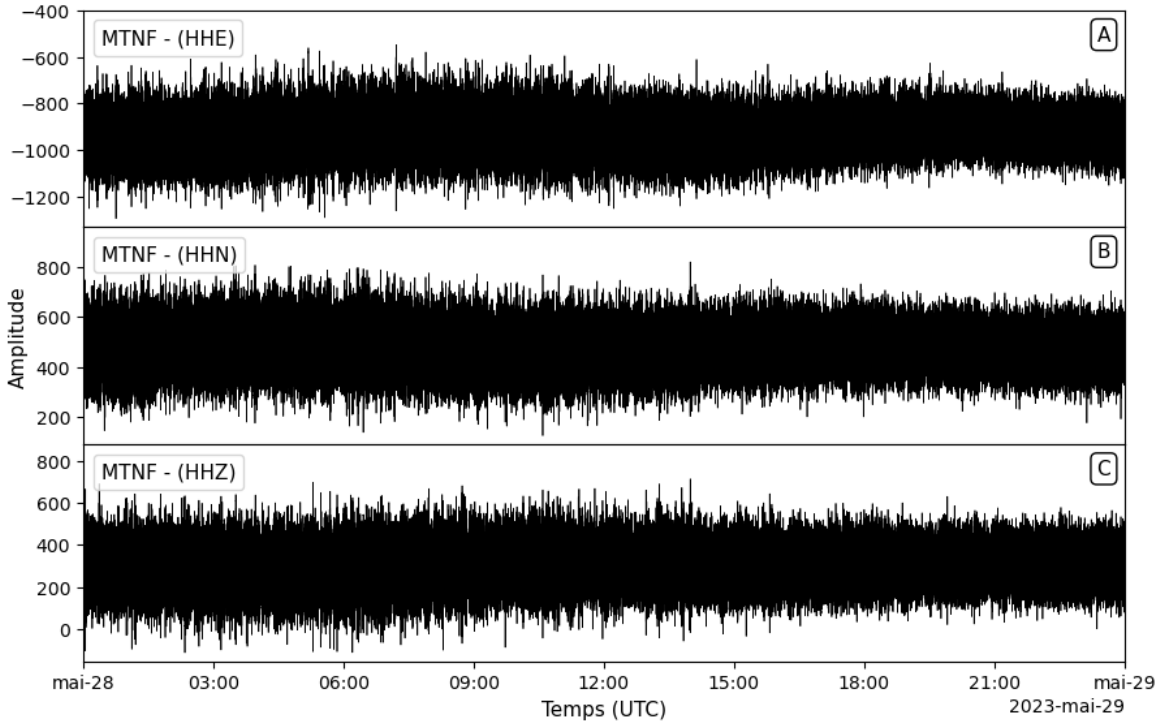

FIGURE 3.2 – Exemple de signal sismique enregistré lors de la journée du 28 mai 2023, sur les trois composantes de la station MTNF (Montenay, France).

La figure 3.3 présente le spectrogramme de la station MTNF (HHN), au cours de l'année 2022. Une caractéristique marquante de ce spectrogramme est la présence d'une forte énergie, entre 0,05 et 1 Hz. Cette arrivée d'énergie, connue sous le nom de « pics microsismiques », est commune à toute les stations terrestres, résultant de l'influence de l'océan sur le signal sismique (voir par exemple Ebeling (2012) et Beucler et coll. (2015) pour une analyse détaillée des pics microsismiques). Par ailleurs, le contenu fréquentiel se trouve également régulièrement influencé par les tremblements de Terre. Bien que les séismes affectent une large gamme de périodes, ces derniers sont notamment visibles sur les basses fréquences (pour les plus puissants), se démarquant alors de la faible énergie enregistrée en dessous de 0,05 Hz (voir par exemple la flèche noire dans la

figure 3.3). Finalement, l'emplacement de la station joue également un rôle clef sur la qualité du signal enregistré. En effet, l'activité humaine génère diverses sources de vibrations pouvant influencer le signal sismique, en particulier sur les hautes fréquences ($\geq$ 10 Hz). Parmi celles-ci, on peut citer l'activité agricole, la circulation routière, les constructions d'infrastructures, ou encore les rassemblements de grandes ampleurs.

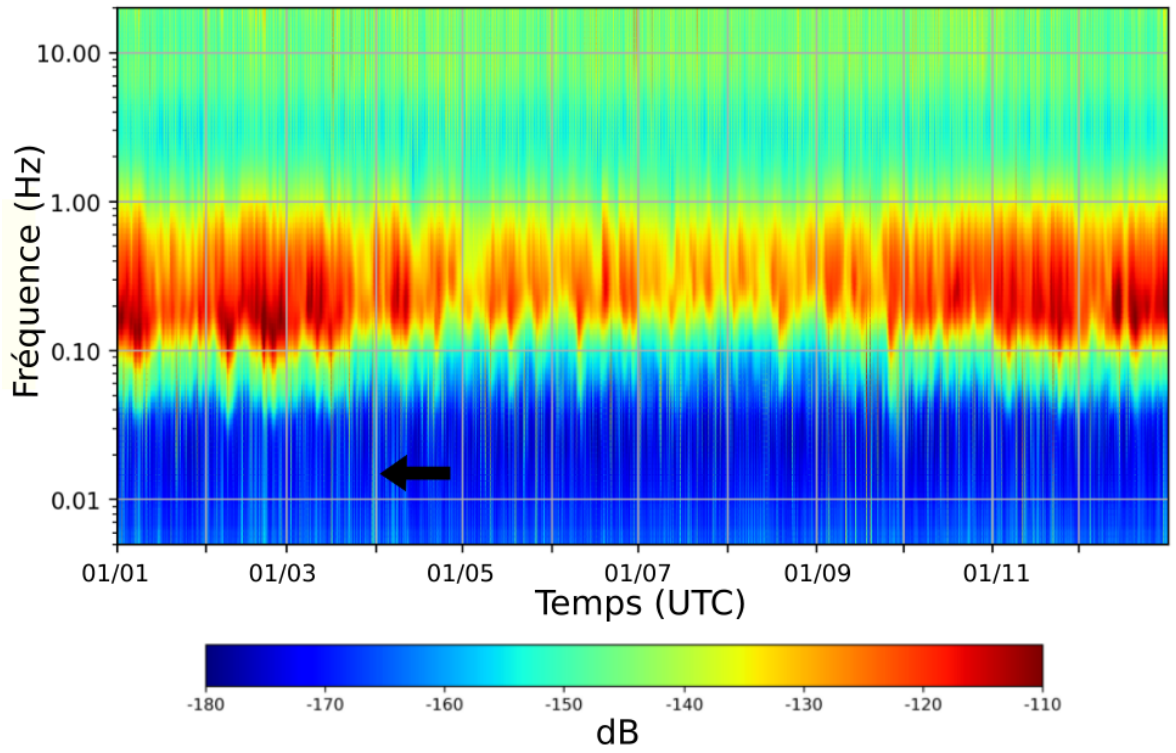

FIGURE 3.3 – Spectrogramme de la station MTNF lors de l'année 2022 (composante nord).

L'objectif de cette courte introduction étant ici de proposer un bref aperçu des données sismiques, nous n'abordons évidemment pas tous les détails constituant ce signal. On peut toutefois se diriger vers les études de Bormann et Wielandt (2013); Peterson et coll. (1993); Scales et Snieder (1998); Weaver (2005) ou Webb (2007) pour une description détaillée des différentes propriétés du « bruit » sismique, ainsi que Bormann et coll. (2012) pour une analyse approfondie sur les tremblements de Terre et les phénomènes de propagation des ondes.

3.1.2 Gaussianité du signal de fond sismique

Dans la quête de la compréhension des propriétés du signal sismique, un effort particulier a été apporté ces dernières décennies à l'étude de ses différentes caractéristiques statistiques. Dans ce contexte, les contributions de Groos et Ritter (2009) et Aggarwal et coll. (2020) ont révélé la nature Gaussienne de la distribution de nombreux signaux sismiques. Cette hypothèse a été confirmée quelques années plus tard par Zhong et coll. (2015a,b), aboutissant également à la conclusion d'un bruit sismique Gaussien, et précisant la nature stationnaire de ce dernier (moyenne et écart type constants au cours du temps). Par ailleurs, cette hypothèse de normalité est également régulièrement exploitée lors de simulations numériques du bruit sismique, générés par un tirage aléatoire Gaussien (Asgedom et coll., 2012; Tang et Ma, 2010). Cette conjecture demeure toutefois débattue dans la communauté scientifique, comme le montre l'étude de Wang et coll. (2014), qui ne permet pas de conclure à la gaussianité des données sismiques analysées via un test d'adéquation à la normalité de Shapiro-Wilk. Ces différentes conclusions semblant paradoxales, il est alors important d'apporter un éclaircissement sur la nature Gaussienne, ou non, du signal sismique. Une nuance intéressante est apportée par Groos et Ritter (2009), indiquant que la distribution normale des données sismiques se trouve régulièrement altérée par des signaux transitoires de grandes amplitudes (séismes, influence anthropique, *glitchs*, ...). Cette conclusion, plus mesurée, met alors en avant une normalité du signal de fond sismique, communément appelé « bruit ».

La figure 3.4 illustre la distribution Gaussienne des données sismiques enregistrées sur les stations CLEV, SDOF et MELF (France), le 12 juin 2022. Celle-ci présente sur sa partie haute, le signal sismique brut enregistré par la station CLEV. Les oscillations régulières sont semblables aux simulations Gaussiennes étudiées lors du chapitre 2 (voir par exemple la figure 2.4, A1). Ce signal n'étant pas altéré par une quelconque perturbation notable, la représentation de son histogramme normalisé (en rouge sur la partie basse de la figure) affiche une forte correspondance avec une courbe Gaussienne, représentée par une ligne rouge discontinue. Bien que cette correspondance ne soit évidemment pas suffisante pour aboutir à une conclusion forte quant à la gaussianité des

données sismiques, celle-ci est toutefois intrigante. De plus, cette forte similitude avec une distribution normale des données est également observée sur les histogrammes des signaux sismiques enregistrés sur les stations SDOF et MELF, en bleu et vert, respectivement. Il est aussi important de noter que la gaussianité de ces signaux est constatée lors de l'étude de trois composantes différentes dans chaque cas : Z, N et E pour SDOF, CLEV et MELF, respectivement. Bien que chaque signal semble suivre une distribution Gaussienne, leurs propriétés de moyennes et d'écarts types sont toutefois propres à chacun d'entre eux, comme indiqué par l'aplatissement et la position horizontale différente de chaque histogramme. Ces caractéristiques statistiques dépendent de l'amplitude du déplacement du sol enregistré par chaque station, ainsi que de la position de la masse du sismomètre (position d'équilibre autour de laquelle oscille le signal sismique brut).

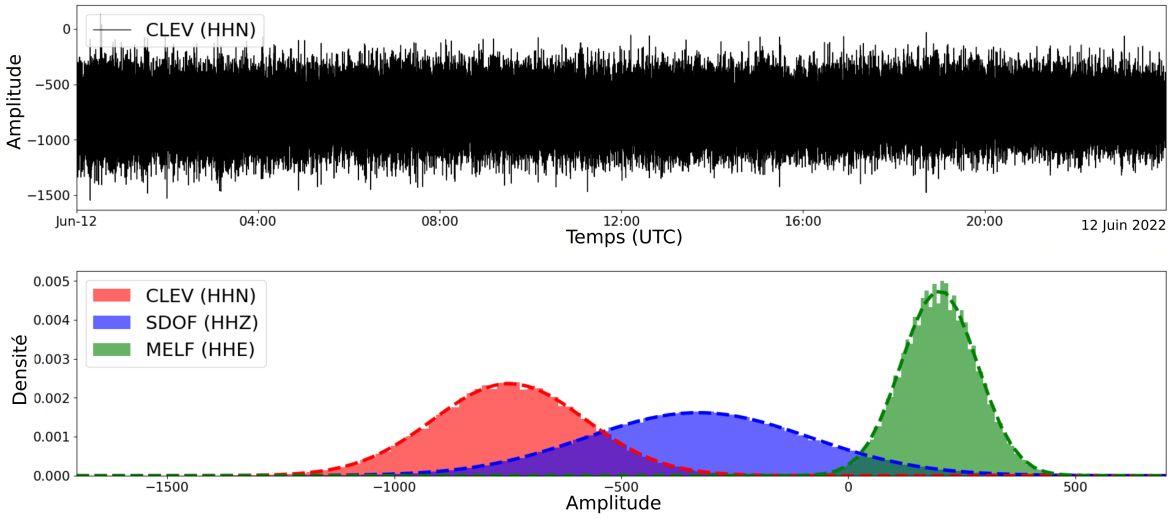

FIGURE 3.4 – Illustration de la distribution Gaussienne du signal sismique, enregistré sur des stations/composantes différentes. Le signal sismique (en haut) représente les données enregistrées sur la composante nord de station CLEV, le 12 juin 2022. La distribution de ce signal est représentée par un histogramme rouge (en bas), correspondant fortement à la densité de probabilité d'une loi normale, illustrée par une ligne rouge discontinue. Au cours de cette même journée, les histogrammes des signaux enregistrés par les stations SDOF et MELF (sur les composantes Z et E, respectivement) semblent également suivre une distribution Gaussienne de moyennes/écarts type différents.

La figure 3.5 propose une étude de la normalité locale d'une section du signal sismique étudié précédemment (partie haute de la figure 3.4), se restreignant ici à une longueur de 10 minutes (3.5 A). Une fois ce signal sismique trié par ordre croissant, on constate que celui-ci correspond parfaitement à la fonction probit modifiée associée

(3.5 B). Par conséquent, ceci indique que la distribution des données du bruit sismique semble Gaussienne lors d'une analyse d'une longue période de temps, mais aussi sur des intervalles temporels plus réduits. Suite à ces observations, il convient de s'intéresser à

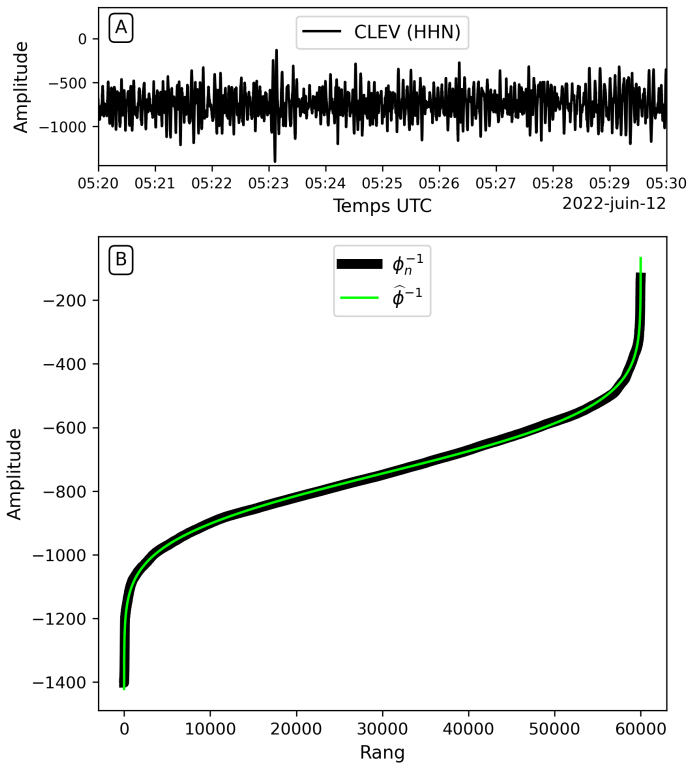

FIGURE 3.5 – Distribution normale du signal sismique sur une période de 10 minutes. (A) : Signal sismique de la station CLEV, composante nord. (B) : Signal sismique trié par ordre croissant ϕ_n^{-1} et la fonction probit modifiée associée $\widehat{\phi}^{-1}$.

l'origine de la distribution normale de ces données sismiques. Cette distribution particulière, est probablement une conséquence directe du théorème limite central (1.2.1), responsable de la gaussianité de nombreux phénomènes physiques. Cette même conclusion fut établie par Groos et Ritter (2009) et Zhong et coll. (2015b), considérant alors le signal sismique comme étant le résultat de multiples contributions indépendantes. Finalement, la nature Gaussienne du signal de fond sismique est rapportée dans la littérature, mais également cohérente avec nos différentes observations présentées dans les figures 3.4 et 3.5. Par conséquent, nous considérerons dans la suite de ce manuscrit que le signal de fond sismique suit une distribution normale.

3.2 Article : Seismic Station Quality Control using Deviation from the Gaussianity

Le travail effectué dans cette section, présente l'article intitulé *Seismic Station Quality Control using Deviation from the Gaussianity*, soumis au journal Seismological Research Letters (accepté avec révisions mineures). Celui-ci propose une application simple de la méthode NG-loc au signal sismique, utilisant la gaussianité de ce dernier en tant qu'estimateur de sa qualité. Cette étude permet de mettre en évidence des dégradations de la qualité de différents signaux sismiques, affectant des périodes temporelles, composantes, et gammes de fréquences spécifiques.

De manière plus précise, l'approche présentée dans cet article repose sur la comparaison entre l'écart type classique σ d'un signal donné, et celui de sa partie Gaussienne σ_G (éléments compris dans l'intervalle $[Q_A, Q_B]$). Une différence entre ces deux écarts types témoigne alors de la présence d'une perturbation non Gaussienne dans le signal. Les signaux sismiques continus de 4 stations (ECH, CAMF, CARF et VIEF) sont analysés au cours de cette étude, via une approche par fenêtre glissante, permettant d'estimer, la qualité des signaux sur de longues périodes temporelles. Bien que la station ECH se démarque de par la qualité remarquable de son signal, les stations CAMF, CARF et VIEF présentent quant à elles d'importantes déviations à la gaussianité sur certaines gammes de fréquences spécifiques. La cause la plus probable des dégradations observées dans ces données sismiques semble être la forte augmentation du niveau d'humidité environnant, ayant pour conséquence des phénomènes de corrosions de la connectique du sismomètre, entraînant finalement ce type d'altérations.

En conclusion, le travail présenté dans cette article propose une approche efficace d'analyse de la qualité du signal sismique, permettant de mettre en lumière des dégradations, soudaines ou progressives, sur certaines stations. Finalement, cette méthode d'analyse de la qualité des signaux sismiques ouvre naturellement la voie vers une application de cette dernière, à un plus large panel de stations.

Seismic Station Monitoring using Deviation from the Gaussianity

Arthur Cuvier[1], Éric Beucler[1,2], Mickaël Bonnin[1,2] and Raphaël F. Garcia[3]

[1] Laboratoire de planétologie et géosciences, Nantes Université, Univ. d'Angers, Le Mans Univ., CNRS UMR-6112, Nantes, France

[2] Observatoire des sciences de l'Univers de Nantes Atlantique, CNRS UAR-3281, Nantes, France.

[3] Institut Supérieur de l'Aéronautique et de l'Espace (ISAE-SUPAERO), Toulouse, France

E-mail: arthur.cuvier@etu.univ-nantes.fr

Declaration of Competing Interests

The authors acknowledge there are no conflicts of interest recorded.

Abstract

Degradation of the seismic signal quality sometimes occurs at permanent and temporary stations. Although the most likely cause is a high level of humidity, leading to corrosion of the connectors, environmental changes can also alter recording conditions in different frequency ranges and not necessarily for all three components in the same way. Assuming that the continuous seismic signal can be described by a normal distribution, we present a new approach to quantify the seismogram quality and to point out any time sample that deviates from this Gaussian assumption. We introduce the notion of background Gaussian signal (BGS) to characterize a set of samples that follows a normal distribution. The discrete function obtained by sorting the samples in ascending order of amplitudes is compared to a modified probit function to retrieve the elements composing the BGS, and its statistical properties (mostly its standard deviation σ_G). As soon as there is any amplitude perturbation, σ_G deviates from the standard deviation of all samples composing the time window (σ). Hence, the parameter $\log(\frac{\sigma}{\sigma_G})$ directly quantifies the alteration level. For a given frequency range and a given component, the median of all $\log(\frac{\sigma}{\sigma_G})$ that can be computed using short time windows, reflects the overall gaussianity of the continuous seismic signal. We demonstrate that it can be used to efficiently monitor the quality of seismic traces by using this approach at four broadband permanent stations. We show that the daily $\log(\frac{\sigma}{\sigma_G})$ is sensitive to both subtle changes on one or two components as well as the signal signature of a sensor's degradation. Finally, we suggest that $\log(\frac{\sigma}{\sigma_G})$ and other parameters that are computed from the BGS bring useful information for station monitoring in addition to existing methods.

1 Introduction

Both permanent and temporary deployed seismometers can be degraded during their operating time (*e. g.* Ekstrom et al., 2006; Davis and Berger, 2007). Visual inspection of the daily signal at each station allows any alteration of the signal to be detected quickly, but is incompatible with limited observatory staff that can operate more than 50 stations. On the other hand, as the continuous seismic signal varies as a function of time and frequency, and not necessarily in the same way for the three components, a decision of physical intervention on site driven by an AI based on observables such as spectrograms is, to our knowledge, not fully operational yet. There is thus a need for simple but reliable parameters to efficiently monitor the seismic signal quality.

Though the noise level depends on location and installation conditions, a number of issues such as mass-centering failures, glitches, increases in instrument self-noise, or corroded components can alter the continuous seismic signal. It may also sometimes happen that the failure disappears and the signal returns to a satisfactory quality, so no one will know that a problem ever occurred. One of the well known origin of recording condition degradation can be found in a high level of humidity, leading to corrosion of the internal electronic system. Hutt and Ringler (2011) indicate that i) high humidity conditions can modify the response of the instrument and ii) water vapor and moisture in the electronics appears to explain many of the observed anomalies.

In the field of quality control which aims to rapidly detect any deterioration, progress have been made during the last years (*e. g.* McNamara and Boaz, 2010). One can note the emergence of several automatic methods for monitoring stations, as presented in Ringler et al. (2015) and Casey et al. (2018) but those approaches are mostly dedicated to the detection of other issues than a degradation of the seismic signal quality (signal continuity, data availability). Probability of power spectral densities (PPSD) can provide very useful information, but require sufficiently large time windows to detect changes over time. To evaluate the seismic data quality, a strategy

consists in comparing signals recorded at collocated sensors (Tasič, 2018), or at stations in close proximity (Kimura et al., 2015). This generally cannot be used for a permanent array with station inter-distances of about 50 km.

Pedersen et al. (2020) present an innovative way to measure the quality of a single station, by comparing the standard deviation of the signal between the different components. Although this method appears to be efficient to detect malfunctions, it is not suitable for detecting signal degradation affecting all components simultaneously, and it seems difficult to define common thresholds that works for all stations.

In this article, we propose a novel approach based on the study of the seismic signal gaussianity to detect possible degradation of its quality. In section 2, we present a method allowing to discriminate, in any data set, the samples that can be considered as Gaussian, from the others (*i.e.* perturbed samples). Assuming that the seismic signal is intrinsically Gaussian (Groos and Ritter, 2009; Zhong et al., 2015b,a; Aggarwal et al., 2020), we perform in section 3 an analysis of the signal quality of the stations G.ECH, FR.CAMF, FR.CARF and FR.VIEF. Finally, we propose in section 4 a comparison between our approach and the method described in Pedersen et al. (2020).

2 Detection of non-Gaussian samples in an ensemble

Let us consider a set of samples whose distribution follows a Gaussian law, hereafter referred to as "background Gaussian signal" (BGS). This ensemble is often written $X \sim \mathcal{N}(\mu_0, \sigma_0)$, where μ_0 and σ_0 are the mean and the standard deviation, respectively. Such a distribution can be characterised by a bell-shaped histogram (*e. g.* DeGroot, 2002) or a kernel density estimate as well as the Cumulative Distribution Function (CDF) in order to avoid any arbitrary choice of discretisation (bin). For a real-valued random variable X, the CDF (ϕ) is defined as the probability that X takes a value less than or equal to a given real x. One can also use the

quantile function (*i. e.* the inverse of the CDF), called the Probit function (Bliss, 1934) in the special case of the standard normal distribution: $\mu_0 = 0$ and $\sigma_0 = 1$ (see eq. (A4)). In practical, the Probit function (hereafter denoted as ϕ^{-1}), can be approximated by sorting, according to increasing values, any set of n samples $(X_i)_{0 \leq i \leq n-1}$ which follows the standard normal distribution (see theorem 2 in the appendix). The result of this sorting operation is hereafter called empirical Probit function, noted as ϕ_n^{-1}, which is represented as a function of quantiles.

In a general case, if the BGS follows a given Gaussian law (μ_0, σ_0), the Probit function (ϕ^{-1}) can no longer describe the distribution of the ensemble, we then introduce the modified probit function, denoted as $\hat{\phi}^{-1}$, by a translation/homothety of μ_0 and σ_0,

$$\hat{\phi}^{-1} = \mu_0 + \sigma_0 \phi^{-1}. \tag{1}$$

At this stage, μ and σ, the arithmetic average and the standard deviation, respectively (*e. g.* Feller et al., 1971) describe entirely both the BGS statistical properties and $\hat{\phi}^{-1}$ ($\mu = \mu_0$ and $\sigma = \sigma_0$).

If the sample set is now altered by a perturbation, which means presence of elements with large variations in amplitudes which significantly differ from the BGS, the classical estimators are biased ($\mu \neq \mu_0$ and $\sigma \neq \sigma_0$). The idea behind our method is to extract the subset of points composing the BGS from the complete ensemble. This can be done, once the signal is sorted according to increasing values, because deviant samples are located at the edges of ϕ_n^{-1}. Consequently, it exists a given quantile interval $[Q_A, Q_B]$, separating the samples composing the BGS from those of the perturbations which can be located through a full exploration of the sorted sample space. In practical, ϕ_n^{-1} is extracted for each tested quantile interval, its mean and standard deviation define the local $\hat{\phi}^{-1}$ (eq. 1) over the same amount of samples. According to theorem 2, the misfit between ϕ_n^{-1} and $\hat{\phi}^{-1}$ is measured by the difference at the sense of the

L^∞-norm. The interval finally selected, hereafter denoted as $[Q_A, Q_B]$, defines the subset of samples which achieve the lowest misfit. In the following, the mean and the standard deviation of samples within $[Q_A, Q_B]$ are denoted as μ_G and σ_G, respectively, as they define the statistical properties of the BGS.

The theory presented above is illustrated through three synthetic experiments (Fig. 1). The BGS (A1) is obtained by a random draw of $n = 2,000$ points, with $\mu_0 = 314$ and $\sigma_0 = 16$, which are the parameters to retrieve for all cases. The classical arithmetic mean and standard deviation (μ and σ) of the three sample sets, are displayed in A1, B1, C1.

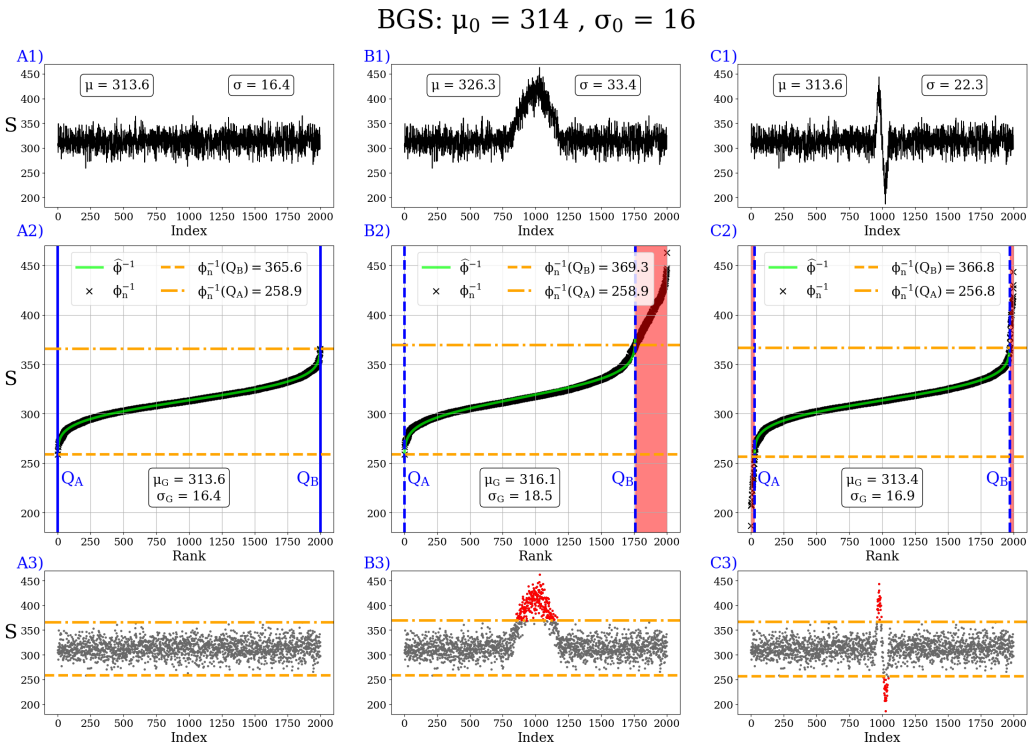

Figure 1. Illustration of how retrieving the Gaussian samples in three synthetic data set. The same BGS is imposed for each case (A, B and C) with $\mu_0 = 314$ and $\sigma_0 = 16$. A wide and a narrow perturbations are added in B and C, respectively. The second line (A2, B2 and C2) presents these signals (black crosses), once sorted by increasing order of amplitude, noted ϕ_n^{-1}. For each case, the interval $[Q_A, Q_B]$ is given by the best fit between ϕ_n^{-1} and $\hat{\phi}^{-1}$ (green), defining μ_G and σ_G, approaching the properties of the BGS.

Let's start with the pure BGS case (A1, A2, A3). The samples shown in A1 are sorted by

ascending order of amplitudes to generate ϕ_n^{-1} (black crosses in A2). The best fit between ϕ_n^{-1} and $\hat{\phi}^{-1}$ (green curve in A2) is obtained for the interval $[Q_A, Q_B] = [0, 1999]$, indicating that all samples follow a Gaussian law with $\mu_G = 313.6$ and $\sigma_G = 16.4$. Obviously, since all the samples are considered as Gaussian here, $\mu_G = \mu$ and $\sigma_G = \sigma$, and are relatively close to μ_0 and σ_0.

In the second column of Fig. 1, a perturbation is added to the BGS. We can first notice that, obviously, μ and σ now differ from the values to be recovered (μ_0, σ_0). The exploration of all possible quantile intervals gives $[Q_A, Q_B] = [0, 1759]$, which efficiently excludes the outlayer samples (red area in B2). This interval is associated with values of $\mu_G = 316.1$ and $\sigma_G = 18.5$ which are much closer to μ_0 and σ_0 compared to μ and σ. The values of $\phi_n^{-1}(Q_A)$ and $\phi_n^{-1}(Q_B)$ are of 258.9 and 369.3, respectively (horizontal dashed/dotted orange lines), which allow to separate anomalous samples (red points in B3) from the BGS.

For the narrow anomaly case (C1), μ is not affected due to the symmetric shape of the perturbation but the σ is biased since all the elements are taken into account. The exploration of the sorted data space returns here $[Q_A, Q_B] = [27, 1971]$, excluding outlayer samples composing the perturbation (red areas in C2). Back to the index domain (C3), the orange lines, given by $\phi_n^{-1}(Q_A)$ and $\phi_n^{-1}(Q_B)$, define the amplitude domain composing the BGS. Any sample above or below these two limits can be considered as perturbations. Once again, the value of $\sigma_G = 16.9$ is closer to the value of $\sigma_0 = 16$ compared to $\sigma = 22.3$. For all cases, the two horizontal orange lines are very similar, which is consistent with the fact that the same BGS is imposed in the three synthetic signals.

Finally, this approach allows to efficiently retrieve $[Q_A, Q_B]$ and thus the statistical characteristics of a BGS: μ_G and σ_G. As soon as an amplitude perturbation alters the data set, there is a mismatch between σ_G and σ. For the analysis of real signals, as μ_0 and σ_0 are unknown, any deviation from the gaussianity of a given data set can be measured by $\log\left(\frac{\sigma}{\sigma_G}\right)$, in order not to depend on amplitudes and to reflect possible large variations from the reference state ($\sigma = \sigma_G$).

For instance, in Fig. 1, the values of $\log(\frac{\sigma}{\sigma_G})$ is 0 exactly in (A) while it reaches values of 0.26 and 0.12 in (B) and (C), respectively, which correspond to significant deviations. A difference between μ and μ_G can also point out non-Gaussian features but can suffer from special cases such as a zero mean signals and/or symmetrical perturbations (Fig. 1 C). In the following, the word "perturbation" is used to describe any deviation from the Gaussian hypothesis (BGS), characterised by values of $\log(\frac{\sigma}{\sigma_G})$ greater than 0.

3 Application to the seismic station monitoring

In this section, we propose to analyse the continuous seismic signal recorded at four permanent broadband stations, using the method presented in section 2. In the following, it is assumed that the continuous seismic signal follows a Gaussian distribution (*e.g.* Groos and Ritter, 2009; Zhong et al., 2015b,a; Aggarwal et al., 2020).

3.1 Methodology

The gaussianity of the continuous seismic signal recorded during 24 h can be quantified by multiple analysis of short time windows. Results are shown in Fig. 2, using 1 h time windows, sliding with an overlap of $\frac{2}{3}$. Hence, each sample is analysed three times. In order to investigate the frequency dependence of the gaussianity, the signal is analysed through four period ranges: LF ($T > 80$ s), BP1 (20 s $< T <$ 80 s), BP2 (1 s $< T <$ 20 s) and HF ($T < 1$ s). In order to allow a reliable comparison between the different period bands, all signals are decimated at 20 samples per second in order to have the same amount of samples in each analysed window. The instrument response is removed in the period range [0.1, 160] s and the signal is converted into ground velocity.

For each time window σ is computed using all samples whereas σ_G is defined after the computation of $[Q_A, Q_B]$. Although we mostly focus on $\log(\frac{\sigma}{\sigma_G})$ to quantify the gaussianity and to

detect anomalous behaviour of seismic stations, three other parameters can also be investigated:

- μ_G, the Gaussian mean of the ranked samples within $[Q_A, Q_B]$. Since the arithmetic average is subtracted from the signal amplitude before each filtering operation of a given 1 h time window, μ_G must be compared to zero;

- $\mathcal{G}$, the Gaussian point ratio, defined by the amount of selected samples in $[Q_A, Q_B]$ divided by the total amount of points of the sliding short time window;

- M_{L^2}, the misfit between ϕ_n^{-1} and $\hat{\phi}^{-1}$ (Fig. 1, second row), using the L^2-norm,

$$M_{L^2} = \frac{1}{(Q_B - Q_A)} \sqrt{\sum_{i=Q_A}^{Q_B} \left(\hat{\phi}^{-1}(i) - \phi_n^{-1}(i) \right)^2}. \tag{2}$$

  A low value of M_{L^2} then reflects a high degree of gaussianity of the subset of samples selected in $[Q_A, Q_B]$.

3.2 Single day analysis of the gaussianity

Fig. 2 exhibits the four parameters defined above for a signal duration of 24 h (June 1, 2019), recorded at FR.CAMF (North component) and filtered in two frequency ranges: BP1 and HF. The sensor (Nanometrics T120QA) of this broadband permanent station is installed on the ground in a WWII blockhaus, in Brittany (France), and located at the top of a cliff facing the Atlantic Ocean (Fig. 3). The rock basement is composed of Armorican sandstone. Although the quality of the installation is standard and made with great care, the continuous seismic signal is altered for different reasons: at high frequency, the proximity of the village and the energy of breaking waves on the cliff and at longer periods, temperature and pressure variations in addition to tidal modulations (*e. g.* Beucler et al., 2015).

The signal filtered in the BP1 frequency domain (red in Fig. 2) is less energetic than the HF

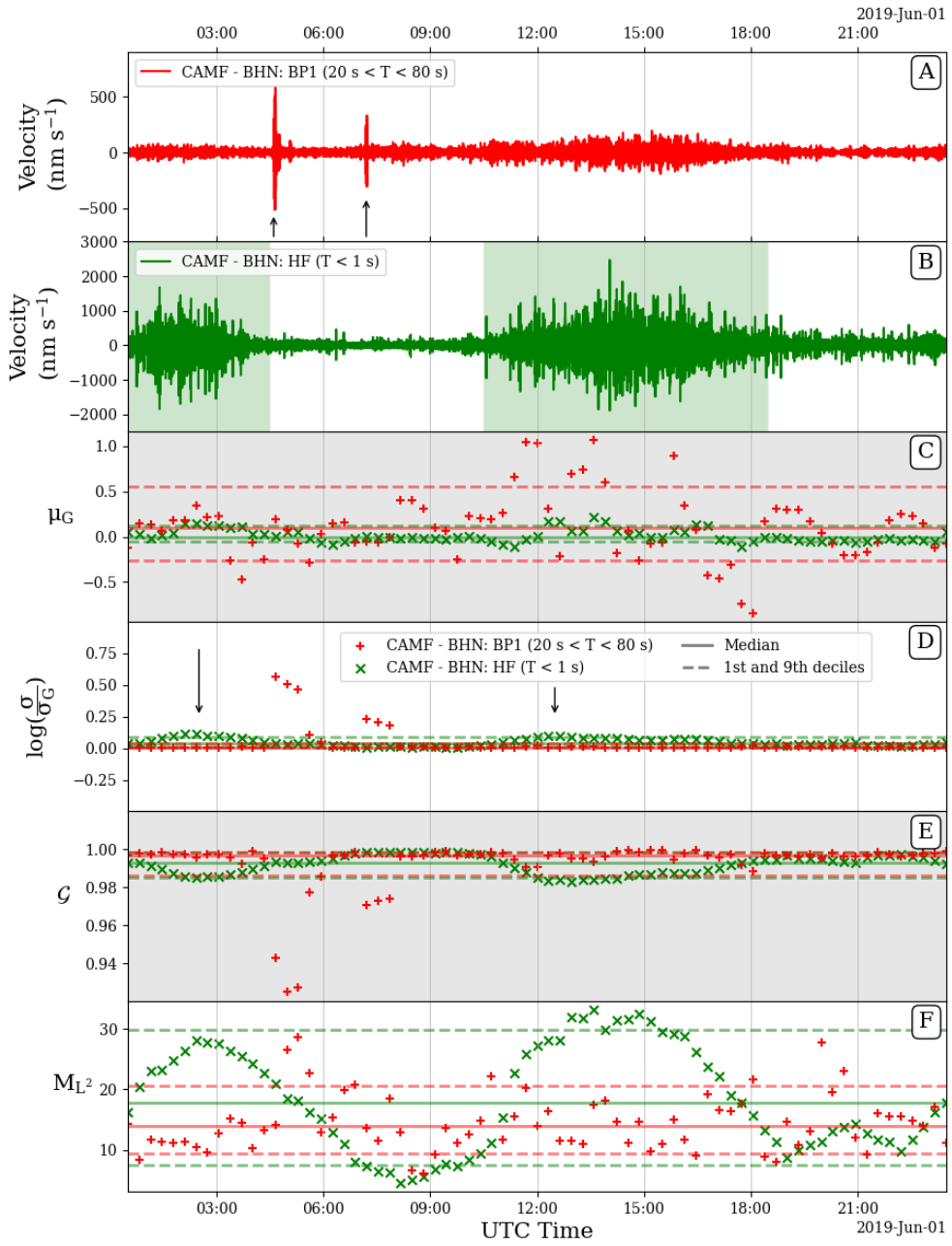

Figure 2. Analysis of a continuous seismic signal during a full day, using a sliding window approach. (A and B): Seismic signals from the FR.CAMF station on June 1, 2019 (BHN), deconvolued and filtered from 20 to 80 s (A) and below 1 s (B). (C): Mean of the BGS. (D): Logarithm of the ratio between the classical and the BGS standard deviation. (D): Proportion of Gaussian points in the $[Q_A, Q_B]$ interval. (E): Misfit between $\hat{\phi}^{-1}$ and ϕ_n^{-1} in $[Q_A, Q_B]$, using the L^2-norm.

filtered trace (green) but contains some similarities. A diffuse extra energy is visible on two ∼3 h windows, centered around 2:20 and 14:45 UTC, respectively (green areas in Fig. 2 B). They both coincide with the high tides occurring twice a day. The seismic signal is then modulated in the HF range due to the breaking waves on the cliff but also to a lesser degree in BP1 since this frequency domain comprises the edge of the primary microseismic peak and a part of the infragravity wave period range (*e. g.* Nawa et al., 1998; Ardhuin et al., 2011; Stutzmann et al., 2012). In addition, the surface waves of two $M_{\mathrm{W}} \simeq 5$ earthquakes that occurred in Greece (epicentral distances of approximately 2,200 km) are well visible in BP1 trace (indicated by the two vertical arrows in A) but are less obvious for HF.

For both BP1 and HF domains the values of the BGS mean (μ_G) lie between -0.89 and 1.18 (Fig. 2 C) and $\log(\frac{\sigma}{\sigma_G})$ is very stable around the value of 0. For the HF case (green crosses) two $\log(\frac{\sigma}{\sigma_G})$ deviations up to 0.15 are observed at the times of high tides (pointed out by the two black arrows in D) indicating that, locally, the samples that composed a 1 h window are less Gaussian than the rest of the day. The consequence is a decrease of $\mathcal{G}$ ($\sim 98.2\%$ for both high tide windows) and large increases of M_{L^2} (up to 32.5) which leads to conclude that even in the $[Q_A, Q_B]$ interval the fit to $\hat{\phi}^{-1}$ is not as good as for quieter parts of the day.

The BP1 frequency range analysis for the same day (red pluses in Fig. 2) shows a very stable behaviour all over the 24 h except during the two earthquakes. Those impulsive transient energies do not affect μ_G, which is consistent with surface wave wavetrains that make the ground oscillating symmetrically around an equilibrium position, but they are well visible on $\log(\frac{\sigma}{\sigma_G})$ with values up to 0.6. For the corresponding time windows, $\mathcal{G}$ decreases down to 0.925.

Finally, it is important to notice that these parameters are sensitive only to amplitude variations and not to the level of the seismic energy. This allows to propose that such a study can be performed for any component of any seismic station and for different ranges of periods. In the following, since $\log(\frac{\sigma}{\sigma_G})$ reflects both mean translation and sample dispersion around this latter, we will mainly use this parameter to quantify the Gaussianity of a single day. This is

realised using the median value of the 74 one hour windows (solid lines in Fig. 2) that composed a day (with an overlap of $\frac{2}{3}$). As shown in Fig. 2 D, the median is not affected by transient waveforms such as earthquakes and/or spurious signals.

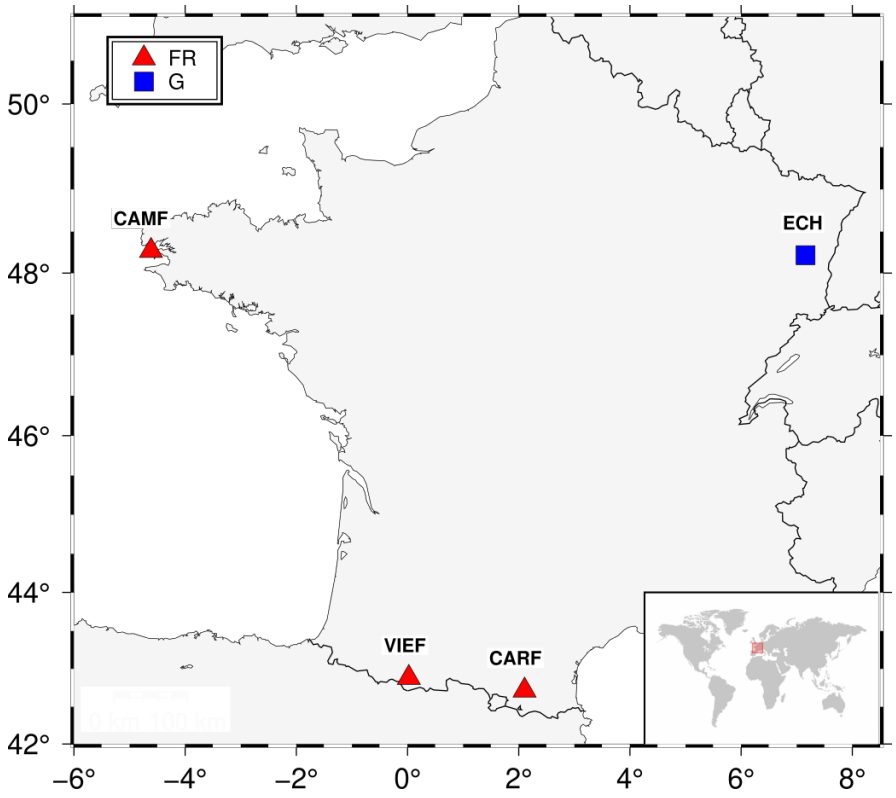

Figure 3. Locations of seismic stations used in this study. They are all belonging to the French permanent broad-band array, from the RESIF (1995)(FR) and the GEOSCOPE (G) (Institut de physique du globe de Paris (IPGP) and École et Observatoire des Sciences de la Terre de Strasbourg (EOST), 1982) networks.

3.3 Daily analyses of the seismic signal gaussianity at four permanent stations

In order to analyse the behaviour of a permanent station in terms of deviation from gaussianity day by day, we focus hereafter on four broadband seismic stations (Fig. 3). Let us start with G.ECH, located in Echery (eastern France), that we consider as the reference station in terms

of signal quality.

3.3.1 ECH

The sensor (STS1) is installed on a concrete pavement in a 250 m long tunnel inside an abandoned silver mine. The site geology is mostly composed of gneiss. This station is running for more than 22 years and is known for the stability of its quality over the years. In a few words, this station is of high quality at short periods (PSD lower than 150 dB for $T < 1$ s) and exhibits a vertical component energy close to the low noise model (Peterson, 1993) between 20 and 200 s period. The horizontal components are noisier for periods greater than 40 s and the North component is more affected than the East one.

The analysis of G.ECH in terms of $\log(\frac{\sigma}{\sigma_G})$ variations, in four frequency ranges (see section 3.1), and for the whole year 2019 is presented in Fig. 4. For each day, the medians of $\log(\frac{\sigma}{\sigma_G})$, computed for the 74 time windows, are displayed for the three components. Compared to other station analyses (Figs. 5, 6 and 7), the values are so close to 0 that we propose a vertical scale between 0.01 and 0.06. For the HF and BP2 frequency ranges, the values are very stable around 0 for the whole year and reach maxima of 0.014 and 0.001, respectively. This implies that, the continuous seismic signal at periods lower than 20 s are in very good agreement with a Gaussian distribution. This is particularly true for BP2 which comprised the frequency band of the microseismic peaks (*e. g.* Ebeling, 2012).

In contrast, for BP1 and LF, we observe a greater dispersion, for instance, it is 10 times larger for BP1 than for HF. For the BP1 frequency range (Fig. 4 C), the mean of all North $\log(\frac{\sigma}{\sigma_G})$ (red crosses) is 0.015 whereas they are of 0.06 for the two other components. This could indicate that the extra energy which makes this component noisier (as indicated by power spectral densities that can be computed for this station) with respect to others, also alters the gaussianity of the signal. This phenomenon can be observed to a lesser degree in the LF domain (Fig. 4 D), for which the three components exhibit however a more stable behaviour over the

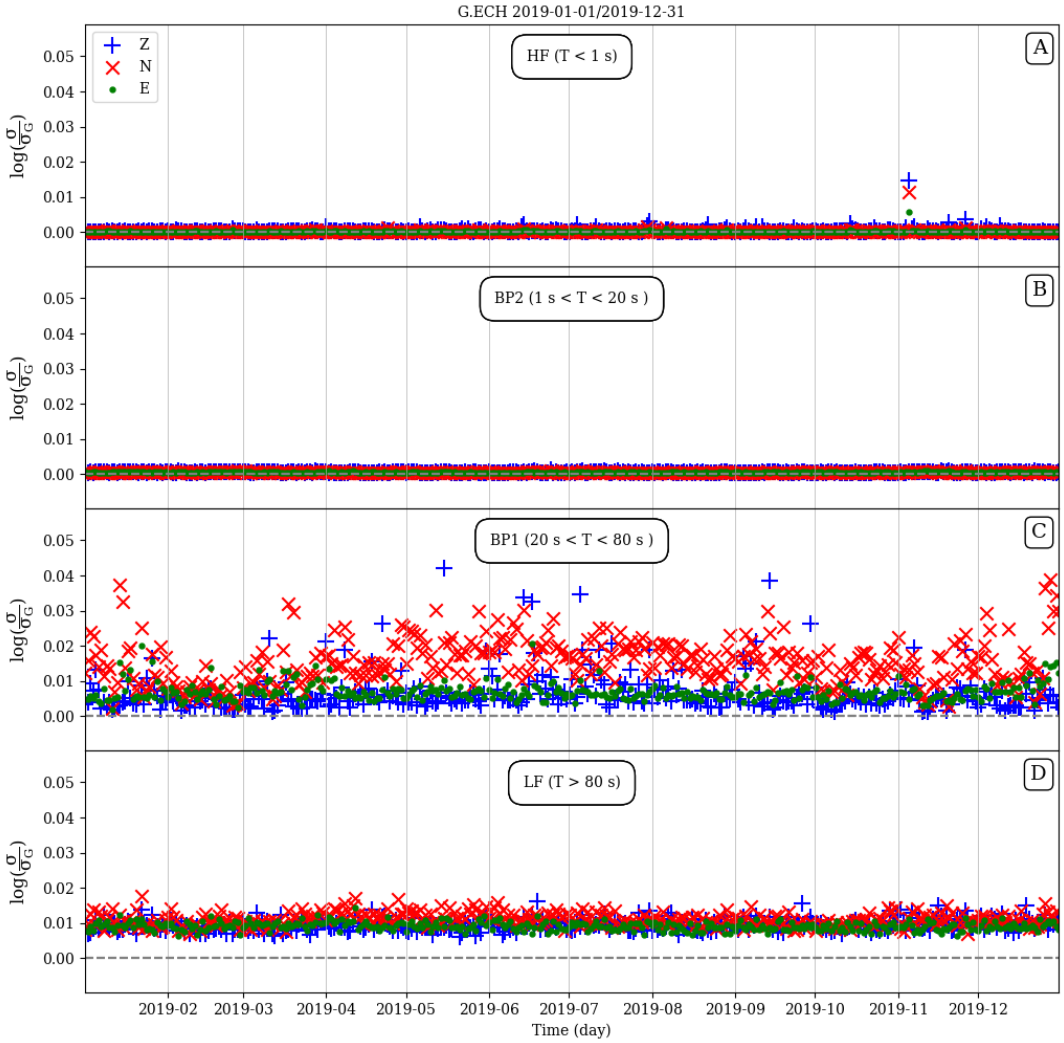

Figure 4. Analysis of the continuous seismic signal recorded at G.ECH in 2019. The medians of the daily $\log(\frac{\sigma}{\sigma_G})$ are displayed for the three components (plusses, crosses and dots for the vertical, north and east, respectively) and the four frequency bands (LF, BP1, BP2 and HF), defined in section 3.1.

year. One can notice that the mean of the $\log(\frac{\sigma}{\sigma_G})$ oscillates here around of 0.01, ant not exactly 0, which is only a side effect due to the length of 1 h for all analysed windows, allowing less oscillations of the signal than for the highest frequencies. To avoid any misinterpretation, only values greater than 0.1 are considered as noticeable deviations from the Gaussian case (BGS). One can notice a $\log(\frac{\sigma}{\sigma_G})$ variation during November, 5 on the HF frequency bands (Fig. 4 A),

which is caused by a surprisingly large occurrence of earthquakes and quarry blasts (more than 50 events).

Finally, since we are interested mostly in the time variations of $\log(\frac{\sigma}{\sigma_G})$, we consider hereafter that the maximum values at ECH (for each frequency range) can be used as reference thresholds for other stations (orange lines in Fig. 5).

3.3.2 CAMF

The site conditions of FR.CAMF are already detailed in section 3.2. As for G.ECH, the analysis of $\log(\frac{\sigma}{\sigma_G})$ variations of the continuous seismic signal recorded at FR.CAMF in 2019 are displayed in the four frequency ranges in Fig. 5. For each frequency band, horizontal orange line is shown to indicate the maximum value of all medians (of all components) measured at G.ECH. They can be considered as threshold references to point out any alteration of the signal.

We choose the year 2019 because the recording conditions of FR.CAMF have been modified between the beginning of July and mid-October (period highlighted by the grey and magenta vertical dashed lines, respectively in Fig. 5). Due to high humidity at this time, the sand that insulates the sensor gradually became waterlogged. This led to a deterioration of the long period signal quality of horizontal components that can be seen on the spectrograms in Fig. 9.

Considering the signal before July 7, 2019, all components in the HF frequency bands are much more dispersed than for BP2, BP1 and LF. The variations of $\log(\frac{\sigma}{\sigma_G})$ are of the same order of magnitude as the single day analysis presented in Fig. 2. They are due to the seismic extra energy, caused by breaking waves on the cliff. Since the degradation of the recording conditions does not affect HF (Fig. 9), it is not possible to detect any noticeable modification in this frequency range. For the same reason and because the microseismic peak energy is obviously very large at FR.CAMF, $\log(\frac{\sigma}{\sigma_G})$ is always close to 0 in BP2 (with mean equals to 0.0004). This contrasts with the values observed in BP1 and LF bands (Fig. 5 C, and D), where the daily $\log(\frac{\sigma}{\sigma_G})$ reach 0.38 and 0.23, respectively. These large deviations are only visible for

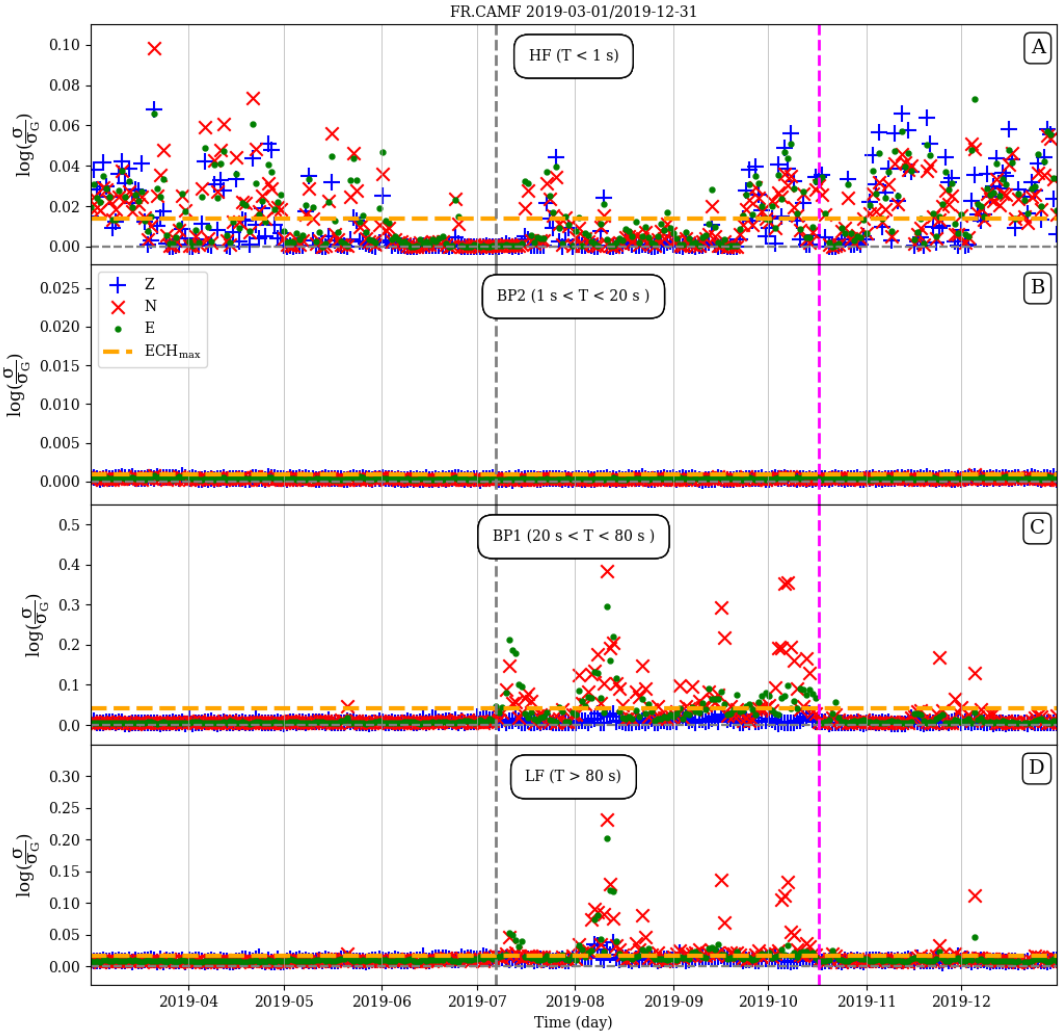

Figure 5. Same legend as Fig. 4, but for FR.CAMF. Horizontal orange lines are indicating the maximum $\log(\frac{\sigma}{\sigma_G})$ value (for all components), computed for G.ECH in each frequency band (see section 3.3.1). Due to the high level of humidity, the recording conditions are degraded during the time window defined by the two vertical dashed lines.

the horizontal components which is consistent with Fig. 9. However, while a classical energy analysis, such as PPSD or spectrograms, do not yet show any significant changes, the $\log(\frac{\sigma}{\sigma_G})$ turns to anomalous values (up to 0.21 for the East component) as early as the July, 7 (grey dashed line). This deteriorated recorded conditions ended on October, 17 (magenta dashed line) when the wet sand has been replaced by dry one. This intervention brought back the sensor

into the normal operating conditions, resulting in $\log(\frac{\sigma}{\sigma_G})$ values that rapidly return to 0. One can notice few anomalous values of $\log(\frac{\sigma}{\sigma_G})$ for both BP1 and LF (Fig. 5 C, and D) between November, 21 to December, 5. After visual inspection, it appears that three days have been perturbed by long period glitches that mostly affect the north component.

3.3.3 CARF and VIEF

FR.CARF and FR.VIEF are both located in the Pyrénées mountains (France) at altitudes of 1,200 and 1,000 m, respectively. The geology of FR.CARF is composed of limestones while FR.VIEF is installed in a shale massif. Their sensors (T120QA for FR.CARF and T120PA for FR.VIEF) are installed in a $\sim$ 1 m depth vault and insulated with sand. FR.VIEF is located about 30 m of a village, making it theoretically more exposed to anthropic activity than FR.CARF, although this is not that obvious in the HF frequency band of Figs. 6 and 7 neither in the spectrograms shown in Fig. 9. The choice of these two stations is motivated because both of their recorded signals have been suddenly deteriorated by humidity that corroded connections. The insulation was realised using sandbags arranged around the sensors and the water that seeped in was guided to the connectors. This appears between September 13–30, 2021 for FR.CARF and between February 9–17, 2022 for FR.VIEF, as indicated by the grey and magenta vertical dashed lines in Figs. 6 and 7.

For both stations and before degradation, the signals of the three components have a high degree of gaussianity characterised by values of $\log(\frac{\sigma}{\sigma_G})$ very close to 0. This is particularly true at high frequency and in the microseismic bandwidth (BP2) while few variations are observed for the signal at long periods (BP1 and LF), mostly on the East component for FR.CARF and on the North component for FR.VIEF (although it is not obvious in Fig. 7 C due to the vertical scale). These descriptions can be linked to the fact that FR.CARF and FR.VIEF are located on the eastern and southern flanks of mountains, respectively.

The two stations have encountered a degradation of their operating conditions when large

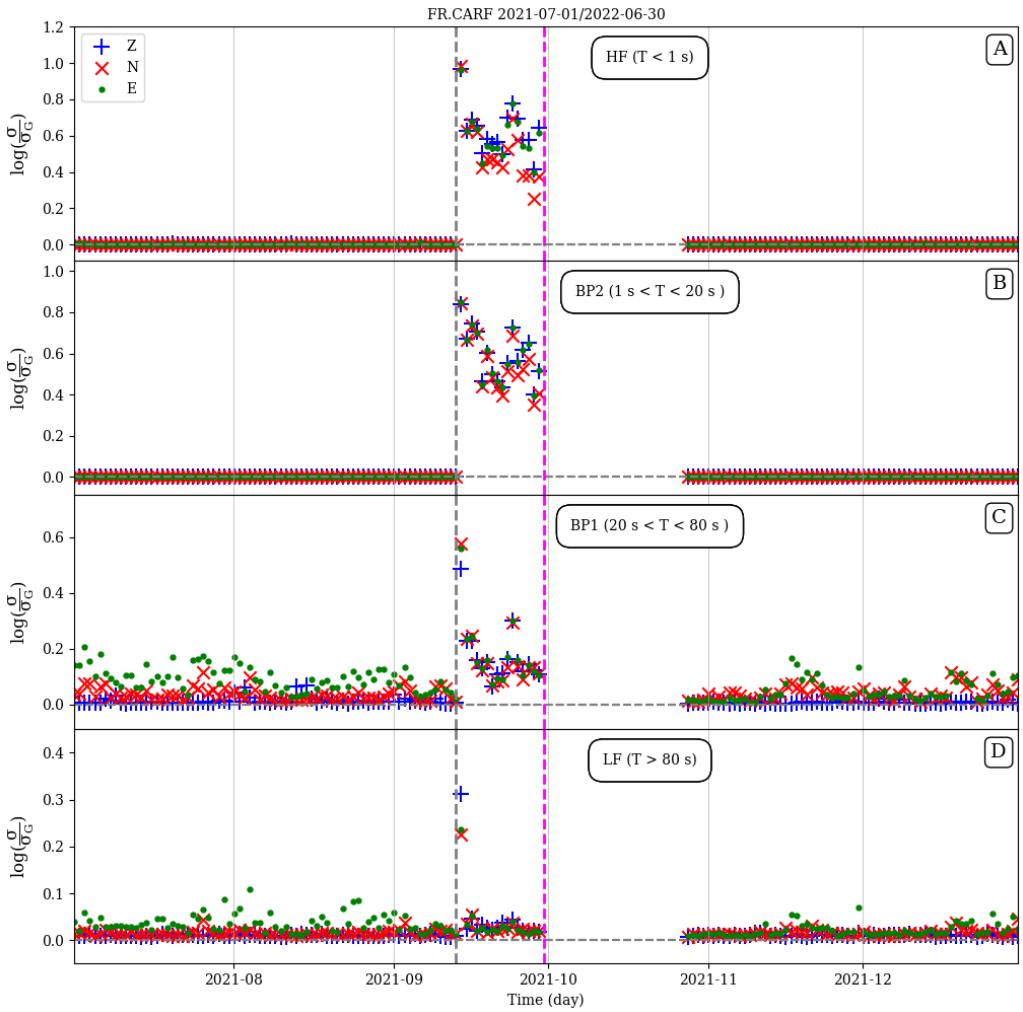

Figure 6. Same legend as Fig. 4, but for FR.CARF. Due to the high level of humidity, the recording conditions are degraded during the time window defined by the two vertical dashed lines.

modifications of $\log(\frac{\sigma}{\sigma_G})$ are observed. In both cases, the $\log(\frac{\sigma}{\sigma_G})$ signatures differ as a function of the frequency. For instance, the LF domain although largely affected in terms of the signal energy (see FR.VIEF spectrograms in Fig. 9) is not obvious in Fig. 6(D) and 7(D).

In the HF and BP2 frequency domains at FR.CARF, the daily $\log(\frac{\sigma}{\sigma_G})$ values are remarkably stable and never exceeds 0.01, before and after the degradation time (Fig. 6 A and B). *A contrario*, as soon as the recording conditions are degraded, they become very large (up to 0.98

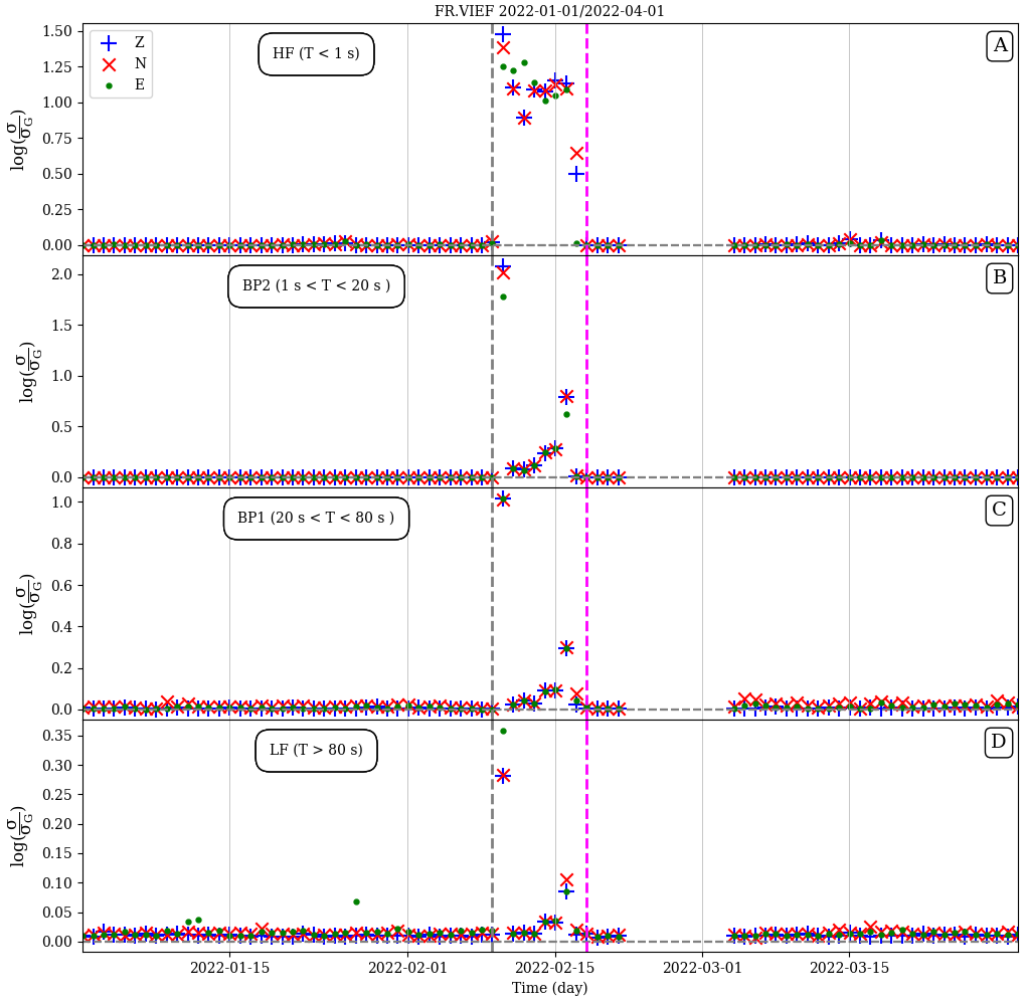

Figure 7. Same legend as Fig. 6, but for FR.VIEF.

and never lower than 0.25). At longer periods, a modification of $\log(\frac{\sigma}{\sigma_G})$ is also observed but to a lesser degree, except for the 1st day (September 14, 2021), where it reaches values of 0.57 and 0.31 for BP1 and LF, respectively. The station operators removed the corroded sensor on September 30 and installed a new one on October, 27 (explaining the data gap). The gaussianity in the different frequency domains returns to the same level as before the degradation.

A more detailed study is realised for FR.VIEF (Figs. 7, 8 and 9). The same HF and BP2 $\log(\frac{\sigma}{\sigma_G})$ signatures as for FR.CARF are observed during the degradation time, but in this case

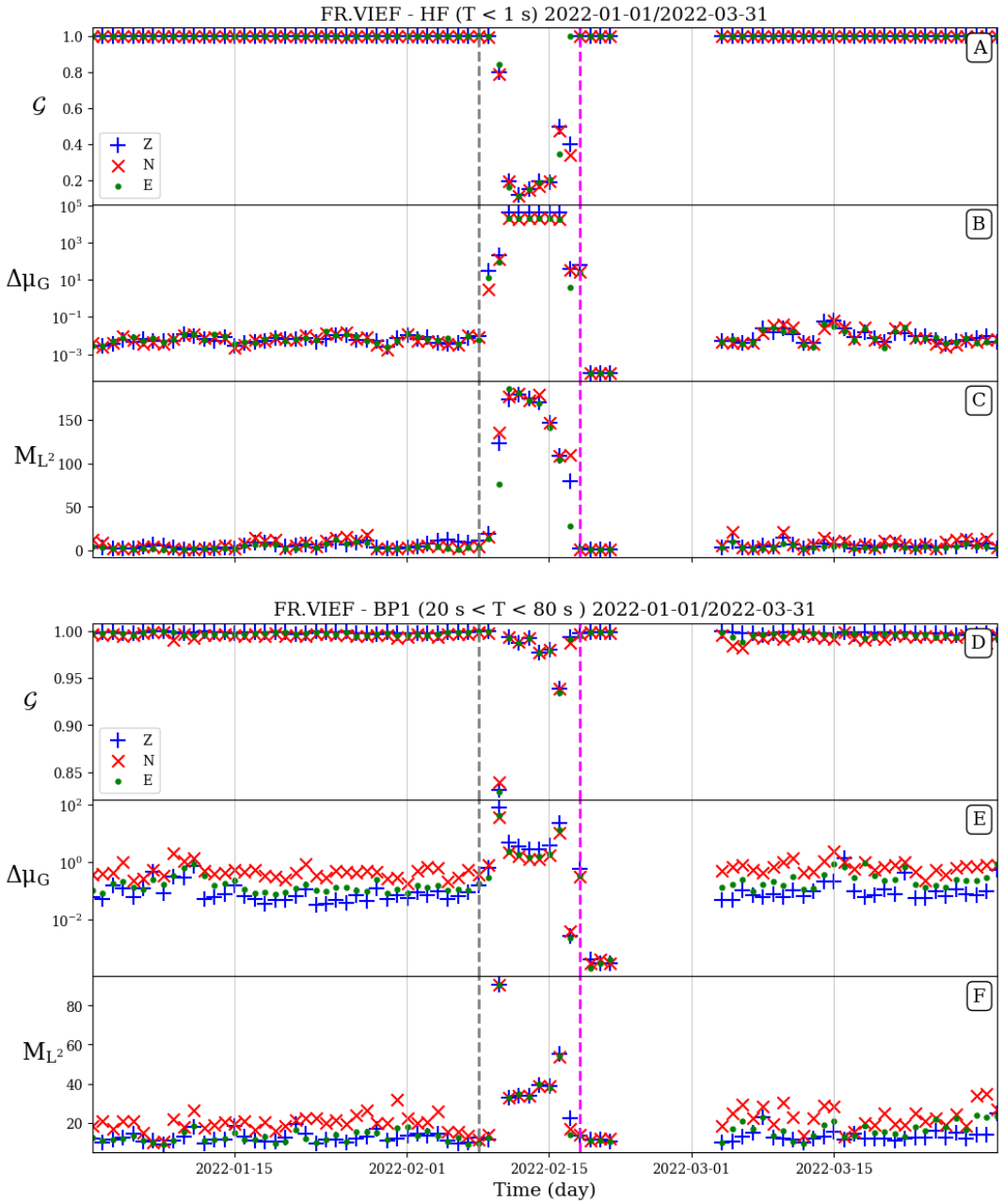

Figure 8. Gaussianity analysis of FR.VIEF during the same time period than in Fig. 7. For both frequency bands, $\mathcal{G}$ and M_{L^2} are the Gaussian point ratio and the misfit at the least-squares sense, respectively. $\Delta\mu_G$ is the difference between the 9th and the 1st deciles of all daily μ_G values.

with values larger than 1.4 for HF and 2 for BP2 (Fig. 7 A and B).

Supplementary information are given in Fig. 8, where three other parameters are shown for HF and BP1. $\mathcal{G}$ and M_{L^2} are detailed in section 3.1 and $\Delta\mu_G$ represent here the difference between the 9th and the 1st deciles of the set of all one hour μ_G values computed every day (as indicated for instance by the horizontal dashed lines in Fig. 2 C). This parameter quantifies the stability of μ_G for a given day and, for the sake of comparison, low values are bounded to 10^{-4}.

As for $\log(\frac{\sigma}{\sigma_G})$, in the HF domain, $\mathcal{G}$, $\Delta\mu_G$ and M_{L^2} exhibit large variations during the degradation time. One can notice that $\mathcal{G}$ reaches values of 0.15, indicating that only 15% of samples are selected to belong to $[Q_A, Q_B]$, which is consistent with the large values of $\log(\frac{\sigma}{\sigma_G})$ shown in Fig. 7 A. Such $\mathcal{G}$ values are very close to the minimal proportion of Gaussian samples that is authorized in our method ($\mathcal{G} = 0.1$). In addition, M_{L^2} values are the largest, telling that even the 15% of selected samples are much less Gaussian than outside the degradation time. Plus, very large values of $\Delta\mu_G$ ($\sim$ 43,000) are observed, confirming that huge μ_G variations are occurring within a day. Finally, all these parameters are converging toward the same diagnostic of an ill-sensor with very large energy fluctuations and dramatically different signal quality compared to before, as also shown by the spectrograms (Fig. 9).

At longer periods (BP1 and LF in Fig. 7), the $\log(\frac{\sigma}{\sigma_G})$ values are less affected by the signal degradation. This can be due to a long period feedback deterioration which could decrease the sensor sensitivity as shown by a slightly different behaviour of all other parameters (Fig. 8 D, E and F).

One can notice a sudden return of $\log(\frac{\sigma}{\sigma_G})$ to 0, just after the end of the degradation (magenta line), for all the components and frequency bands. It is simply due to the numerical noise of the digitizer, which continued to operate even once the sensor have been removed. The channels have been officially closed three days after sensor removal producing the data gap.

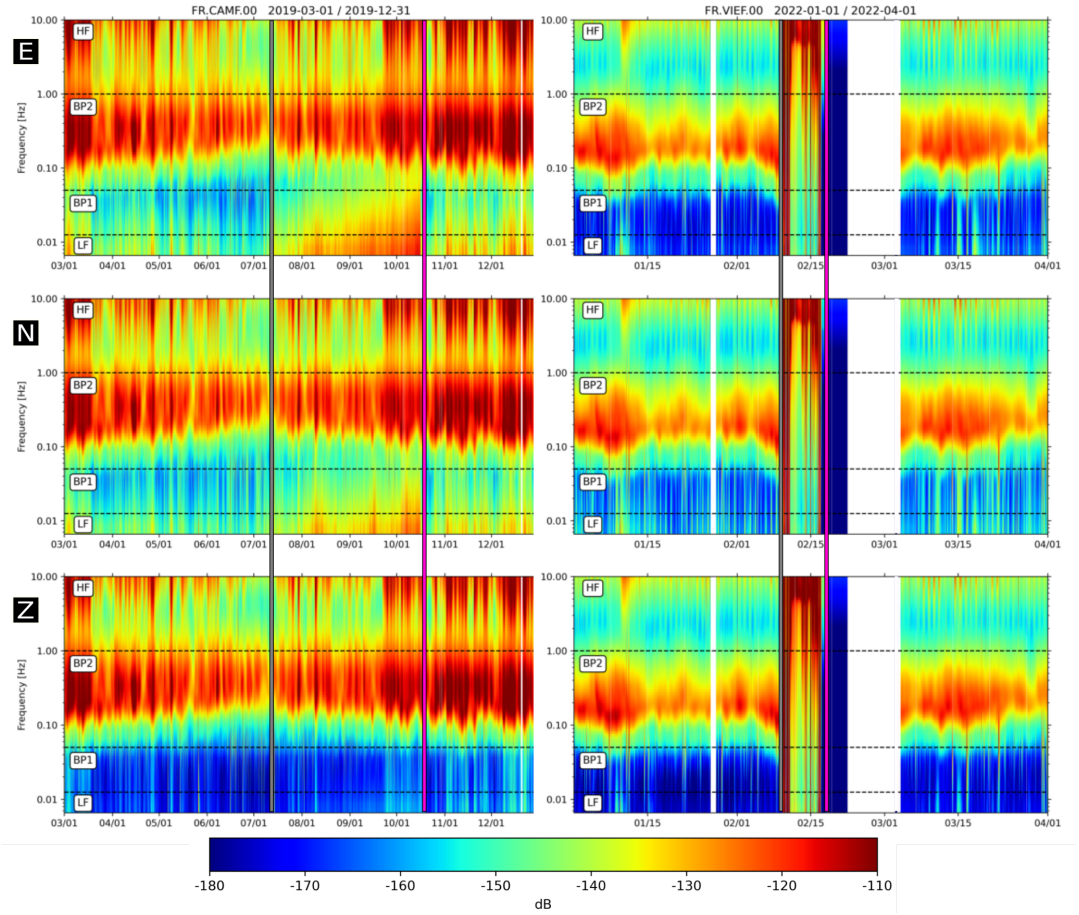

Figure 9. Spectrograms for FR.CAMF (left) and FR.VIEF (right). They are computed using 3600 s length windows with no overlap. The grey and magenta vertical lines correspond to the edges of the signal degradation time windows and plotted as dashed lines in Figs. 5, 7 and 8. For each spectrogram, the horizontal black dashed lines bound the four frequency domains.

4 Discussion and conclusion

The method presented in this article aims to point out anomalous features in the continuous seismic signals using different gaussianity estimators. As shown in the previous figures, we focus mainly on one of them, which is the ratio of the classical standard deviation σ and the BGS standard deviation σ_G. It can be compared to a method which aims to monitor the seismic signal quality using the ratio of the classical standard deviations for two components (Pedersen et al., 2020).

4.1 Comparison to a component ratio approach

In their approach, Pedersen et al. (2020) compute, for each component and 8 frequency bands, the classical standard deviation in 5-minute time windows of the continuous seismic signal, recorded at various Geoscope stations, with no overlap. For all the short time windows of a given day, the energy ratio is quantified by the ratio of the standard deviations for each pair of the three components (E/N, E/Z and N/Z). The estimate of the component daily energy is then defined using the median of all ratios.

In order to illustrate the difference between this method and the one presented in this paper, we present in Fig. 10 a focus on the HF domain around the degradation time for FR.VIEF as already studied in Figs. 7, 8 and 9. In addition to the $\log(\frac{\sigma}{\sigma_G})$ median values shown in Fig. 7, we display in Fig. 10 the decile interval comprised between the 1st and the 9th deciles of all daily $\log(\frac{\sigma}{\sigma_G})$ values (using the same colors as their median: blue for Z, red for N and green for E). For a given day and a given component, when all $\log(\frac{\sigma}{\sigma_G})$ values are very close, the dispersion is so small that it cannot be seen on Fig. 10 A. It is nevertheless possible to obtain a median value close to 0 with a large decile interval such as for January 26, 2022. On this day, all components are particularly affected by 60 local earthquakes ($0.4 \leq M_l \leq 2.8$) occurring around FR.VIEF and within an epicentral distance range of 100 km.

As soon as the sensor is corroded enough to affect the recording conditions at 2022-02-09T17:29 UTC (grey vertical dashed line), the decile interval suddenly increases up to 2.2 while the median is not yet modified. This is due to the fact that more than 50% of this day recorded a clean and Gaussian signal, as shown in the daily seismogram (Z component, HF filter) inserted in Fig. 10 A. The following days are characterized by both large values of the median of $\log(\frac{\sigma}{\sigma_G})$ and large width of the decile intervals. Finally, when the sensor has been disconnected at 2022-02-18T09:56 UTC (magenta dashed line) while the digitizer continued to operate, the recorded signal (pure numeric noise) has very low values in terms of median and deciles.

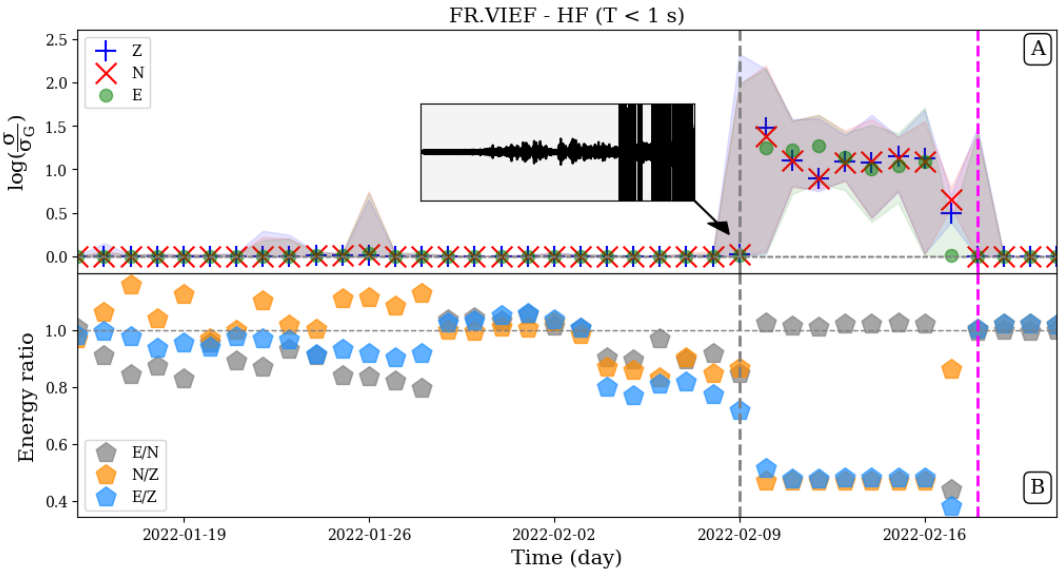

Figure 10. Comparison between our approach and a method based on energy ratios for each pair of components (Pedersen et al., 2020). The period of interest focuses on the FR.VIEF sensor degradation as previously shown in Figs. 7, 8 and 9. (A) For each component, the values of the 1st and the 9th deciles of all daily $\log(\frac{\sigma}{\sigma_G})$ are displayed with the same color as the associated median. (B) For each pair of components, the energy ratio (represented by colored pentagons) are given by the median of all daily ratio of standard deviation of the two considered components.

The same methodology as in Pedersen et al. (2020) is followed for this station. The three energy ratios for each pair of components are displayed in Fig. 10 B. Before the beginning of the signal degradation they are all characterised by a quite large discrepancy. Values are ranging between 0.8 and 1.16 although the daily signal is very clean. Indeed, visual inspection of the whole signal during these 25 days did not allow to spot any precursor of the alteration of the sensor connection which is *a contrario* well reflected by the very low $\log(\frac{\sigma}{\sigma_G})$ values that do not exceed 0.03 (A).

After the vertical grey dashed line, the variations of the daily energy ratios suddenly decrease to converge towards values of 0.47 for N/Z and E/Z and 1.02 for E/N which attest of the seismic signal modification. These values testify that, once the recording conditions have been degraded, the vertical component is about twice more energetic than the two others which are similar. The

comparison of one component with respect to another (B) can thus bring fruitful information on the actions to be taken (even if it is not the case here) although the estimator of the signal quality before the degradation is more stable in (A) than in (B).

4.2 Concluding remarks

The method presented in this article introduces a new approach to point out all samples of a given data set that do not agree the dominant gaussianity, referred to as BGS. For a given time window, means a set of n samples (and we estimate that n must be greater than 1,000, as shown in Fig. A), our approach relies on four parameters to characterize the gaussianity: M_{L^2}, $\mathcal{G}$, μ_G and σ_G. Using the classical definition of the standard deviation, $\log(\frac{\sigma}{\sigma_G})$ therefore measures the non gaussianity of a given data set. Although the M_{L^2}, $\mathcal{G}$ and μ_G bring useful information, $\log(\frac{\sigma}{\sigma_G})$ alone can efficiently estimate whether the considered data set follows a normal distribution. At the scale of a single day, since many time windows can be processed following a sliding strategy, the median of all $\log(\frac{\sigma}{\sigma_G})$ gives a good quantification of the daily overall gaussianity without giving too much weight to transient waveforms such as earthquakes. Thus, it could be used as a new estimator to reliably monitor the continuous seismic signal assuming that any modification in the recording conditions affects the gaussianity of the signal. As shown in this article, $\log(\frac{\sigma}{\sigma_G})$ is sensitive to both subtle changes on one or two components (Fig. 5) but also major degradations of sensors altering all of them (Figs. 6 and 7). It appears that to seize any kind of temporal modification, it is necessary to process various frequency ranges.

Although spectrogram analyses bring fruitful information they face two difficulties for monitoring purposes: i) for a given frequency range, the seismic energy vary a lot as function of days/months/years and ii) the detection of anomalous behaviour of the station needs long time series. *A contrario*, $\log(\frac{\sigma}{\sigma_G})$ includes in few values any statistical deviation from normal seismograph operation and does not depend on the signal energy. We consider therefore $\log(\frac{\sigma}{\sigma_G})$ as a simple and meaningful parameter to monitor seismic station quality. We propose that, for

a given frequency range, any daily $\log(\frac{\sigma}{\sigma_G})$ value greater than 0.1 requires a visual inspection of the signal since it corresponds to a σ value greater than 30% of σ_G. Finally, we think that this approach can bring useful information for seismic station monitoring purposes and then can be in line with methods that already exist. It can be used for permanent stations transmitting data in real time, as well as for identifying problems that occurred in the past.

Data and Resources

The Python code underlying this article will be shared on reasonable request to arthur.cuvier@etu.univ-nantes.fr. In this study we used data from networks with FDSN code FR (RESIF, 1995a) and G (Institut De Physique Du Globe De Paris (IPGP) and Ecole Et Observatoire Des Sciences De La Terre De Strasbourg (EOST), 1982). The seismic data set used in this study can be accessed at https://doi.org/10.15778/RESIF.FR and https://doi:10.18715/GEOSCOPE.G.

Acknowledgments

This project is funded by ANR-MAGIS-19-CE31-0008-02. Résif-Epos is a Research Infrastructure (RI) managed by the CNRS-INSU. Authors warmly thank Hélène Pauchet (IRAP-OMP) and Damien Fligiel (OSUNA) for there explanations about sensor failures. The work presented in this study was done with a Python program vanRossum (1995) using in particular the NumPy Harris et al. (2020), Scipy Virtanen et al. (2020) and Obspy Beyreuther et al. (2010) libraries for the signal processing. The figures presented in this study were generated with the Matplotlib library Hunter (2007).

References

Aggarwal, K., Mukhopadhyay, S., and Tangirala, A. K. (2020). Statistical characterization and time-series modeling of seismic noise. *arXiv preprint arXiv:2009.01549*.

Alu, K. I. (2011). *Solving the Differential Equation for the Probit Function Using a Variant of the Carleman Embedding Technique*. PhD thesis, East Tennessee State University.

Ardhuin, F., Stutzmann, E., Schimmel, M., and Mangeney, A. (2011). Ocean wave sources of seismic noise. *J. Geophys. Res.: Oceans*, 116(C9).

Beucler, É., Mocquet, A., Schimmel, M., Chevrot, S., Quillard, O., Vergne, J., and Sylvander, M. (2015). Observation of deep water microseisms in the north atlantic ocean using tide modulations. *Geophys. Res. Lett.*, 42(2):316–322.

Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., and Wassermann, J. (2010). Obspy: A python toolbox for seismology. *Seismological Research Letters*, 81(3):530–533.

Blair, J., Edwards, C., and Johnson, J. H. (1976). Rational chebyshev approximations for the inverse of the error function. *Mathematics of Computation*, 30(136):827–830.

Bliss, C. I. (1934). The method of probits. *Science*, 79(2037):38–39.

Casey, R., Templeton, M. E., Sharer, G., Keyson, L., Weertman, B. R., and Ahern, T. (2018). Assuring the quality of iris data with mustang. *Seismological Research Letters*, 89(2A):630–639.

Davis, P. and Berger, J. (2007). Calibration of the global seismographic network using tides. *Seismological Research Letters*, 78(4):454–459.

DeGroot, M. H. (2002). Probability and statistics.

Ebeling, C. W. (2012). Chapter one - inferring ocean storm characteristics from ambient seismic noise: A historical perspective. In Dmowska, R., editor, *Advances in Geophysics*, volume 53 of *Advances in Geophysics*, pages 1 – 33. Elsevier.

Ekstrom, G., Dalton, C. A., and Nettles, M. (2006). Observations of time-dependent errors in long-period instrument gain at global seismic stations. *Seismological Research Letters*, 77(1):12–22.

Feller, W. et al. (1971). An introduction to probability theory and its applications.

Finney, D. J. (1971). Probit analysis, cambridge university press. *Cambridge, UK*.

Francinou, S., Gianella, H., and Nicolas, S. (2013). *Exercices de Mathématiques (oraux X-ENS): analyse 2*. Cassini.

Glivenko, V. (1933). Sulla determinazione empirica delle leggi di probabilita. *Gion. Ist. Ital. Attauri.*, 4:92–99.

Groos, J. C. and Ritter, J. R. R. (2009). Time domain classification and quantification of seismic noise in an urban environment. *Geophys. J. Int.*, 179(2):1213–1231.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

Hoffman, D. L. and Low, S. A. (1981). An application of the probit transformation to tourism survey data. *Journal of Travel Research*, 20(2):35–38.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(03):90–95.

Hutt, C. and Ringler, A. (2011). Some possible causes of and corrections for sts-1 response changes in the global seismographic network. *Seismological Research Letters*, 82(4):560–571.

Institut de physique du globe de Paris (IPGP), & École et Observatoire des Sciences de la Terre de Strasbourg (EOST). (1982). *GEOSCOPE, French Global Network of broad band seismic stations*. Institut de physique du globe de Paris (IPGP), Université de Paris. https://doi.org/10.18715/GEOSCOPE.G.

Kimura, T., Murakami, H., and Matsumoto, T. (2015). Systematic monitoring of instrumentation health in high-density broadband seismic networks. *Earth, Planets and Space*, 67(1):1–15.

Kockelman, K. M. and Kweon, Y.-J. (2002). Driver injury severity: an application of ordered probit models. *Accident Analysis & Prevention*, 34(3):313–321.

McNamara, D. E. and Boaz, R. I. (2010). Pqlx: A seismic data quality control system description, applications, and users manual. *US Geol. Surv. Open-File Rept*, 1292:41.

Nawa, K., Suda, N., Fukao, Y., Sato, T., Aoyama, Y., and Shibuya, K. (1998). Incessant excitation of the earth's free oscillations. *Earth, planets and space*, 50(1):3–8.

Pedersen, H. A., Leroy, N., Zigone, D., Vallée, M., Ringler, A. T., and Wilson, D. C. (2020). Using component ratios to detect metadata and instrument problems of seismic stations: Examples from 18 yr of geoscope data. *Seismological Research Letters*, 91(1):272–286.

Peterson, J. (1993). Observations and modelling of seismic background noise. *US Geological Survey, open-file report*, 93 -322:1–94.

Pourhoseingholi, A., Pourhoseingholi, M. A., Vahedi, M., Safaee, A., Moghimi-Dehkordi, B., Ghafarnejad, F., and Zali, M. R. (2008). Relation between demographic factors and type of gastrointestinal cancer using probit and logit regression. *Asian Pac J Cancer Prev*, 9(4):753–5.

RESIF. (1995). *RESIF-RLBP French Broad-band network, RESIF-RAP strong motion network and other seismic stations in metropolitan France [Data set]*. RESIF - Réseau Sismologique et géodésique Français. https://doi.org/10.15778/RESIF.FR.

Ringler, A. T., Hagerty, M., Holland, J., Gonzales, A., Gee, L. S., Edwards, J., Wilson, D., and Baker, A. M. (2015). The data quality analyzer: A quality control program for seismic data. *Computers & Geosciences*, 76:96–111.

Stutzmann, E., Ardhuin, F., Schimmel, M., Mangeney, A., and Patau, G. (2012). Modelling long-term seismic noise in various environments. *Geophys. J. Int.*, 191(2):707–722.

Tasič, I. (2018). Interdependent quality control of collocated seismometer and accelerometer. *Journal of Seismology*, 22(6):1595–1612.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

vanRossum, G. (1995). Python reference manual. *Department of Computer Science [CS]*, (R 9525).

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Zhong, T., Li, Y., Wu, N., Nie, P., and Yang, B. (2015a). Statistical properties of the random noise in seismic data. *Journal of applied Geophysics*, 118:84–91.

Zhong, T., Li, Y., Wu, N., Nie, P., and Yang, B. (2015b). A study on the stationarity and gaussianity of the background noise in land-seismic prospecting. *Geophysics*, 80(4):V67–V82.

Contacts information

Arthur Cuvier: arthur.cuvier@etu.univ-nantes.fr

Éric Beucler: eric.beucler@univ-nantes.fr

Mickaël Bonnin: Mickael.Bonnin@univ-nantes.fr

Raphaël Garcia: raphael.garcia@isae-supaero.fr

A The Probit function

The so-called *Probit function* was first introduced by Bliss (1934). This probabilistic function was originally developped to measure the effectiveness of a poison used in the fight of insect pests. However, it turns out that the Probit function goes beyond the scope of Biology and concerns many fields (*e.g.* Hoffman and Low, 1981; Kockelman and Kweon, 2002; Pourhoseingholi et al., 2008). The wide range of applications is logically due to the fact that the distribution of any standard Gaussian law converges toward the Probit function. Moreover, the mathematical progress during the past decades allowed a better understanding of the Probit function and its properties (Finney, 1971; Alu, 2011). We present hereafter the mathematical theory of the Probit function. We focus on the analytical expression of the Probit function and prove the link between any sorted standard Gaussian set of samples and the Probit function through a convergence theorem.

A.1 Definitions

The cumulative distributive function (CDF) of a random real-value variable X is a function (not necessarily continuous), defined as

$$F(t) = \mathbb{P}(X \leq t), \quad \forall t \in \mathbb{R}. \tag{A1}$$

For any CDF named F, we can define its related quantile function,

$$Q(u) = \inf\{x \in \mathbb{R} \ ; \ F(x) \geq u\}, \quad \forall u \in [0,1]. \tag{A2}$$

Hence, Q is the left inverse of F. In the special case of a continuous CDF, we have then $Q = F^{-1}$.

A.2 The Probit function

We denote as ϕ the CDF in the special case of the standard Gaussian law ($\mu = 0$ and $\sigma = 1$). The related quantile function Q is now called the *Probit function*, and since ϕ is continous, $Q = \phi^{-1}$.

It is well known that ϕ can be expressed as

$$\phi(x) = \frac{1}{2}\left(1 + \mathrm{erf}\Big(\frac{x}{\sqrt{2}}\Big)\right), \tag{A3}$$

where erf denotes the error function. Consequently, computing the inverse function of ϕ, the analytic expression of the Probit function is thus given by

$$\phi^{-1}(u) = \sqrt{2}\,\mathrm{erf}^{-1}(2u - 1), \tag{A4}$$

where erf^{-1} could be, in practical, approximated by a Mac Laurin expansion (*e.g.*, Blair et al., 1976). The representative curve of the Probit function is plotted in green in Fig. A.

A.3 Empirical quantile function

This section is devoted to the link between the discrete equivalents of the CDF and the quantile functions, obtained from a given statistical sample $(X_1, .., X_n)$. This leads to the definition of both empirical CDF and empirical quantile function.

For any set of n samples $(X_1, .., X_n)$, we define the empirical CDF,

$$F_n(t) = \frac{1}{n}\mathrm{Card}\big(\{X_i\,;\, Xi \leq t\}\big), \tag{A5}$$

where $\mathrm{Card}(X)$ represents the cardinal function. Among the n values, $F_n(t)$ thus represents the proportion of points lower than t in a given set of samples.

Following eq. (A2), the empirical quantile function can be defined as

$$Q_n(u) = \inf\{x \in (X_1, .., X_n) \ ; \ F_n(x) \geq u\}, \quad \forall u \in [0, 1]. \tag{A6}$$

The empirical quantile function represents, for a given sample $(X_1, ..., X_n)$, its values sorted by increasing order of amplitudes. Indeed, $Q_n(u)$ represents the u-th quantile of a dataset $(X_1, .., X_n)$ as its smallest value for which the empirical CDF $F_n(x)$ is greater than or equal to u, effectively sorting the samples by increasing order of amplitudes.

For a set of random values, the convergence between the CDF and the empirical CDF can be found in the Glivenko-Cantelli theorem (Glivenko, 1933).

Theorem 1 ***Glivenko-Cantelli theorem***

Assuming that $(X_1, ..., X_n)$ are independent and identically-distributed random variables in $\mathbb{R}$ with common cumulative distribution function F. Then, we have an uniform convergence almost surely of F_n toward F, i.e.

$$\|F_n - F\|_\infty = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n \to +\infty]{} 0 \ \textit{almost surely.} \tag{A7}$$

Theorem 2

Assume that $(X_1, ..., X_n)$ are independent and identically-distributed random variables in $\mathbb{R}$ with common cumulative distribution function F and quantile function Q. Noting F_n the empirical CDF and Q_n the empirical quantile function, we have the following equivalence:

$$|F_n - F| \xrightarrow[n \to +\infty]{} 0 \iff |Q_n - Q| \xrightarrow[n \to +\infty]{} 0. \tag{A8}$$

Plus, in the special case of the the standard normal distribution, this convergence is uniform.

The proof is detailed in (Van der Vaart, 2000, chapter 21, lemma 21.2) and the uniform

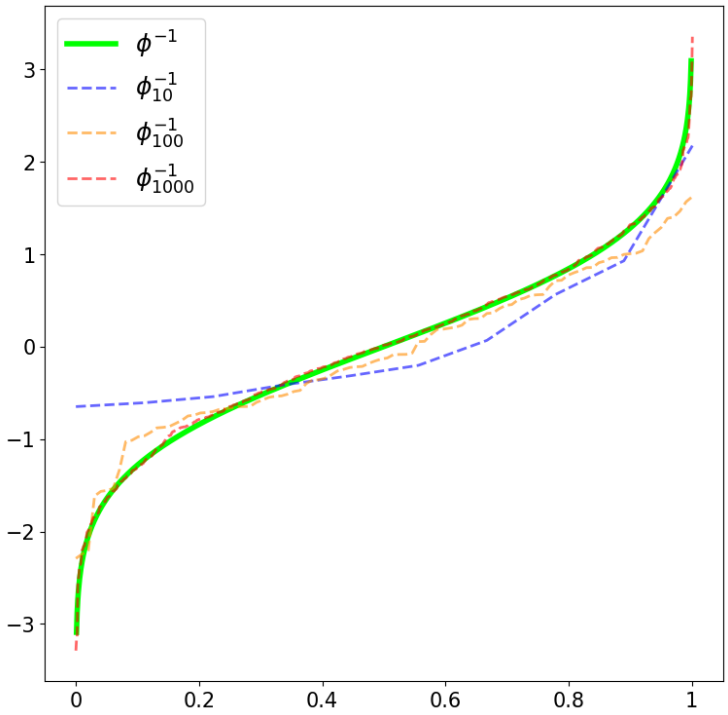

Figure A. Illustration of the convergence of the empirical discrete Probit functions ϕ_n^{-1} towards the Probit function ϕ^{-1} (theorem 2). Examples for $n = 10$ (blue), 100 (orange) and $1,000$ (red).

convergence in the particular case of the standard normal distribution is deduce by the Dini's theorem (Francinou et al., 2013). In the special case of the standard Gaussian law, the theorem 1 demonstrates the convergence of F_n towards ϕ, where F_n is the empirical CDF obtained from a random draw. Consequently, the theorem 2 ensures as well the convergence between ϕ_n^{-1} and ϕ^{-1}, where ϕ_n^{-1} denotes the empirical discrete Probit function. In order to illustrate this convergence, the result of a numerical experiment is presented in Fig. A. Three random draws of n elements ($n = 10$, $n = 100$ and $n = 1,000$) are realised to obtain $(X_1, ...X_n)$, where $X_i \sim \mathcal{N}(0,1), \forall i \in [1,n]$. Once data sets are sorted by increasing order of amplitude, they can be compared to the Probit function defined by eq. (A4), displayed in green. Each sorted data set thus is an empirical discrete Probit function, and we observe a reliable convergence since n is sufficently large.

3.3 Applications diverses de NG-loc sur le signal sismique

3.3.1 Estimation de l'hétérogénéité du signal sismique des stations du quart Nord-Ouest de la France.

Dans le cadre de l'acquisition du signal sismique en France métropolitaine, le laboratoire de planétologie et géosciences et l'observatoire des sciences de l'Univers Nantes Atlantique sont tous deux en charge de l'installation des stations sismiques du quart nord ouest de la France ainsi que la surveillance de la bonne qualité des données enregistrées. Dans ce contexte et au vu des résultats encourageants obtenus dans l'article présenté dans la précédente section 3.2, nous proposons désormais d'établir un classement de la qualité du signal enregistré par 26 stations sismiques du quart nord ouest de la France (réseau FR). Le terme de qualité, au sens de la sismologie, pouvant toutefois différer de notre résultat obtenu via l'analyse de NG-loc, nous préférons dans cette sous-section le terme, plus adapté, d'hétérogénéité statistique. De plus, ce choix est également motivé par la nature de NG-loc, indépendant de l'amplitude des signaux analysées. De la même manière que lors de la précédente section 3.2, chaque station sera ici examinée au sens de la distribution Gaussienne de son signal. L'étude des signaux par fenêtre glissante (d'une longueur de 10 minutes) s'effectuera sans *overlap*, impliquant une unique analyse de chaque portion du signal. Afin de garantir la fiabilité des résultats, l'étude portera sur l'analyse du signal enregistrée par les 26 stations durant l'année 2022.

Pour chaque station, l'étude d'une journée aboutit à un total de 144 fenêtres glissantes analysées, permettant d'extraire pour chacune d'entre elles, la valeur $\log(\frac{\sigma}{\sigma_G})$ associée (de la même manière que dans l'article présenté dans la section 3.2). Afin de quantifier l'hétérogénéité du signal au i-ème jour ($1 \leq i \leq 365$), on calcule la moyenne m_i de ces 144 valeurs de $\log(\frac{\sigma}{\sigma_G})$. À noter que le choix se porte ici sur le calcul de la moyenne, et non de la médiane comme présenté dans l'article. La moyenne est ici préférée à la médiane car celle-ci est sensible aux perturbations ayant affecté moins de

50% du signal journalier, comme par exemple l'influence anthropique (bruit de pas, circulation routière, etc...). Le choix de la médiane est principalement adapté aux analyses présentées dans l'article, car étudiant des stations éloignées les unes aux autres, ce qui permet alors de s'affranchir de la présence de larges séismes, et altère le signal de certaines stations. Toutefois, l'étude des stations effectuée dans cette sous-section étant limitée à une même zone géographique (voir figure 3.6), les séismes d'importantes amplitudes affectent donc de manière similaire chaque signal analysé. L'étude de l'hétérogénéité du signal lors d'une année nous permet d'obtenir un vecteur $[m_1, ..., m_{365}]$ où chaque élément témoigne de la l'hétérogénéité journalière du signal analysé. Afin d'estimer l'hétérogénéité annuelle du signal sismique, nous proposons d'étudier le paramètre M, représentant la moyenne de cette série de données.

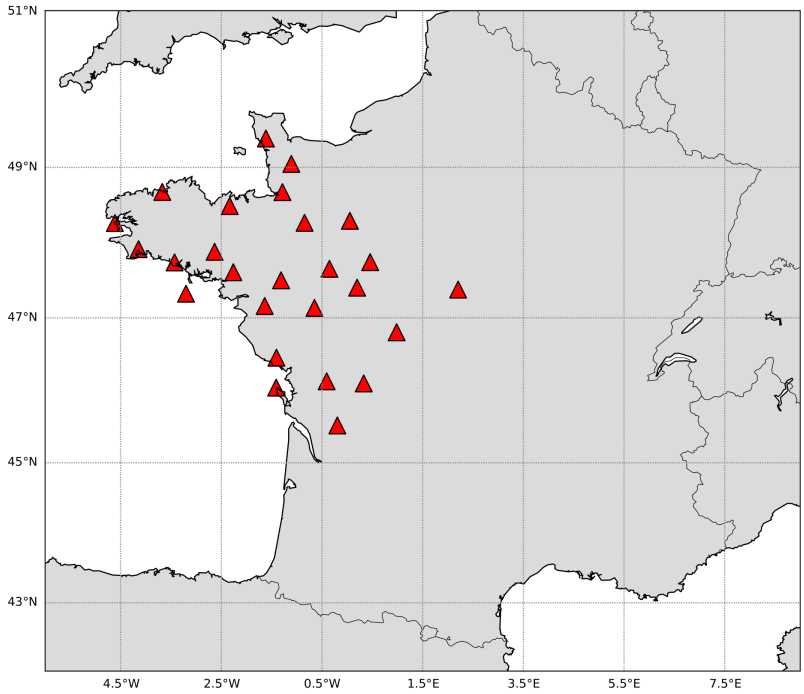

FIGURE 3.6 – Localisation des 26 stations sismiques analysées (réseau FR).

La figure 3.7 propose cet estimateur annuel M pour chacune des 26 stations analysées. Afin de proposer un classement de l'hétérogénéité des stations, celles-ci sont ordonnées par ordre croissant, par rapport à leurs valeurs de M associées. On distingue d'importantes différences entre ces 26 stations, avec des valeurs de M cinq fois plus élevées pour GIZF que pour SDOF valant 0,0065 et 0,0012, respectivement. Au vu

de notre estimateur d'hétérogénéité, SDOF est donc la station avec le signal le moins perturbé tandis que GIZF semble être celle constituée du plus grand nombre d'altérations. Afin de comprendre de tels écarts, nous proposons d'inspecter les signaux sismiques enregistrés sur ces deux stations.

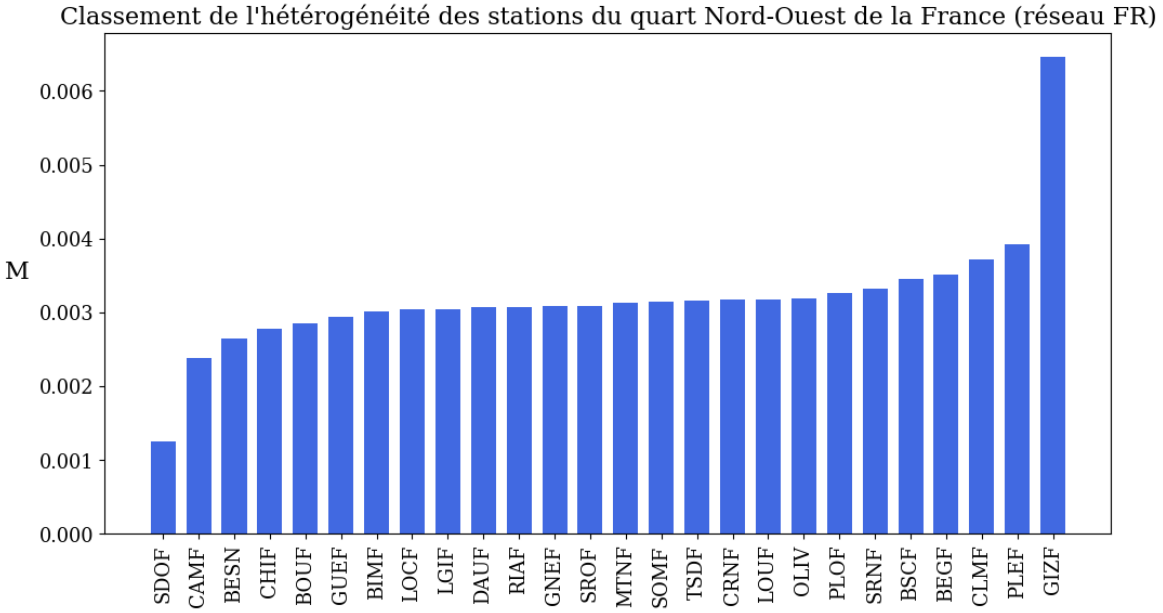

FIGURE 3.7 – Classement de l'hétérogénéité des stations du quart nord ouest de la France au cours de l'année 2022 (voir le texte pour la définition de l'estimateur M).

La figure 3.8 présente le signal des stations SDOF (A) et GIZF (B) enregistrés sur la composante verticale le 19 juillet 2022. En accord avec notre classement (figure 3.7), on distingue de très importantes différences entre la nature des signaux de SDOF, composé d'un bruit sismique régulier et GIZF, sujet à de très nombreuses altérations. Afin de comprendre l'origine de ces grandes altérations dans le signal de GIZF, il est nécessaire de comprendre le contexte environnemental autour de cette station. La station GIZF est située au milieu d'une zone agricole, entourée de nombreux champs, enregistrant alors un signal sismique fortement influencé par le passage de véhicules lourds (tracteurs, moissonneuses-batteuses, etc). Ceci explique la présence de ces très nombreuses perturbations hautes fréquences dans le signal (A), confirmée par une accalmie lors de la nuit, et autour de midi. À la vue des importantes perturbations anthropiques enregistrées par GIZF, il pourrait par exemple être intéressant d'envisager une relocalisation de cette station.

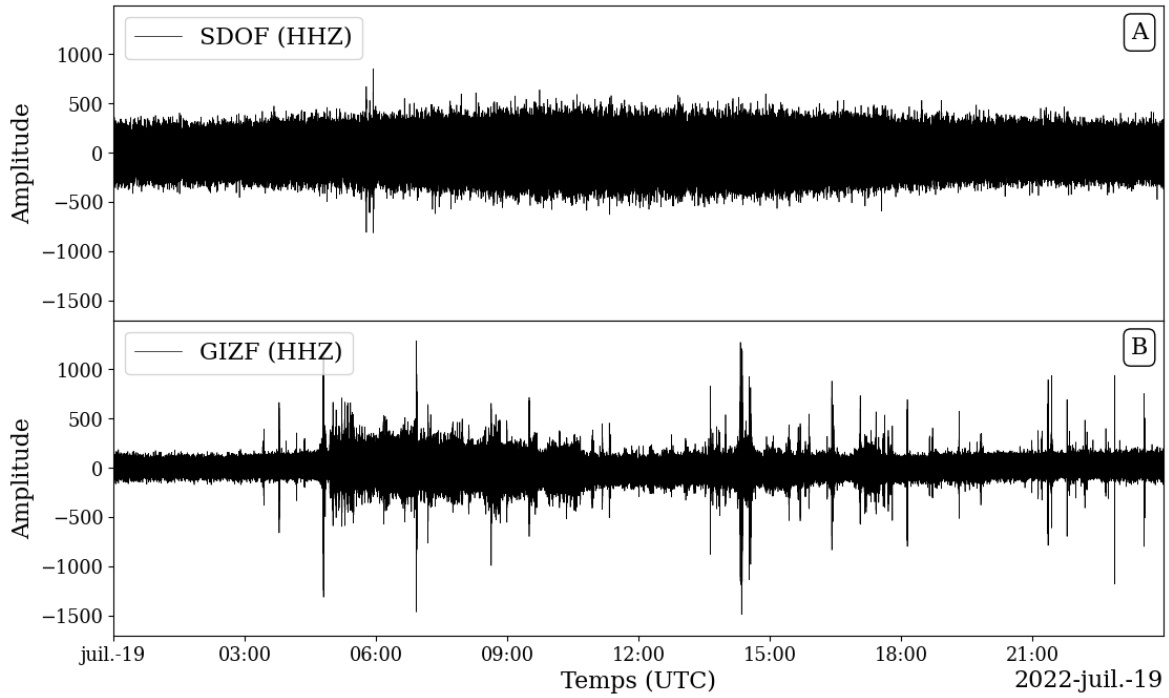

FIGURE 3.8 – Illustration de la différence d'hétérogénéité entre les stations SDOF (A) et GIZF (B), lors de la journée du 19 juillet 2022.

Au vu de ces fortes altérations anthropiques pouvant, comme dans le cas de GIZF, pouvant affecter l'hétérogénéité des signaux, nous effectuons désormais une étude similaire sur le signal haute fréquence (supérieur à 1 Hz.) La figure 3.9 présente un classement de l'hétérogénéité du signal sismique haute fréquence des 26 stations du quart Nord-Ouest de la France. On distingue dans cette gamme de fréquence d'importants écarts entre les scores obtenus pour ces différentes stations, variant de 0,003 pour SDOF à 0,06 pour la station PLEF. La comparaison entre les figures 3.7 et 3.9 est intéressante, nous révélant par exemple une grande hétérogénéité du signal brut de CAMF (deuxième meilleure station) alors que celle-ci est classée avant-dernière lors de l'analyse des ses données hautes fréquences. Cette différence de classement s'explique par le fait que le signal brut enregistré par la station CAMF (placé à quelques centaines de mètre de l'océan Atlantique) est très sensible au pics micro sismiques, dissimulant alors la majorité des perturbations hautes fréquences.

Nous proposons dans la figure 3.10 une comparaison entre le signal haute fréquence de CAMF et celui de la station SDOF, présentant de très grandes différences de clas-

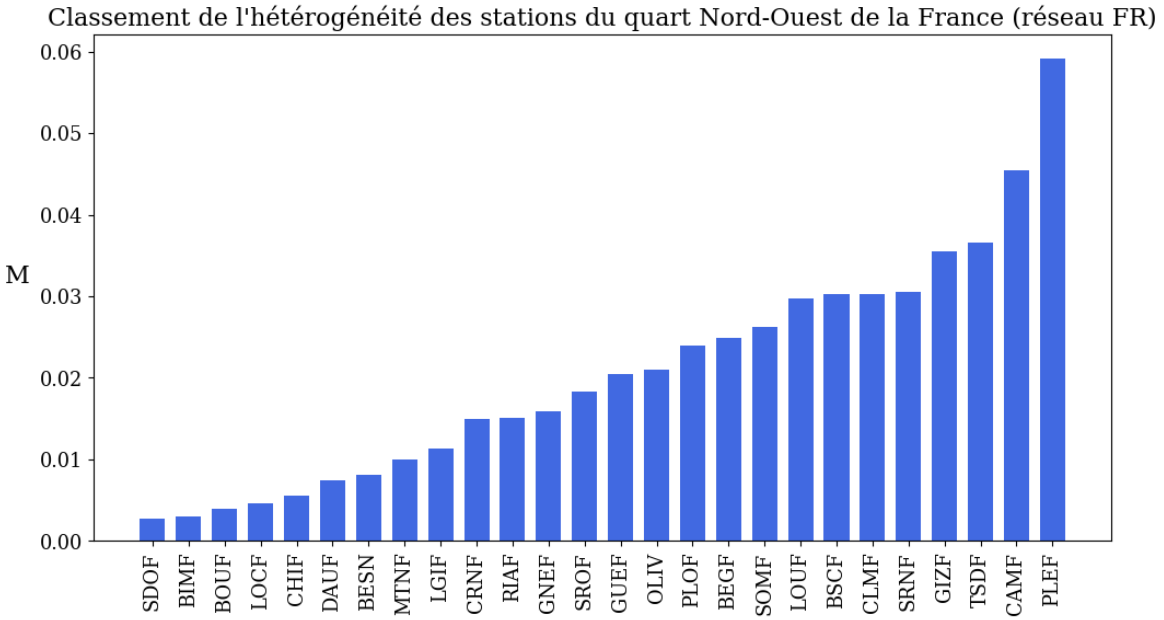

FIGURE 3.9 – Même légende que pour la figure 3.7, pour le signal haute fréquence.

sement dans la figure 3.9. Les différences mesurées par le classement sont bel et bien observées sur ces différents signaux. En effet, le signal SDOF (A) se caractérise par des oscillations régulières, présentant également un niveau de bruit très stable (autour des amplitudes -300 et 300 environ), même si de très rares altérations peuvent parfois être observées (à 6h00 par exemple). Ces légères altérations sont toutefois incomparables à celles observés sur CAMF (B), se produisant quasiment toute la journée, induisant irrémédiablement de forts écarts statistiques lors de l'analyse de ce type de signal par NG-loc. L'origine de ces altérations, peuvent s'expliquer par la localisation de la station CAMF, située à proximité d'un chemin et d'une route, enregistrant alors de nombreuses perturbations anthropiques liées à la circulation.

Bien que l'étude présentée dans la figure 3.8 se concentre sur les stations SDOF et GIZF, une analyse similaire peut également être effectuée sur d'autres stations, présentant des fluctuations d'hétérogénéité statistique cohérente avec nos classements des figures 3.7 et 3.9. Même si ce classement apporte des informations intéressantes quant aux perturbations statistiques des ces données sismiques, celui-ci ne constitue toutefois pas une vérité absolue et générale quant à la qualité des signaux analysés. La qualité du signal ne se définit pas uniquement par sa gaussianité et peut être analysée

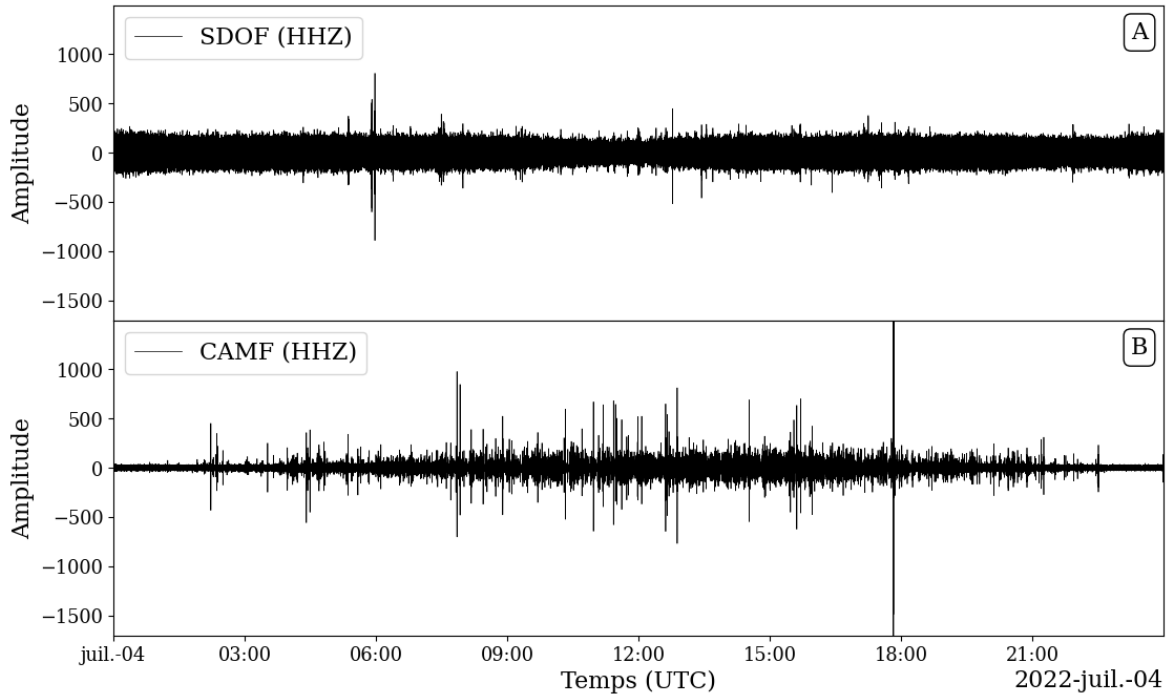

FIGURE 3.10 – Illustration de la différence d'hétérogénéité entre les signaux sismiques hautes fréquences des stations SDOF (A) et CAMF (B), lors de la journée du 4 juillet 2022.

via le prisme de différents critères (voir par exemple Pedersen et coll. (2020)). En effet, on remarque sur la figure 3.8 que l'amplitude du bruit sismique enregistré au cours des trois premières heures de la journée sur la station GIZF (B) est environ deux fois plus faible que celui de SDOF (A). Cette forte différence d'amplitude du bruit est également observée en comparant les deux premières heures des signaux hautes fréquences des stations SDOF (A) et CAMF (B) sur la figure 3.10. Par conséquent, il serait alors intéressant d'étudier le paramètre σ_G obtenu par NG-loc, témoignant de l'amplitude du signal de fond Gaussien.

3.3.2 Détection des séismes par NG-loc

La sismologie est un vaste domaine de recherche, se définissant par l'étude de la propagation des ondes sismiques dans le sol. Celle-ci se concentre principalement sur la détection de signaux impulsifs causés par des séismes ce qui permet 1) de les localiser et 2) d'inférer la structure interne de la Terre. En pratique, un séisme se caractérise par

une arrivée d'énergie soudaine dans le signal, dont l'amplitude se démarque du niveau de bruit enregistré précédemment. Ceci provoque alors une modification significative de la distribution statistique, altérant la gaussianité du signal de fond. Par conséquent, la méthode NG-loc est sensible à cette modification statistique et peut donc détecter ces formes d'ondes particulières venant altérer le signal de fond.

Nous choisissons de comparer, dans la figure 3.11, le résultat de NG-loc lors de l'analyse d'un séisme, avec une méthode classiquement utilisée en sismologie afin de détecter les arrivées d'ondes impulsives associées à des tremblements de Terre : le STA/LTA (Allen, 1978, 1982). Cette approche s'appuie sur l'étude du rapport de deux moyennes d'amplitudes, calculées sur une petite et une longue période temporelle. Lorsque ce rapport devient supérieur à une valeur fixée (seuil « ON »), un séisme est alors détecté dans le signal. Une fois ce rapport suffisamment faible (c'est-à-dire inférieur à un certain seuil « OFF »), ceci marque alors la fin de cet événement sismique.

L'événement analysé dans notre cas (A) correspond à un séisme de faible magnitude ($M_W = 1{,}6$), enregistré le 17 février 2021 près de la ville de Vannes (France). Celui-ci se caractérise par une arrivée d'onde P, d'amplitude relativement faible, à peine visible dans le signal (l'onde S se démarque quant à elle nettement du niveau de bruit). Les courbes de STA/LTA (B) sont calculées en utilisant deux couples de fenêtres de longueurs différentes : 1 s/4 s (bleu clair) et 0,5 s/2 s (bleu foncé). Les seuils ON/OFF de début/fin de détection sont ici fixés à 2,5 (trait rouge) et 2,0 (trait orange) respectivement, correspondant à des valeurs utilisées par les agences nationales. Bien que les deux courbes de STA/LTA aient détecté l'arrivée de l'onde S, comme indiqué par des valeurs supérieures au seuil ON à cet instant, aucune d'entre elle n'a été capable de localiser l'arrivée d'énergie de faible amplitude associée à l'onde P. NG-loc (C) localise toutes les amplitudes anormales lors de ces deux arrivées d'ondes. Cette sensibilité remarquable permet aussi de localiser des points perturbés appartenant à la coda, à la suite de l'onde S. Bien que ce séisme soit de faible magnitude, NG-loc s'avère être un outil efficace permettant de localiser ces points d'amplitudes anormalement élevés par rapport au signal de fond Gaussien.

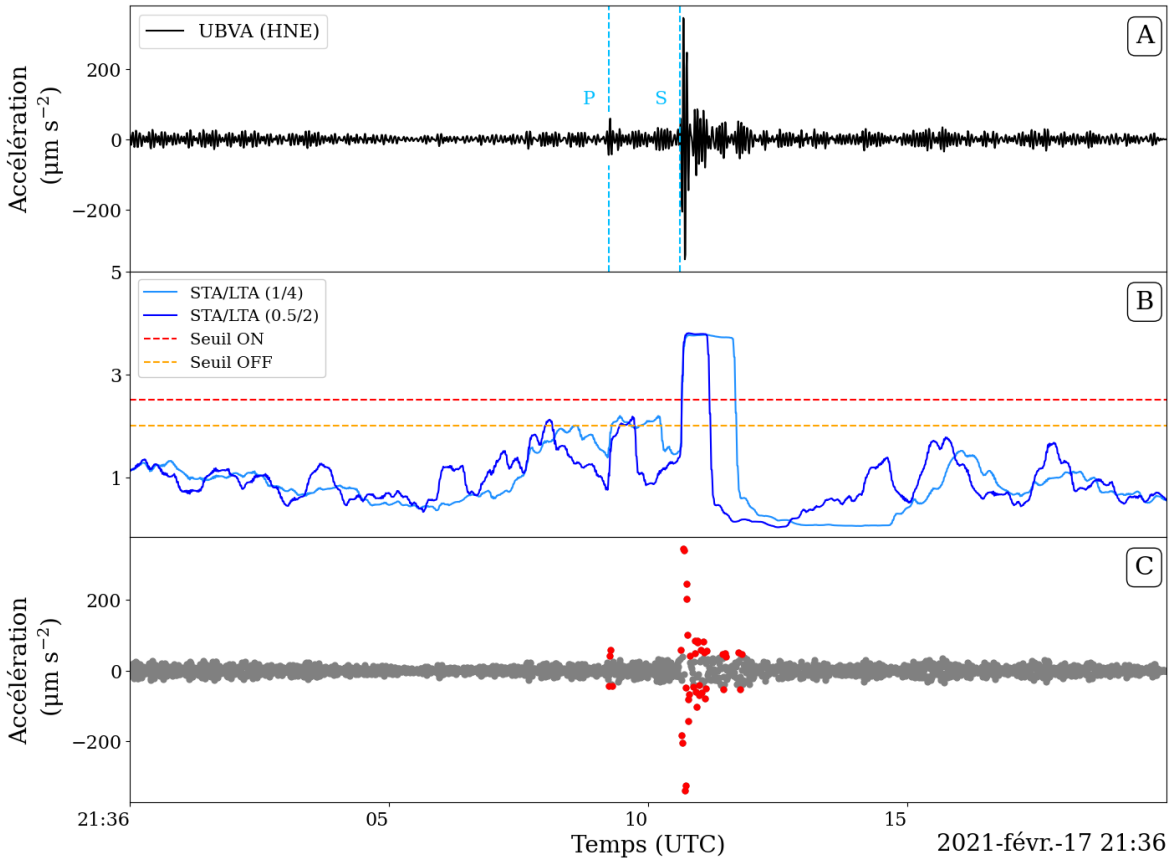

FIGURE 3.11 – Détection de séismes : comparaison entre NG-loc et STA/LTA. (A) : Signal sismique analysé (station UBVA ,Vannes, France). (B) : Résultat du STA/LTA, pour deux couples de fenêtres de longueurs différentes : 1/4 secondes (bleu clair) et 0,5/2 secondes (bleu foncé). Les seuils ON et OFF, utilisés pour détecter le début/la fin d'un séisme sont ici fixés à 2,5 (ligne rouge) et 2 (ligne orange), respectivement. (C) : Points non Gaussiens, détectés par NG-loc.

En conclusion, NG-loc se révèle capable de détecter des tremblements de Terre dans le signal sismique, s'inscrivant comme une alternative efficace aux méthodes traditionnellement utilisées. De plus, NG-loc se distingue de par sa simplicité d'utilisation, ne nécessitant aucun choix de seuil, se révélant bien souvent déterminant dans les processus de détections de type STA/LTA. Cette simplicité se traduit également par un unique choix de taille de fenêtre glissante pour NG-loc tandis que deux sont requis dans le cadre de la méthode STA/LTA. Au vu de ces différents éléments, nous pensons donc que NG-loc peut apporter d'intéressantes contributions dans le cadre de la détection d'événements sismiques de faibles amplitudes. Finalement, NG-loc peut également se révéler utile afin de pointer précisément les ondes P et S, de par sa capacité à localiser

au points près, chaque échantillon s'écartant de la distribution statistique Gaussienne du signal de fond.

3.3.3 Estimation de la durée de la coda : Application à la magnitude de durée

En complément de la magnitude de moment M_W et de la magnitude locale M_l, la magnitude de durée M_d est également utilisée pour mesurer l'énergie des tremblements de Terre, en général pour les faibles magnitudes $M_l \leq 5$. La magnitude de durée est déterminée via une estimation de la longueur de la coda d'un événement sismique. La coda d'un séisme se définit par l'énergie enregistrée après une arrivée impulsive de l'onde S, caractérisée par de fortes oscillations dans le signal, s'atténuant progressivement jusqu'à un retour à l'équilibre similaire à celui constaté avant le tremblement de Terre. La magnitude de durée est définie par Lee et coll. (1972),

$$M_d = -0{,}87 + 2\log_{10}(\tau) + 0{,}0035d \tag{3.1}$$

où d représente la distance séparant l'événement sismique de la station et τ la durée de la coda de l'onde S. Pour une station fixée, la détermination de la magnitude de durée repose donc entièrement sur une estimation de la longueur de la coda.

L'estimation de la durée de la coda est généralement obtenue par l'analyse de l'enveloppe de Hilbert du signal (Taner et coll., 1979), lorsque celle-ci revient à un niveau d'énergie similaire à celui observé avant le séisme. Cependant, cette technique d'estimation est parfois imprécise et peut entraîner une sous-estimation de la durée de la coda, pouvant être causée par un bref retour à l'équilibre n'étant toutefois pas associée à la fin de l'énergie du séisme. Nous proposons une méthode alternative, exploitant la sensibilité de NG-loc aux petits tremblements de Terre (voir section 3.3.2).

La figure 3.12 présente le signal sismique d'un séisme de faible magnitude locale ($M_l = 1{,}8$), dont les ondes P et S sont représentées par des lignes verticales oranges et vertes, respectivement (A). L'analyse de l'enveloppe du signal aboutit dans ce cas

à une estimation de la durée de la coda d'une longueur de $\tau = 8{,}24$ s, dont la fin est représentée par la ligne bleue verticale (A). Cette valeur de durée de la coda donne lieu à un calcul de magnitude de durée de $M_d = 0{,}986$, presque deux fois plus faible que la magnitude locale ($M_l = 1{,}8$).

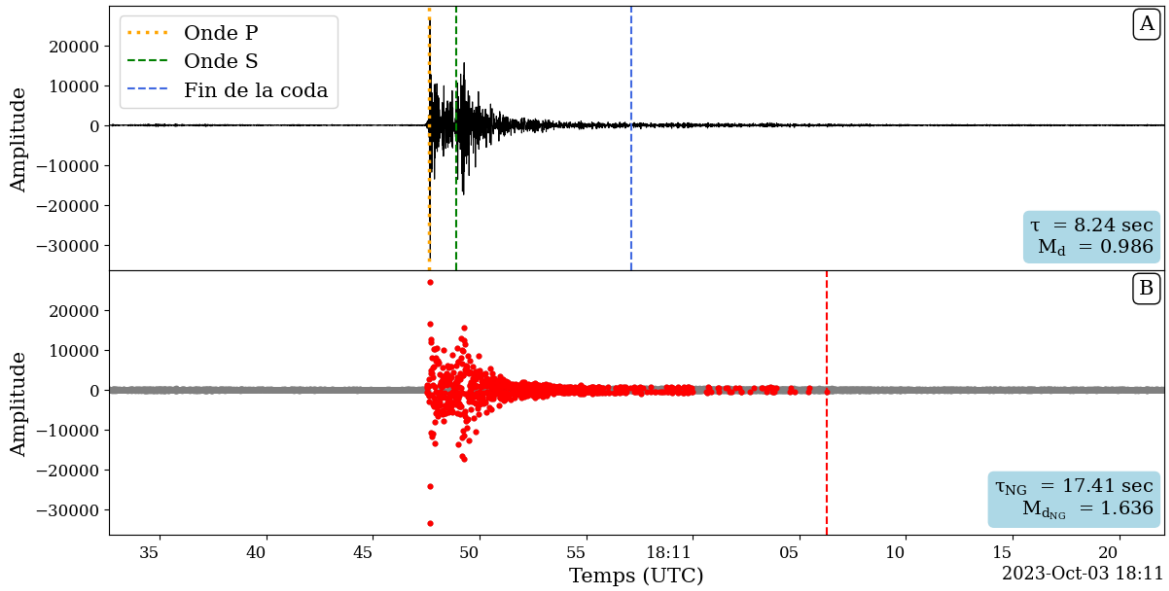

FIGURE 3.12 – Comparaison d'estimations de la durée de la coda d'un séisme via deux approches (station temporaire PSIS2). (A) : Détermination de la fin de la coda par l'étude de l'enveloppe de Hilbert du signal. (B) : Estimation de la durée de la coda en utilisant NG-loc. Le résultat de la fin de la coda obtenue par chaque approche est représentée par une ligne verticale bleue pour (A) et rouge pour (B).

La partie B de la figure présente le résultat de l'analyse de ce signal par NG-loc. Chaque point rouge indique la présence d'éléments s'écartant de la distribution Gaussienne générale. La durée de la coda est dans ce cas estimée via l'analyse de ces points non Gaussiens, et correspond au dernier élément perturbé conduisant à une valeur de $\tau_{NG} = 17{,}41$ secondes (ligne rouge verticale). En utilisant cette nouvelle valeur, on obtient alors une nouvelle magnitude de durée $M_{d_{NG}} = 1{,}636$, désormais bien plus proche que la magnitude locale de cet événement ($M_l = 1,8$).

En pratique, quelques irrégularités statistiques dans le signal sur des exemples de séismes différents peuvent causer la présence de quelques points étant classifiés comme perturbés, mais faisant tout de même partie du bruit sismique (voir le cas synthétique 2.7 L3). Cet artefact statistique ne constitue toutefois pas un problème et peut trouver

sa résolution dans le calcul de τ_{NG} pouvant être défini par le dernier point perturbé après l'onde S, précédent par exemple une période d'une seconde d'éléments non altérés. Dans le cas de la figure 3.12(B), une telle modification de la définition de τ_{NG} ne change pas le résultat obtenu.

Bien que la figure 3.12 présente une approche innovante d'estimation de la durée de la coda sur un signal donné, la magnitude de durée d'un événement se calcule en pratique par la moyenne des magnitudes de durées calculées sur plusieurs stations. La figure 3.13 présente les estimations de la durée de la coda du séisme présenté dans la figure 3.12, pour trois autres stations. Pour chaque exemple, la fin de la coda est déterminée par l'étude de l'enveloppe du signal (trait vertical bleu), mais aussi par la méthode NG-loc (trait vertical rouge). Des différences d'estimations sont observées sur chaque exemple, avec $\tau = 36{,}69$ s et $\tau_{NG} = 40{,}46$ s pour PSIS1 (A), ainsi que $\tau = 28{,}02$ s et $\tau_{NG} = 19{,}92$ s pour LGIF (B). L'exemple C est caractérisé par une arrivée d'énergie moins importante que dans les cas A et B. Par conséquent, l'amplitude du signal lors du séisme ne diffère pas nettement de celle enregistré par le bruit sismique (voir les 30 premières secondes du signal) et mène à une estimation de la fin de la coda immédiatement après l'onde S avec $\tau = 0{,}07$ s. *A contrario*, NG-loc étant extrêmement sensible aux faibles changements d'amplitudes dans le signal, l'estimation de la coda est donc ici de $\tau_{NG} = 1{,}95$ s.

En utilisant ces valeurs de τ et τ_{NG}, les magnitudes de durées sont ensuite calculées pour chaque station, et présentées dans le tableau 3.1. Une conséquence directe de l'estimation de $\tau = 0{,}07$ s pour la station BEGF résulte en l'obtention d'une magnitude négative $M_D = -2{,}494$ (qui ne sera pas exploitée par la suite). On remarque que les valeurs de magnitude de durées M_D obtenus par analyse de l'enveloppe du signal varient fortement par rapport à celles calculés via NG-loc. D'autre part, l'approche NG-loc présente dans l'exemple de BEGF un intérêt évident, ayant été la seule approche ayant permis un calcul cohérent de la magnitude de durée. En effectuant une moyenne des magnitudes de durées calculées pour chaque méthode, on obtient alors des valeurs de $M_D = 1{,}826$ pour la méthode utilisant l'enveloppe du signal et $M_{D_{NG}} = 1{,}966$ pour celle utilisant NG-loc.

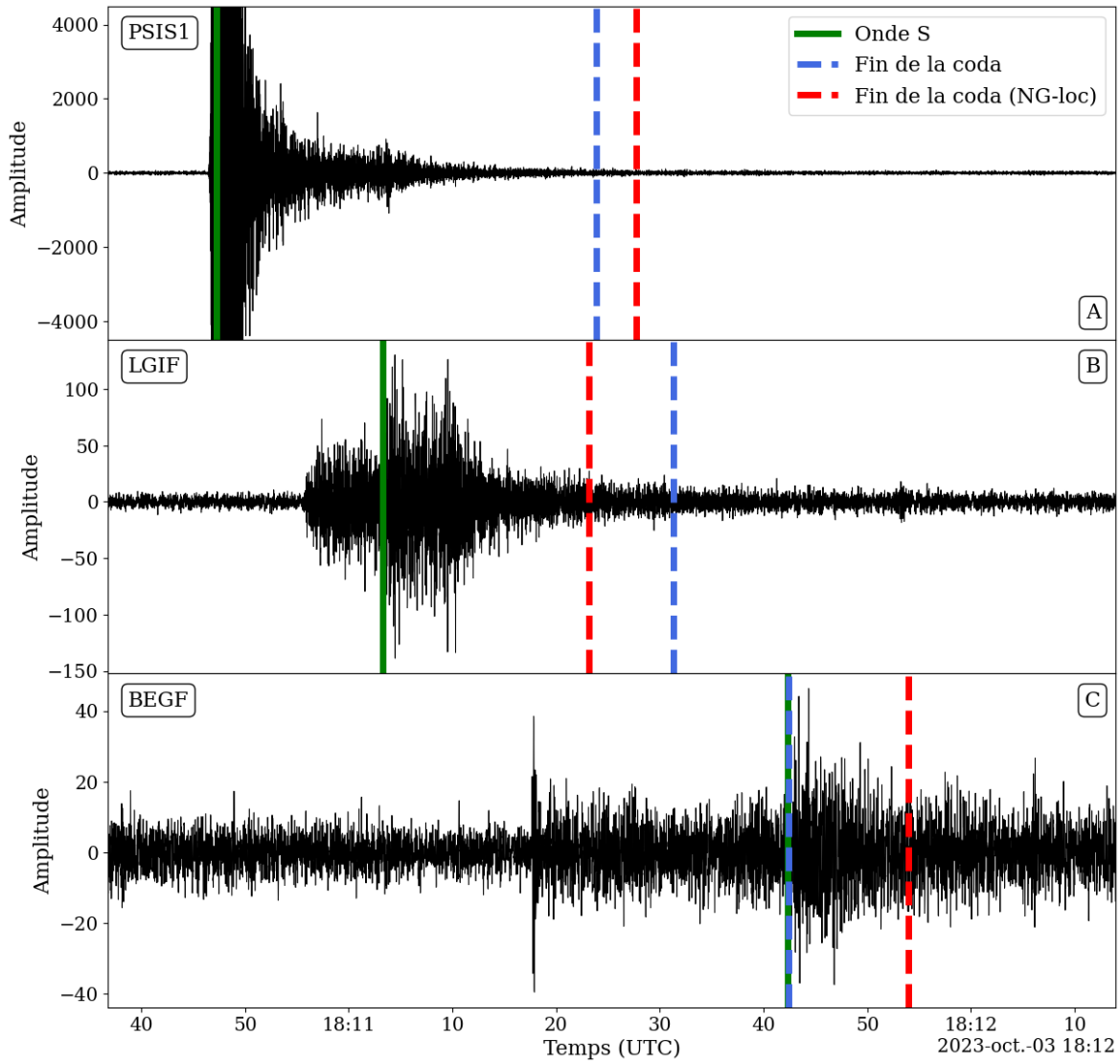

FIGURE 3.13 – Estimation de la durée de la coda du séisme présenté sur la figure 3.12 pour les stations PSIS1 (A), LGIF (B) et BEGF (C). Pour chaque exemple, l'onde S est représentée par un trait vertical vert. L'estimation de la fin de la coda via l'étude de l'enveloppe du signal et par NG-loc sont représentés par des traits verticaux bleus et rouges, respectivement.

TABLEAU 3.1 – Magnitude de durée obtenue sur chaque station via l'étude de l'enveloppe du signal et par NG-loc.

	M_D	$M_{D_{NG}}$
PSIS2	0,986	1,636
PSIS1	2,262	2,347
LIGF	2,231	1,935
BEGF	×	1,945

Il est important de noter que les résultats présentés dans cette section se concentrent uniquement sur l'analyse d'un seul événement. Dans le but de porter un jugement définitif sur l'intérêt présenté par NG-loc afin d'estimer la durée de la coda, il pourrait être intéressant de procéder à l'analyse d'un grand nombre de séismes.

Chapitre 4

Applications au signal sismique martien de la Mission InSight

Sommaire

4.1 Préambule

Les différentes applications présentées dans le chapitre 3 nous ont permis de justifier l'efficacité de NG-loc, se révélant capable d'estimer la qualité du signal sismique, mais aussi de mettre en évidence d'éventuelles dégradations. Nous proposons dans ce chapitre, de nous intéresser à l'analyse d'un signal sismique particulier, ayant été enregistré sur la planète Mars, dans le cadre de la mission spatiale InSight (2018-2022).

Nous introduisons tout d'abord dans la section 4.2 un aperçu des différentes missions spatiales ayant mené à l'acquisition de données sismiques sur des corps extraterrestres. La section 4.3 propose une présentation détaillée de la mission InSight, dévoilant les principaux objectifs de la mission ainsi que quelques caractéristiques typiques du signal sismique enregistré. Nous décrivons ensuite dans la section 4.4, l'application de NG-loc au signal sismique enregistré lors des quatre années d'acquisition des données sur le sol martien. Les résultats de cette analyse complète nous permettent par la suite de développer deux grand axes. Le premier axe, présenté lors la section 4.5, consiste en une analyse détaillée de la qualité du signal enregistré au cours de la mission, avec une étude minutieuse de plusieurs zones d'intérêts. Le second axe, quant à lui développé dans la section 4.6, présente une méthode de discrimination automatique par *machine learning* d'un certain type de perturbations sismiques, correspondant à des tornades de poussière.

4.2 Sismologie planétaire

Au cours du XXe siècle, la sismologie s'est révélée être un outil incontournable permettant la caractérisation de la structure interne de la Terre. Parmis les principaux résultats obtenu, on pourra par exemple citer les contributions de Jeffreys et Bullen, décrivant les modèles de vitesses des ondes sismiques à l'intérieur de la Terre, ayant finalement abouti au **PREM** (*Preliminary Reference Earth Model*) (Dziewonski et Anderson, 1981), mais aussi la découverte de la graine (Lehmann, 1936), ou encore les discontinuités de Gutenberg (Gutenberg, 2016) et de Mohorovicic (Mohorovicic, 1910). Au vu de ces résultats, dont la liste est évidemment loin d'être exhaustive, il est donc généralement accepté de considérer la sismologie comme étant le meilleur outil géophysique permettant la détermination des propriétés, et de la structure interne d'une planète. Par conséquent, il est alors naturel d'étendre son application à l'étude d'autres corps de notre système solaire dans le but de caractériser leurs structures internes, mais également de comprendre les processus de formation complexes de ces derniers. Afin de parvenir à cet objectif, un effort conséquent a été effectué à la fin du XXe siècle lors de différentes missions spatiales, en commençant par le programme Apollo (1969 - 1977), ouvrant alors la voie aux missions Vénéra (1982) et Viking (1976 - 1978) ayant toutes les trois étudié les ondes sismiques se propageant à travers la Lune, Vénus et Mars, respectivement.

Lors du programme Apollo, 4 sismomètres furent installés sur la surface de la Lune au cours de l'expérience **LSPE** (*Lunar Seismic Profiling Experiment*), formant ainsi le premier réseau de stations sismiques sur un corps extraterrestre. L'analyse de ces données sismiques, acquises sur une période de 8 ans, a mené à la détection de plus de 12 000 événements sismiques (Nakamura et coll., 1982; Latham et coll., 1969, 1970b,a, 1971), classés en 4 catégories : profonds, peu profonds, thermiques et impacts. L'exploitation de cette importante base de données de séismes a permis notamment d'analyser la vitesse de propagation des ondes sismiques, rendant alors possible d'obtenir les premières caractéristiques de la structure interne de la Lune (Nakamura et coll., 1976; Goins et coll., 1981).

En 1976, Viking 1 et Viking 2 furent les premières missions spatiales à envoyer des sismomètres sur la planète Mars, dans le cadre de l'étude de sa sismicité (Anderson et coll., 1976, 1977). Cependant, ces deux atterrisseurs se sont tous deux retrouvés face à des contraintes majeures empêchant toute mesure sismique fiable au cours de leurs missions respectives. Viking 1 fut sujet à un problème technique lors du déverrouillage de son sismomètre, peu après son atterrissage sur Mars, rendant ce dernier inutilisable. Le second sismomètre embarqué à bord de Viking 2 fut quant à lui incapable d'enregistrer des données sismiques exploitable, étant fixé au sommet de l'atterrisseur, rendant ce dernier extrêmement sensible aux vents martiens. Il faut ajouter à ça la difficulté de la transmission des ondes sismiques par la base de l'atterrisseur (muni d'absorbeurs de chocs), ainsi que la pollution des données causée par l'extrême sensibilité du sismomètre aux moindres mouvements mécaniques. Ces différents obstacles ne permirent pas de conclure à une quelconque activité sismique de Mars.

L'étude de la sismicité de Vénus relève quant à elle d'un défi autrement plus ardu que celle de la planète Mars. En effet, l'acquisition de données sismiques est rendue complexe par les conditions extrêmes régnant sur sa surface, avec une température moyenne de 460 °C, et une pression atmosphérique de l'ordre de 9 MPa (90 bar). Malgré ces conditions défavorables, une courte tentative d'acquisition de signal sismique fut toutefois menée à son terme sur Vénus, lors de la mission Venera 13 en 1982, ayant alors survécu quelques heures sur la surface de la planète. Dans le cadre de cette opération, le sismomètre embarqué à bord permit alors d'enregistrer environ 1h30 de signal, dévoilant pour la première fois le bruit de fond sismique de Vénus ainsi qu'une possible activité micro-sismique (Ksanfomaliti et coll., 1982). L'acquisition des données *in situ* étant extrêmement laborieuse sur Vénus, il sembla naturel de se tourner vers l'étude de la planète Mars, l'autre voisine de la Terre, se distinguant par des conditions de surfaces moins extrêmes, avec une température de l'ordre de -60 °C et une pression atmosphérique d'environ 600 Pascal (6 mbar) (Leovy, 2001; Barth, 1974).

En conclusion, bien que le programme Apollo ait été un succès incontestable ayant permis pour la première fois de caractériser l'intérieur d'un corps extraterrestre, le programme Venera et les missions Vikings n'ont eux pas abouti à l'interprétation de

résultats concernant les structures internes de Vénus et Mars, respectivement. C'est donc dans ce contexte, visant à la compréhension des caractéristiques physiques des planètes telluriques, que s'inscrit alors la mission spatiale InSight (2018-2022), ayant pour objectif de caractériser l'intérieur de la planète Mars, via l'acquisition d'ondes sismiques, pour la première fois directement depuis le sol de la planète.

4.3 La mission InSight

4.3.1 Présentation de la mission InSight

La mission spatiale **InSight** (*Interior Exploration using Seismic Investigations, Geodesy and Heat Transport*) a été sélectionnée par le programme Discovery de la NASA afin d'étudier les processus fondamentaux de la formation et de l'évolution des planètes telluriques en réalisant une étude géophysique *in situ* de la planète Mars. Suite à son décollage en mai 2018, InSight a atterri dans la région d'Elysium Planitia le 26 novembre 2018 (figure 4.1), déployant son instrument principal, **SEIS** (*Seismic Experiment for Internal Structure*) (Lognonné et coll., 2019), permettant d'enregistrer, pour la première fois, des ondes sismiques directement depuis le sol martien.

Bien que l'objectif principal de la mission consiste à caractériser la structure interne de Mars (croûte, manteau, noyau) via l'analyse de ces signaux sismiques inédits, InSight dispose également de nombreux autres instruments, permettant de mener à bien une étude géophysique complète des caractéristiques de son environnement (voir la figure 4.2). Parmi ces derniers, on pourra citer :

- la station météorologique **APSS** (*Auxiliary Payload Sensor Suite*) (Banfield et coll., 2019, 2020), incluant un magnétomètre, un capteur de pression atmosphérique ainsi que le capteur de vent et de température **TWINS** (*Temperature and WINd Sensor*),
- la sonde de flux de chaleur **HP^3** (*Heat flow and Physical Properties Package*) (Spohn et coll., 2018),

- le magnétomètre **IFG** (*InSight FluxGate*),
- l'instrument géodésique **RISE** (*Rotation and Interior Structure Experiment*) (Folkner et coll., 2018),
- le bras robotique **IDA** (Instrument Deployement Arm) (Trebi-Ollennu et coll., 2018), muni de la camera **IDC** (Instrument Deployement Camera) (Maki et coll., 2018).

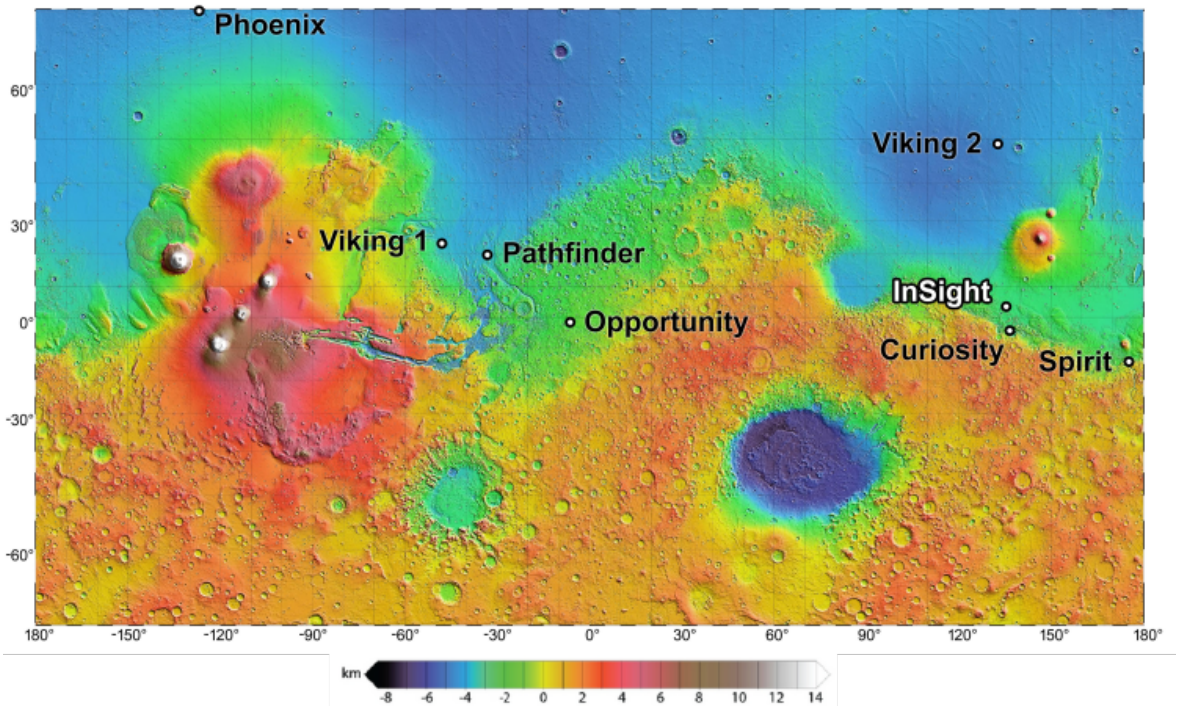

FIGURE 4.1 – Carte topographique de Mars, indiquant la localisation de l'atterrisseur InSight, en comparaison de quelques autres missions spatiales (crédits : NASA/JPL-Caltech). Voir Golombek et coll. (2017) pour des informations détaillées concernant le choix du site d'atterrissage.

L'instrument SEIS est équipé de deux sismomètres, l'un large-bande **VBB** (*Very Broad Band*) et l'autre courte période **SP** (*Short Period*), comportant chacun trois composantes et couvrant une gamme de fréquences étendue allant de 0,01 Hz à 50 Hz. La conception de SEIS relève d'une prouesse technologique et scientifique, de par sa capacité à enregistrer des déplacements du sol extrêmement faibles, inférieurs au rayon de l'atome d'hydrogène. Par ailleurs, plusieurs stratégies ont également été mises au points, afin de procéder à l'analyse et la polarisation des séismes martiens en utilisant

une unique station (Khan et coll., 2016; Böse et coll., 2017; Lognonné et coll., 2019; Drilleau et coll., 2020).

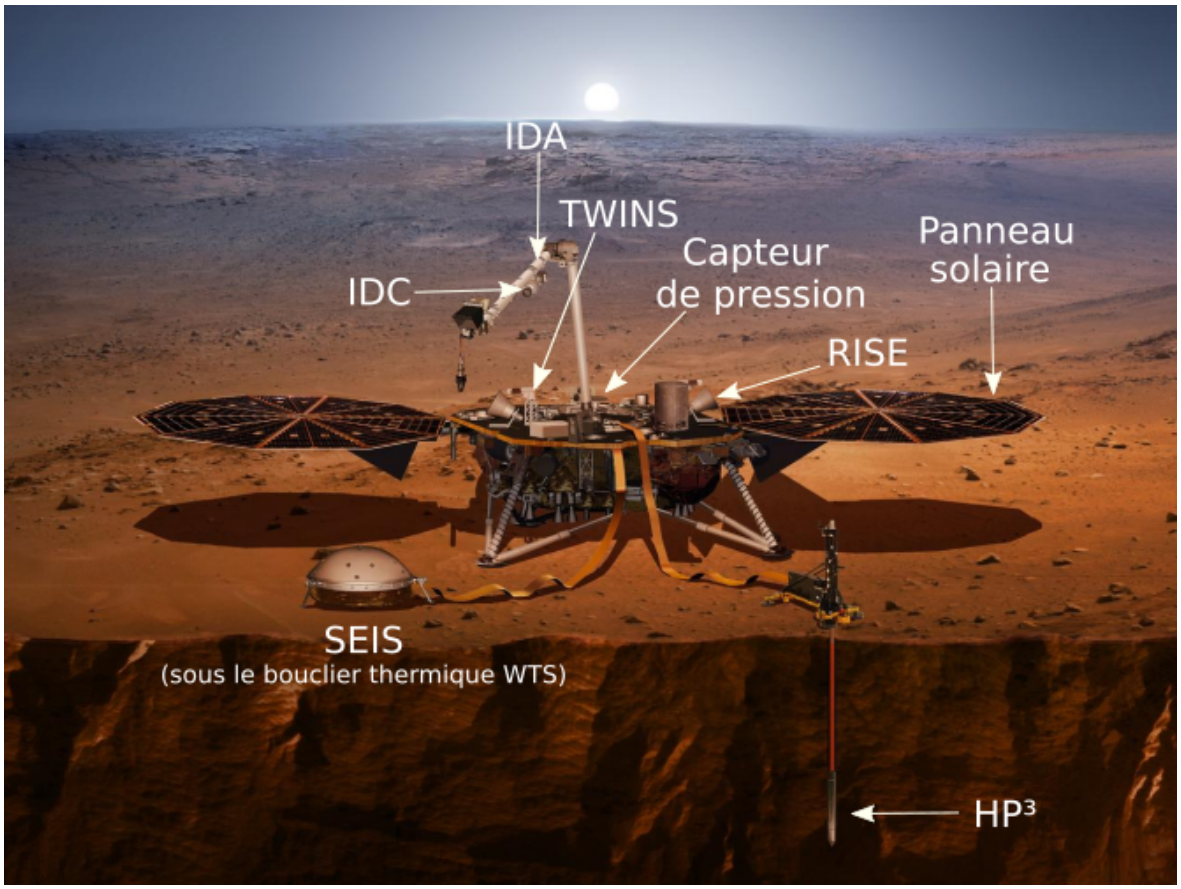

FIGURE 4.2 – Illustration artistique de l'atterrisseur InSight et de ses instruments (crédits : NASA/JPL-Caltech).

Le déploiement de SEIS sur le sol martien a été effectué à l'aide du bras mécanique IDA, quelques mois après l'atterrissage d'InSight, en février 2019, permettant alors d'enregistrer les premiers signaux sismiques directement depuis le sol de Mars, 42 ans après les missions Vikings (voir la figure 4.3 pour une photo/illustration de l'installation de SEIS). Afin de protéger SEIS des fortes variations de températures et des vents se produisant sur la surface martienne, celui-ci est recouvert par le bouclier thermique et éolien **WTS** (*Wind and Thermal Shield*), permettant ainsi de minimiser l'influence environnementale sur le signal sismique enregistré.

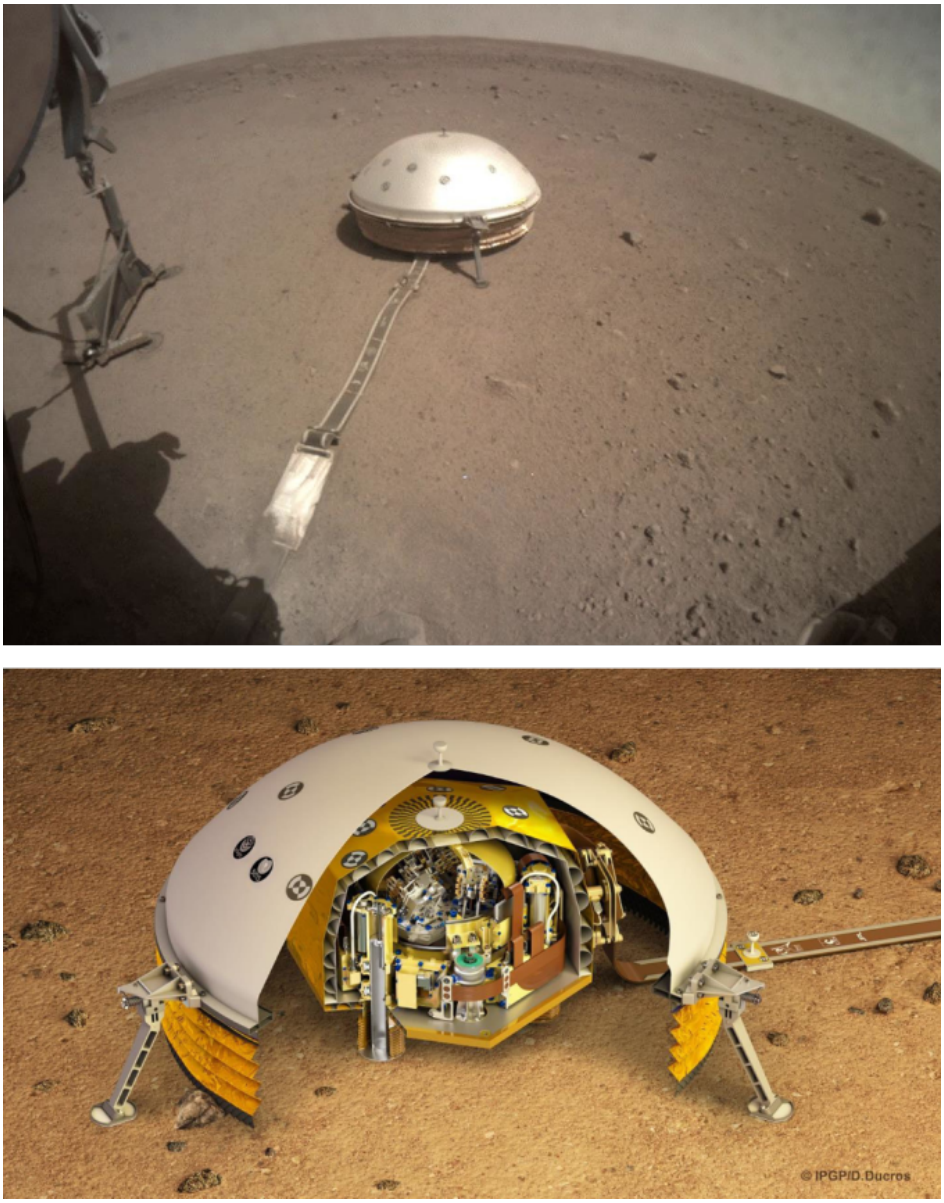

FIGURE 4.3 – (Haut) : Photo de l'instrument SEIS sur le sol martien (relié à l'atterrisseur par un câble), prise par la caméra IDC (crédits : NASA/JPL-Caltech). (Bas) : Illustration artistique de SEIS, à l'intérieur du bouclier thermique/éolien WTS (crédit : © IPGP/David Ducros).

La collecte des données *in situ* lors de la mission Insight s'est déroulée sur plus de 4 années terrestres, correspondant à 1446 jours martiens, appelés « sols » (d'une durée de 24 heures et 39 minutes environ). Ceci a permis, pour la première fois, d'enregistrer des séismes provenant de la planète Mars, avec un total de plus de 1300 événements détectés au cours de la mission (Ceylan et coll., 2022; Clinton et coll., 2021; Giardini et coll., 2020). Par ailleurs, parmi les nombreux résultats scientifiques obtenus par

InSight (Banerdt et coll., 2020), on pourra par exemple citer la première caractérisation de la sismicité de Mars (Giardini et coll., 2020), la détection de nombreuses tornades de poussière autour de l'atterrisseur (Spiga et coll., 2021; Lorenz et coll., 2021), mais aussi de multiples impacts de météorites ayant provoqués de large événements sismiques (Garcia et coll., 2022; Posiolova et coll., 2022).

La seconde moitié de la mission (2020-2022) a été marquée par une accumulation importante de poussières sur les panneaux solaires d'InSight, constituant son unique source d'énergie (voir la figure 4.4). Ceci a donc entraîné d'importantes restrictions énergétiques, imposant la mise en veille progressive de nombreux instruments et aboutissant finalement à la perte de contact avec l'atterrisseur le 21 décembre 2022.

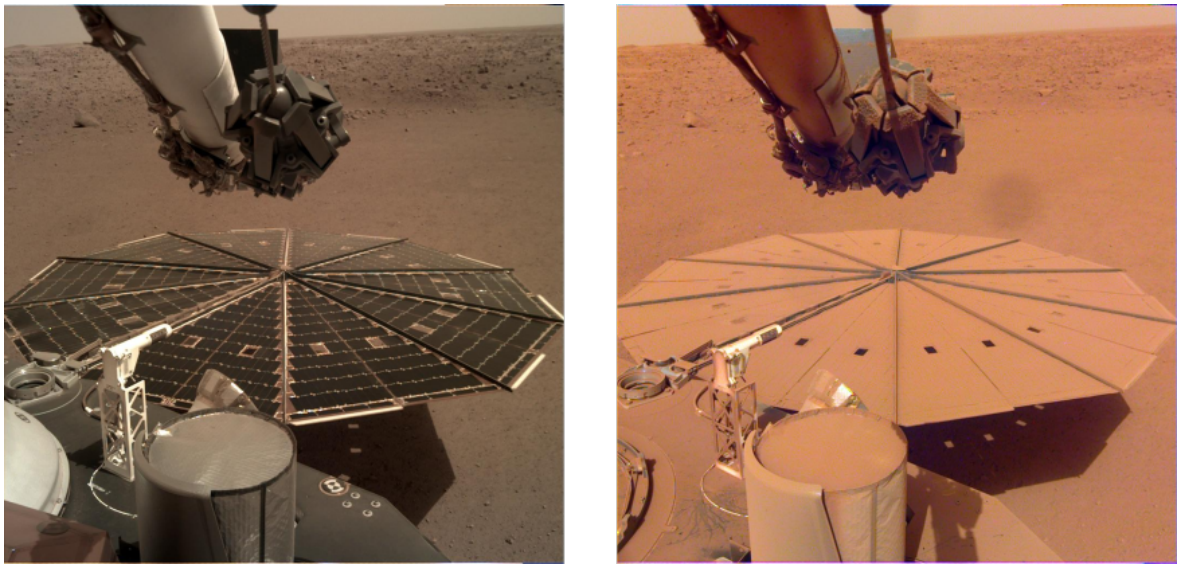

FIGURE 4.4 – Accumulation de poussières sur les panneaux solaires d'InSight au cours de la mission : en décembre 2018 (à gauche) et avril 2022 (à droite) ayant finalement conduit à la fin de la mission le 21 décembre 2022 (crédits : NASA/JPL-Caltech).

Les données sismiques SEIS enregistrées au cours de la mission InSight et présentées dans ce chapitre ont été obtenues via le InSight Mars SEIS Data Service. (2019). SEIS raw data, Insight Mission. IPGP, JPL, CNES, ETHZ, ICL, MPS, ISAE-Supaero, LPG, MFSC (https://doi.org/10.18715/SEIS.INSIGHT.XB_2016). Nous remercions la NASA, le CNES, leurs agences et institutions partenaires (UKSA, SSO, DLR, JPL, IPGP-CNRS, ETHZ, IC, MPS-MPG) et l'équipe chargée des opérations de vol au JPL, SISMOC, MSDS, IRIS-DMC et PDS pour avoir fourni les données SEED SEIS.

4.3.2 Caractéristiques du signal sismique martien

4.3.2.1 Signal sismique journalier

Bien que le signal sismique martien enregistré au cours de la mission présente, évidemment des similarités évidentes avec celui obtenu sur Terre, celui-ci possède toutefois de nombreuses caractéristiques qui lui sont propres. En effet, bien que le bouclier thermique WTS atténue l'influence du vent et des fortes variations de température au sein d'une même journée martienne (de l'ordre de 60 °C), SEIS reste tout de même sensible à ces conditions environnementales extrêmes, bien différente de celles observées généralement sur Terre. Par ailleurs, la position de SEIS, directement sur le sol martien, est également atypique en comparaison de celles des stations terrestres, bien souvent installées à l'abri dans des puits de forages/mines. De plus, le contenu fréquentiel du signal martien est, nécessairement et intrinsèquement différent de celui observé sur Terre, influencé majoritairement par les pics microsismiques, provoqués par l'activité océanique (Ebeling, 2012; Beucler et coll., 2015). La proximité directe de l'atterrisseur InSight, provoque inévitablement une influence directe de ce dernier sur la qualité du signal sismique enregistré. Au vu de ces différences notables, nous proposons dans cette section de présenter quelques caractéristiques propres du signal sismique enregistré lors de la mission InSight, ainsi que de son contenu fréquentiel.

La figure 4.5 présente le signal sismique brut (vitesse du sol) d'une journée martienne typique, enregistré par le sismomètre VBB. On propose ici d'étudier le sol 319, c'est-à-dire le 319e jour martien depuis le début de la mission, débutant au sol 0. Trois signaux sont présentés dans la figure 4.5, correspondant aux trois composantes du sismomètre, nommées U (en haut), V (au milieu) et W (en bas) (voir Lognonné et coll. (2019) pour une description complète de l'instrument). Le signal affiché ici est échantillonné à 20 points par seconde. On remarquera que l'axe temporel est ici représenté en temps LMST (*Local Mean Solar Time*, voir Allison et McEwen (2000)), correspondant à l'heure martienne à la position sur Mars de l'atterrisseur InSight. De la même manière que l'heure terrestre habituellement utilisée, un jour martien est donc composé naturellement de 24 heures LMST.

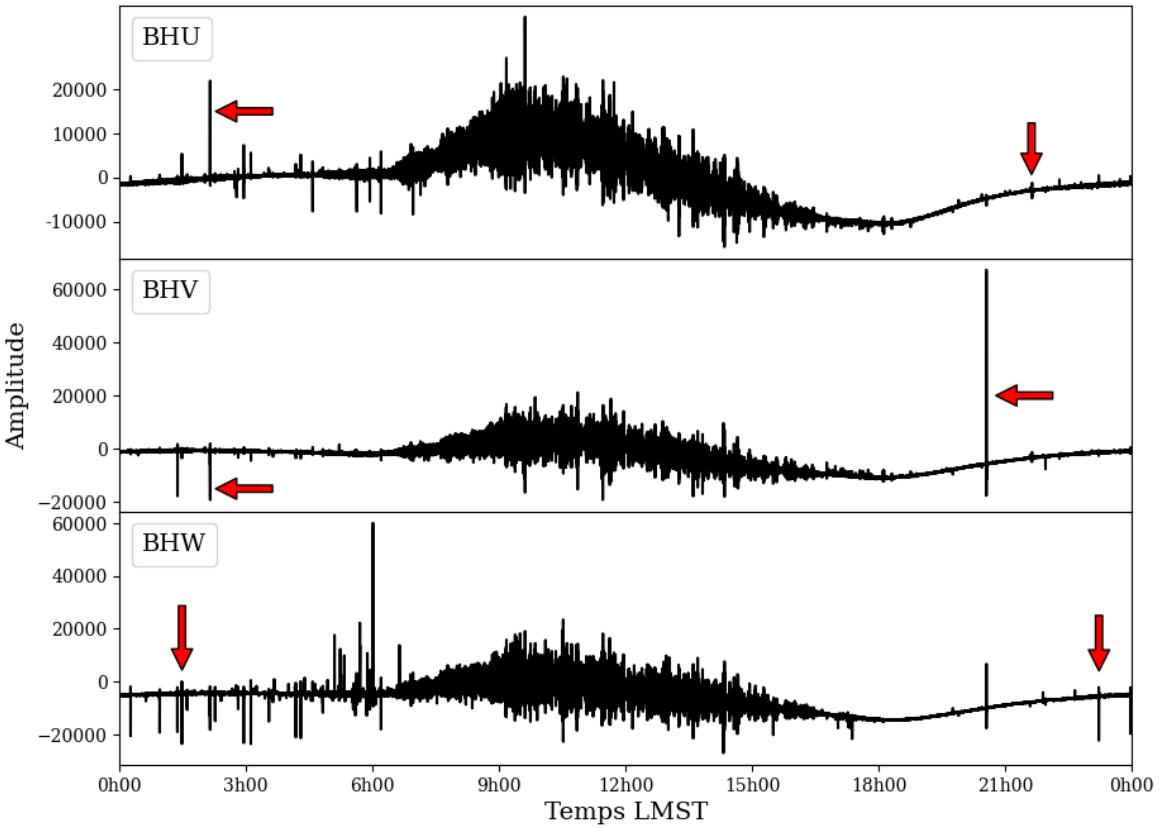

FIGURE 4.5 – Signal sismique brut enregistré par le sismomètre très large bande VBB, lors du sol 319 de la mission InSight, sur les composantes U (en haut), V (au milieu) et W (en bas).

Une première observation de ces données sismiques nous permet de constater une augmentation considérable de l'amplitude du signal au milieu de la journée martienne, entre 6h00 et 18h00 LMST environ, ainsi que de fortes variations du niveau de base du signal (c'est-à-dire sa moyenne) au cours d'un même sol. Ces caractéristiques du signal sismique enregistré par InSight, visible sur tous les sols, est une conséquence de l'installation de SEIS, directement sur le sol martien, enregistrant un bruit sismique fortement influencé par le vent au cœur de la journée ainsi que de fortes fluctuations de température (Lognonné et coll., 2020; Charalambous et coll., 2021).

La figure 4.6 propose une illustration du contenu fréquentiel composant le signal sismique martien lors du sol 319. Une différence notable avec les spectrogrammes associés au signal sismique terrestre (voir figure 3.3) est, bien évidemment, l'absence d'énergie associée au pic microsismique (entre 0,05 et 0,5 Hz), causée sur Terre par l'influence des

océans. Lors de la journée martienne, on observe une forte amplification de l'énergie du signal, excitant toutes les fréquences entre 6h00 et 18h00 LMST, correspondant à la période temporelle où l'amplitude du signal sismique augmente dans la figure 4.5.

Par ailleurs, on distingue également sur la figure 4.6, de nombreuses altérations du contenu fréquentiel sur des gammes de fréquence variées, représentées par des lignes vertes. On remarque que la plupart de ces modes de résonance sont constants au cours du sol, même si la fréquence de certains d'entre eux diminuent sensiblement au milieu de la journée. Ces altérations sont causées, pour la plupart d'entre elles, par la résonance de l'atterrisseur InSight, dont la proximité avec le sismomètre SEIS altère inévitablement le signal sismique enregistré. Parmi ces modes altérés, on pourra cependant noter quelques cas particuliers, comme par exemple la résonance à 2,4 Hz, probablement naturelle, dont l'origine a été localisée sous l'atterrisseur (Giardini et coll., 2020) ou encore celle à 1 Hz exactement, nommé « tick noise », causée par un artefact électronique (Ceylan et coll., 2021).

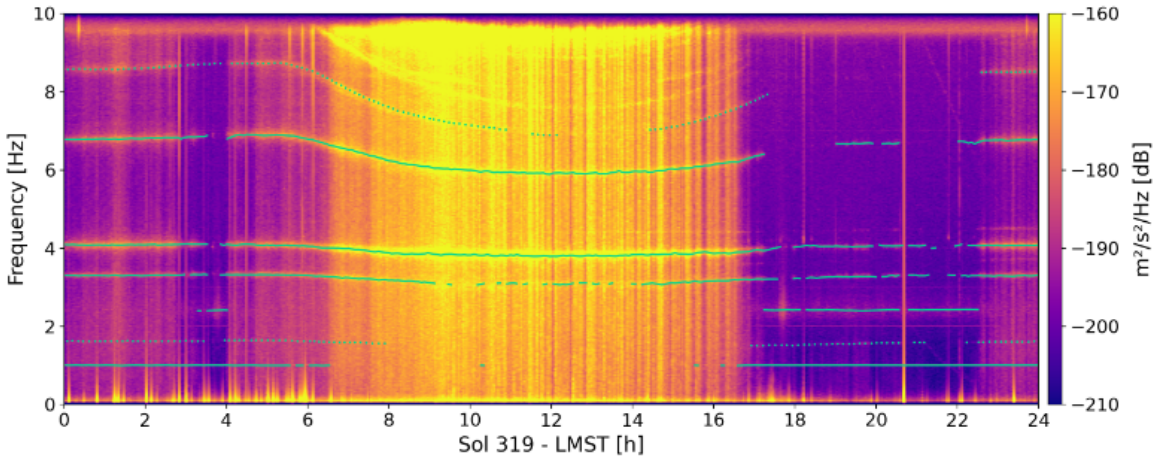

FIGURE 4.6 – Scalogramme du signal sismique (sol 319), construit par somme des spectrogrammes sur les 3 composantes BHU, BHV et BHW, calculés en utilisant une fenêtre glissante de 300 s et un *overlap* de 50% (c'est-à-dire la proportion commune entre deux fenêtres successives). Figure extraite de Dahmen et coll. (2021).

Finalement, on remarque une forte énergie affectant les basses fréquences (inférieures à 1 Hz), particulièrement visible lors de la nuit martienne (voir par exemple entre 4h00 et 6h00 LMST). L'étude de ces artefacts, également visibles dans le signal

sismique (flèches rouges dans la figure 4.5) fera l'objet d'une description détaillée dans la prochaine section.

4.3.2.2 *Glitchs*

Une caractéristique atypique du signal sismique martien est la présence de nombreux « pics transitoires », appelés *glitchs*, altérant fortement la qualité des données enregistrées (voir les flèches rouges dans la figure 4.5). Ces *glitchs*, sont caractérisées par des amplitudes différentes, mais aussi des périodes variées (variant de 1 à 30 secondes environ) et sont particulièrement visibles lors de la nuit martienne, lorsque le bruit enregistré est beaucoup plus faible. La fréquence élevée d'apparition des *glitchs* (100/200 occurrence par sol), ainsi que la difficulté des ces derniers à être facilement détectés a donc entraîné un effort particulier lors du début de la mission ayant pour objectif la caractérisation, détection et suppression de ces derniers (Scholz et coll., 2020). Cet effort est également justifié car la présence des *glitchs* complique fortement l'analyse de la forme d'ondes des séismes (Giardini et coll., 2020), induit des artefacts dans les techniques d'autocorrélation (Kim et coll., 2021), les méthodes des fonctions récepteurs (Knapmeyer-Endrun et coll., 2021) et se révèle également hautement problématique lors de la détection des marées induites par Phobos (Pou et coll., 2021), l'un des deux satellites naturels de Mars.

Bien qu'il soit établi que les *glitchs* ne constituent pas une source de signal sismique, mais plutôt un artefact mécanique, l'origine exacte de ces derniers (possiblement multiples) est toutefois discutée (Scholz et coll., 2020). Parmi les causes les plus probables, on pourra par exemple citer celle d'une relaxation thermique de SEIS, induite par les fortes variations de température sur Mars, d'infimes mouvements de torsion du câble reliant SEIS à l'atterrisseur (voir la figure 4.3), ou encore une légère instabilité du sismomètre sur le sol martien.

La figure 4.7 propose une vue détaillée des formes d'ondes de quelques *glitchs* altérant le signal sismique martien au cours de la mission. Le premier exemple (A1, A2, A3) présente trois *glitchs*, (numérotés 1, 2 et 3) ayant perturbé le niveau de base du

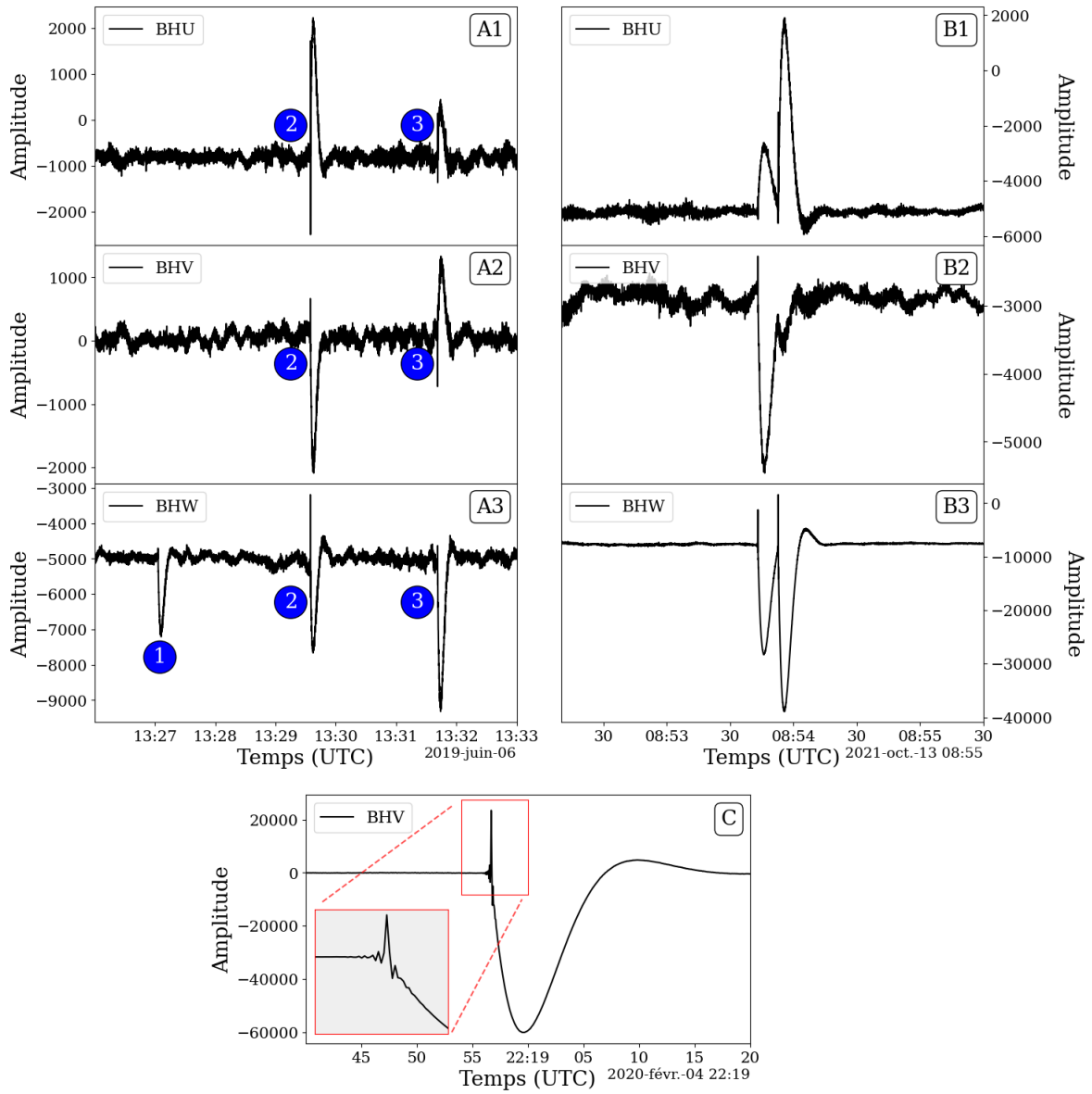

FIGURE 4.7 – Quelques exemples de *glitchs* altérant le signal sismique martien. (A1,A2,A3) : Trois *glitchs* provenant du sol 187 (numérotés 1,2 et 3), affectant une seule (1), ou plusieurs composantes (2 et 3). (B1,B2,B3) : Double *glitchs*, enregistrés lors du sol 1024, affectant simultanément les composantes U, V et W. (C) : Zoom sur un *glitch* et son précurseur (sol 424), altérant le signal sismique de la composante V.

signal. En zoomant pour la première fois sur ces *glitchs*, on remarque que le contenu fréquentiel de ces derniers est alors principalement constitué d'une altération longue période. La forme d'onde des *glitchs* suit une certaine polarité, étant orientés soit vers le haut ou vers le bas (c'est-à-dire supérieur ou inférieur à la moyenne locale), avant de revenir lentement au niveau de base du signal. Cette polarité des *glitchs* est extrême-

ment variable et dépend de la nature de ces derniers. On constate en effet, que même si la polarité des *glitchs* est toujours la même sur les composantes U et W, elle change cependant sur la composante V : les *glitchs* 2 et 3 étant « tournés » vers le bas, et le haut, respectivement. Par ailleurs, on remarque également que l'altération n'est pas systématiquement multi-composantes, comme pour le *glitch* 1, affectant uniquement le signal sismique de BHW.

Le second exemple (B1, B2, B3) présente un cas classique de double *glitchs*, observés régulièrement sur le signal sismique. Leur présence régulière dans le signal se révèle problématique pour la détection des *glitchs*. En effet, certaines techniques de détections, basées sur la cross-corrélation des signaux (en modélisant la forme d'onde du *glitch* recherché en amont), sont alors inefficaces dans ce genre de cas (voir Scholz et coll. (2020) pour un inventaire complet des méthodes de détections des *glitchs*).

Dans l'exemple C, on propose d'observer, sur une fenêtre de temps encore plus réduite (40 s ici contre quelques minutes dans les exemples précédents) la forme d'onde typique d'un *glitch* dans le signal sismique. On observe que celui-ci n'est pas uniquement composé d'un signal longue période, mais est également constitué d'un « choc impulsif », quasi instantané, et haute fréquence, appelé précurseur (rectangle rouge). Les précurseurs sont visibles sur de nombreux *glitchs*, témoignant de leur origine, et sont probablement causés par de brefs chocs mécaniques. La partie longue période du *glitch* suivant le précurseur ne constitue que la réponse instrumentale du sismomètre à ce choc, altérant pendant quelques secondes la moyenne du signal. Pour finir, il est également important de noter que, bien que le signal sismique présenté ici est celui obtenu par le sismomètre large bande, le capteur courte période est également affecté par ces *glitchs* (Scholz et coll., 2020).

4.3.2.3 Tornades de poussière

L'étude d'un phénomène atmosphérique intéressant est celui de l'occurrence de multiples tornades de poussière sur la surface de Mars. Ces tornades de poussière se forment au milieu de la journée (entre 8h00 et 18h00 LMST), engendrées par l'énergie

solaire réchauffant la surface de Mars, et provoquant un tourbillon convectif ascendant. Un intérêt tout particulier est porté sur l'étude de ces phénomènes, jouant un rôle clef dans le climat martien, de par la grande quantité de poussière soulevée dans l'atmosphère. Dans ce contexte, de nombreuses tornades de poussière ont régulièrement été observées et analysées lors des missions spatiales Vikings (Ryan et Lucich, 1983; Hess et coll., 1977), Phoenix (Ellehoj et coll., 2010), Curiosity (Kahanpää et coll., 2016; Ordonez-Etxeberria et coll., 2018; Steakley et Murphy, 2016) ainsi que, plus récemment, Perseverance (Jackson, 2022). En pratique, un moyen efficace de détecter des tornades de poussière s'effectue par la localisation temporelle de baisses de pression, associées au phénomène d'aspiration induit par les tornades de poussière.

Dans le cadre de la mission InSight, le capteur de pression se révèle particulièrement adapté pour procéder à la détection de ces tornades de poussière (Spiga et coll., 2018). L'étude de ces baisses de pressions soudaines, a été largement exploitée, permettant la détection d'un grand nombre de tornades de poussière (Lorenz et coll., 2021; Banerdt et coll., 2020; Kenda et coll., 2020) aboutissant finalement à la création d'un catalogue de 12 000 événements (Spiga et coll., 2021; Lorenz et coll., 2021), associés à des baisses de pression inférieures à -0,35 Pa. En complément de la détection de ces baisses de pression, de multiples traces sur le sol martien furent également observées, témoignant du passage de ces tornades de poussière à proximité de l'atterrisseur (Perrin et coll., 2020).

Bien que le capteur de pression représente indéniablement l'outil le plus efficace pour détecter ces tornades de poussière, une grande cohérence temporelle entre les données sismiques et celles de pressions lors du passage de tornades de poussière de forte intensité a également été observé (Garcia et coll., 2020; Kenda et coll., 2020). Les tornades de poussière associées aux plus grandes baisses de pression sont régulièrement visibles sur le signal sismique, comme illustré par la figure 4.8. On observe dans cette illustration, trois différentes tornades de poussière, issues du catalogue de Spiga et coll. (2021), ayant été détectées par les chutes de pression soudaines associées (voir les courbes bleues). Les cas étudiés présentent des décroissances de pression de différentes intensités : -3,55 Pa, -0,740 Pa et seulement -0,397 Pa pour les exemples A1, B1 et

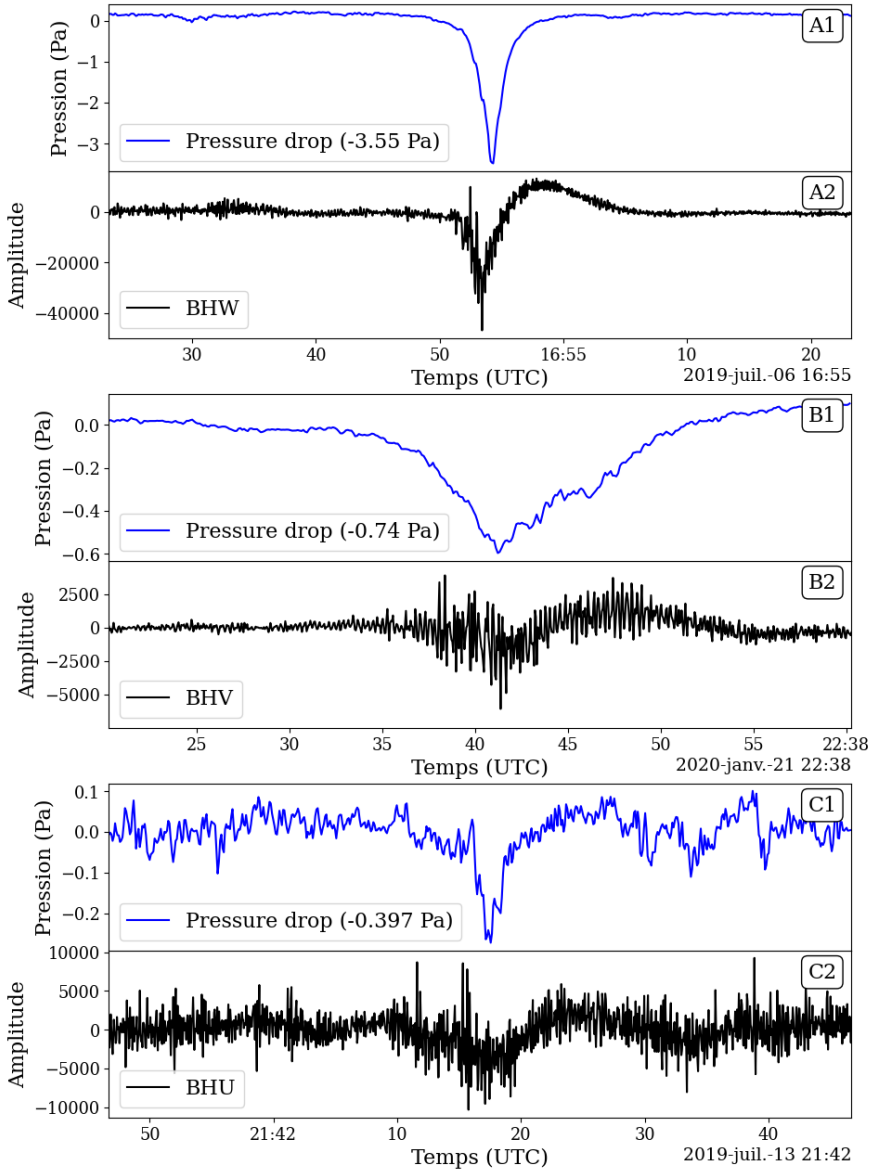

FIGURE 4.8 – Corrélation temporelle entre les baisses de pressions liées aux tornades de poussière et quelques arrivées d'énergies sur le signal sismique. Les tornades de poussière associées aux plus grosses baisses de pressions (A1 et B1 à -3,55 Pa et -0,74 Pa, respectivement) sont également visibles sur les signaux sismiques correspondants (A2 et B2). *À contrario*, rien d'évident n'apparaît sur le signal sismique dans le dernier cas (C2), associé à une chute de pression relativement faible (C1 à -0,397 Pa).

C1, respectivement. Les deux tornades de poussière associées aux baisses de pressions de plus fortes intensités correspondent à des perturbations évidentes dans les signaux sismiques associés (A2 et B2). Cependant, on observe que cette correspondance n'est pas systématique, comme le montre l'exemple C1, dont la décroissance de pression,

d'amplitude relativement faible, n'a pas affecté de manière évidente le signal sismique C2.

4.3.2.4 Séismes martiens

Plus de 1300 séismes martiens ont été enregistrés au cours des 4 ans de la mission InSight (Ceylan et coll., 2022; Clinton et coll., 2021; InSight Marsquake Service, 2023), ayant naturellement abouti à la première caractérisation de la sismicité observée sur Mars (Giardini et coll., 2020; Ceylan et coll., 2023).

La figure 4.9 présente les signaux sismiques de quelques tremblements de terre martiens enregistrés au cours de la mission, lors des sols 173 (A), 1000 (B) et 1222 (C), dont les arrivées des ondes P et S sont représentées par des lignes horizontales rouges et bleues, respectivement. L'exemple A dévoile le signal sismique d'un des premiers séismes martiens ($M_W = 3{,}6$) jugé de bonne qualité, c'est-à-dire avec des ondes P et S clairement identifiables, ayant alors permis une localisation de sa source (voir Giardini et coll. (2020) pour ces résultats ainsi que pour l'estimation de la magnitude). On pourra toutefois noter que l'énergie libérée par ce séisme est tout de même relativement faible et que cet événement est altéré par un *glitch*, quelques secondes seulement après l'arrivée de l'onde P. Les exemples B et C présentent des séismes associés à des arrivées d'ondes bien plus énergétiques. Le tremblement de terre martien associé au signal B ($M_W = 4{,}1$), fut provoqué par l'impact d'une météorite sur la surface de la planète (Posiolova et coll., 2022), ce qui a permis la première détection d'ondes de surface sur Mars (Kim et coll., 2022). Le séisme C présente l'événement sismique de plus grande intensité enregistré lors de la mission InSight, associé à une magnitude $M_W = 4{,}6$ (Kawamura et coll., 2023). L'origine de cet événement majeur demeure à ce jour débattue, pouvant être causée par l'impact d'une météorite (Kawamura et coll., 2023) ou provoquée par une source tectonique (Fernando et coll., 2023).

L'opportunité sans précédent offerte par la mesure continue du signal sismique martien au cours des 1446 sols de la mission InSight a permis l'enregistrement d'une très grande variété de séismes, divisés en plusieurs catégories selon leur contenu fréquentiel

(Clinton et coll., 2021). Une attention toute particulière fut portée sur l'étude de ces tremblements de terre martiens, l'interprétation de leur origine, ainsi que la localisation de leur foyer (Perrin et coll., 2022). L'analyse de ces séismes a notamment permis d'obtenir de nombreux résultats permettant de caractériser la structure interne de la planète Mars, répondant ainsi à l'objectif principal de la mission InSight. Parmi ces résultats, on pourra citer les premières contraintes sur l'épaisseur de la croûte martienne entre 30 et 72 km (Lognonné et coll., 2020; Knapmeyer-Endrun et coll., 2021; Li et coll., 2023; Wieczorek et coll., 2022), des modèles de vitesses de propagation des ondes dans le manteau (Drilleau et coll., 2022; Lognonné et coll., 2023), ou encore l'estimation du rayon du noyau de la planète à 1830 ± 40 km (Stähler et coll., 2021; Samuel et coll., 2021). Finalement, une analyse récente des données sismiques enregistrées lors de la mission InSight, via l'utilisation des techniques de *machine learning* (Dahmen et coll., 2022a), a permis d'augmenter considérablement la base de données des séismes martiens, incluant désormais un total d'environ 2000 événements.

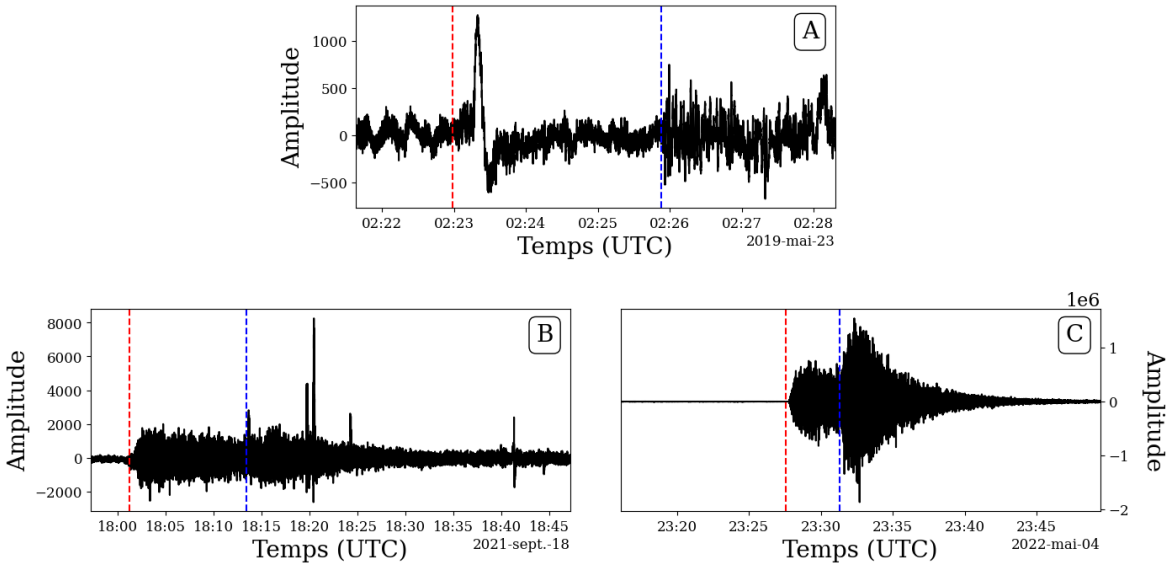

FIGURE 4.9 – Quelques exemples de tremblements de terre martiens ayant été détectés au cours de la mission, lors des sols 173 (A), 1000 (B) et 1222 (C). Pour chaque séisme, les arrivées des ondes P et S sont représentées par des lignes verticales rouges et bleues, respectivement.

4.3.3 Gaussianité du signal sismique martien

Les différents artefacts altérant le signal sismique martien (*glitchs*, résonances de l'atterrisseur,...), l'absence des pics microsismiques, ainsi que les effets atmosphériques (vents, tornades de poussière,...) constituent de nombreuses différences notables en comparaison du signal habituellement enregistré sur Terre. À ces différences viennent également s'ajouter un niveau de bruit, lors de la nuit martienne, jusqu'à 500 fois plus faible (Stutzmann et coll., 2021), que le modèle de référence terrestre (Peterson et coll., 1993). En dépit de ces particularités, le signal sismique enregistré lors de la mission InSight comporte de nombreuses caractéristiques communes avec ses équivalents terrestres. De la même manière que sur Terre, ce signal de fond semble suivre une distribution Gaussienne.

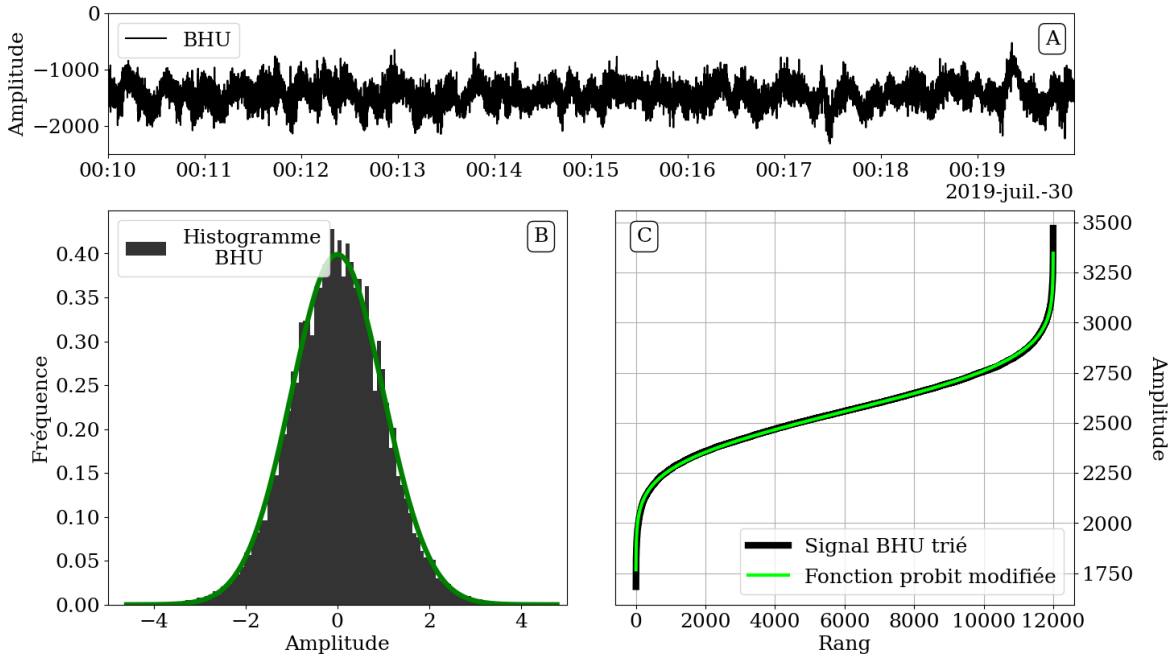

FIGURE 4.10 – Distribution Gaussienne du signal sismique martien. Exemple sur une fenêtre de longueur 10 minutes (12 000 points) extraite du sol 239. La distribution de ces données est proposée via son histogramme normalisé (B), soulignant une bonne correspondance avec la densité de probabilité d'une loi normale centrée réduite (en vert), mais aussi par ses données triées, ayant la même distribution que la fonction probit modifiée (courbe verte).

La figure 4.10 propose l'étude de la distribution du signal sismique martien, sur un intervalle de 10 minutes, enregistré lors du sol 239. Une première visualisation rapide

de la gaussianité de ces données est effectuée (B) par l'observation de son histogramme (normalisé) coïncidant avec la fonction de densité de la loi normale centrée réduite. En regardant désormais, de manière plus précise (car ne nécessitant pas de choix d'un nombre de bins), la distribution des données triées, celle-ci correspond parfaitement à celle de la fonction probit modifiée, (c'est-à-dire la fonction Probit à laquelle on a attribué la moyenne et l'écart type du signal sismique).

En complément de cette observation, de multiples analyses nous révèlent que ce comportement Gaussien présenté dans la figure 4.10 ne constitue non pas un cas isolé, mais la distribution habituelle du signal de fond, lorsque celui-ci n'est pas affecté par une quelconque perturbation (*glitch*, tornade de poussière, etc). En conclusion, de la même manière que lors de l'étude des données terrestres, nous considérons par la suite que le signal sismique enregistré lors de la mission InSight suit une distribution Gaussienne.

4.4 Application de NG-loc au signal sismique de SEIS

4.4.1 Analyse du signal par fenêtres glissantes

De manière similaire à l'étude du signal sismique terrestre étudié dans le chapitre 3, nous procédons à une analyse du signal sismique martien via une approche par fenêtre glissante afin d'appliquer NG-loc aux données enregistrées sur la totalité de la mission. Nous privilégions au cours de cette analyse, et dans la mesure du possible, une application aux signaux échantillonnés à 20 points par seconde, correspondant à la grande majorité des données enregistrées, du sol 181 jusqu'à la fin de la mission (InSight Marsquake Service, 2022). Afin d'obtenir une analyse la plus complète possible à partir de l'installation de SEIS sur la surface de Mars, les données échantillonnées à 10 points par seconde (enregistrées avant le sol 181) seront également analysées. La durée des altérations susceptibles de perturber la gaussianité du signal martien étant, dans la très grande majorité des cas, inférieures à 30 s, nous choisissons donc une fenêtre glissante d'une longueur de 10 minutes. Un tel choix de taille de fenêtre, bien plus grande que celle

de ces altérations (voir les figures 4.7 et 4.8), nous permet de conserver une proportion d'éléments non-Gaussien inférieure à 90% du signal étudié, en accord avec les pré-requis indispensable au bon fonctionnement de NG-loc (voir figure 2.12). Par ailleurs, une application de la méthode NG-loc sur un grand nombre de points ($n = 12\,000$ pour une fenêtre de 10 minutes comportant des données échantillonnées à 20 points par secondes) entraîne également une plus grande stabilité des résultats, comme démontré lors de l'étude des cas synthétiques dans la figure 2.11. Le pas sélectionné entre les fenêtres glissantes (c'est-à-dire la distance séparant chacune d'entre elles) est fixé à 2 minutes, permettant une analyse fine de chacun des sols étudiés. Chaque échantillon sera analysé 5 fois, ce qui nous permettra par la suite de confirmer ou d'infirmer l'appartenance de chaque point au signal de fond Gaussien.

Afin d'analyser la gaussianité du signal sismique martien sur différentes gammes de fréquence, nous choisissons d'effectuer notre analyse sur deux domaines distincts : les basses fréquences (BF), comprises entre 0,05 et 0,9 Hz et les hautes fréquences (HF), entre 1,1 et 4,5 Hz. Le choix de ces domaines permet d'exclure le *tick noise* (Ceylan et coll., 2021), qui altère le signal à une fréquence de 1 Hz exactement. La limitation à 4,5 Hz des HF permet d'utiliser les données échantillonnées à 10 points par seconde, limitée par une leur fréquence de Nyquist à 5 Hz. De la même manière que pour l'analyse du signal terrestre présenté dans la section 3.2, les opérations de filtres/suppressions de la réponse instrumentale seront effectuées sur des fenêtres locales légèrement plus grande (14 minutes) afin d'éviter les effets de bords engendrés par ces derniers. De manière détaillée, le traitement du signal analysé pour chaque sol/composante s'effectue de la façon suivante :

1. Sélection du signal sismique correspondant au sol que l'on souhaite analyser, sur les trois composantes U, V et W.
2. Extraction du signal sismique obtenu par lecture de la première fenêtre glissante (c'est-à-dire grande fenêtre locale) de 14 min .
3. Suppression de la réponse instrumentale de la grande fenêtre locale.
4. Rotation des signaux afin d'obtenir les données sismiques sur les composantes Z, N et E.

5. Application du filtre passe-bande en fonction de la gamme de fréquence étudiée : $[0{,}05 - 0{,}9]$ Hz pour BF et $[1{,}1 - 4{,}5]$ Hz pour HF.
6. On coupe le signal obtenu 2 minutes à gauche et à droite afin d'obtenir la fenêtre que l'on souhaite analyser (de longueur 10 minutes).
7. Application de NG-loc sur le signal obtenu.
8. On revient à l'étape 2 en incrémentant le début de la fenêtre de 2 min.

Bien que l'algorithme ci-dessus propose une méthode servant à analyser le signal sismique enregistré par SEIS sur un sol donné, son application en pratique est toutefois un peu plus complexe. En effet, il est important de noter que l'analyse d'une journée martienne se doit de commencer 10 minutes avant le début du sol (correspondant au minuit LMST), afin de garantir que chacun des points composant cette journée soit analysé 5 fois exactement. Pour les mêmes raisons, il est nécessaire que l'analyse à la fin d'une journée martienne finisse exactement 10 minutes plus tard. Par ailleurs cette étude du signal sismique martien se trouve également complexifiée par la présence de trous dans les données, ce qui s'explique notamment par des problèmes de communications avec l'atterrisseur InSight. D'autre part, les nombreuses restrictions énergétiques, survenant à la fin de la mission, ont finalement abouti à la mise en veille régulière et répétée de l'instrument SEIS, engendrant alors une indisponibilité conséquente des données sismiques après le sol 1200. Afin de pallier le problème technique engendré par la présence de ces données manquantes, l'analyse pourra se porter, dans le cas d'un sol incomplet, seulement sur une portion réduite du signal. L'étude de chacun des sols par NG-loc aboutit à un total de 743 fenêtres glissantes analysées (par composante/-gamme de fréquence), où chaque valeur de l'intervalle des quantiles $[Q_A, Q_B]$ est alors sauvegardée.

4.4.2 Extraction des perturbations majeures

Nous détaillons désormais, une approche visant à extraire les perturbations BF significatives détectées par NG-loc dans le signal martien. Cette base de données sera exploitée par la suite dans la section 4.6 afin de procéder à une discrimination des

perturbations associées aux tornades de poussière.

La totalité des points étudiés étant analysée exactement 5 fois par NG-loc, on peut donc attribuer, pour chacun d'entre eux, une classe, permettant de compter le nombre de fois où chaque point a été catalogué comme étant non Gaussien par notre approche, car exclu de la fenêtre $[Q_A, Q_B]$. On nommera alors ces classes $(C_i)_{0\leq i\leq 5}$, où un point du signal appartient à C_i si celui-ci a été considéré exactement i fois comme un élément perturbé par NG-loc. Dans le tableau 4.1, nous présentons pour chaque composante, le nombre de points constituant les classes $(C_i)_{1\leq i\leq 5}$, correspondant donc à ceux ayant été vus 1, 2, 3, 4 ou 5 fois comme non-Gaussiens par NG-loc.

TABLEAU 4.1 – Nombre de points constituant les classes $(C_i)_{1\leq i\leq 5}$, pour chaque composante suite à l'analyse complète de la mission.

	C_1	C_2	C_3	C_4	C_5
BHZ	9 773 055	7 053 881	5 778 525	5 276 868	57 997 904
BHN	10 830 111	6 866 291	5 286 343	4 625 720	30 893 749
BHE	9 808 676	6 246 800	4 860 567	4 355 780	43 256 469

L'analyse de ce tableau nous permet d'observer un nombre d'éléments perturbés bien plus important sur la composante Z que sur les deux autres, déjà constaté dans la sous section 4.5.2. D'autre part, le nombre d'éléments dans chaque classe est totalement déséquilibré, avec une grande majorité de points appartenant à la classe C_5 (pour toutes les composantes), c'est-à-dire ayant été caractérisés comme perturbés lors de chacune des 5 analyses effectuées par NG-loc. La quantité non négligeable d'éléments dans les classes C_1, C_2, C_3 et C_4 indiquent que certains points ont été classés comme perturbés, seulement 1,2,3 ou 4 fois lors des 5 analyses de NG-loc. Cette décision incertaine peut être causée, par exemple, par certains points ayant une amplitude légèrement au dessus du signal de fond Gaussien, introduisant alors un doute sur leur appartenance ou non à la distribution normale (pouvant être levé en analysant une autre fenêtre temporelle).

Afin de discriminer, de manière binaire, les éléments altérés de ceux Gaussiens, nous considérons par la suite qu'un point est perturbé si celui-ci a été classé comme non Gaussien lors d'au moins 3 analyses sur 5 par NG-loc (c'est-à-dire ceux apparte-

nant aux classes C_3, C_4 et C_5). Par conséquent, on obtient alors un total de nombre de points perturbés pour les composantes Z, N et E de 69 053 297, 40 380 096 et 52 472 816, respectivement. Bien que ces chiffres soient relativement élevées, ceux-ci sont évidemment à remettre en perspective avec le nombre total de points ayant été analysés au cours de la mission InSight, s'élevant alors approximativement à plus de 5,6 milliards de points (pour les seules données BHU, BHV et BHW).

Exploitons désormais ces ensembles de points non Gaussiens afin de construire une base de données fiable de perturbations. Chaque perturbation importante dans le signal sismique martien (tornade de poussière, *glitchs*, etc...) étant nécessairement formée d'un nombre important de points consécutifs non Gaussiens, nous procédons alors à un regroupement des éléments perturbés. Celui-ci se traduit sous la forme d'un rassemblement des points altérés, si ceux-ci définissent un intervalle temporel continue de points non Gaussiens. Par conséquent, ceci nous permet d'obtenir un nombre total de perturbations continues au cours de la mission sur les composantes Z, N et E de l'ordre de 2,5 millions, 828 000 et 839 000, respectivement.

Par ailleurs, nous procédons également à une fusion des perturbations, lorsque celles-ci sont suffisamment proches, dans le but de rassembler les perturbations ayant la même origine. Afin de déterminer la distance de fusion (DF) la plus adaptée à notre base de données, nous étudions alors dans la figure 4.11 quatre perturbations différentes, avec et sans cette opération de fusion (pour $DF = 0$, 2 et 5 secondes). Les perturbations étudiées représentent deux *glitchs* (A et B) et deux tornades de poussière (C et D) détectés par le catalogue de Spiga et coll. (2021) sur la composante BHE (une étude similaire pouvant être effectuée de la même manière sur les autres composantes). La forme d'onde de ces perturbations est ici différente de celles étudiées dans les figures précédentes car correspondant ici au signal filtré et déconvolué, tel qu'analysé par NG-loc (l'amplitude est alors affichée en nm s^{-1}). On remarque que ces perturbations déconvoluées se caractérisent sous la forme d'un pic entouré d'oscillations longues périodes autour du niveau d'équilibre d'amplitude (c'est-à-dire zéro). Ces oscillations autour de zéro entraînent cependant un retour du signal à sa position d'équilibre, c'est-à-dire autour du signal de fond Gaussien. Par conséquent, les alté-

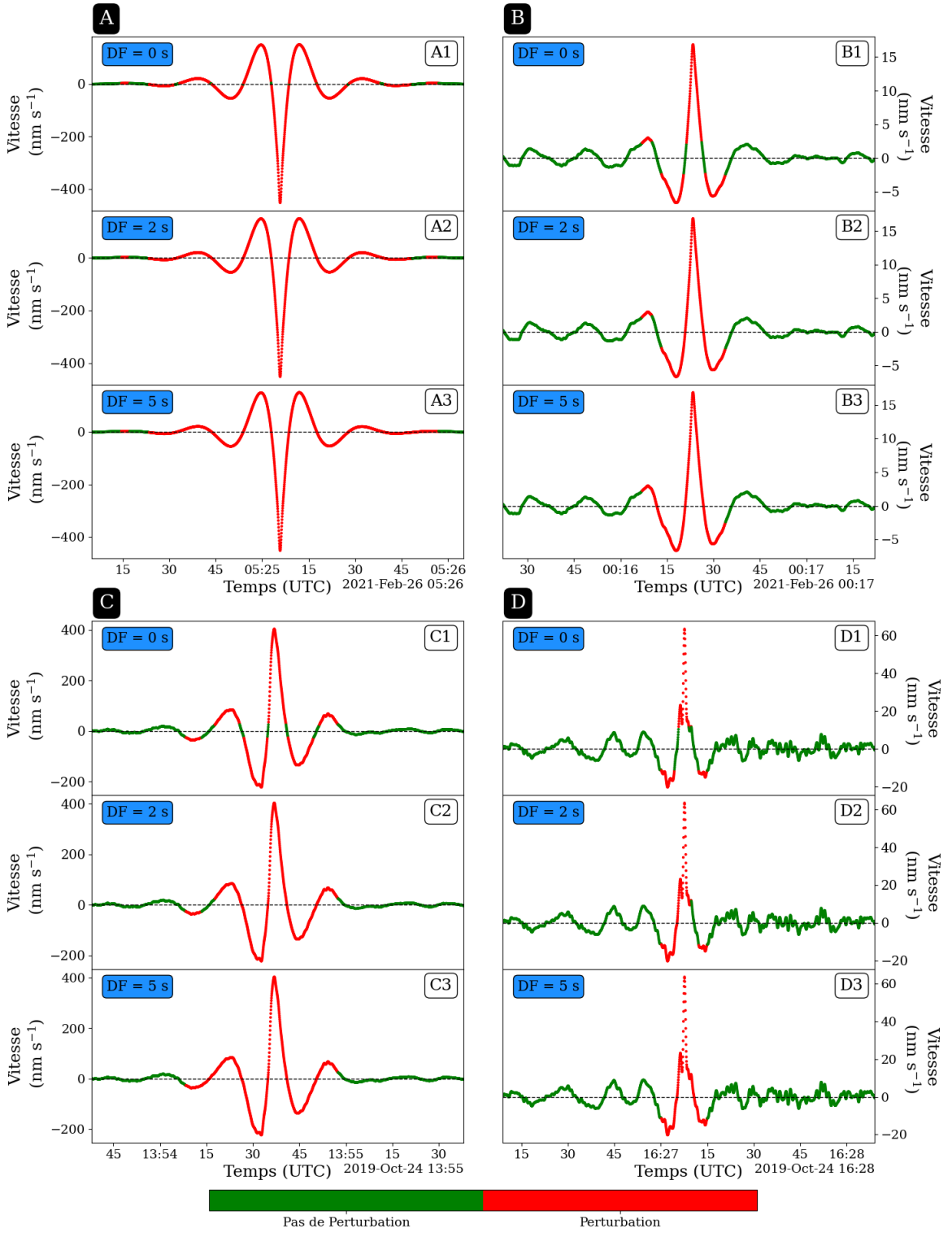

FIGURE 4.11 – Application de la fusion des perturbations continues sur notre base de données pour des distances de fusions différentes (DF = 0, 2 et 5). Illustration sur deux *glitchs* (A et B) et deux tornades de poussière (C et D). Chaque point est affiché en rouge si celui-ci fait partie d'une perturbation continue et en vert sinon.

rations A1, B1, C1 et D1 sont alors composées de plusieurs perturbations continues, comme indiqué par le code couleur de chaque point (en rouge si celui-ci fait partie d'une perturbation continue et en vert sinon).

Si l'on fusionne désormais les perturbations séparées par une distance de deux secondes ou moins (A2, B2, C2 et D2), un bref retour à l'équilibre n'est donc plus suffisant pour que notre approche considère que le signal n'est plus perturbé. Bien que ceci justifie alors l'intérêt de la fusion sur notre base de donnée, celle-ci ne semble pas suffisante pour $DF = 2$ s, comme constaté par les observations de (A2, B2, C2 et D2), toujours scindés en plusieurs perturbations. Le choix de $DF = 5$ s paraît quant à lui plus intéressant, car permettant d'obtenir une unique perturbation continue dans les cas B3, C3 et D3, même si celui-ci n'inclut pas un léger écart de l'amplitude dans l'exemple A3 (vers 05h24 et 15 secondes). Finalement, ceci justifie alors le choix d'une distance de fusion d'au moins 5 secondes.

La figure 4.12 propose une analyse du nombre de perturbations continues, répertoriées au cours de la mission sur les composantes BHZ (en haut), BHN (au milieu) et BHE (en bas), en fonction de la distance de fusion DF. On constate la même tendance sur les 3 composantes, avec une décroissance rapide du nombre de perturbations pour $0 \leq DF \leq 10$ (illustrant le constat effectué lors de l'analyse de la figure 4.11), puis une relative linéarité des courbes ensuite ($DF \geq 10$). Au vu de ce résultat, nous choisissons alors finalement d'appliquer une distance de fusion $DF = 10$ s à nos perturbations (indiqué par la ligne verticale orange sur la figure 4.12).

Nous choisissons désormais de nous intéresser aux perturbations de longueurs supérieures à quelques secondes. Ceci nous permet, de sélectionner uniquement les altérations ayant alors provoquées une déviation de la gaussianité du signal sur une période conséquente, en excluant celles n'ayant affectées que quelques points. La figure 4.13 présente la distribution des longueurs des perturbations continues sur la composante Z. La répartition des distributions est extrêmement inégale, avec une énorme majorité de petite perturbations de longueurs relativement faible ≤ 4 s, suivi ensuite d'une relative stabilité après ce seuil (ligne orange verticale dans la figure 4.13). Par conséquent, nous choisissons ici d'exclure ces petites perturbations, d'une longueur inférieure à 4 se-

condes, ne correspondant pas à des altérations significatives du signal. Par ailleurs, cet argument est également étayé par l'inspection visuel des perturbations associées aux *glitchs*/tornades de poussière étudiées dans la précédente figure 4.11, toutes de longueur bien supérieur à 4 s. Bien que la figure 4.13 présente uniquement les résultats sur la composante Z, une observation similaire est également effectuée sur les deux autres composantes aboutissant à la même conclusion.

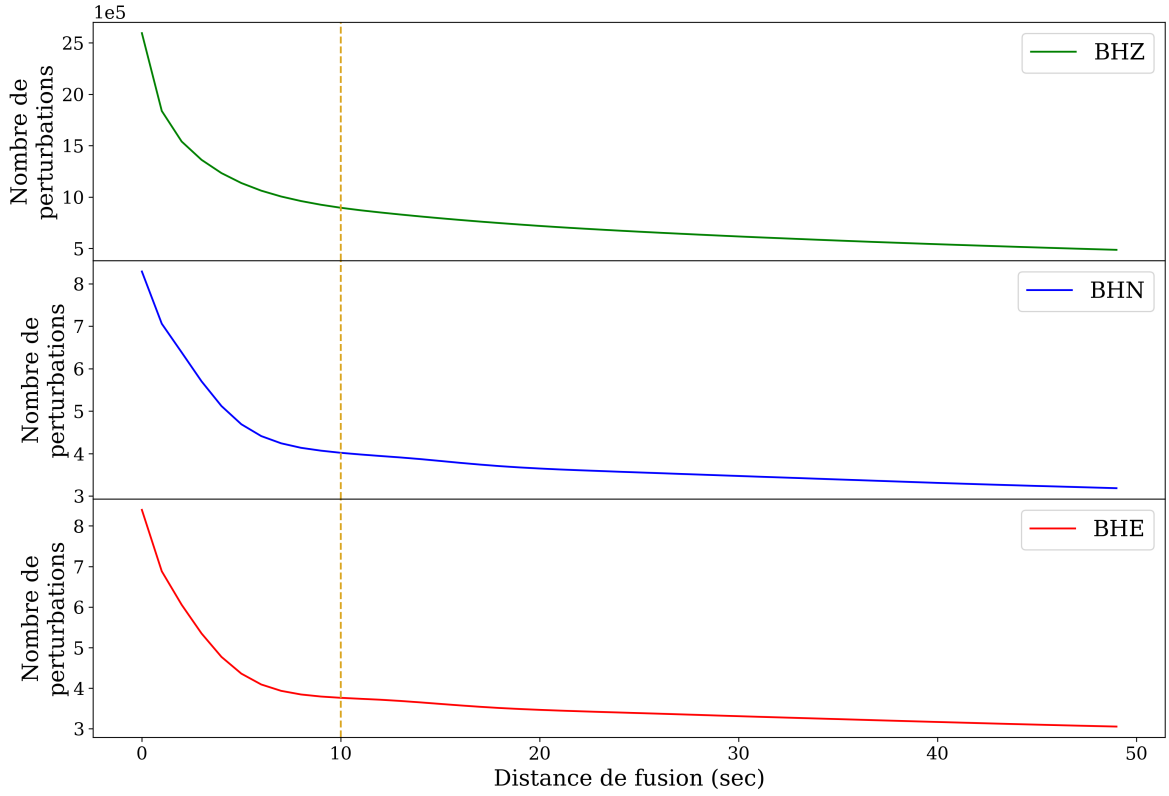

FIGURE 4.12 – Nombre de perturbations continues comptabilisées au cours de la mission InSight en fonction de la distance de fusion sur les trois composantes, Z (en haut), N (au milieu), et E (en bas).

Suite aux traitement des données présentées dans cette sous-section, nous obtenons finalement un nombre de total de perturbations détectées au cours de la mission de 306 823 pour BHZ, 153 403 pour BHN et 145 361 pour BHE. Nous proposons dans la figure 4.14 la localisation temporelle de chacune de ces perturbations au cours de la mission. Bien que l'analyse fine des principales tendances des perturbations ne constitue pas ici l'objet de notre étude (et sera effectuée lors de la section 4.5), celles-ci illustrent cependant bien le grand nombre de perturbations détectées au cours de la

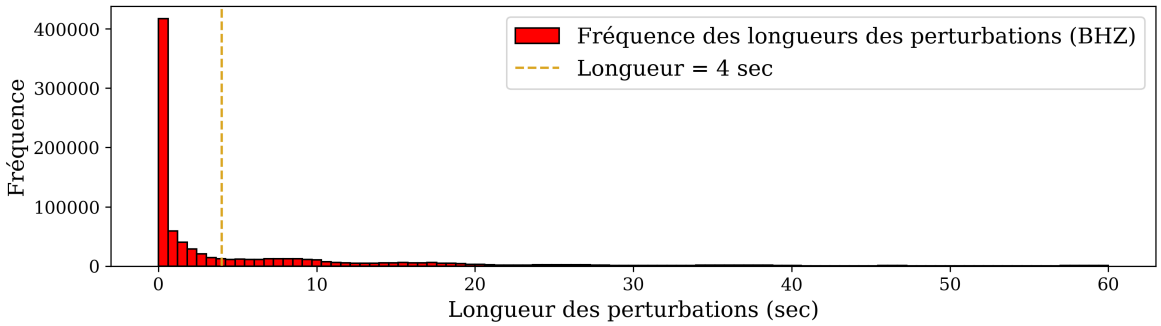

FIGURE 4.13 – Distribution de la longueur des répartitions des perturbations continues détectées sur la composante BHZ.

mission, aussi bien le jour que la nuit martienne. Même si la figure 4.14 semble indiquer un énorme nombre d'altérations (notamment sur BHZ), celles-ci ne représentent cependant en moyenne que 223, 111 et 105 perturbations par sol pour BHZ, BHN et BHE, respectivement.

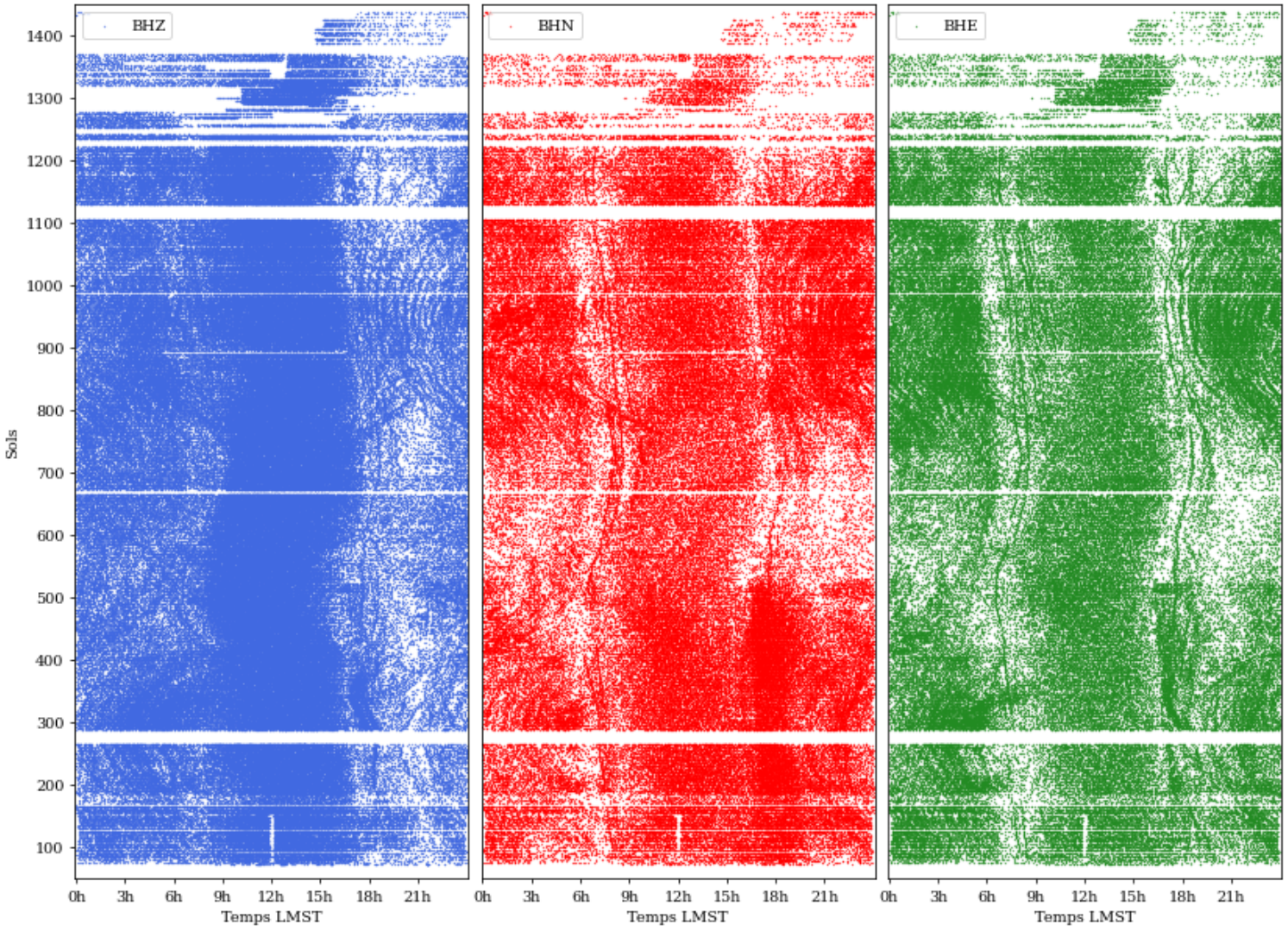

FIGURE 4.14 – Perturbations NG-loc observées sur BHZ (à gauche), BHN (au milieu) et BHE (à droite), au cours de la mission InSight.

4.5 Analyse de la qualité du signal sismique enregistré au cours de la mission InSight

Nous proposons dans cette section, une analyse détaillée des perturbations détectées par NG-loc au cours de la mission. Plus précisément, l'étude se concentre sur le pourcentage de points non Gaussien ayant été détectés par application de NG-loc sur le signal sismique continu enregistré lors de la mission InSight, comportant un total de 1446 sols. L'interprétation de ces résultats nous permet d'obtenir des informations

intéressantes quant à la qualité du signal, mais aussi d'effectuer une étude détaillée des périodes temporelles correspondant à un signal altéré.

Nous détaillons dans la sous-section 4.5.1, l'analyse minutieuse des points non Gaussiens détectés par NG-loc, sur la partie basse fréquence de la composante verticale du signal sismique. La sous-section 4.5.2 fera l'objet d'une comparaison de ces résultats, avec ceux obtenus via l'analyse des perturbations BF détectées sur les deux composantes horizontales. Suite à cette étude, l'analyse des éléments non Gaussien sur le signal HF sera alors effectué dans la sous-section 4.5.3. L'estimation de la qualité du signal sismique nous permettra ensuite de mettre en lumière dans la sous-section 4.5.4, une corrélation évidente entre l'apparition de certains *glitchs* à des seuils spécifiques de température enregistrée sur l'instrument SEIS. Finalement, nous effectuerons dans la sous-section 4.5.5 une analyse détaillée de la qualité du signal autour de trois événement précis ayant eu lieu au cours de la mission, ayant pu, ou non, affecter celui-ci. Ces derniers étant 1/ la mise en route du chauffage de SEIS au sol 168, 2/ un refroidissement de l'instrument causé par une perte de contact prolongée entre les sols 268 et 287 et 3/ une manœuvre entre les sols 816 et 877, visant à ensevelir le câble reliant SEIS à l'atterrisseur, dont la torsion est susceptible d'être responsable de nombreux *glitchs*.

4.5.1 Signal basse fréquence - Composante verticale

Nous effectuons tout d'abord une description détaillée des résultats de NG-loc, sur le signal basse fréquence, de la composante Z, présenté dans la figure 4.15. Celle-ci illustre par un code couleur le pourcentage de points perturbés de chacune des fenêtres glissantes analysées au cours de la mission. La taille des fenêtres glissantes choisie au cours de notre analyse étant relativement grande (10 minutes) en comparaison des différentes perturbations, nous choisissons alors de restreindre l'intensité du code couleurs entre 0,1 %, en bleu foncé, et 5 %, en jaune. Par ailleurs, les points verts sur cette figure correspondent aux fenêtres glissantes constituées de moins de 0,1 % d'éléments perturbés. L'axe des abscisses propose la position temporelle (en LMST) du centre de chaque fenêtre glissante, tandis que l'axe des ordonnées indique le sol sur

lequel celle-ci se situe. À noter que l'acquisition des données sismiques ne commence pas au sol 0, correspondant à l'atterrissage de InSight, mais au sol 73, suite au long processus de déploiement de l'instrument SEIS sur le sol martien (Trebi-Ollennu et coll., 2018).

De nombreuses zones blanches sont présentes dans la figure 4.15. Elles correspondent aux périodes temporelles où le signal sismique est indisponible. Parmi les plus notables, on pourra par exemple citer un problème de transmission des données entre les sols 268 et 287 suite à la période de conjonction (c'est-à-dire alignement) entre Mars, le Soleil et la Terre. Par ailleurs, une forte période d'absence de données sismiques, entre les sols 1107 et 1129 est également observée, causée par les restrictions énergétiques lors de la fin de la mission imposant une mise en veille de SEIS. Ces fortes restrictions énergétiques, provoquées par l'accumulation progressive de poussière sur les panneaux solaires (voir figure 4.4) sont également responsables d'importantes indisponibilités du signal sismique, causés par les mises en veilles régulières et prolongés de SEIS. Cette indisponibilité des données est clairement visible sur la figure 4.15, après le sol 1200, et jusqu'au sol 1446, correspondant à la fin de la mission (perte de contact avec l'atterrisseur engendrée par une énergie insuffisante). Quelques périodes d'indisponibilité des données sismiques, impactant parfois un sol entier, ou simplement quelques heures, sont également observées au cours de la mission, pouvant alors être provoquées, entre autre, par des problèmes de communication avec l'atterrisseur InSight.

La première caractéristique notable visible sur la figure 4.15 est la différence marquée de la distribution des perturbations entre la nuit et la journée martienne. En effet, la journée martienne, comprise entre 7h00 et 18h00 LMST, se caractérise par une répartition bien plus uniforme des perturbations. De la même manière que sur Terre, on observe une légère variation de l'heure du lever/coucher de Soleil marquant le début/la fin de la journée au cours d'une année martienne (composée d'environ 668 sols). *À contrario*, les perturbations détectées lors de la nuit martienne obéissent à une distribution bien plus singulière. En effet, on observe de nombreuses « courbes jaunes continues » dans la figure 4.15 qui feront l'objet d'une étude détaillée dans la suite de cette sous-section. Toutefois, bien que de nombreuses perturbations soient observées

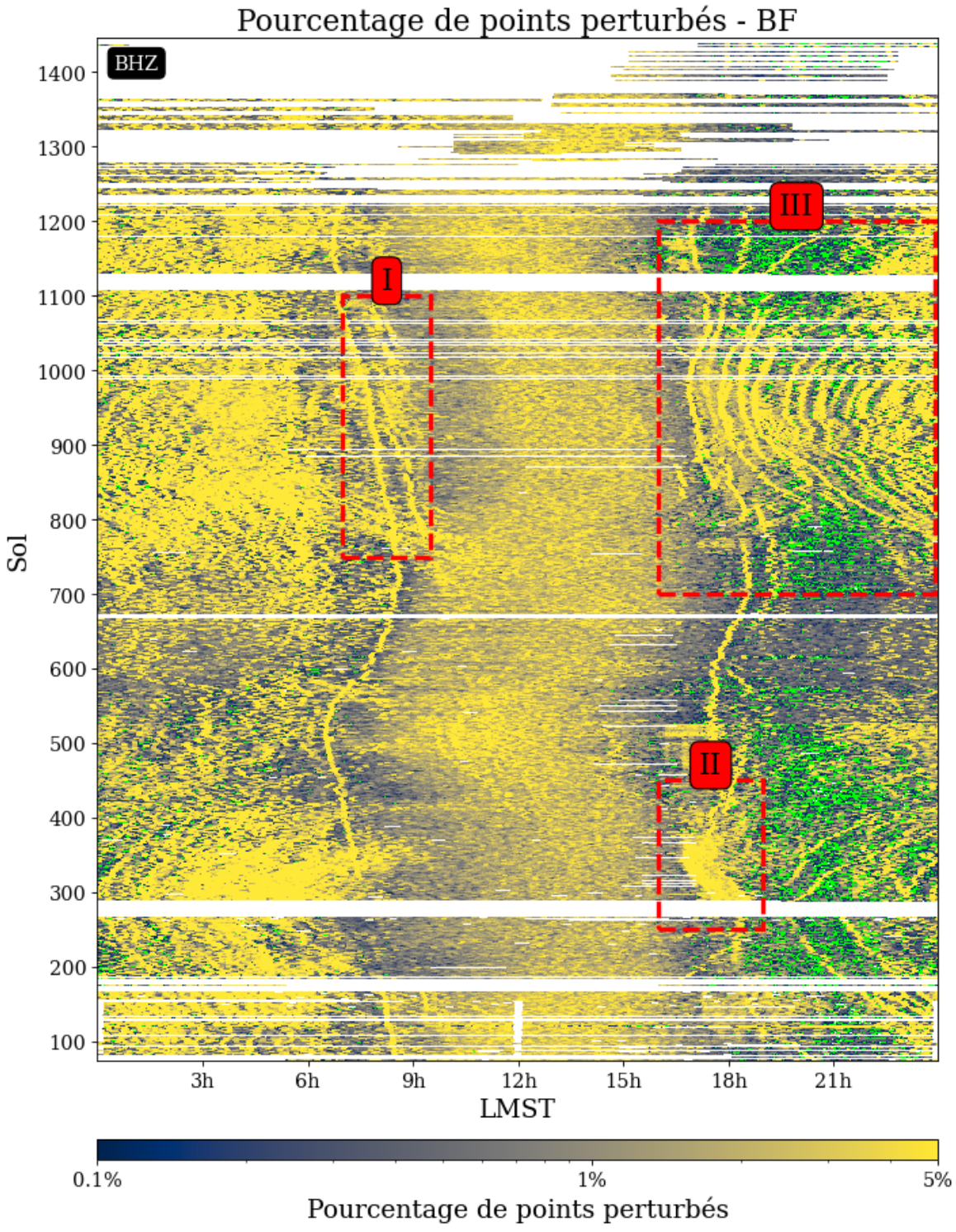

FIGURE 4.15 – Pourcentage de points perturbés de chaque fenêtre glissante suite à l'analyse de NG-loc (composante Z, basses fréquences). La présence d'un point vert indique une fenêtre glissante composée de moins de 0,1 % de points altérés.

en dehors de la journée, c'est pourtant au cours de la soirée, entre 18h00 et minuit LMST, que le signal sismique est également le moins altéré (présence de points verts dans la figure). Le signal sur cette période n'étant par définition pas perturbé par les vents survenant au cœur de la journée, celui-ci se trouve alors particulièrement régulier et Gaussien lorsqu'il n'est pas altéré par les glitchs ou une autre perturbation.

D'autre part, bien que cette figure 4.15 dévoile une analyse sur une longueur totale d'environ deux années martiennes, on observe cependant un motif légèrement différent entre sa partie basse et haute (avant et après le sol 741). Ceci peut s'expliquer notamment par l'activation du chauffage de SEIS, sur une bonne partie de la première année martienne analysée, entre les sols 168 et 584, ayant un fort impact sur la qualité du signal. En effet, les *glitchs* altérant fortement le signal sismique basse fréquence sont sensibles aux variations de température et sont alors nettement moins nombreux lorsque le chauffage est activé Scholz et coll. (2020). Cette caractéristique semble également visible sur la figure 4.15 au cours de la soirée, après 18h00 LMST. Sur cette période, assez peu de perturbations sont visibles lors du chauffage de SEIS (sol 168 à 584), en comparaison de la période correspondante une année martienne plus tard, entre les sols 836 et 1252. Par conséquent, ceci nous indique que ces perturbations, sensibles à la température, sont alors probablement des *glitchs* (une visualisation des traces sismiques permettra, par la suite, de confirmer cette hypothèse).

Une autre caractéristique étonnante présente sur la figure 4.15 est la présence d'une atténuation du nombre de perturbations détectées, particulièrement visible après 18h00 LMST, entre les sols 525 et 775. Cette baisse soudaine des perturbations est causée par une diminution du nombre de *glitchs* lors de la tempête de sable, ayant provoqué des vents violents, même pendant la nuit martienne. Il convient désormais de comprendre les origines d'une telle diminution lors de la tempête.

La figure 4.16 présente une comparaison des signaux sismiques BF enregistrés pendant la tempête de sable (A) et en dehors de celle-ci (B), lors des sols 650 et 950 respectivement. Un nombre bien plus important de *glitchs* est observé au cours du sol 950 par rapport au signal enregistré lors du sol 650 (grand traits verticaux noirs). Par ailleurs, on remarque un niveau de bruit bien plus élevé lors de la période de tempête,

particulièrement au cours de la nuit. Bien que ce bruit de forte intensité dissimule sûrement de nombreuses perturbations, ceci n'explique pas les différences notables entre ces deux signaux, où celui du sol 950 (B) semble comporter un nombre bien plus important de *glitchs* de grande amplitudes. Une hypothèse pouvant expliquer cette différence provient de l'origine elle-même des *glitchs*. En supposant que les *glitchs* soient causés par une relaxation thermique de SEIS, cette dernière est alors le fruit d'une accumulation de contraintes mécaniques au cours du temps. L'influence du vent lors de la tempête pourrait alors provoquer le déclenchement, de manière précoce, de cette relaxation mécanique. Les contraintes relâchées étant alors moins intenses, ceci pourrait alors expliquer le plus faible nombre de *glitchs* observés lors de la tempête, mais aussi leurs plus faibles amplitudes. On pourra noter que cette hypothèse d'un déclenchement précoce des *glitchs* fut déjà avancée par Scholz et coll. (2020), soulignant que ces derniers peuvent être provoqués par des événements mécaniques extérieurs telles que les tornades de poussière, le mouvement du bras mécanique d'InSight ou encore l'arrivée d'ondes sismiques (voir la figure 4.9 A).

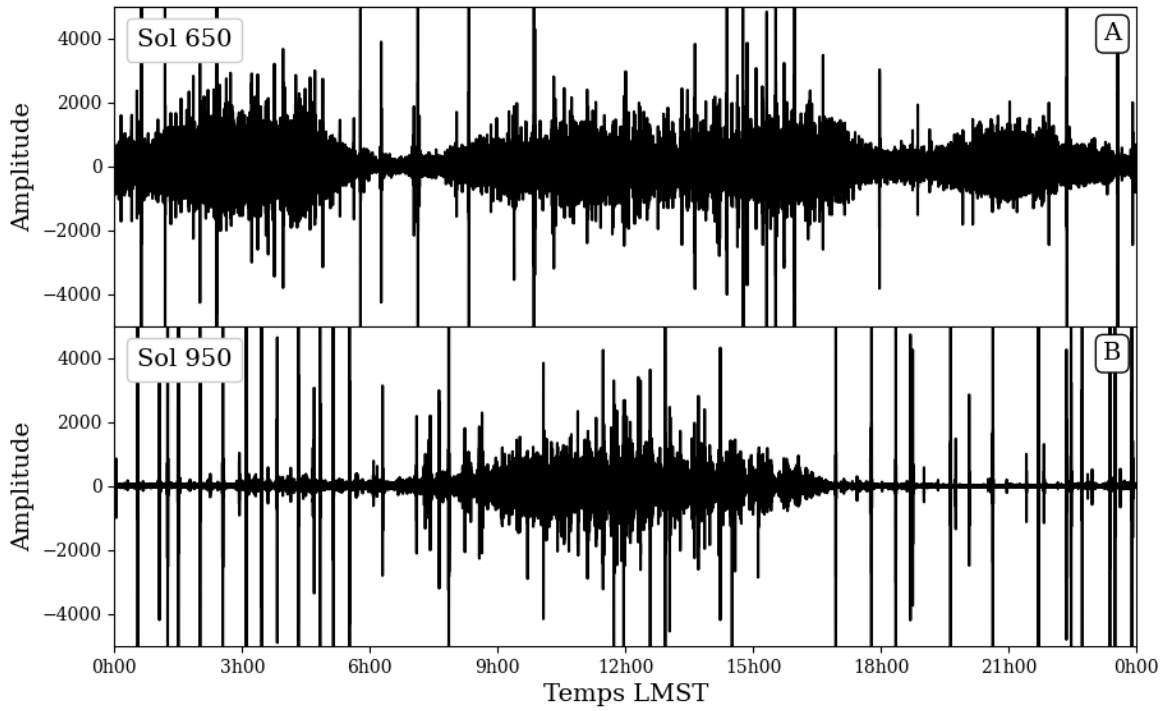

FIGURE 4.16 – Comparaison entre le signal sismique BF enregistré en dehors et au cours de la tempête de sable sur la composante BHZ. (A) : Signal enregistré au sol 650, lors de la tempête de sable. (B) : Signal enregistré au sol 950, après la tempête.

Suite à cette analyse globale de la figure 4.15, nous choisissons désormais de nous intéresser en détail à trois zones d'intérêts, représentées par des rectangles rouges, labellisés I, II et III. Nous proposons dans la figure 4.17, l'étude de la « zone I », entre les sols 750 et 1100, et située dans l'intervalle compris entre 7h00 et 9h30 LMST, correspondant à la période du lever de soleil. On observe sur cette période de nombreuses lignes jaunes continues, associées à la présence d'un fort pourcentage de points perturbés. Cette continuité des perturbations entre chaque sol est représentée par des traits rouges discontinus dans la zone I (labellisés A1, A2, A3, A4 et A5). Ce critère de continuité nous permet de définir dans ce cas, plusieurs groupes de perturbations, dont les éléments ont toutes une continuité temporelle évidente entre chaque sol. On remarque toutefois que cette continuité temporelle est soumise à une légère variation qui est semblable pour les 5 groupes. Chaque perturbation qui les composent se produit légèrement plus tôt à chaque sol. Cette translation régulière coïncide avec celle du lever de soleil, ce qui peut expliquer l'origine de ces perturbations.

Intéressons-nous désormais aux périodes temporelles A, B et C, représentées par les 3 traits horizontaux cyans sur la figure 4.17, lors des sols 875, 825 et 775, respectivement (entre 7h00 et 9h30 LMST). On constate que ces périodes temporelles représentées par les traits cyans intersectent les lignes rouges discontinues A1 et A2, indiquant la présence d'altérations, confirmée par l'observation des données sismiques basses fréquences A, B et C présentées dans la figure 4.17 (correspondant aux périodes temporelles de même noms), comportant des perturbations de grandes amplitudes, associées aux groupes d'altérations A1 et A2. Une caractéristique étonnante de ces perturbations est la très forte similarité de l'amplitude des altérations appartenant à un même groupe, valant par exemple 30 900, 33 200 et 27 500 pour les trois *glitchs* associés à A1 lors des sols 875, 825 et 775, respectivement. Par ailleurs, une observation détaillée du signal brut autour de ces perturbations (non présenté ici) nous permet de conclure que ces dernières sont en réalité des *glitchs*, et que chaque groupe se caractérise par une forme d'onde spécifique dans le signal sismique.

En revenant désormais à une observation plus globale de la mission sur la figure 4.15, on constate alors que ce phénomène se prolonge en dehors de la zone I, et ce

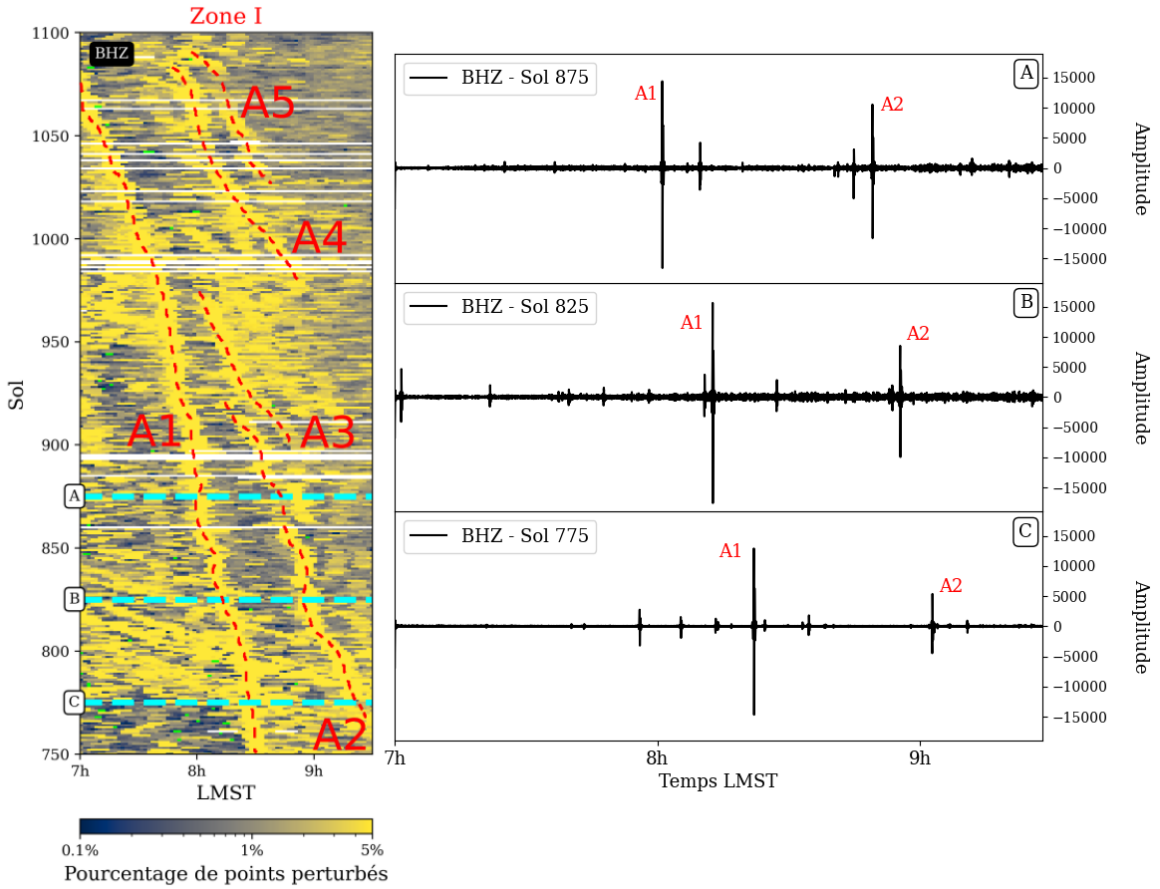

FIGURE 4.17 – Étude de la zone I (à gauche), extraite de la figure 4.15. Chaque ligne rouge pointillée dans la carte de couleur (A1, A2, A3, A4 et A5) correspond à un *glitch* majeur se répétant de manière régulière à chaque sol, comme illustré par A, B, C présentant le signal sismique (BHZ, basse fréquence) et l'occurrence des *glitchs* appartenant aux groupes A1 et A2.

pendant la quasi-totalité de la mission. En effet, une ligne jaune continue traversant la zone I est observée (correspondant au lever du soleil), indiquant la présence d'un *glitch* de grande amplitude. Par ailleurs, une observation équivalente est effectuée lors du coucher de soleil, coïncidant avec une ligne jaune continue, particulièrement visible au dessus de la zone II dans la figure 4.15. La présence régulière de ces *glitchs* peut être expliquée par une relaxation thermique de l'instrument SEIS, causée par une augmentation/diminution soudaine de la température lors du lever/coucher de soleil.

Nous étudions dans la figure 4.18 les sols 250 à 450 entre 16h00 et 19h00 LMST. Notre choix se porte sur cette zone d'intérêt car elle contient une importante quantité

de points perturbés, juste après la période de conjonction, marquée par une absence de données sismiques entre les sols 268 et 287. Cette période de conjonction fut associée à la mise en veille temporaire du chauffage de SEIS, causant un refroidissement durable de SEIS, perdurant même après la fin de celle-ci. Une conséquence directe de ce refroidissement de l'instrument fut la détection d'une forte augmentation du nombre de *glitchs* dans le signal sismique pour une durée d'environ 100 sols après la fin de la conjonction (Scholz et coll., 2020). Nos observations semblent également en accord avec ce résultat, montrant un important nombre de perturbations (points jaunes) dans la zone II, s'estompant après le sol 375.

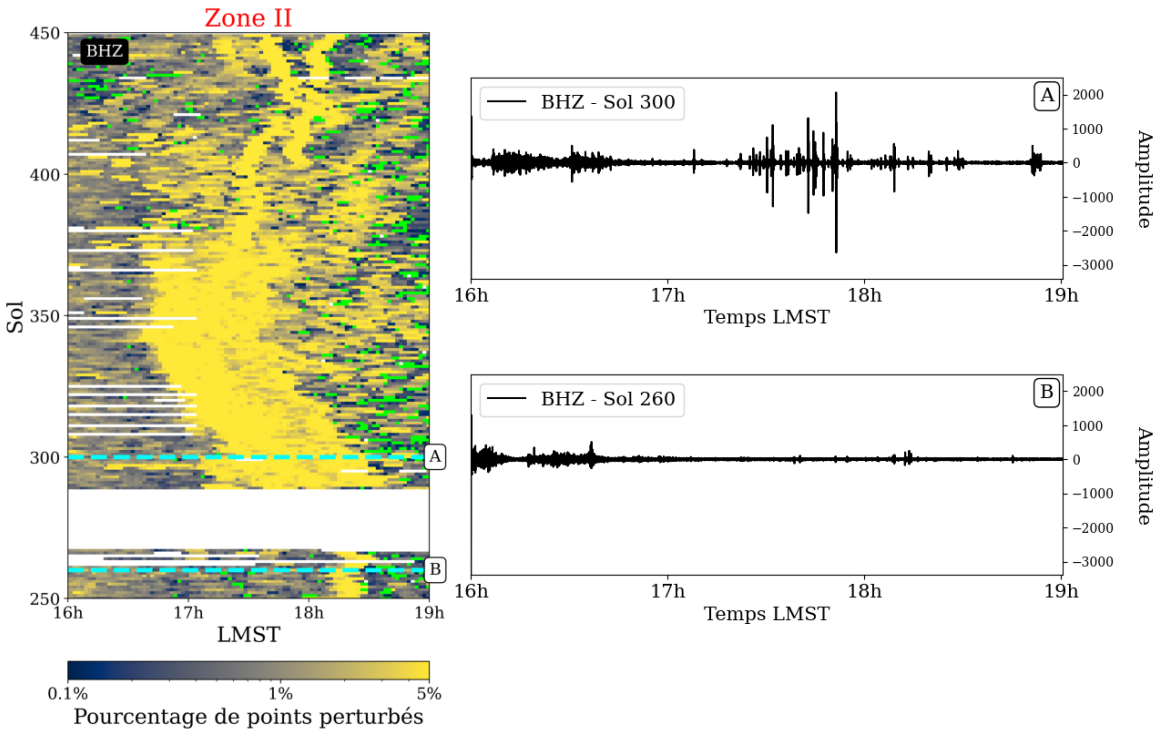

FIGURE 4.18 – Étude de la zone II (à gauche), extraite de la figure 4.15. Chaque trait cyan dans la zone II correspond à un signal sismique affiché (A et B) sur les mêmes périodes temporelles, lors des sols 300 et 260, respectivement.

L'étude du signal sismique confirme également une nette différence en terme de nombre de *glitchs* observés avant et après la conjonction sur cette période. En effet, A (correspondant à la ligne cyan A dans la zone II) montre un signal fortement altéré par de très nombreux *glitchs*, particulièrement après 17h00 LMST. Cette observation se caractérise par une forte augmentation de la moyenne du pourcentage de points

perturbés, entre 17h30 et 18h30 entre les sols 260 (B) et 300 (A), passant de 1,90% à 7,83%, respectivement. On notera également la présence d'un signal haute fréquence plus intense avant 17h00 LMST, mais correspondant quant à lui à l'influence du vent lors de la fin de la journée (voir figure 4.5). L'étude des données sismiques B, enregistrées lors du sol 260 (ligne cyan dans la zone II), nous permet d'observer un signal très différent de celui étudié dans l'exemple A. En effet, celui-ci ne contient pas les *glitchs* de grandes amplitudes, et semble de bien meilleure qualité.

La figure 4.19 se concentre sur l'étude de la zone III, entre les sols 700 et 1200, lors de la soirée, entre 16h00 et minuit LMST. La répartition des altérations dans cette zone est ici très particulière et semble former des groupes de perturbations, respectant un motif très régulier, quasi circulaire, comme indiqué par les lignes rouges discontinues. Cette répartition atypique des perturbations est d'ailleurs d'autant plus intrigante que celle-ci survient lors d'une très grande partie de la mission InSight (la zone III regroupant 500 sols). Ces importantes perturbations sont, comme souvent lors de l'analyse du signal BF, causées par la présence de *glitchs* de grandes amplitudes. La compréhension de l'origine de ces *glitchs* dégradant fortement la qualité du signal sur une longue durée est primordiale et sera détaillée par la suite dans la sous-section 4.5.4.

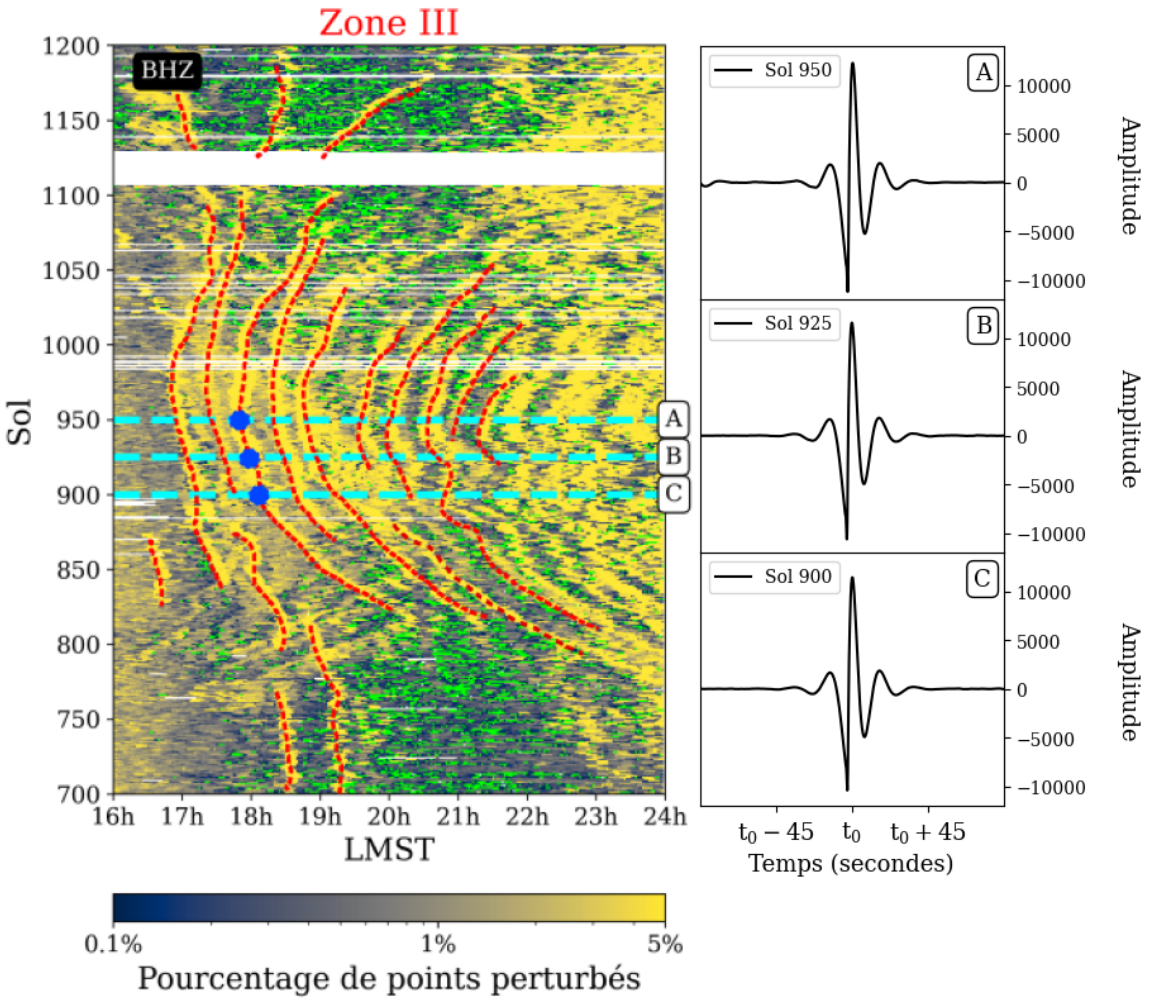

FIGURE 4.19 – Étude de la zone III (à gauche), extraite de la figure 4.15. De nombreux groupes de perturbations sont affichés dans la zone III, indiqués par de multiples lignes rouges discontinues. Les trois points bleus représentent la position temporelle de perturbations appartenant à un même groupe, lors des sols 950, 925 et 900, représentés par les lignes horizontales cyans A, B et C, respectivement. Les trois formes d'ondes affichées sur la droite de la figure (A, B et C) correspondent aux signaux sismiques autour des localisations temporelles t_0, indiquées par les points bleus (Zone III).

Étudions désormais trois exemples de signaux sismiques (A, B et C) appartenant au même groupe de *glitchs*, enregistrés lors des sols 950, 925 et 900, respectivement. La position temporelle de chaque signal t_0 a été obtenue par l'intersection des lignes cyans dans la zone III (représentant les trois sols cités) et l'un des groupes de *glitchs*, représenté par une ligne rouge discontinue. Bien que l'heure et le sol soient bien différents dans chacun de ces trois cas, nous choisissons d'afficher, par soucis de comparaison, les

trois signaux sismiques A, B et C autour de ces temps t_0 ($\pm$90 s). Nous constatons une similarité flagrante de ces signaux, ayant tous les trois une forme d'onde quasiment identique, se traduisant par des résultats de cross-corrélations toujours supérieurs à 96% en comparant ces derniers entre eux. Il est important de noter que cette similarité des formes d'ondes n'est pas un cas particulier propre à ce groupe de *glitch*, mais une caractéristique également observée sur les différents groupes, tous associés à une forme d'onde spécifique. La forte continuité temporelle de ces perturbations au cours de la mission, ainsi que les similarités évidentes des formes d'ondes associées nous permettent alors d'émettre l'hypothèse que chaque groupe de *glitch* a une origine commune. Bien que notre étude se concentre ici sur la zone III, cette continuité temporelle n'est pas uniquement spécifique à cette période. Des tendances similaires peuvent également être observées, mais dans une moindre mesure, lors d'autres périodes, comme par exemple entre minuit et 6h00 entre les sols 200 et 400.

4.5.2 Signal basse fréquence - Composantes horizontales

Suite à l'analyse des perturbations basses fréquences détectées sur la composante Z dans la précédente sous-section, nous présentons désormais dans la figure 4.20 les résultats obtenus sur les composantes Nord et Est, à gauche et à droite, respectivement. Bien que l'objectif visé ici ne consiste pas en une analyse aussi minutieuse que celle effectuée pour BHZ, nous discutons ici de quelques similarités et différences entre les points perturbés observés sur les composantes horizontales de la figure 4.20 et celles de la composante verticale étudiée précédemment (figure 4.15).

De nombreuses tendances globales en termes de répartition des perturbations, déjà observées sur la composante Z (figure 4.15) se retrouvent également sur les deux composantes horizontales de la figure 4.20. On constate par exemple la même répartition, presque uniforme, des perturbations au cœur de la journée, ainsi qu'une nette décroissance des altérations détectées au cours de la tempête de sable, entre les sols 525 et 775, particulièrement après 18h00 LMST. On distingue également des lignes jaunes continues lors du lever/coucher du soleil, comme sur la composante Z. En ce qui concerne les tendances locales, on retrouve les mêmes motifs atypiques, déjà étudiés dans la

zone III de la précédente section (entre les sols 700 et 1200, après 16h00). Comme l'on pouvait s'y attendre, de très grandes similarités sont observées entre les perturbations présentes sur les trois composantes. Ce comportement n'est pas surprenant car les *glitchs*, représentant très probablement la grande majorité de ces perturbations basses fréquences, affectent très régulièrement le signal sismique sur plusieurs composantes (voir la figure 4.7, mais également Scholz et coll. (2020) pour une statistique détaillée des *glitchs* affectant plusieurs composantes).

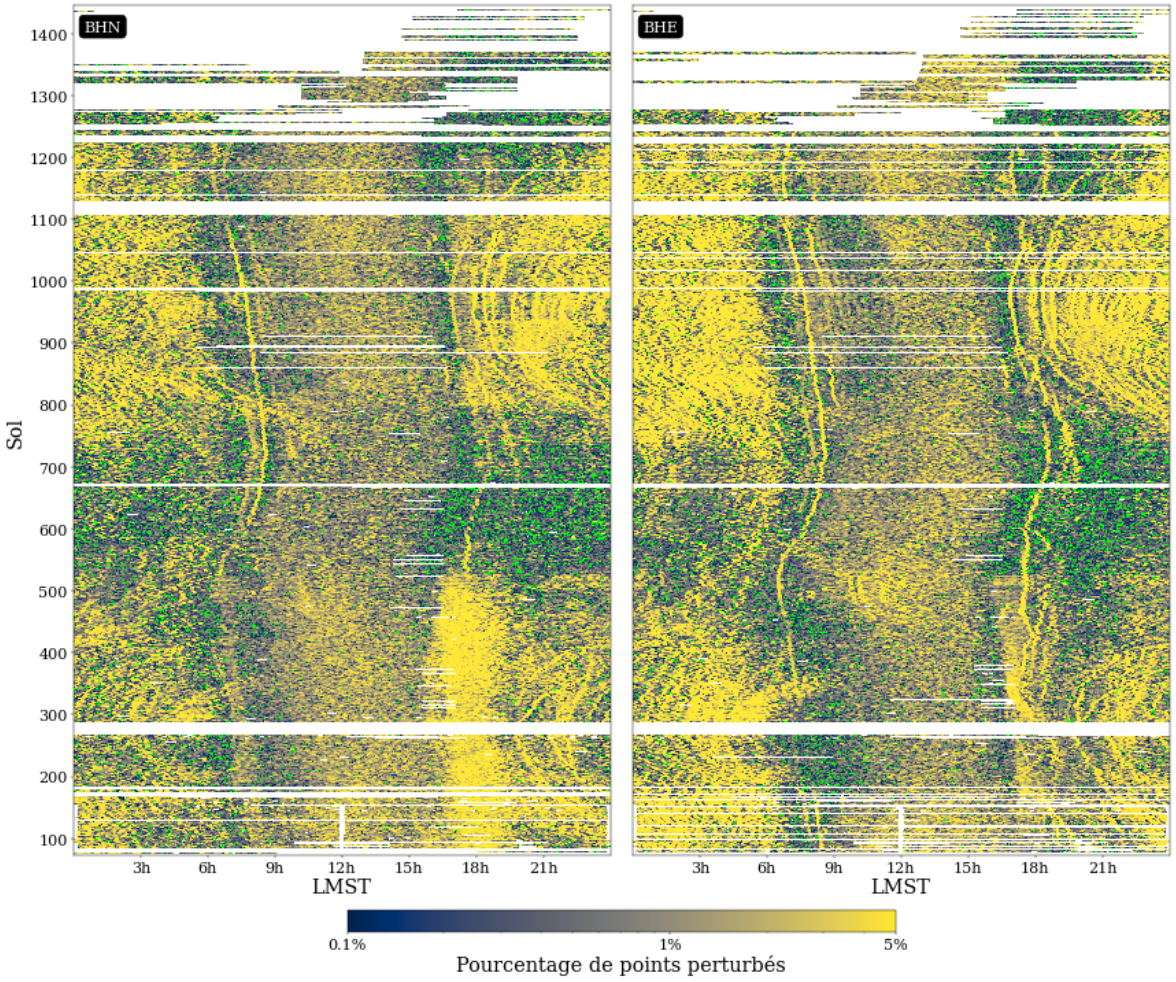

FIGURE 4.20 – Même légende que pour la figure 4.15, mais pour les composantes Nord et Est, à gauche et à droite, respectivement

Par ailleurs, quelques différences notables sont néanmoins observées entre les perturbations provenant des composantes Est et Nord, en comparaison avec la composante verticale. On remarque par exemple un nombre global de fortes perturbations (c'est-à-dire points jaunes) qui semble moins important dans le cas de composantes

horizontales. Ceci est en effet confirmé par l'inspection des médianes des pourcentages de points perturbés au cours de la mission, valant 1,10% et 1,16% pour les composantes Nord et Est, respectivement, alors que celle-ci atteint 1,91% pour BHZ. Bien que cette forte inégalité ne fut jamais mesurée au cours de la mission entière, celle-ci est toutefois cohérente avec les résultats de Scholz et coll. (2020), pointant du doigt une forte disparité entre le nombre de *glitchs* détectés sur les trois composantes au cours des 400 premiers sols. Par ailleurs, certaines distributions de perturbations semblent également propres à chacune des composantes, comme par exemple le fort pourcentage de points perturbés observé sur BHN, entre les sols 73 et 500, autour de 18h00 LMST (figure 4.20). Cette forte augmentation des perturbations est causée par de multiples *glitchs* survenant en fin de journée, causée probablement par une relaxation thermique de SEIS, dû à son refroidissement soudain. Les différences de polarisations (c'est-à-dire la provenance) de ces *glitchs* peut alors expliquer pourquoi ceux-ci affectent principalement la composante BHN (voir Scholz et coll. 2020 pour une analyse de la polarisation des *glitchs*).

4.5.3 Signal haute fréquence

Nous proposons désormais de nous intéresser aux perturbations détectées lors de l'analyse de la gamme des hautes fréquences (HF), par NG-loc. La figure 4.21 présente le pourcentage de points perturbés de chaque fenêtre glissante analysée, pour les trois composantes Z, N et E du signal HF, à gauche, au milieu et à droite, respectivement. On remarque sur toutes les composantes, un pourcentage de points perturbés bien plus faible que ceux ayant été observés lors de l'analyse des basses fréquences dans les précédentes sous-sections 4.5.1et 4.5.2. Ceci est particulièrement visible lors de la nuit martienne, entre 18h00 et 6h00 LMST, période durant laquelle on observe de nombreux points verts, indiquant des pourcentages d'éléments perturbés inférieurs à 0,1%. Cette observation est confirmée par le calcul des médianes de pourcentage de points perturbés, valant ici 0,41%, 0,40% et 0,30% pour les composantes Z, N et E, respectivement, alors que celui-ci fut toujours supérieur à 1,10% dans le cas de l'étude des basses fréquences. L'écart mesuré entre les médianes associées aux données HF

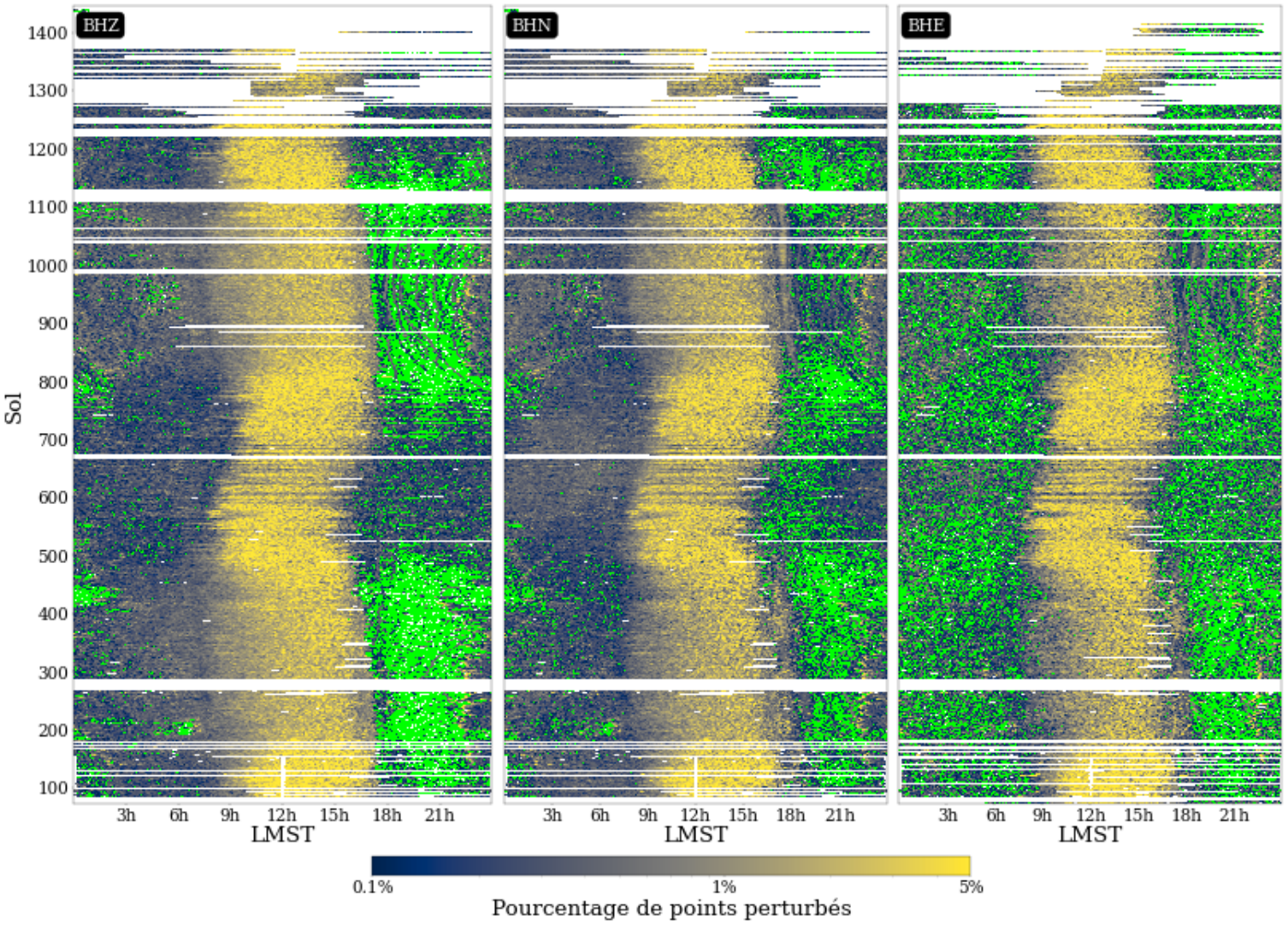

FIGURE 4.21 – Même légende que la figure 4.15, pour les trois composantes du signal hautes fréquences : BHZ, BHN et BHE, à gauche, au milieu et à droite, respectivement.

nous permet également de conclure que le signal enregistré sur BHE est moins sujet aux perturbations que ceux associés aux composante BHN et BHZ. Ceci nous permet de déduire que le signal haute fréquence enregistré lors de la mission InSight est bien moins altéré que celui associé aux basses fréquences. Cette forte différence s'explique principalement par la présence des nombreux *glitchs*, altérant principalement les basses fréquences du signal enregistré durant l'intégralité de la mission.

La grande majorité des perturbations hautes fréquences observées dans la figure 4.21 surviennent au milieu de la journée martienne. Cette période se caractérise par l'apparition de nombreuses tornades de poussière, mais aussi par une augmentation

significative de la puissance du vent (Banfield et coll., 2020), venant tous deux influencer le signal sismique. Bien que ces figures semblent contenir moins d'informations que l'analyse des basses fréquences, on distingue tout de même des motifs similaires à ceux étudiés précédemment dans la zone III (figure 4.19), particulièrement sur BHZ, lors de la soirée, entre les sols 700 et 1200. Ces motifs ne sont toutefois plus marqués par des points jaunes, mais par des lignes noires, se démarquant du fond vert, indiquant un signal très peu altéré. Ces altérations proviennent des précurseurs hautes fréquences des *glitchs*, provoquant des perturbations soudaines sur quelques points du signal (voir la figure C.1 en annexe). Bien que ces précurseurs n'altèrent pas le signal sur une longue durée (inférieure à 1 s), celle-ci est parfois suffisante pour perturber plus de 0,1% des données analysées sur une fenêtre glissante. En effet, chaque fenêtre glissante étant d'une longueur de 600 s, un tel pourcentage représente alors une altération de seulement 0,6 s du signal sismique.

4.5.4 *Glitchs* et seuils de température

Nous cherchons désormais à expliquer la répartition temporelle particulière des *glitchs* observés dans la zone III (figure 4.19). L'un des paramètres évoluant au cours du temps est celui de la température, mesurée de manière continue, à l'intérieur du bouclier thermique de SEIS. Des premières contributions importantes concernant la compréhension de l'origine des *glitchs* furent apportées par Scholz et coll. (2020), pointant du doigt une atténuation du nombre de *glitchs* lorsque la température de SEIS augmente. En effet, il fut constaté que la mise en route du chauffage au sol 168 permit de réduire leurs présences dans le signal sismique (voir la figure C.2 en annexe). Une fois ce lien mis en lumière, il fut également proposé dans cette étude que la cause la plus probable des *glitchs* est celle d'une relaxation thermique de l'instrument SEIS. Bien que ce dernier soit partiellement protégé par son bouclier thermique, il subit néanmoins de fortes variations de température au cours d'une même journée, ce qui pourrait engendrer alors de nombreux *glitchs*.

La figure 4.22 présente la température de l'instrument SEIS (canal VKI), au cours de l'intégralité de la mission, en commençant au sol 73, correspondant au déploiement du

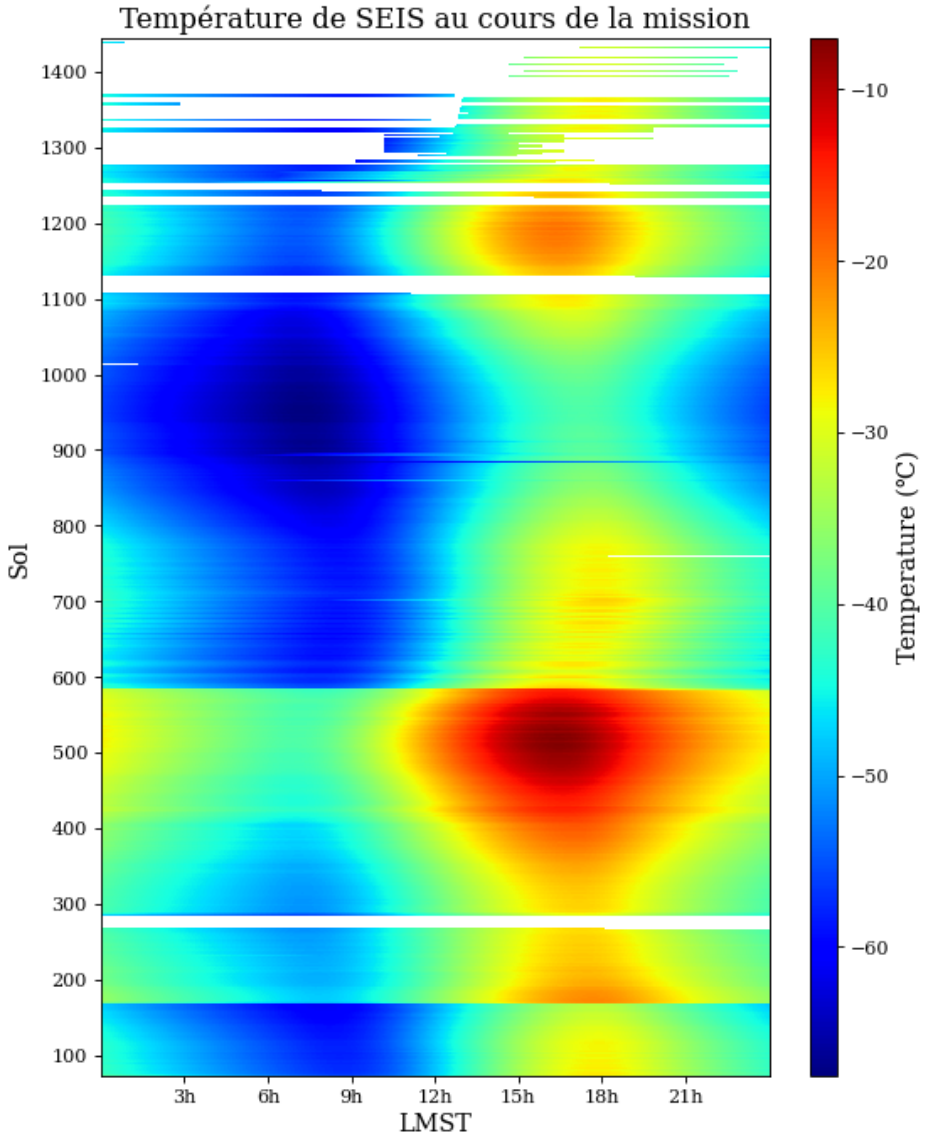

FIGURE 4.22 – Température de l'instrument SEIS (°C) au cours de la mission InSight.

sismomètre sur le sol martien. L'observation de l'échelle de couleur nous permet d'apprécier la grande variabilité d'amplitude des températures de SEIS lors de la mission (-68 °C à -8 °C). Par ailleurs, la période de fonctionnement du chauffage est également nettement visible sur cette figure, correspondant alors en une forte augmentation de la température entre les sol 168 et 584.

Cette figure semble présenter quelques tendances similaires, à celle de certaines perturbations basses fréquences détectées lors de notre analyse de la figure 4.15. Afin de vérifier ceci, nous comparons dans la figure 4.23 les variations de température de

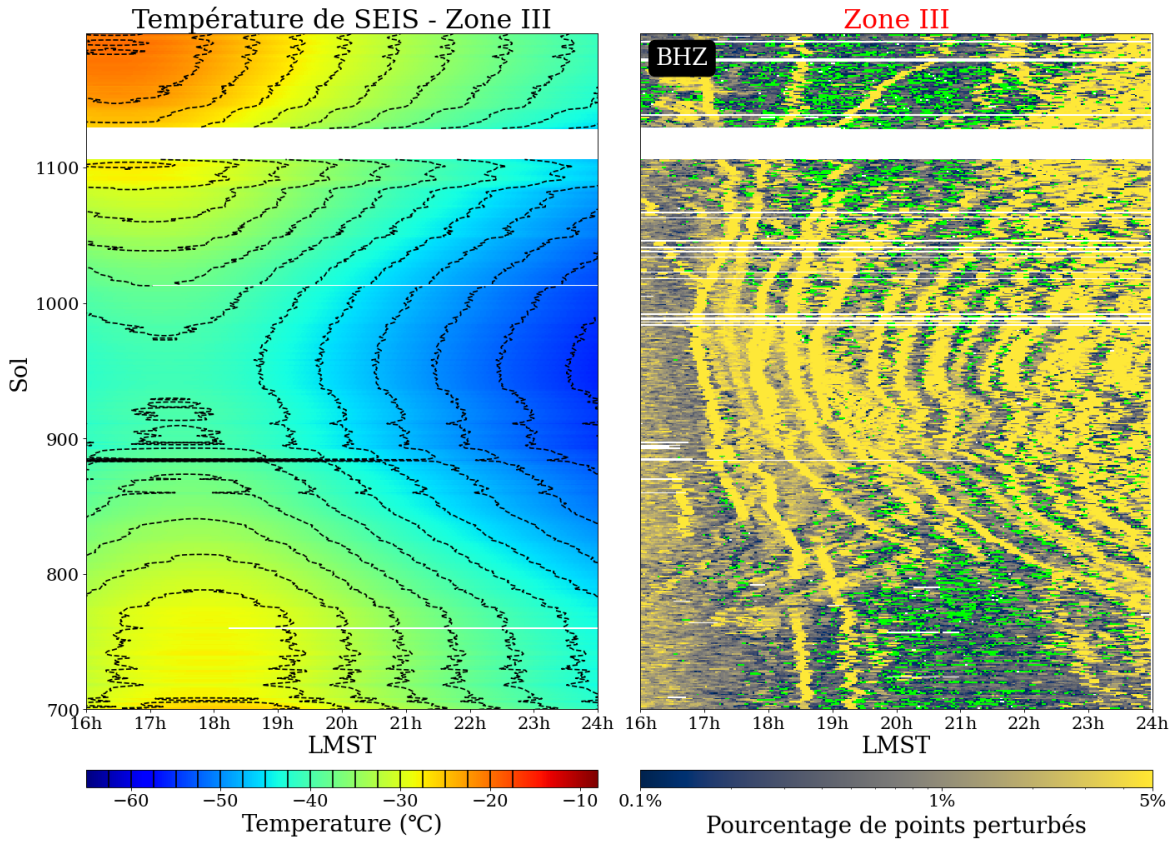

FIGURE 4.23 – Similarités entre la température de SEIS (à gauche) et les groupes de *glitchs* détectés lors de l'analyse des basses fréquences (à droite) sur la zone III.

SEIS et les perturbations basses fréquences détectées dans la zone III (entre les sols 700 et 1200, en fin de journée). On remarque généralement une bonne correspondance entre les variations thermiques de SEIS et la distribution des fenêtres glissantes associées à un fort pourcentage de points perturbés (points jaunes). Par conséquent, cette observation semble indiquer que certains groupes de *glitchs* détectés lors de cette période sont directement associés à des seuils spécifiques de températures atteints par SEIS. Une fois l'un de ces seuils atteint, ceci engendre alors une relaxation thermique spécifique, se répétant de manière régulière lors de chaque sol.

4.5.5 Qualité du signal sismique martien au cours du temps

Nous proposons désormais de nous intéresser à l'étude de la qualité du signal sismique basse fréquence au cours de la mission InSight. Une telle étude nous permet alors

de mieux comprendre l'origine des *glitchs*, constituant la grande majorité des perturbations basses fréquences. Afin de caractériser la qualité du signal enregistré sur chaque sol, nous étudierons ici un nouvel estimateur, nommé Med_{PPP}, représentant la médiane journalière du pourcentage de points perturbés. La figure 4.24 présente cet estimateur, pour chaque sol, sur les composantes BHZ, BHN, et BHE (en haut, au milieu, et en bas, respectivement). Afin que Med_{PPP} ne soit pas biaisée par une quantité insuffisante de signal enregistré au cours d'un même sol, celle-ci est uniquement calculée sur les journées où au moins 90% des données sismiques sont disponibles. La période étudiée se limite donc au sol 1250, à partir duquel la quantité de signaux sismiques disponibles au cours d'une même journée dépasse très rarement ce seuil de 90% (voir les zones blanches en haut de la figure 4.15).

La zone rouge pâle dans la figure 4.24 représente la période de fonctionnement du chauffage de SEIS, entre les sols 168 et 584. La mise en route du chauffage correspond à une très nette décroissance de notre estimateur Med_{PPP} sur les trois composantes, indiquant un lien évident entre une augmentation de la température de SEIS et une forte diminution du nombre de *glitchs*. Cette décroissance remarquable se mesure aisément, en calculant par exemple la moyenne de notre estimateur Med_{PPP}, 50 sols avant et après cette mise en route du chauffage, nous permettant d'obtenir des valeurs passant de 3,44% à 1,43% pour BHZ, 2,30% à 1,65% pour BHN et 2,31% à 1,24% pour BHE.

La date de mise en veille du chauffage lors du sol 584 (par soucis d'économie d'énergie), ne correspond quant à elle pas à une nette variation de notre estimateur Med_{PPP}. Ceci peut s'expliquer par le fait que cette mise en veille survient dans un contexte où le signal sismique est affecté par une forte tempête de poussière (approximativement entre les sols 525 et 775), qui a augmenté l'amplitude du niveau de bruit (plus de vent), et diminuant le nombre de *glitchs* observés (voir la figure 4.16). Par conséquent, ceci nous empêche de tirer une quelconque conclusion quant à une dégradation éventuelle de la qualité du signal lors de la mise en veille du chauffage.

La période de conjonction, entre les sols 268 et 287 est représentée par une zone grise hachurée dans la figure 4.24. Cette période d'absence de données sismiques fut également marquée par la mise en veille du chauffage, entraînant un refroidissement

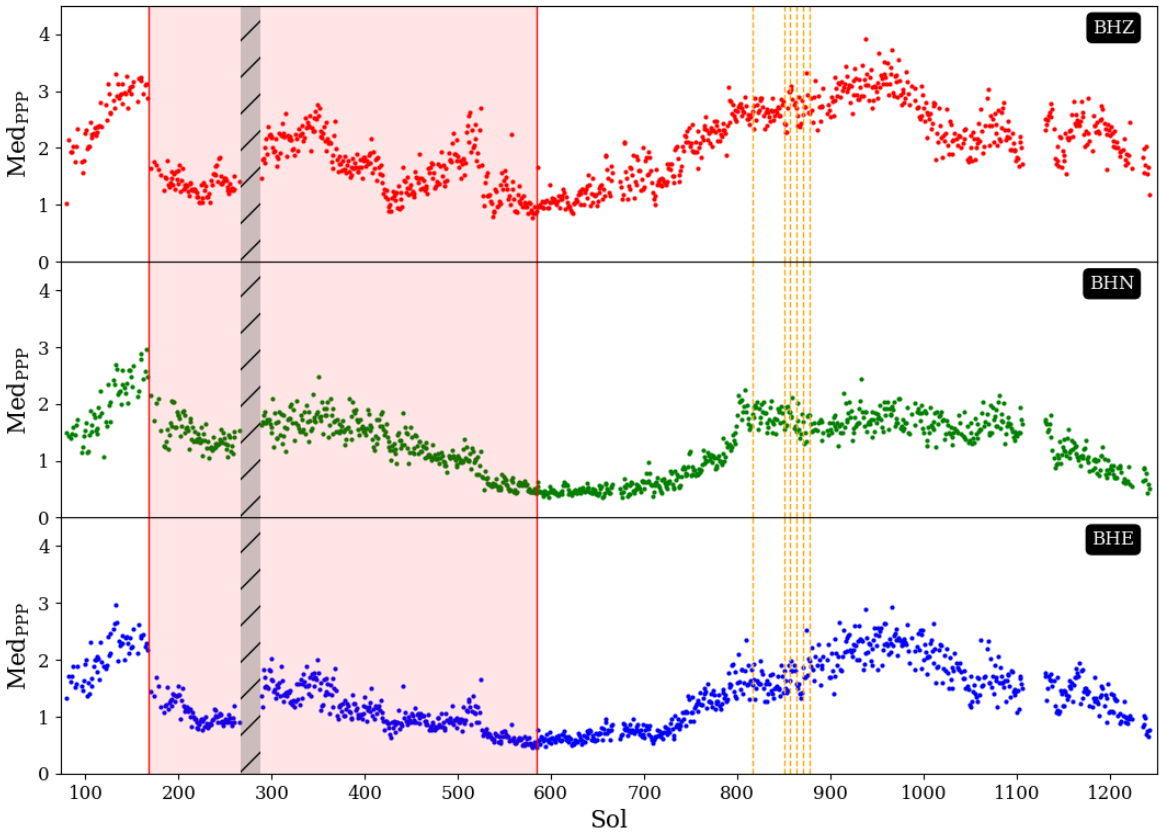

FIGURE 4.24 – Pourcentage de points perturbés médian, mesuré sur le signal sismique basse fréquence au cours de la mission, sur chacune des composantes Z, N et E (en haut, au milieu, et en bas, respectivement). La période de fonctionnement du chauffage est représentée en rouge, entre les sols 168 et 584 et la zone hachurée grise correspond à la conjonction (sols 268 à 287). Les localisations temporelles de plusieurs manœuvres d'ensevelissement du câble reliant SEIS à l'atterrisseur, ayant eu lieu entre les sols 816 et 877 sont indiquées par des lignes oranges verticales.

brutal de SEIS durant 19 sols. Le chauffage fut ensuite réactivé, ce qui semble avoir induit de nombreux *glitchs* supplémentaires dans le signal (Scholz et coll., 2020). Cette forte augmentation du nombre de *glitchs* est également nettement visible via l'étude de notre estimateur Med_{PPP}, notamment sur les composantes BHZ et BHE. Le niveau moyen de Med_{PPP} varie fortement avant et après la conjonction, de 1,34% à 2,11% pour BHZ et 0,93% à 1,50% pour BHE.

Finalement, nous choisissons d'analyser la qualité du signal basse fréquence autour d'une troisième période d'intérêt, celle marquée par les manœuvres d'ensevelissement du câble reliant SEIS à l'atterrisseur. L'objectif de cette opération fut de tenter de

diminuer le nombre de *glitchs*, dont l'origine de certains d'entre eux furent localisée en direction de ce câble (Scholz et coll., 2020). Cette observation aboutit alors à une nouvelle hypothèse de provenance de certains *glitchs*, susceptible d'être la conséquence physique d'infimes phénomènes de torsions du câble. Par conséquent, l'idée consista à utiliser la pelle du bras mécanique IDA d'InSight (Trebi-Ollennu et coll., 2018) afin de déposer du régolithe sur le câble, lors des sols 816, 850, 856, 863, 870 et 877, représentés par les lignes verticales oranges dans la figure 4.24. Cette manœuvre ayant pour objectif une diminution du nombre de *glitchs* dans le signal, celle-ci est donc alors susceptible d'être visible via notre estimateur Med_{PPP}. Cependant, nous ne distinguons, à priori, aucune tendance nette en faveur de l'amélioration de la qualité du signal à la suite de cette opération. En conclusion, cette opération d'ensevelissement ne semble pas avoir provoqué une nette diminution du nombre de *glitchs* altérant le signal sismique.

4.6 Discrimination automatique des tornades de poussière détectées par NG-loc via le *machine learning*

4.6.1 Motivations

Le capteur de pression embarqué à bord d'InSight constitue le moyen le plus fiable afin de détecter la présence de tornades de poussière à proximité de l'atterrisseur. Cependant, cette capacité de détection a été limitée par la faible quantité de données enregistrées lors de la fin de la mission, conséquence directe de la mise en veille prolongée du capteur de pression en raison d'importantes restrictions énergétiques engendrées par l'accumulation de poussières sur les panneaux solaires (voir figure 4.4). Par conséquent, une importante différence en termes de quantité de données disponibles est observée entre les signaux sismiques et de pressions, illustrée dans la figure 4.25 (BDO représentant le signal de pression et BHU, BHV et BHW les données sismiques). En effet, après le sol 900, le signal sismique enregistré (échantillonné à 20 points par se-

conde) correspond alors à une période totale cumulée de 351 sols de données, alors qu'il ne représente que 29 sols seulement pour le signal de pression. Ceci conduit alors à une détection extrêmement limitée des tornades de poussière sur cette période temporelle.

La détection classique via le capteur de pression étant impossible lors de la fin de la mission, nous avons tenté d'utiliser le signal sismique, lui-même sensible à ces phénomènes atmosphériques (Garcia et coll., 2020; Murdoch et coll., 2021; Onodera et coll., 2023), illustré par la précédente figure 4.8. En effet, il est également démontré que les tornades de poussière de fortes intensités provoquent une « signature sismique » spécifique dans le signal (Lorenz et coll., 2021), déjà observée sur la planète Terre, quelques années auparavant (Lorenz et coll., 2015).

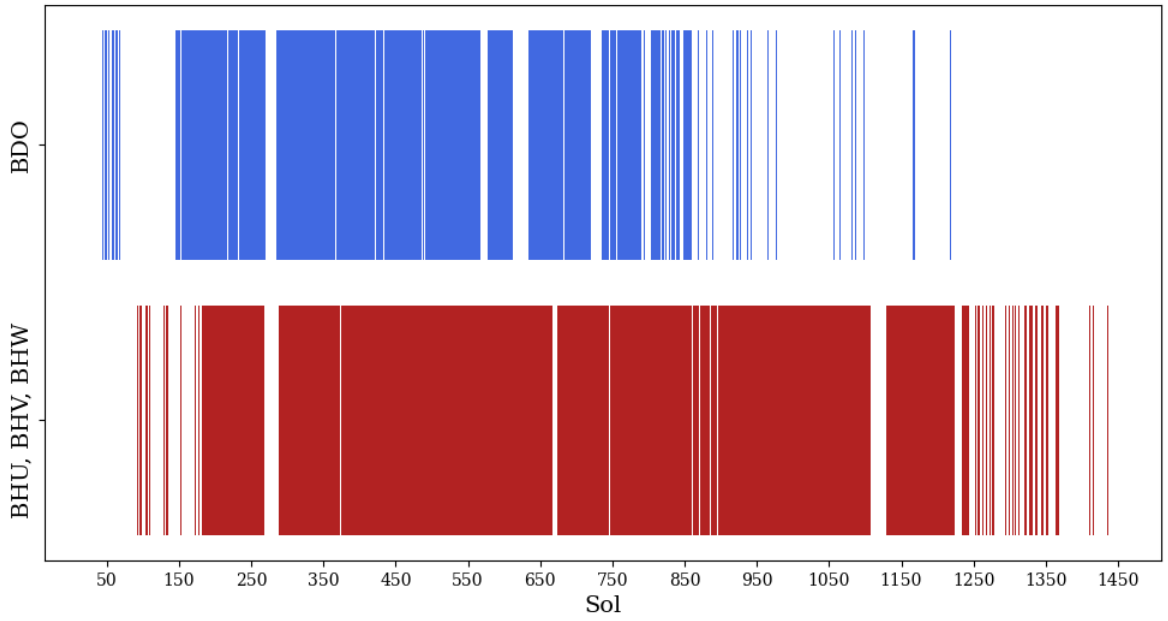

FIGURE 4.25 – Comparaison de la disponibilité des données entre le signal de pression BDO (en bleu), et les données sismiques BHU, BHV et BHW (en rouge), au cours de la mission InSight.

La figure 4.26 présente quelques heures de signal sismique brut, influencé par une tornade de poussière ayant provoqué un signal impulsif nettement visibles dans les données sismiques (labellisée « TdP » dans la figure). Bien que le signal présenté ici soit celui du sismomètre, la détection de cette tornade de poussière fut effectuée via l'analyse du capteur de pression. On distingue autour de cette tornade de poussière de nombreux autres pics impulsifs dans le signal sismique, particulièrement visibles sur la composante

BHV. Cependant, ces perturbations sont quant à elle des *glitchs* (labellisés « G » dans la figure). L'objectif de cette section consiste à tirer profit des détections de perturbations effectuées par NG-loc, afin de proposer une méthode efficace visant à discriminer celles étant associées à des tornades de poussière. Pour répondre à ce besoin, la stratégie sélectionnée repose sur une approche de discrimination automatique par *deep learning*, outil particulièrement bien adapté afin de procéder à une telle classification.

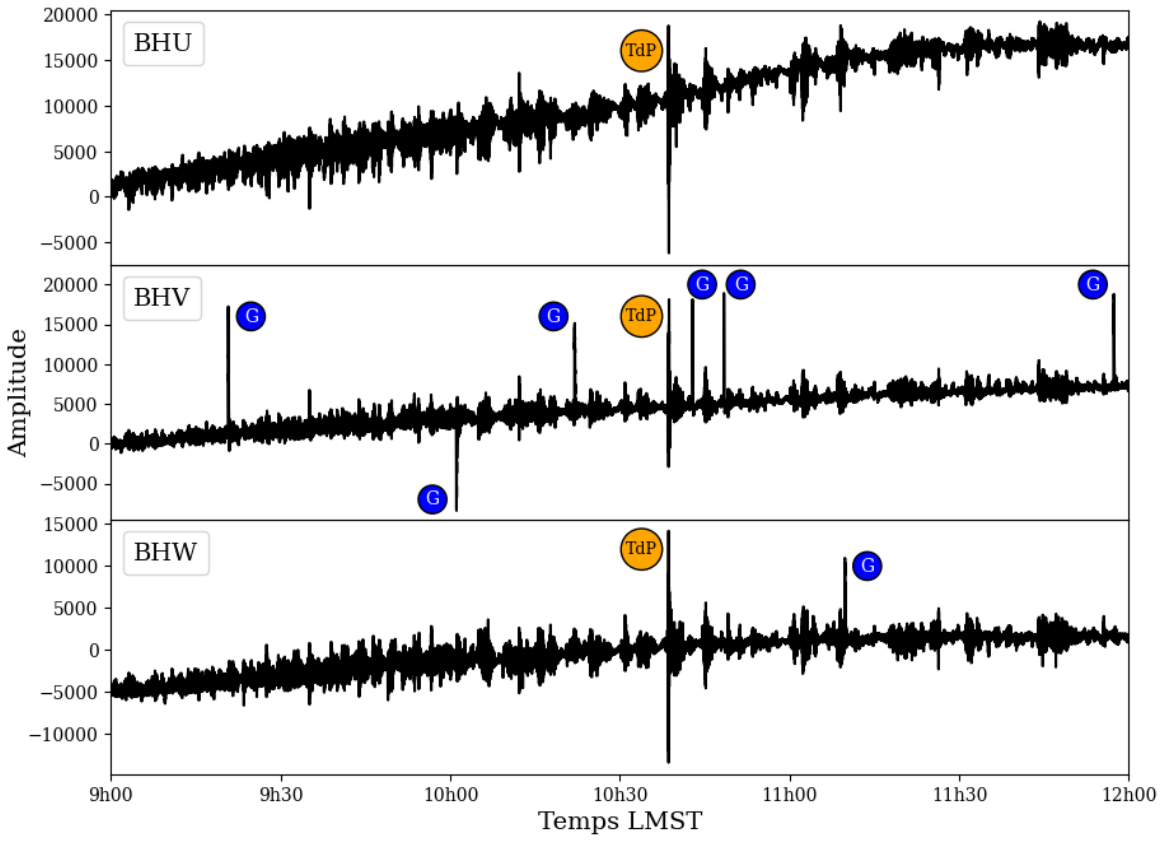

FIGURE 4.26 – Influence d'une tornade de poussière sur le signal sismique lors de la journée martienne du sol 771 (TdP). Les autres altérations, particulièrement visibles sur BHV sont quant à elles des *glitchs* (G).

4.6.2 Machine Learning

Ces dernières décennies ont notamment été marquées par une augmentation significative et rapide du volume de données numériques, causée en partie par de nombreuses innovations technologiques ainsi que l'optimisation des techniques de stockage informatique. Par conséquent, il est essentiel de développer des méthodes performantes et

efficaces afin de traiter et d'analyser ces grands jeux de données. Parmi les outils permettant un tel traitement, un moyen adapté réside dans l'utilisation de techniques de *machine learning* (apprentissage machine), utilisant l'intelligence artificielle. Le *machine learning* se définit par une approche permettant à un algorithme d'effectuer un apprentissage automatique des caractéristiques d'une base de données, sans avoir été au préalable programmé spécifiquement à cet effet. Les résultats prometteurs des techniques de *machine learning*, ont abouti, au cours de ces dernières années, à son utilisation massive sur un très large panel d'applications, comme par exemple la reconnaissance d'images, la prévisions des fluctuations du marché financier ainsi que la conception de véhicules autonomes (*e.g.* Jordan et Mitchell, 2015; Eibe et coll., 2016).

En pratique, l'idée générale du *machine learning* consiste à enseigner à un ordinateur la reconnaissance de caractéristiques statistiques (appelées *features*) associées aux éléments d'une base de données. Par la suite, les informations obtenues lors de cet entraînement sont exploitées dans le but d'effectuer une tâche spécifique sur une nouvelle base de données (prise de décision, classification, etc). À noter que la notion de base de données est ici extrêmement large et peut alors aussi bien être constituée de chiffres que d'images, de textes ou même de contenus audios ou vidéos.

Les algorithmes de *machine learning* sont divisés en deux catégories, présentées dans la figure 4.27. La première étant l'apprentissage supervisé, lorsque chaque élément de la base de donnée d'entraînement est attribué à un certain label (en fonction de sa nature). La seconde catégorie est l'apprentissage non supervisé, lorsque qu'aucune information n'est indiquée en amont. Au cours de cette section, notre étude se concentrera sur l'apprentissage supervisé, dans le cas particulier d'une méthode de classification, permettant de prédire le label de chaque élément. De manière plus spécifique, nous nous intéresserons au cas du *deep learning*, s'inspirant du fonctionnement du cerveau humain, utilisant un « réseau de neurones », permettant d'extraire des caractéristiques spécifiques d'une base de donnée.

Un exemple simple d'application d'apprentissage supervisé de classification par réseau de neurones est celui de la reconnaissance d'images, permettant de discerner, par exemple, la présence d'un chat dans une photo. Afin d'arriver à un tel résultat, l'ordi-

nateur doit en amont, analyser une grande base de données d'images labellisées, chaque label indiquant si un chat est présent, ou non. Cette première étape, appelée « entraînement », permet à l'algorithme d'apprendre à reconnaître les caractéristiques visuelles associées à un chat (moustaches, grandes oreilles, coussinets, etc...). En utilisant ces nouvelles informations, l'algorithme est alors capable de déterminer, sur de nouvelles images, si celles-ci contiennent ou non un chat.

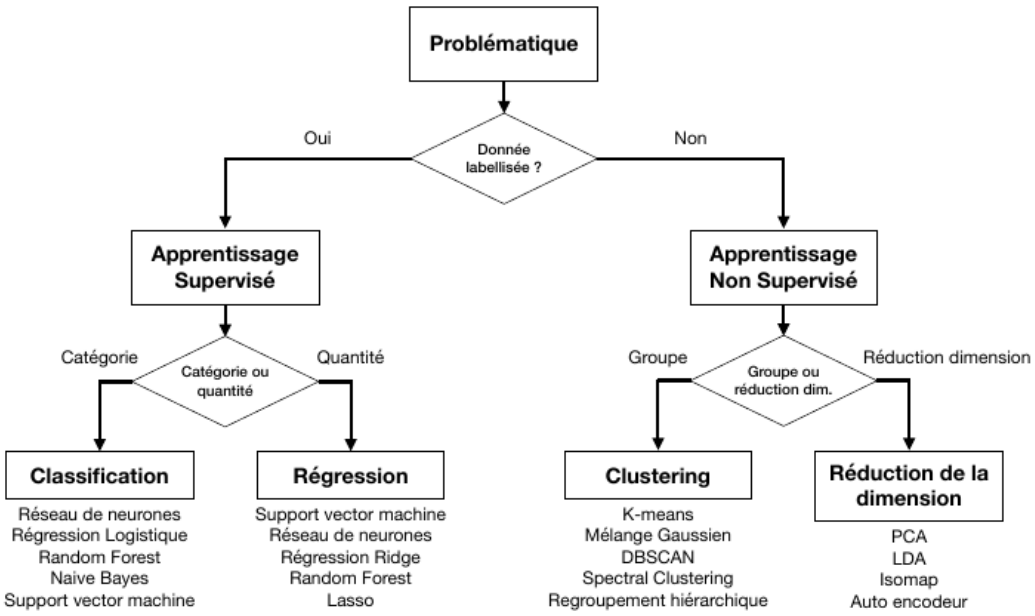

FIGURE 4.27 – Aperçu des différents types d'algorithmes de *machine learning*. Figure traduite depuis l'article de Kong et coll. (2019).

L'utilisation de ces techniques de *machine learning* se révèle particulièrement adaptée au domaine de la sismologie, nécessitant le traitement de grands volumes de données enregistrés chaque jour lors de l'acquisition du signal sismique continu. Par conséquent, de nombreuses techniques de *machine learning* ont alors été appliquées aux données sismiques terrestres au cours de ces dernières années (Kong et coll., 2019; Mousavi et Beroza, 2022). Parmi le large panel de ces applications, on peut citer par exemple la détection/classification de séismes (Seydoux et coll., 2020; Hourcade et coll., 2023), l'identification des temps d'arrivés des ondes P et S (Zhu et Beroza, 2019; Mousavi et coll., 2020), ou encore les techniques de discriminations entre les tremblements de terre et le bruit sismique (Meier et coll., 2019).

Dans le contexte de la mission InSight, de multiples contributions apportées par l'application des méthodes de *machine learning* ont également permis de prouver l'efficacité de ces techniques sur l'analyse du signal sismique martien. Parmi les résultats obtenus, on peut citer par exemple la détection supplémentaire de 700 nouveaux séismes venant s'ajouter au catalogue initial (Dahmen et coll., 2022b) ou encore la prédiction du bruit atmosphérique en utilisant les données sismiques (Stott et coll., 2023). Par ailleurs, des contributions intéressantes ont notamment été initiées par Barkaoui et coll. (2021), proposant une classification automatique des *glitchs*, mais aussi des tornades de poussière, en exploitant notamment la forte correspondance temporelle entre les perturbations détectées sur le capteur de pression et le signal sismique. Cette méthode innovante de localisation des tornades de poussière en utilisant les données sismiques se démarque alors de la stratégie habituelle, reposant uniquement sur une analyse du signal de pression (Spiga et coll., 2021). L'approche de Barkaoui et coll. (2021) tire profit de la grande sensibilité de l'instrument SEIS, se révélant capable d'enregistrer des perturbations atmosphériques, impliquant alors une grande cohérence entre les données sismiques et le signal de pression, constatée au cours de la mission (Garcia et coll., 2020; Kenda et coll., 2020).

Nous proposons ici une approche de discrimination des signaux sismiques, en nous basant sur la méthode développée par Hourcade et coll. (2023), utilisant une technique de reconnaissance d'images par « réseau de neurones convolutifs », appliquée aux spectrogrammes des trois composantes d'une station. L'étude de ces spectrogrammes se révèle particulièrement adaptée à notre objectif de discrimination des tornades de poussière, permettant alors de déceler leur signature sismique particulière. Pour proposer une discrimination performante, il est alors nécessaire de construire deux bases de données labellisées, la première, nommée **TdP**, contenant uniquement des tornades de poussière, et la seconde, nommée **Autres**, constituée de perturbations d'origines différentes. Une attention toute particulière est portée à la construction de ces deux bases de données, dont la qualité détermine l'efficacité de la méthode de *machine learning*.

4.6.3 Architecture du réseau de neurones convolutifs

La figure 4.28 présente l'architecture du réseau de neurones convolutifs utilisé dans l'article de Hourcade et coll. (2023). La première étape de ce réseau de neurones consiste à réduire la dimensions des spectrogrammes (3 spectrogrammes de dimensions 237 × 50) afin d'extraire les caractéristiques de ces derniers. En pratique, cette réduction de dimension s'effectue via l'application d'une série de filtres aux images de chaque spectrogramme analysé, correspondant à la partie *feature extraction* dans la figure 4.28 Une fois cette première étape effectuée, ces informations sont ensuite injectées dans le réseau de neurone, dont l'entraînement déterminera ensuite la qualité de la discrimination finale. Chaque perturbation analysée se voit finalement attribuer une probabilité témoignant de son appartenance à une classe spécifique. Ces différentes étapes sont scindées en plusieurs « couches » (*e.g. layers*), dans la figure 4.28, nommées L1, L2, L3 et L4 pour la partie d'extraction des caractéristiques et L5, L6 et L7 pour celles associées au réseau de neurone. La présentation de tous les aspects techniques de ce réseau de neurones n'étant pas tous détaillés dans cette section, on pourra se diriger vers l'article de Hourcade et coll. (2023) pour d'avantages d'informations.

Dans le cas de notre étude, l'architecture du réseau de neurones convolutif sera légèrement simplifiée par rapport à celle présentée dans la figure 4.28. Les dimensions des spectrogrammes seront, dans notre cas, de tailles moins importantes (3 spectrogrammes de dimensions 116 × 20), seules les deux premières couches L1 et L2 seront conservées. Par ailleurs, notre méthode de discrimination ne consistera pas dans notre cas à effectuer une discrimination entre des évènements naturels ou anthropiques, mais en une classification des perturbations associées, ou non, à des tornades de poussière.

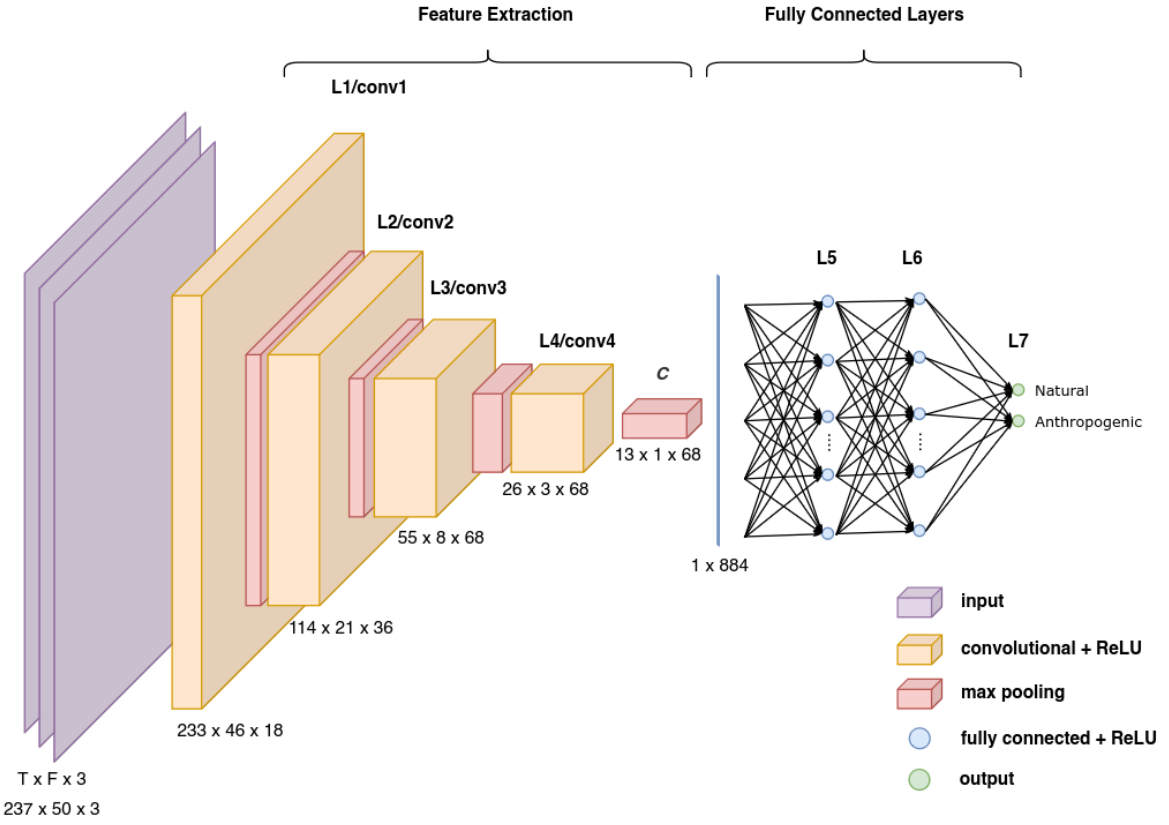

FIGURE 4.28 – Architecture du réseau de neurones convolutifs. Les trois spectrogrammes analysés sont soumis à une réduction de dimension (couches L1 à L4) avant d'intégrer le réseau de neurone (couches L5 à L6), aboutissant finalement à l'obtention d'une probabilité de discrimination. Figure extraite de l'article de Hourcade et coll. (2023).

4.6.4 Construction des bases de données

4.6.4.1 Base de données (TdP) : Perturbations associées à des tornades de poussière

Nous présentons ici, la construction de la base de données **TdP**, composée de signaux sismiques associés à des tornades de poussière, qui sera utilisée par la suite afin de générer les spectrogrammes, constituant la base d'entraînement de notre réseau de neurones convolutifs. La stratégie adoptée ici consiste à utiliser le catalogue de tornades de poussière fourni par Spiga et coll. (2021), obtenu par une analyse du signal de pression. Lorsque la valeur de celui-ci devient plus faible qu'un seuil fixé à -0,35 Pa, le signal est alors interprété comme étant associé à une tornade de poussière. Ce cata-

logue recense environ 12 000 tornades de poussière, dont l'amplitude en pression atteint jusqu'à -9,183 Pa, bien que la majorité d'entre elles correspondent à des décroissances de pression bien plus faibles (la médiane du catalogue complet étant de -0,45 Pa). Ce catalogue fournit la localisation temporelle de nombreuses tornades de poussière détectées au cours de la mission mais se restreint donc naturellement aux périodes où le signal de pression est disponible (figure 4.25).

Comme illustré par la figure 4.29, on observe une forte arrivée d'énergie dans le signal sismique, au moment de la détection d'une baisse de pression par Spiga et coll. (2021). Bien que cette énergie intense soit nettement visible dans l'exemple A, associé à une baisse de pression de -1,122 Pa, celle-ci n'est pas évidente dans l'exemple B, correspondant à une décroissance de seulement -0,361 Pa. Par conséquent, cette observation souligne qu'une cohérence temporelle est possible entre la détection d'une tornade de poussière et une perturbation des données sismiques, bien que celle-ci ne soit pas systématique, et semble dépendre notamment, de l'intensité de la chute de pression associée. Au vu de la faible influence de la tornade de poussière B dans le signal sismique, il ne semble pas judicieux d'extraire l'intégralité des 12 000 signaux sismiques associées aux tornades du catalogue de Spiga et coll. (2021) afin de construire notre base de données **TdP** (certaines tornades n'influençant pas de manière significative les données sismiques).

La figure 4.30 présente deux exemples atypiques de tornades de poussière, le premier (A) étant associé à une forte décroissance de pression (-1,144 Pa) mais étant invisible sur le signal sismique, alors que le second (B), correspond quant à lui à une baisse de pression de faible intensité (-0,631 Pa) mais une arrivée d'énergie évidente sur le signal sismique. À la vue de ce résultat, nous choisissons donc, pour construire notre base de données, de procéder à une sélection manuelle des signaux sismiques associés aux tornades de poussière détectées par Spiga et coll. (2021). Pour cela, nous inspectons seulement les signaux correspondant à des seuils de pression inférieurs à -0,6 Pa. Cette restriction permet 1) d'effectuer une observation sur un sous-ensemble de tornades étant plus susceptible d'influencer le signal sismique de manière évidente mais aussi 2) de procéder à une inspection manuelle des signaux sur une base de données de taille

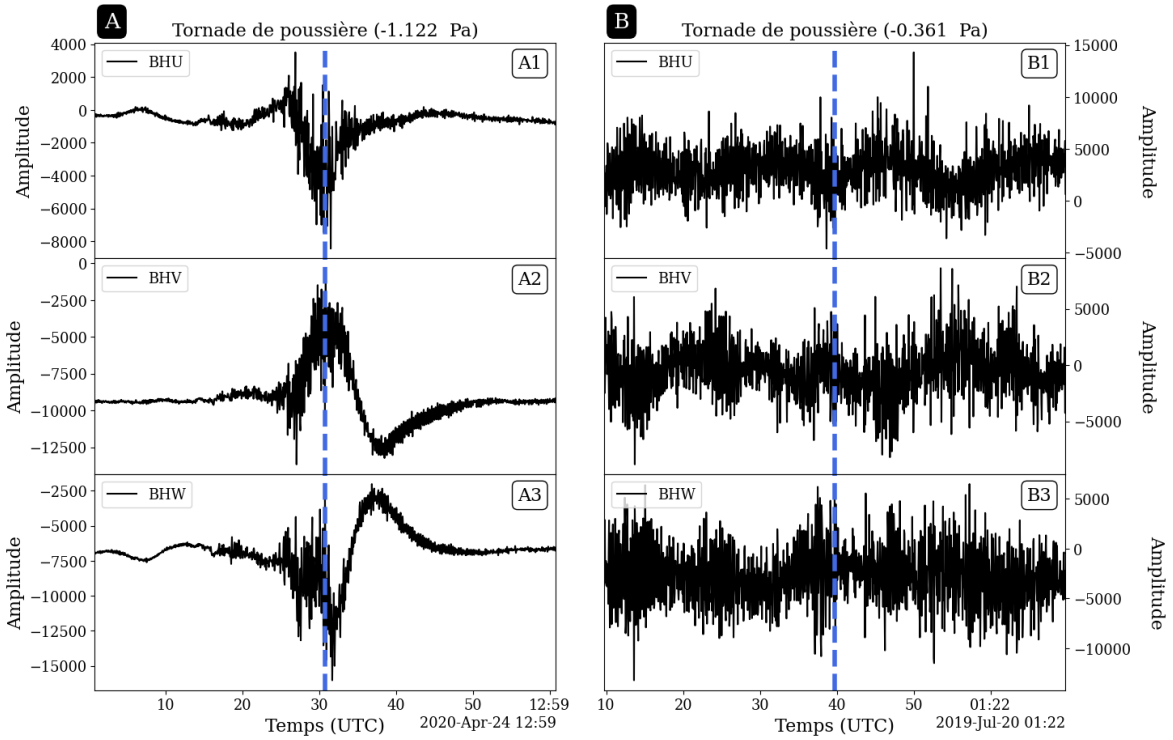

FIGURE 4.29 – Influence des tornades de poussière sur les trois composantes du signal sismique. L'exemple A présente une tornade de poussière associée à une forte baisse de pression de -1,122 Pa, tandis que B correspond à une décroissance de seulement -0,361 Pa. Chaque exemple propose les trois composantes du signal sismique BHU (A1 et B1), BHV (A2 et B2) et BHW (A3 et B3). La localisation temporelle des tornades de poussière, telle que proposée par le catalogue de Spiga et coll. (2021) est représentée, sur chaque exemple, par une droite bleue verticale.

réduite (passant de 12 000 à exactement 3230 tornades de poussière).

L'objectif de cette sélection manuelle, consiste alors à conserver, parmi les 3230 tornades de poussière, celles étant observées, de manière évidente sur le signal sismique (comme par exemples les tornades A de la figure 4.29 et B de la figure 4.30). Bien que cette notion soit évidemment subjective, nous présentons dans la figure 4.31 quelques exemples de signaux, intégrés ou non à la base de données, au cours de cette sélection manuelle. Les exemples A et B sont conservés pour intégrer la base de données **TdP**, car étant tous deux associés à des signaux sismiques se démarquant significativement du bruit, sur les trois composantes. *A contrario*, les tornades de poussière C et D n'influencent pas, de manière évidente, les données sismiques associées, et ne sont donc pas sélectionnées pour faire partie de notre base de donnée. Bien que le doute soit

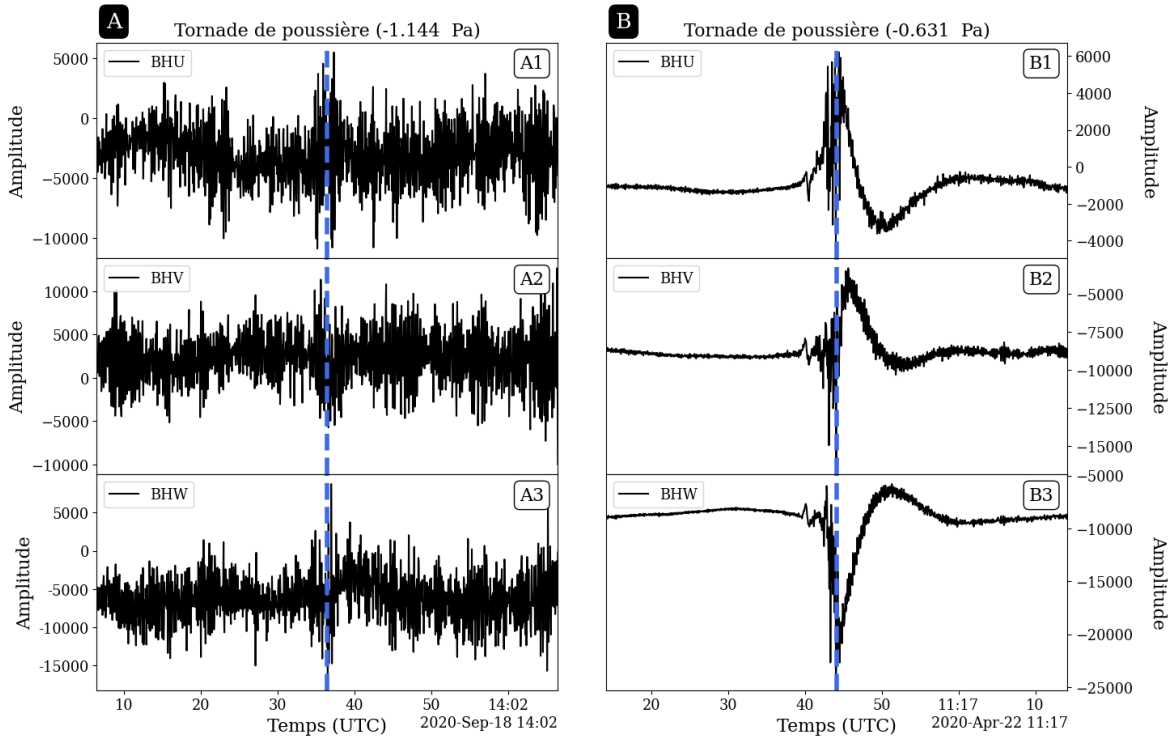

FIGURE 4.30 – Même légende que pour la figure 4.29, pour deux autres exemples de tornades de poussière A et B ayant induit une baisse du signal de pression de -1,144 et -0,631 Pa, respectivement.

permis dans l'exemple D, comportant une légère perturbation longue période en son centre, celle-ci n'est tout de même pas sélectionnée afin d'obtenir au final, une base de donnée **TdP** aussi fiable que possible, permettant d'affirmer avec certitude que les perturbations conservées sont causées par des tornades de poussière.

L'inspection manuelle des données a permis de sélectionner un total de 1326 signaux sismiques (correspondant à un peu moins de la moitié des éléments inspectées) venant constituer la base de données **TdP**. Afin d'adapter la longueur des signaux enregistrés à celle des tornades de poussière dans le signal sismique, nous choisissons de conserver les données sismiques (les trois composantes) sur des fenêtres de 60 s, centrées autour des temps de détection proposées par le catalogue de Spiga et coll. (2021). Finalement, la figure 4.32 présente la position temporelle, de chaque tornade de poussière formant la base de donnée **TdP**.

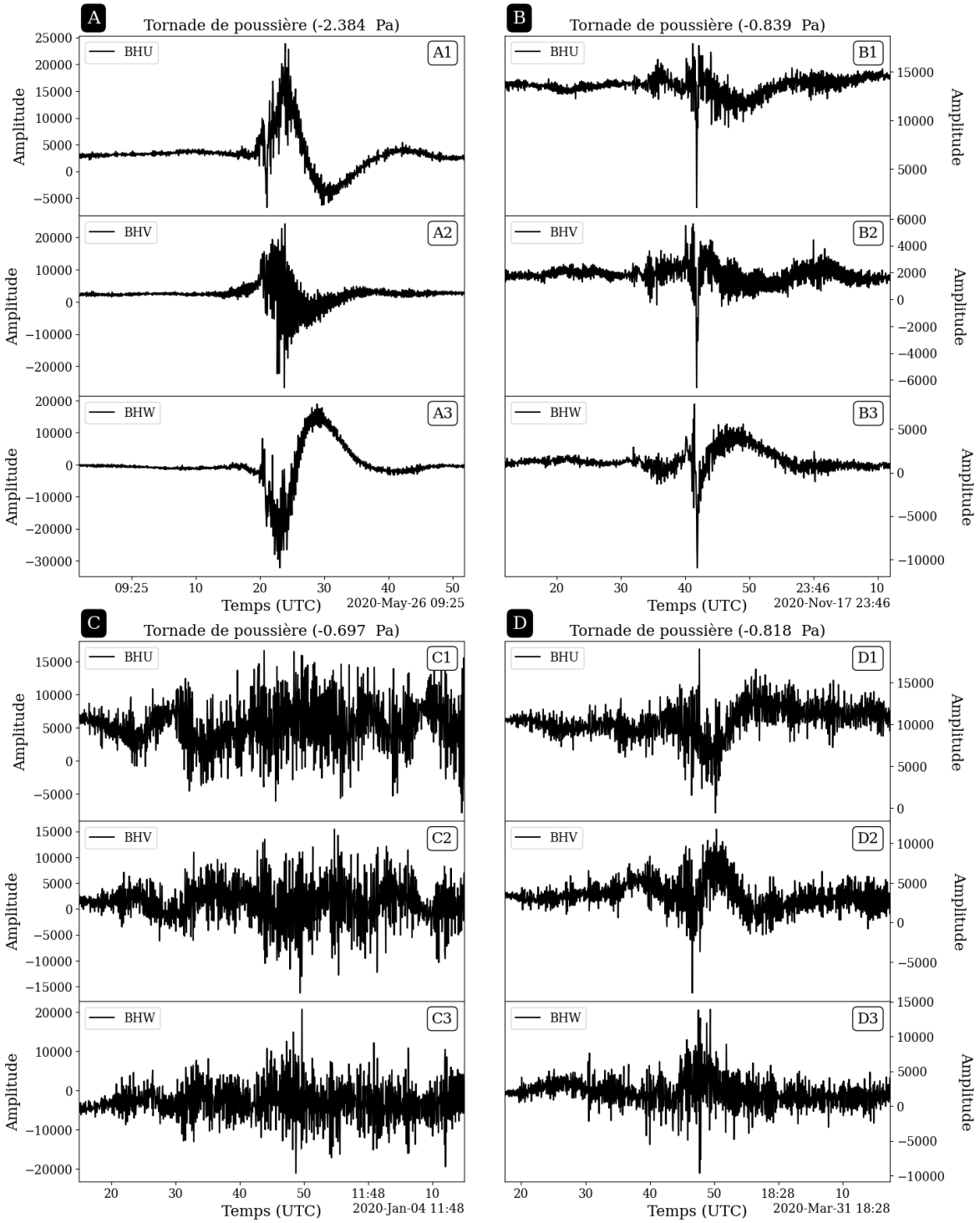

FIGURE 4.31 – Illustration de la sélection manuelle des tornades de poussière : les exemples A et B sont conservés pour être intégrés à la base de données **TdP**, tandis que C et D, en sont exclus, car n'étant pas associés à une énergie évidente sur les signaux sismiques correspondant. Même légende que pour la figure 4.29.

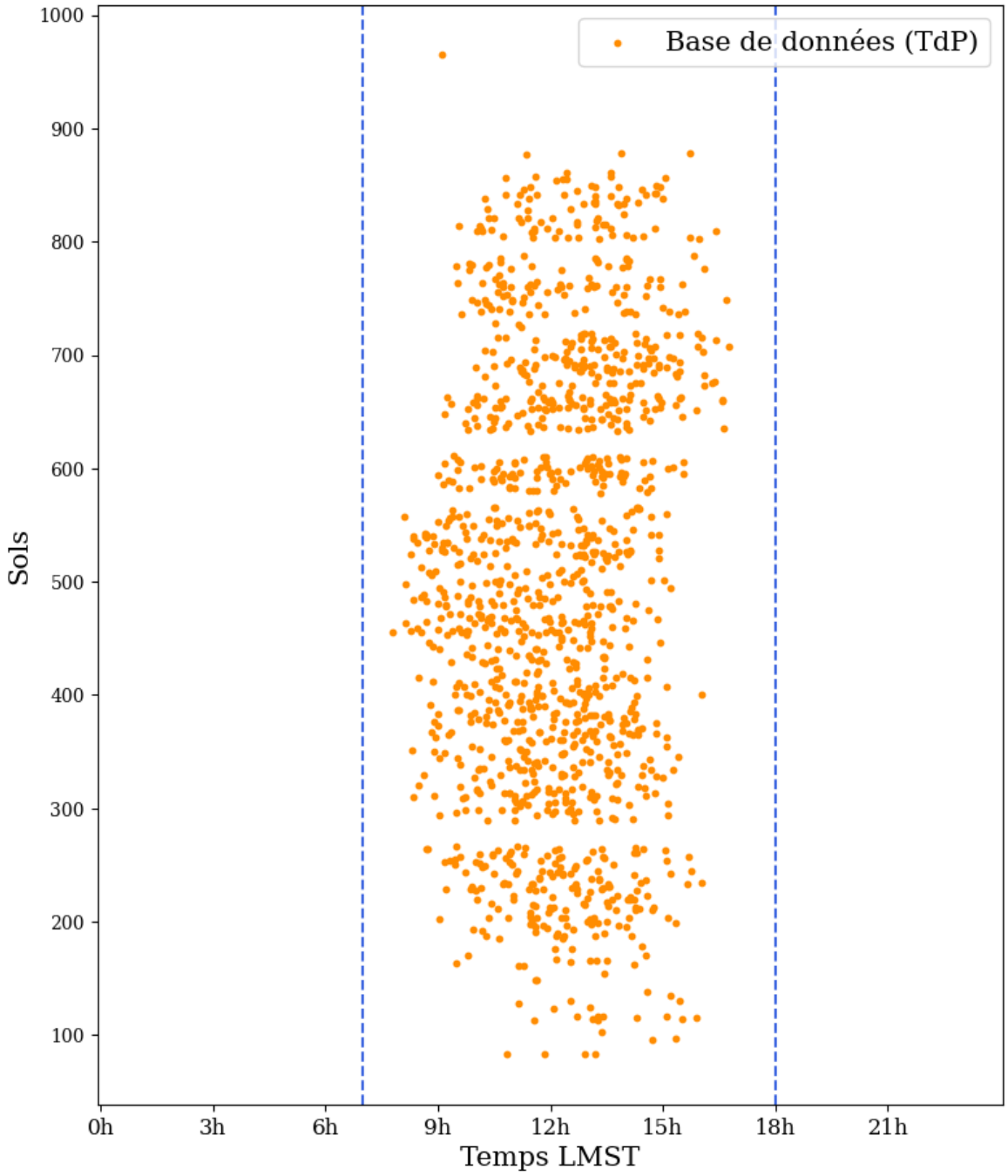

FIGURE 4.32 – Localisation temporelle des tornades de poussière de la base de données **TdP**.

4.6.4.2 Base de données (Autres) : Perturbations non associées à des tornades de poussière

Nous décrivons désormais la construction de notre seconde base de donnée, **Autres**, comportant des perturbations n'étant pas associées à des tornades de poussière. Afin de répondre à cet objectif, la stratégie consiste simplement à utiliser la base de données de perturbations basses fréquences décrite dans la section 4.4.2. Les éléments de la base de données **Autres** devant êtres constituées de perturbations n'étant pas associées à des tornades de poussière, nous excluons celles ayant la même correspondance temporelle avec les tornades de poussière du catalogue de Spiga et coll. (2021). Notre choix s'est ici porté sur une exclusion des perturbations, si celles-ci sont à une distance de moins de 30 secondes d'une tornade de poussière du catalogue. Bien que ce choix de 30 secondes pourrait paraître brutal, il nous permet d'être certains que les 12 000 tornades de poussière détectées au cours de la mission n'appartiennent pas à notre base de données **Autres**.

Finalement, nous parcourons alors nos perturbations, sur chacune des composantes, et sélectionnons, de la même manière que pour la base de données **TdP**, le signal sismique de (BHU, BHV et BHW), sur des fenêtres de 60 secondes, centrées au milieu de chaque altération. Afin d'éviter les doublons dans notre base de données **Autres**, une perturbation détectée sur plusieurs composantes donnera lieu à la conservation d'un unique triplet de signaux sismiques (BHU, BHV et BHW).

4.6.5 Discrimination des tornades de poussière

4.6.5.1 Création des spectrogrammes

Nous utilisons désormais les bases de données **TdP** et **Autres**, dont les constructions furent détaillées dans la sous-section précédente. Nous appliquons la même méthode que celle décrite dans Hourcade et coll. (2023), utilisant un réseau de neurones convolutifs, afin de procéder à une reconnaissance d'image sur ces derniers. La reconnaissance d'image appliquée aux spectrogrammes permet alors, de déceler les caracté-

ristiques fréquentielles d'un certain type de perturbation (dans notre cas, les tornades de poussière). Afin de générer nos spectrogrammes, nous proposons d'appliquer les paramètres de tailles de fenêtres glissantes ayant conduit aux meilleurs résultats, correspondant alors à une longueur de 2 s, contre 1 s seulement dans l'article de Hourcade et coll. (2023), en conservant toutefois le même ratio d'*overlap* de 75%. Par ailleurs, de la même manière que dans l'approche de Hourcade et coll. (2023), chaque spectrogramme est normalisé par son maximum, permettant de mettre en valeur le pic fréquentiel de chacun d'entre eux. Finalement, chaque perturbation dans les bases de données **TdP** et **Autres** correspondent alors à un triplet de spectrogrammes, sur les composantes U, V et W.

Nous présentons dans la figure 4.33 deux exemples de perturbations, l'un associé à une tornade de poussière (colonne de gauche) provenant de notre base de données **TdP**, et l'autre à un *glitch* (colonne de droite) issue de **Autres**. Pour chaque exemple, les signaux sismiques des trois composantes sont affichés (A/G pour BHU, C/I pour BHV et E/K pour BHW), mais également le triplet de spectrogrammes associé (B,D et F pour la tornade de poussière et H, J et L pour le *glitch*). Chaque spectrogramme étant normalisée par son maximum, leur amplitudes est alors représentée par une unité sans dimension, entre zéro (en blanc) et 1 (en bleu foncé).

De nombreuses différences entre ces deux perturbations, sont visibles directement sur le signal sismique (formes d'ondes, présence de précurseur haute fréquence uniquement dans le cas du *glitch*, etc...), mais également sur les spectrogrammes. En effet, le *glitch* se caractérise tout d'abord par un précurseur haute fréquence ($T = 60$ s exactement) visible sur toutes les composantes (H, J et L de la figure 4.33) puis par une large altération longue période observée, dans ce cas, uniquement sur les composantes U et W (G et K). Ces caractéristiques se démarquent nettement de celles visualisées dans les spectrogrammes des tornades de poussière (B, D et F), qui elles présentent alors une altération multi-fréquentielle se caractérisant alors par une large zone sombre (entre les temps $T = 40$ s et $T = 60$ s). Ces caractéristiques fréquentielles différentes sont déterminantes pour permettre le bon fonctionnement de notre approche de discrimination via *machine learning*. Bien que nous ne puissions évidemment pas présenter ici une très

large gamme de perturbations, les observations effectuées sur les spectrogrammes de la base de données **TdP** révèlent que l'altération multi-fréquentielle est une signature caractéristique, associée au passage d'une tornade de poussière.

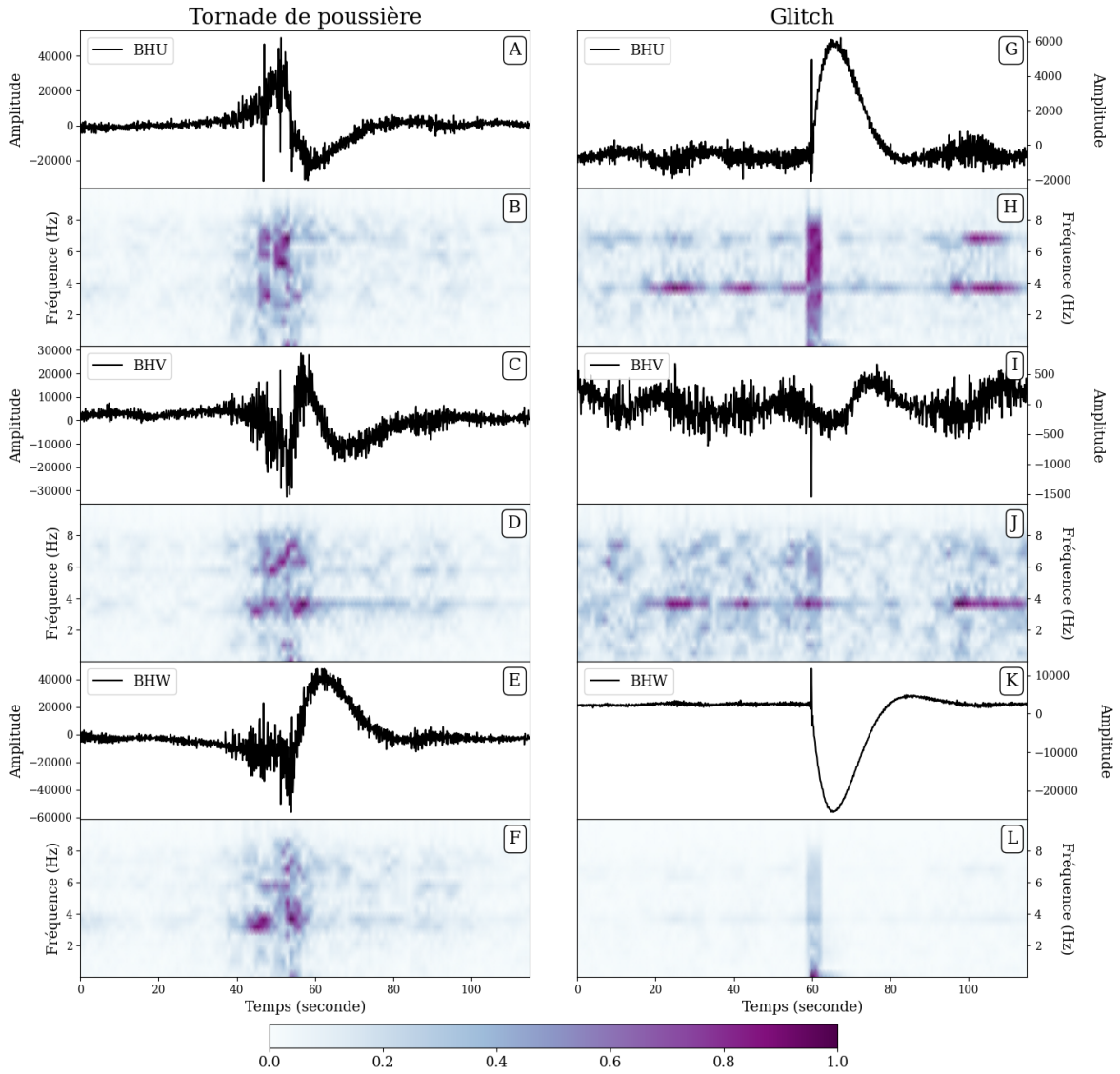

FIGURE 4.33 – Spectrogrammes et signaux sismiques associés à une tornade de poussière provenant du sol 323 (colonne de gauche) et un *glitch*, lors du sol 1340 (colonne de droite). Chaque exemple propose le signal sismique des trois composantes BHU (A et G), BHV (C et I) et BHW (E et K). Sous chaque signal, on affiche le spectrogramme correspondant (B, H, D, J, F et L). Le *glitch* provient de notre base de donnée **Autres** tandis que la tornade de poussière provient de **TdP** et découle donc du catalogue de Spiga et coll. (2021).

On pourra remarquer dans les spectrogrammes une énergie fréquentielle plus intense à environ 4 Hz, causée par la résonance de atterrisseur, déjà illustrée dans la

figure 4.6. Bien que cette bande de fréquence altère le spectrogramme, ce genre de résonance étant observée de manière régulière au cours de la mission, celle-ci ne constitue toutefois pas un critère déterminant dans le choix de la discrimination. À noter que la base de données **Autres** n'est évidemment pas uniquement constituée de *glitchs*, mais aussi de multiples autres altérations d'origines différentes (influence de l'atterrisseur, perturbations causées par des recalibrations du sismomètre, etc). Toutefois, notre choix s'est porté, dans le cas de cette illustration, sur la comparaison avec les spectrogrammes associés à un *glitch* car ces derniers représentent la majorité des perturbations faisant partie de cette base de données.

Finalement, nous proposons dans la figure 4.34 quatre exemples supplémentaires de spectrogrammes de tornades de poussière appartenant à notre base de données **TdP**. Sur tous ces exemples, on distingue sur toutes les composantes cette large zone sombre, qui semble être, entre autres, l'une des caractéristiques fréquentielles des tornades de poussière. Dans chaque cas, on observe approximativement la même position de cette large zone sombre, ainsi que des longueurs similaires d'environ une dizaine de secondes. Par ailleurs, celle-ci se démarque nettement du contenu fréquentiel dans les autres zones du spectrogramme, caractérisée par des couleurs claires. À noter que la normalisation de ces spectrogrammes est ici effectuée entre 0 et 0,5, permettant une meilleure visualisation de ces derniers.

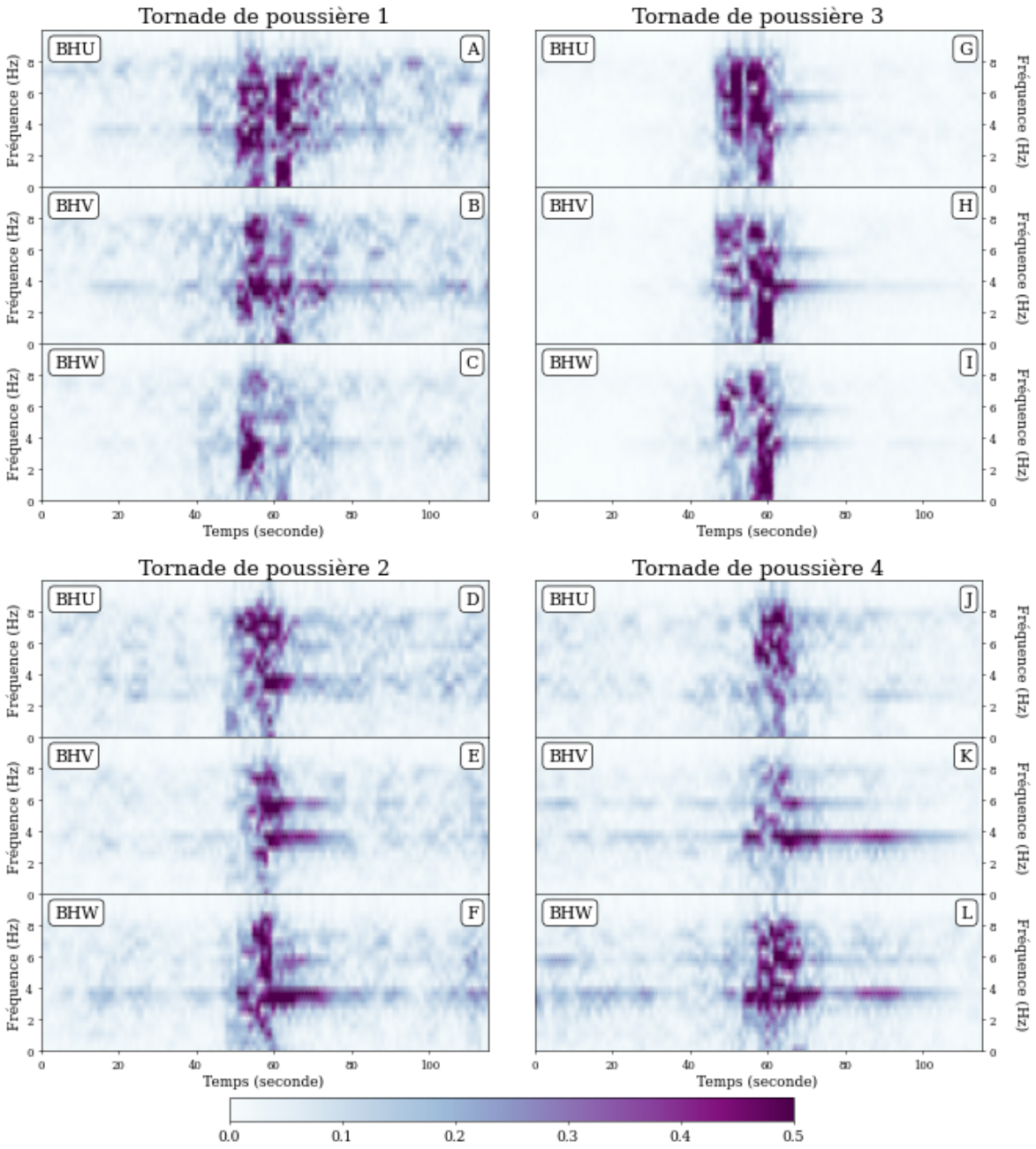

FIGURE 4.34 – Spectrogrammes de quatre tornades de poussière provenant de notre base de données **TdP**. Sur chaque exemple, le spectrogramme sur les composantes U, V et W sont affichés.

4.6.5.2 Entraînement et validation

Nous décrivons ici, la phase d'entraînement de l'approche de *machine learning* utilisant les spectrogrammes (sur les trois composantes) de la totalité des 1326 tornades de poussière de la base de données **TdP** ainsi que 10 000 perturbations, sélectionnées

aléatoirement, provenant de **Autres**. Cette sélection aléatoire est effectuée entre les sols 200 et 500, correspondant à une période temporelle de forte disponibilité du signal de pression (en dehors de la conjonction entre les sols 268 et 284), permettant alors d'exclure les tornades de poussière dans cette base de données. La figure 4.35 illustre la position temporelle de chaque élément appartenant à la base de données **TdP** (points oranges) et **Autres** (points noirs).

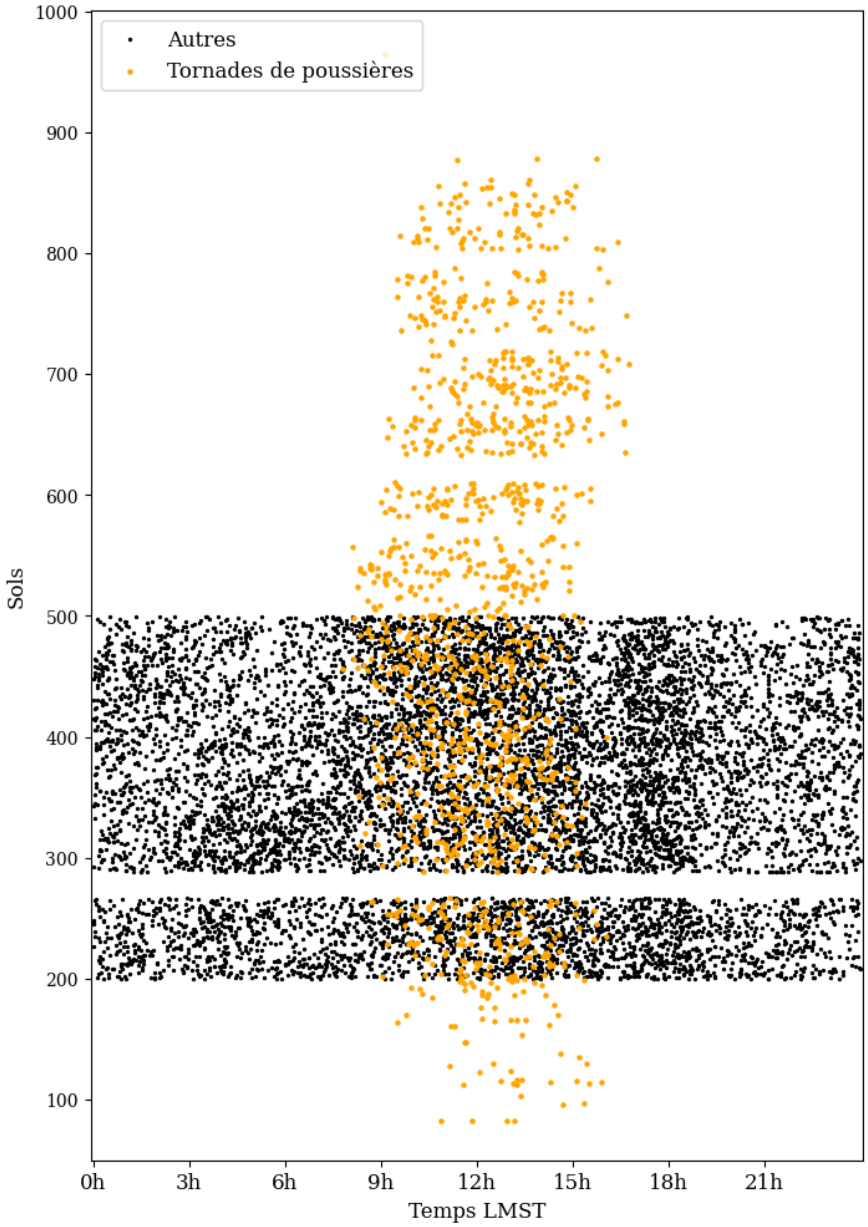

FIGURE 4.35 – Position temporelle des éléments des bases de données **TdP** et **Autres**, utilisées lors de la phase l'entraînement.

Nous procédons à la phase d'entraînement de discrimination des tornades de poussière via une étude des images des spectrogrammes. En pratique, chaque spectrogramme est analysé via une succession de filtres, permettant d'obtenir de multiples informations sur les caractéristiques visuelles associées aux sismogrammes perturbés par les tornades de poussière. Ces informations sont alors conservées via un réseau de neurones, dont les coefficients sont ajustés au cours de l'entraînement pour retrouver au mieux ces caractéristiques (voir Hourcade et coll. (2023) pour des informations détaillées).

Finalement, cette phase d'entraînement est intrinsèquement liée à une seconde étape, dite de validation, permettant de s'assurer du bon déroulement de l'entraînement. En pratique, l'entraînement ne s'effectue que sur 90 % de la base de données complète, permettant ainsi de conserver 10 % des éléments afin de constituer une base de validation. Cette base de données de validation n'intervient pas lors de l'entraînement, et n'influence donc pas l'ajustement des filtres du réseau de neurone. Une fois l'entraînement effectué, cette base de données réduite de validation permet de tester le bon fonctionnement de notre approche de discrimination des tornades de poussière sur une base de données n'ayant pas été analysée lors de la phase d'entraînement. Dans notre cas, notre base de validation est constituée de 124 tornades de poussière et 1009 perturbations d'origines différentes. En pratique, chaque perturbation analysée se voit attribuer une probabilité d'appartenir à la classe 'Tornade de poussière'. Si cette probabilité est supérieure à 50 %, la perturbation sera alors classée en tant que tornade de poussière.

La figure 4.36 présente un tableau appelé matrice de confusion, illustrant les résultats de la discrimination des perturbations de notre phase de validation. La première ligne de cette matrice (ligne du haut) quantifie la qualité de la classification des 124 perturbations ayant été labellisées, en amont, en tant que tornades de poussière dans notre base de données. Suite à la discrimination, ces tornades de poussière sont alors scindées en deux populations, correspondant aux prédictions effectuées par notre approche. La première population étant celle ayant abouti à la bonne classification de ces perturbations en tant que tornades de poussière (en haut à gauche de la matrice de confusion), représentant alors 116 éléments sur 124. La seconde population (case en haut à droite),

correspond aux perturbations labellisées en tant que tornades de poussière mais ayant cependant abouti à une classification dans la catégorie **Autres**, correspondant ici à un nombre réduit de seulement 8 éléments sur 124. L'inspection de cette première ligne nous permet de conclure à des résultats de discrimination satisfaisants, avec 93,5 % de tornades de poussière ayant été bien classées.

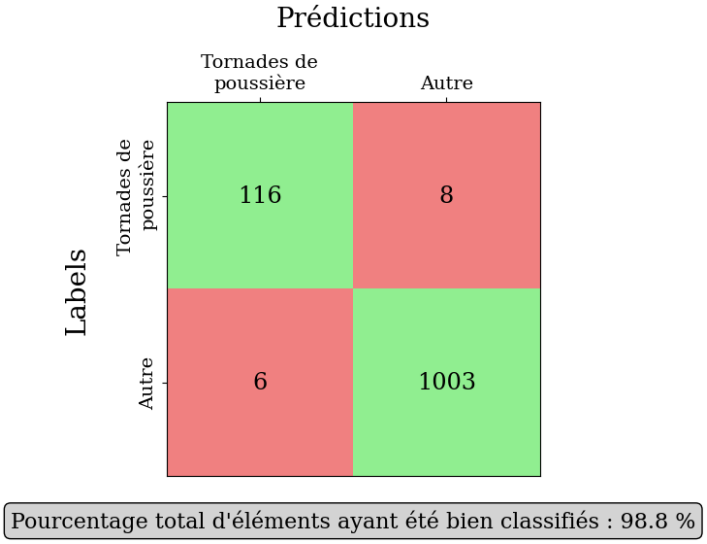

FIGURE 4.36 – Matrice de confusion, illustrant la qualité des résultats obtenus lors de la phase de validation.

La seconde ligne de la matrice de confusion (ligne du bas) représente quant à elle la qualité de la classification des perturbations appartenant à la base de données **Autres**, n'étant pas associées à des tornades de poussière. Parmi les 1009 éléments analysés, seulement 6 d'entre eux ont été (mal) classés en tant que tornades de poussière par notre approche. Cette proportion est faible, au vu des 1003 éléments ayant été classés en tant que perturbation 'Autre', c'est-à-dire non associées à des tornades de poussière. Les résultats de discrimination de cette seconde population sont extrêmement satisfaisants, avec une proportion de 99,4 % d'éléments ayant été bien classés.

Bien que ces résultats soient convaincants, on remarque cependant que la proportion d'élément bien classés n'atteint évidemment pas exactement 100 %. Une telle précision parfaite n'est cependant, en général, pas atteinte dans ce type d'approche de *machine learning* et peut s'expliquer dans notre cas par plusieurs facteurs. Parmi ceux-ci, on peut, par exemple, citer la restriction de notre base de données de tornades de poussière

à seulement 1326 éléments, pouvant alors limiter la qualité du résultat obtenu (une base de données plus grande aurait par exemple permis un meilleur entraînement). Par ailleurs, la fiabilité du catalogue de Spiga et coll. (2021), utilisé pour générer nos bases de données, est ici considérée comme absolue, supposant alors que celui-ci ne loupe aucune tornade de poussière, mais n'est également constituée d'aucune fausse détection.

En conclusion, les résultats obtenus au cours de cette étape de validation sont alors encourageants, avec une pourcentage global de 98,8 % éléments ayant été bien classés. Par conséquent, ceci nous permet de nous assurer du bon déroulement de notre entraînement, ainsi que de la fiabilité de notre approche de discrimination de tornades de poussière par analyse de spectrogrammes. La prochaine étape de notre travail consiste à finalement appliquer cet outil de discrimination efficace à une base de données de perturbations non labellisées, afin de retrouver les tornades de poussière.

4.6.5.3 Résultats

Nous procédons à une discrimination des tornades de poussière, entre les sols 900 et 1446, correspondant à une période d'importante mise en veille du capteur de pression. Les perturbations analysées correspondent ici à celles détectées par notre approche NG-loc sur cette période, représentant alors un nombre total de 106 845 éléments. La figure 4.37 présente les résultats de discrimination de notre approche suite à l'analyse de ces 106 845 perturbations, permettant alors de classer 2256 d'entre elles en tant que tornades de poussière (croix rouges). Bien que de très nombreuses perturbations aient été analysées, et ce, même pendant la nuit martienne (comme indiqué par la répartition quasi uniforme des points bleus dans la figure 4.37), l'immense majorité d'éléments ayant été classés en tant que tornades de poussière sont situées entre 7h00 et 18h00 LMST. Par conséquent, la position temporelle de ces perturbations classées en tant que tornades de poussière est donc un argument convaincant, en faveur de l'efficacité de cette approche de discrimination.

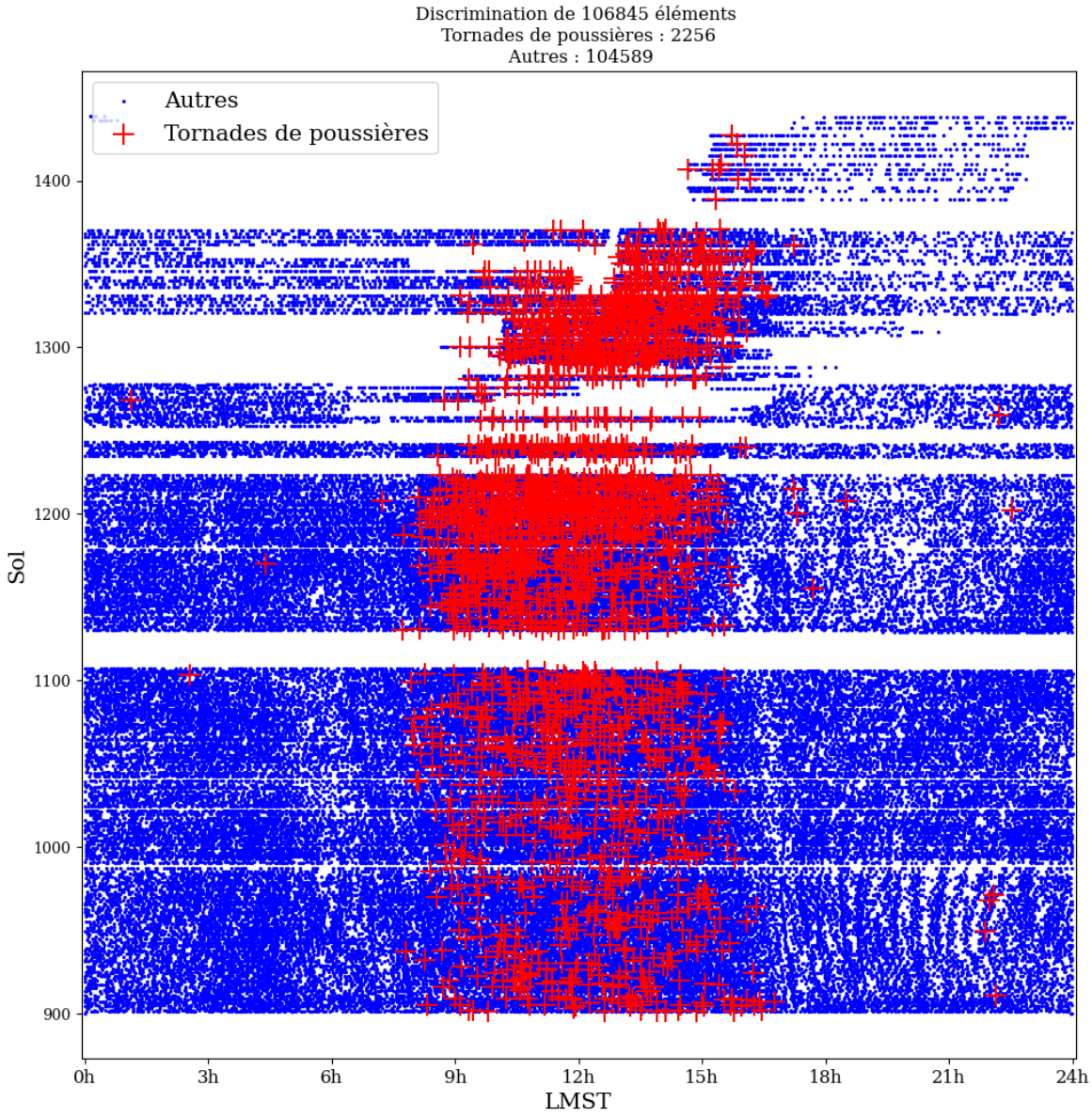

FIGURE 4.37 – Analyse d'un nombre total de 106 845 perturbations entre les sols 900 et 1446 (fin de la mission) par reconnaissance d'image des spectrogrammes associés, aboutissant à une discrimination de 2256 nouvelles tornades de poussière.

Nous observons cependant que, parmi les 2256 perturbations classées en tant que tornades de poussière, dix d'entre elles sont situées lors de la nuit martienne, avant 7h00 ou après 18h00 LMST. Bien que la position temporelle de ces perturbations permettent, à elles seules, de conclure de la mauvaise classification de celles-ci, une rapide observation du signal sismique associé permet également de confirmer cette erreur de discrimination. En effet, la figure 4.38, présente une de ces perturbations

ayant été mal classées (sol 971, 22h05 LMST) nous permettant de comprendre que celle-ci est en réalité un *glitch*, comme indiqué par son précurseur haute fréquence, suivi directement d'une large altération longue période. En complément de ces arguments, la méthode de discrimination a ici classée (par erreur) cette perturbation comme étant une tornade de poussière, avec une probabilité de 53,7 %. Bien que cette probabilité soit supérieure à 50 %, celle-ci demeure toutefois extrêmement faible afin de prétendre à une décision fiable de notre approche. Cette propriété intéressante a également été observée pour les autres perturbations mal classées (points rouges situés lors de la nuit martienne sur la figure 4.37), où ces dernières sont associées à des probabilités de discriminations relativement faibles.

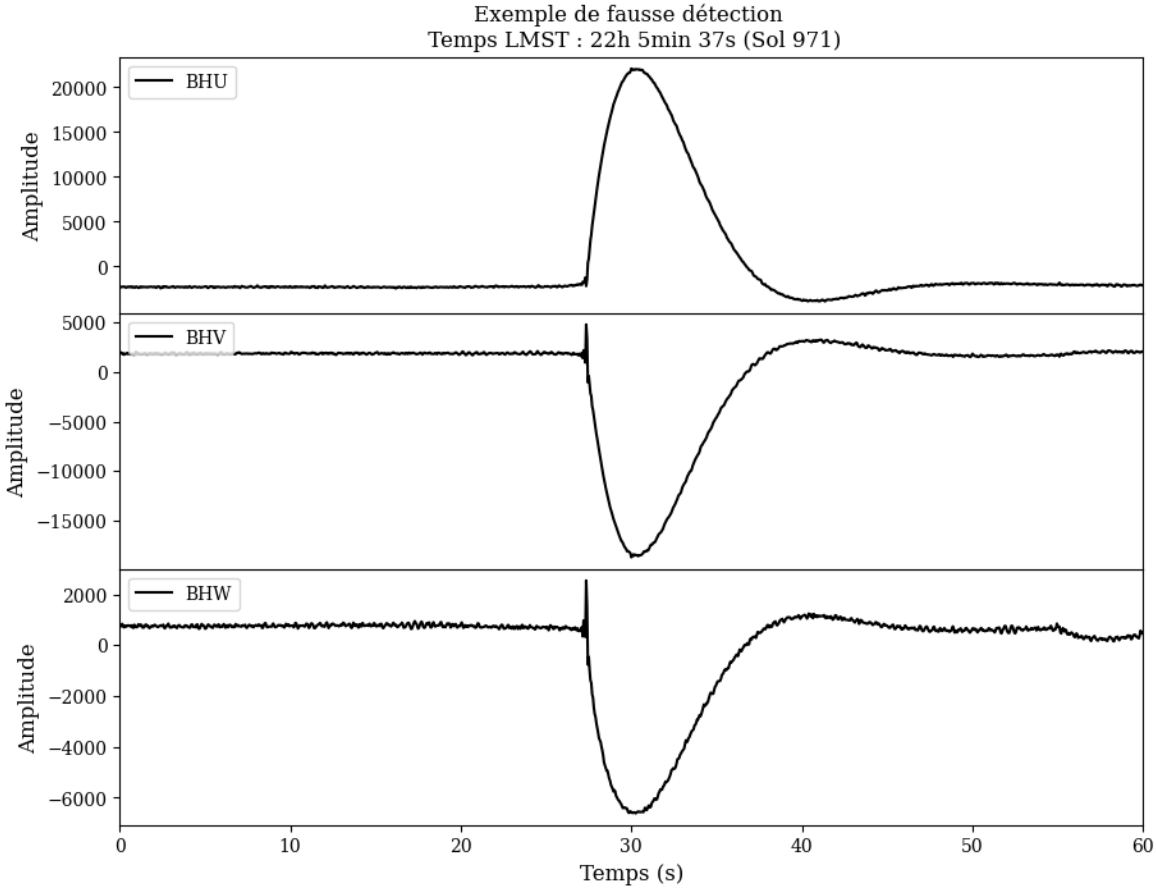

FIGURE 4.38 – Exemple de fausse détection de notre approche de discrimination de tornade de poussière par *machine learning*. *Glitch* ayant été classé en tant que tornade de poussière.

La figure 4.39 présente la localisation temporelle des tornades de poussière associées à des seuils de probabilité supérieurs à 70 %. En comparaison avec la figure 4.37, le nombre de tornades de poussière affiché correspond désormais à un sous-ensemble de 1413 éléments. Ceci nous indique qu'une forte proportion de tornades de poussière (plus de la moitié) sont associées à des fortes probabilités de discrimination. Le nombre de fausses détections en dehors de la journée martienne a quant à lui drastiquement diminué, passant de dix à seulement une seule perturbation classée en tant que tornade lors du sol 1103, vers 3h00 LMST (associée à une probabilité de 82,4 %). Ceci nous indique que la grande majorité des fausses détections évoquées précédemment sont associés à des probabilités de classification relativement faible, en dessous de 70 %.

La figure 4.40 présente une comparaison entre les signaux sismiques associés à deux tornades de poussière détectées par l'approche de *machine learning* à celles faisant partie de la base de donnée **TdP**, utilisée lors de la phase d'entraînement. Les exemples A et C proposent deux tornades de poussière (sol 464 et 459, respectivement), appartenant à la base de données **TdP**, faisant alors partie du catalogue fourni par Spiga et coll. (2021), détectées par une analyse du capteur de pression. En inspectant la forme d'onde de ces signaux sismiques, on remarque de fortes similarités avec ceux obtenus par notre approche de *machine learning* (E et G), ayant été classés en tant que tornades de poussière avec une probabilité de 99,8 % et 99,2 %, respectivement. De plus, cette similarité est également visible sur les spectrogrammes associés, utilisés lors de la discrimination, en comparant B et F, ainsi que D et H.

En conclusion de notre étude de discrimination de tornades de poussière par *machine learning*, nous avons donc de nombreuses raisons convaincantes nous permettant de nous assurer que nos détections sont bien constituées, en très grande majorité, de tornades de poussière. Parmi ces dernières, nous avons tout d'abord la bonne qualité de la phase d'entraînement/validation, dont les résultats encourageants ont été présentés dans la sous-section 4.6.5.2. De plus, la position temporelle des perturbations ayant été classées en tant que tornades de poussière (voir figure 4.37) est également un argument extrêmement convaincant, en faveur de la fiabilité de notre approche. En effet, plus de 99,5 % des tornades de poussière ont été détectées au cours de la

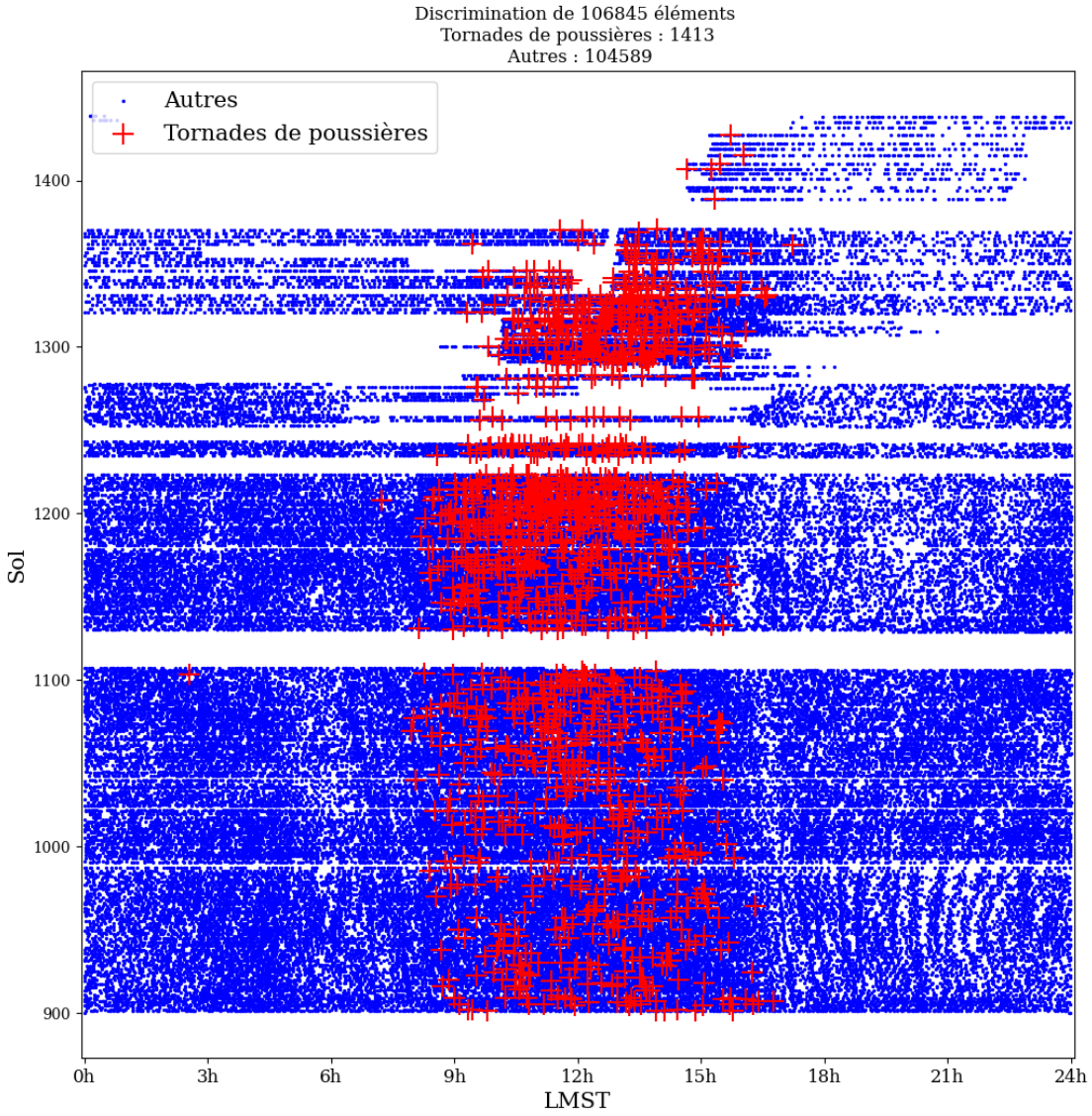

FIGURE 4.39 – Tornades de poussière associées à des seuils de classification supérieurs à 70 %.

journée martienne, alors qu'aucune condition temporelle n'a été intégrée au cours du processus de discrimination. Cette propriété est d'autant plus remarquable au vu de la très grande variété de perturbation ayant été analysée par notre méthode, aussi bien le jour que la nuit martienne (figure 4.37). Finalement, nous constatons dans la figure 4.40 une forte similarité entre les formes d'ondes sismiques des perturbations ayant été classées en tant que tornades de poussière et celles associées au catalogue **TdP** nous permettant une fois de plus de confirmer de l'efficacité de cette discrimination.

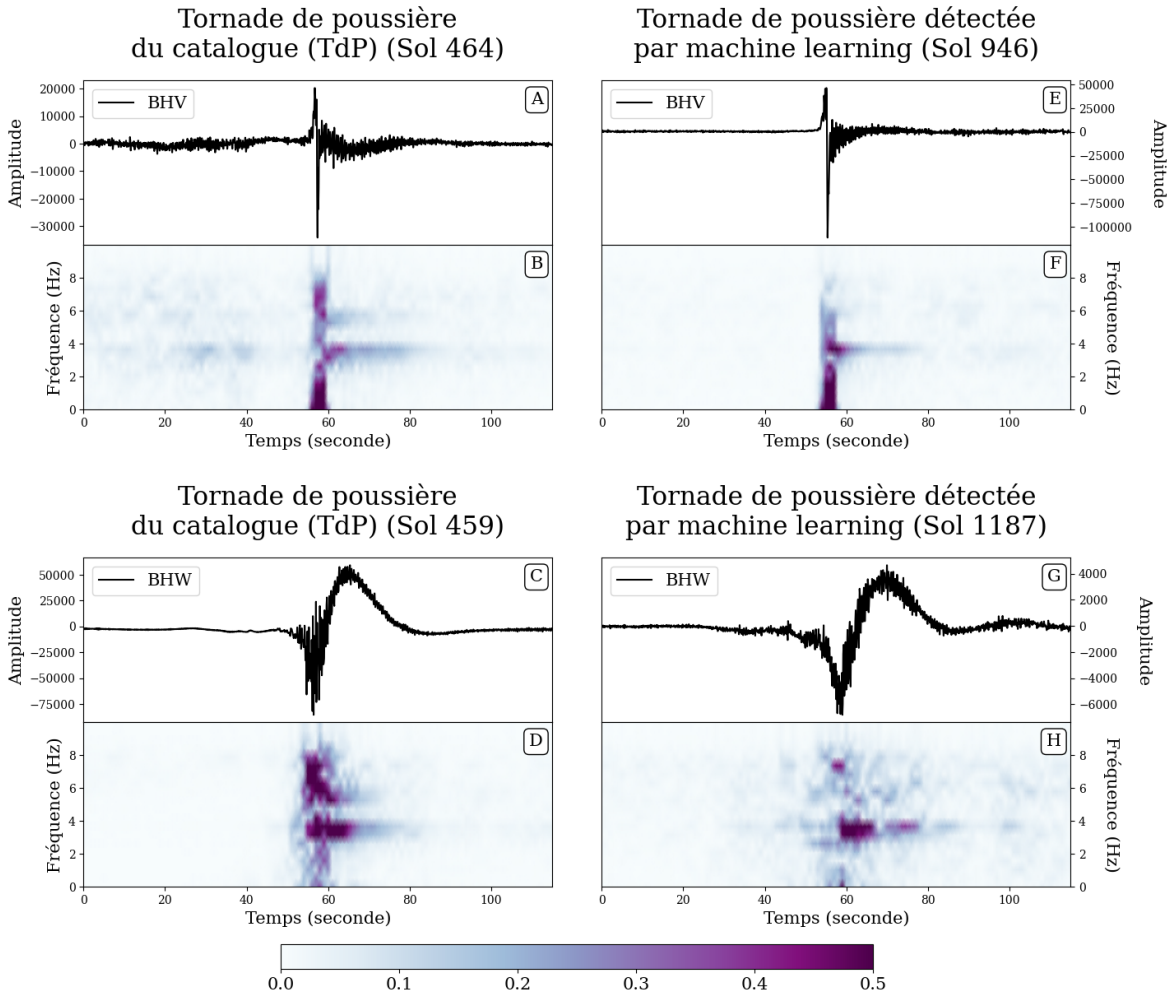

FIGURE 4.40 – Similarités entre les ondes sismiques/spectrogrammes des perturbations ayant été classées en tant que tornades de poussière (E, F, G et H) et celles de notre base de données **TdP** (A, B, C et D).

4.6.5.4 Distribution statistique saisonnière des tornades de poussière détectées

Nous choisissons finalement de nous intéresser à la distribution statistique des tornades de poussière détectées par la discrimination via *machine learning*. la figure 4.41(A) compare la position temporelle au cours de la mission des tornades détectées par Spiga et coll. (2021) (en noir) avec celles que nous avons localisées après le sol 900 (en rouge). On constate une continuité évidente entre la distribution des tornades de poussière provenant du catalogue de la mission et les 2256 tornades que nous avons localisé. Cette continuité se traduit par une tendance similaire de la localisation des

tornades de poussière au cours de la journée, variant sensiblement, en fonction des saisons. En pratique, on observe une variation sinusoïdale du temps moyen LMST des tornades détectées au cours de la journée (dont la période est égale à une année martienne, c'est à dire 668 sols environ). Les tornades que nous avons détectées s'inscrivent dans la continuité de cette tendance sinusoïdale, indiquant donc une bonne cohérence statistique avec celles du catalogue. On remarque également des tendances saisonnières en terme de nombre de tornades de poussière référencées au cours de la mission, comme attesté par la différence de densité des points noirs (A), déjà évoquée par les travaux de Chatain et coll. (2021); Spiga et coll. (2021) et Onodera et coll. (2023). Cette répartition non uniforme des tornades de poussière au cours de la mission est également visible par une inspection de la distribution des points rouges, dont la densité semble augmenter après le sol 1200 environ.

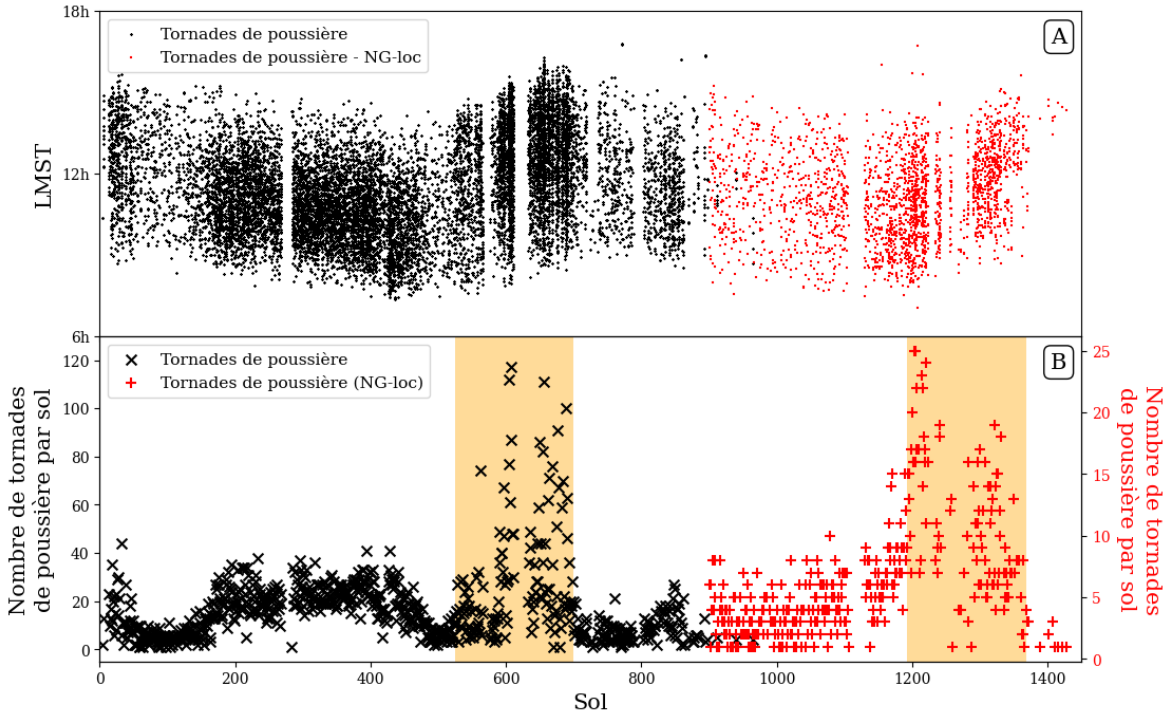

FIGURE 4.41 – Comparaison de la distribution temporelle des tornades de poussière localisées par le catalogue Spiga et coll. (2021) et celles détectées par *machine learning* (après le sol 900), en noir et rouge, respectivement. (A) : Localisation temporelle des tornades au cours de la mission. (B) : Nombre de tornades de poussière journaliers.

La seconde partie de la figure 4.41(B) compare le nombre de tornades de poussière détectées chaque jour par le catalogue de Spiga et coll. (2021) avec celles localisées

par notre approche de discrimination, en conservant le même code couleur (à noter les différentes échelles d'amplitudes). Une comparaison entre ces deux ensembles de points aboutit à des moyennes de 6,3 tornades/jour pour notre catalogue et 17,1 pour celles détectées via le capteur de pression. Cette différence s'explique notamment par le fait qu'un nombre important de tornades de poussière n'entraîne pas d'altérations significatives du signal sismique (voir les figures 4.29 et 4.30). On distingue lors de la zone orangée entre les sols 525 et 700 (B), correspondant à la tempête de poussière ayant eu lieu au milieu de la mission InSight, une nette augmentation du nombre de tornades de poussière détectées. Une année martienne plus tard, lors de la seconde zone orangée entre les sols 1193 et 1368, on observe une augmentation similaire du nombre de tornades détectées par notre approche, pouvant être causée par la seconde tempête de poussière saisonnière au cours de cette période.

Conclusion et perspectives

Au cours de ce travail, nous avons introduit la notion de gaussianité d'un signal, aboutissant au développement de la méthode NG-loc, une approche statistique innovante permettant de localiser les éléments ne respectant pas la distribution normale générale d'une série de données. De la même manière que les tests d'adéquations classiques (Pearson, 1900; Kolmogorov, 1933; Shapiro et Wilk, 1965, 1968; Plackett, 1983; Drezner et coll., 2010), NG-loc se révèle capable de conclure quant à la normalité d'une série de données, mais apporte également une information supplémentaire unique, qui est la discrimination individuelle de chaque point du signal s'écartant de la distribution Gaussienne du signal de fond. Cette opération de discrimination repose sur une méthode d'optimisation, permettant alors de s'affranchir du choix d'un seuil arbitraire, comme souvent utilisé dans d'autres méthodes de détections (Taylor, 2006; Hamamoto et coll., 2018; Lu et Ghorbani, 2008). L'efficacité de NG-loc, combinée à la simplicité de son utilisation et sa vitesse d'exécution ont alors naturellement ouvert la voie à un large champ d'applications.

La méthode NG-loc trouve une application évidente dans l'analyse du signal sismique, particulièrement adapté, de par sa distribution normale, régulièrement altérée par de multiples perturbations (séismes, influence anthropique, *glitchs*,...). Un effort particulier fut apporté à l'estimation de la qualité du signal sismique continu, permettant de localiser des phénomènes de dégradations dans le signal, altérant des composantes/gammes de fréquences spécifiques. Par ailleurs, NG-loc ne se limite toutefois pas à un simple outil de contrôle de qualité des données sismiques et se révèle égale-

ment sensible à toute arrivée d'onde impulsive dans le signal, mettant alors en avant sa capacité à détecter des tremblements de Terre mais également à estimer la longueur de la coda associée.

L'analyse complète du signal sismique martien enregistré au cours de la mission InSight (2018-2022) permet d'extraire d'intéressants résultats quant à l'évaluation de la qualité des données acquises au cours de la mission. L'analyse de ce signal, mit notamment en évidence un certain nombre de tendances particulières concernant la répartition temporelle des *glitchs* détectés au cours de la mission, mais également d'établir une corrélation directe entre certains d'entre eux survenant à des seuils de températures spécifiques. La richesse des informations apportées par une telle analyse fut également exploitée lors de l'étude détaillée des données sismiques autour de quelques dates clés de la mission (activation du chauffage, conjonction, et enfouissement du câble), permettant de mettre en évidence l'influence de ces derniers sur la qualité du signal enregistré.

L'analyse complète par NG-loc des données sismiques enregistrées au cours de la mission a été l'occasion de porter une attention toute particulière aux perturbations associées à des tornades de poussières. Une méthode innovante proposée par Hourcade et coll. (2023) fut alors exploitée, permettant de discriminer les tornades, par l'analyse de leurs spectrogrammes, relatant de la signature sismique particulière de ces événements (Lorenz et coll., 2021). Cette détection directement depuis le signal sismique, trouve particulièrement son intérêt lors de la fin de la mission, marquée par la mise en veille répétée du capteur de pression, utilisé jusqu'à présent pour détecter ces tornades. Un total de 2256 nouvelles tornades de poussière ont finalement été détectées lors de la fin de la mission, venant alors s'ajouter au 12 000 déjà répertoriées dans le catalogue de Spiga et coll. (2021).

Les résultats encourageants obtenus dans le cadre de l'application de NG-loc aux signaux sismiques terrestres et martiens ouvrent naturellement la voie à plusieurs perspectives :

- ***Monitoring* continu des stations sismologiques terrestres.** Une application en temps réel de NG-loc sur les données sismiques pourrait présenter un fort

intérêt, permettant de détecter les éventuelles dégradations des signaux sismiques enregistrés par les stations terrestres, et de procéder à une rapide réinstallation de la station lorsque celle-ci est nécessaire.

- **Caractérisation des différentes propriétés statistiques des signaux sismiques au delà du massif Armoricain.** Une comparaison plus large des propriétés statistiques des signaux sismiques enregistrés par les stations pourrait également être envisagée sur la France entière, voire aux autres pays.
- **Clustering des *glitchs* associés à certains seuils de températures.** Nous avons démontré lors de l'étude du signal sismique martien une corrélation évidente entre certains groupes de *glitchs* survenant à des seuils spécifiques de température. Par conséquent, il serait alors intéressant d'effectuer une classification des *glitchs*, basée sur leurs formes d'ondes combinée à la température enregistrée par SEIS au même moment. Par ailleurs, à ce critère pourrait également s'ajouter celui du gradient de la température enregistrée, indiquant si SEIS est en phase de chauffage ou de refroidissement.
- **Discrimination des tornades de poussière sur l'intégralité de la mission.** Une application généralisée de notre approche automatique de discrimination des tornades de poussière à l'intégralité des données sismiques de la mission pourrait permettre de 1/ confirmer la présence de tornades référencées par le catalogue de Spiga et coll. (2021) mais également 2/ d'en localiser de nouvelles (le capteur de pression étant également régulièrement éteint avant le sol 900).
- **Discrimination automatique des *glitchs*.** Finalement, ces résultats encourageant utilisant le *machine learning* ouvrent également la voie à une discrimination des *glitchs*, tirant profit de leur signature fréquentielle particulière (précurseur haute fréquence suivi d'une altération longue période). La figure A présente une matrice de confusion obtenue lors de l'étape de validation d'un entraînement visant à discriminer les *glitchs*. Cet entraînement fut effectué avec deux bases de données comportant 1000 perturbations chacune (labellisées *Glitchs* et Autres), obtenues via une sélection manuelle des altérations détectées par NG-loc. Bien que les deux bases de données soient ici de taille relativement réduites, un résultat

encourageant de 94,5% d'éléments bien classifiés est toutefois obtenu.

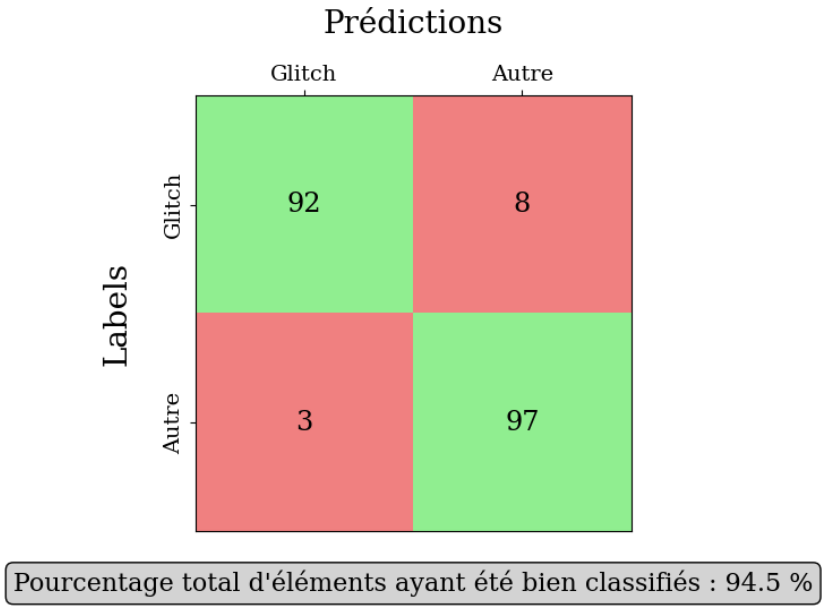

FIGURE A – Matrice de confusion : Discrimination des *glitchs*.

Bien que de nombreux résultats prometteurs de l'application de NG-loc au signal sismique furent présentés dans ce manuscrit, son utilisation dépasse toutefois le simple contexte de la sismologie et peut trouver son intérêt dans l'analyse de n'importe quelle série de donnée suivant une distribution à priori normale. En effet, l'identification d'éléments ne respectant pas la gaussianité générale d'une série de donnée peut trouver une application intéressante dans l'analyse de signaux acoustiques ou même des fluctuations du marché financier. Par ailleurs, NG-loc peut également être utilisée en tant qu'outil efficace de détection de résultats frauduleux à un examen, venant alors altérer la distribution Gaussienne des notes attribuées. Finalement, une application intéressante de NG-loc peut également être trouvée dans le domaine médical, permettant par exemple la détection d'échantillons venant perturber la distribution normale des séries de données temporelles obtenues lors d'électrocardiogrammes.

Annexes

Annexe A

Démonstrations mathématiques

Nous proposons ici les démonstrations mathématiques du Théorème Limite Central 1.2.1, de Glivenko-Cantelli 1.3.1, ainsi que du second théorème de Dini 1.4.1, présentés dans le chapitre 1. Il est toutefois important de noter que ces démonstrations utilisent cependant des outils, propriétés, définitions et théorèmes sortant du cadre mathématique de la simple introduction à la théorie des probabilités présentée dans ce manuscrit.

A.1 Théorème limite central

La démonstration du théorème limite central utilise notamment le théorème de Lévy, les fonctions génératrices, ou encore la notion de développement limité. On pourra trouver des définitions rigoureuses de ces derniers dans Queffélec et Zuily (2002) pour le théorème de Lévy, Escoffier (2020) pour les fonctions génératrices et Rombaldi et Darracq (2020) pour les développements limités. Rappelons ci-dessous l'énoncé du théorème limite central :

Théorème A.1.1 ▶ Théorème Limite Central

Soit $(X_i)_{1\leq i\leq n}$ une suite de variables indépendantes, identiquement distribuées admettant un moment d'ordre 2. On pose :

$$\mu = \mathbb{E}(X_1)\ ,\ \sigma^2 = \mathbb{V}(X_1)\ ,\ S_n = \sum_{i=1}^{n} X_i \ \text{ et } \ Y_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}\ .$$

Alors, $(Y_n)_{1\leq i\leq n}$ converge en loi vers une variable aléatoire de loi $\mathcal{N}(0,1)$.

Démonstration:

L'idée de la démonstration repose sur le théorème de Lévy, indiquant que la convergence en loi d'une suite de v.a.r Y_n vers une v.a.r Y est équivalente à la convergence simple de leurs fonctions génératrices respectives. En posant φ_{Y_n} la fonction génératrice de Y_n (telle qu'énoncée dans le théorème) et φ_Y la fonction génératrice de Y (avec $Y \sim \mathcal{N}(0,1)$), il suffit alors, pour démontrer le théorème limite central, de prouver la convergence de φ_{Y_n} vers φ_Y. De plus, on sait que la fonction génératrice d'une v.a.r Y suivant une loi normale centrée réduite est donnée par

$$\varphi_Y(t) = e^{-\frac{t^2}{2}}, \quad \forall t \in \mathbb{R}.$$

D'autre part, quitte à normaliser les X_i par $\dfrac{X_i - \mu}{\sigma}$, on peut alors supposer que $\mu = 0$ et $\sigma = 1$, ce qui implique alors $Y_n = \dfrac{Sn}{\sqrt{n}}$. Par définition de la fonction génératrice, on a pour tout réel t,

$$\begin{aligned}
\varphi_{Y_n}(t) &= \mathbb{E}(e^{itY_n}) \\
&= \mathbb{E}(e^{i\frac{t}{\sqrt{n}}S_n}) \\
&= \mathbb{E}(e^{i\frac{t}{\sqrt{n}}X_1} \ \ldots\ e^{i\frac{t}{\sqrt{n}}X_n}) \\
&= \mathbb{E}(e^{i\frac{t}{\sqrt{n}}X_1}) \ \ldots\ \mathbb{E}(e^{i\frac{t}{\sqrt{n}}X_n}) \text{ car les } X_i \text{ sont indépendants} \\
&= \left(\mathbb{E}(e^{i\frac{t}{\sqrt{n}}X_1})\right)^n \text{ car les } X_i \text{ sont identiquement distribués} \\
&= \left(\varphi_{X_1}(\frac{t}{\sqrt{n}})\right)^n . \qquad (A.1)
\end{aligned}$$

Par ailleurs, le développement limité de φ_{X_1} au voisinage de 0 est :

$$\varphi_{X_1}(t) = \varphi_{X_1}(0) + t\varphi'_{X_1}(0) + \frac{t^2}{2}\varphi''_{X_1}(0) + o(t^2) \ .$$

Or, on remarque que $\varphi_{X_1}(0) = 1$, et, d'après les propriétés des fonctions génératrices, on a $\varphi'_{X_1}(0) = i\mathbb{E}(X_1) = 0$ et $\varphi''_{X_1}(0) = -\mathbb{E}(X_1^2) = -1$. Par conséquent,

$$\varphi_{X_1}(t) = 1 - \frac{t^2}{2} + o(t^2) \ .$$

Connaissant désormais l'expression de φ_{X1}, A.1 nous permet donc d'obtenir

$$\varphi_{Y_n}(t) = \left(1 - \frac{t^2}{2n} + o\Big(\frac{t^2}{n}\Big)\right)^n \ .$$

Si n est suffisamment grand, on remarque que l'expression $1 - \dfrac{t^2}{2n} + o\Big(\dfrac{t^2}{n}\Big) \in B(1, \frac{1}{2}) \subset \mathbb{C}$. Par conséquent, d'après la détermination principale du logarithme sur $\mathbb{C} \setminus \mathbb{R}^{-*}$,

$$\begin{aligned}\varphi_{Y_n}(t) &= e^{n \ln\left(1 - \frac{t^2}{2n} + o(\frac{t^2}{n})\right)} \\ &= e^{-\frac{t^2}{2} + o(1)} \xrightarrow[n \to +\infty]{} e^{-\frac{t^2}{2}} \ .\end{aligned}$$

Ce qui prouve enfin, par application du théorème de Lévy, le théorème limite central. □

A.2 Théorème de Glivenko-Cantelli

La démonstration du théorème de Glivenko-Cantelli utilise quant à elle la loi forte des grands nombres (voir Ouvrard, 2004), ainsi que la notion de partition d'un ensemble (voir Halmos (1960) pour une référence classique sur la théorie des ensembles). Celle-ci est issue du livre Van der Vaart (2000), dont les arguments ont ici été clarifiés afin de faciliter sa compréhension. On rappelle donc l'énoncé du théorème :

Théorème A.2.1 ▶ Glivenko-Cantelli

Soient $(X_1, ..., X_n)$, un échantillon de variables aléatoires réelles indépendantes et identiquement distribuées ayant la même fonction de répartition F. Alors, F_n converge uniformément, presque sûrement, vers F, *i.e.*

$$\|F_n - F\|_\infty = \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n\to+\infty]{} 0 \ p.s.$$

Démonstration:

L'idée consiste ici tout d'abord à construire la fonction de répartition empirique comme une somme de fonction indicatrice. À chaque X_i, on associe, une certaine v.a.r $Y_i(x) = \mathbb{1}_{]-\infty,x[}(x)$. Comme les X_i sont indépendants et identiquement distribués, il en est donc de même pour les Y_i et on peut donc appliquer la loi forte des grands nombres. Par conséquent,

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = F_n \xrightarrow[n\to+\infty]{p.s} \mathbb{E}(Y_1) \ .$$

En remarquant alors que $\mathbb{E}(Y_1(x)) = \mathbb{P}(X_1 \leq x) = F(x)$, on a donc la convergence presque sûre de Fn vers F. Pour tout t réel, on a donc $F_n(t) \xrightarrow[n\to+\infty]{p.s} F(t)$. De plus, en notant $F(t-) = \lim_{\substack{x\to t \\ x<t}} F(x)$, on a également $F_n(t-) \xrightarrow[n\to+\infty]{p.s} F(t-)$.

Soit désormais $\varepsilon > 0$ fixé, il existe alors une partition $-\infty = t_0 < t_1 < ... < t_k = \infty$ telle que $F(t_i-) - F(t_{i-1}) < \varepsilon$ pour tout i. Si on considère maintenant t, tel que

$t_{i-1} \leq t < t_i$, F et F_n étant croissante, on a alors,

$$F_n(t) - F(t) \leq F_n(t_i-) - F(t_i-) + \varepsilon$$
$$\text{et} \quad F_n(t) - F(t) \geq F_n(t_{i-1}) - F(t_{i-1}) + \varepsilon.$$

En faisant maintenant tendre k vers l'infini, on en déduit alors, pour tout réel t, la convergence uniforme de $F_n(t)$ et $F_n(t-)$ dans l'ensemble fini $(t_1, ..., t_{k-1})$. Par conséquent, $\lim_{n\to\infty} \|F_n - F\| \leq \varepsilon$ pour tout $\varepsilon > 0$. On a donc enfin,

$$\|F_n - F\|_\infty \xrightarrow[n\to+\infty]{} 0 \ p.s. \ ,$$

ce qui prouve le théorème. □

A.3 Second théorème de Dini

La démonstration du second théorème de Dini utilise notamment le théorème de Heine (Heine, 1872). La preuve est issue de Francinou et coll. (2013).

Théorème A.3.1 ▶ Second théorème de Dini

Soient $a, b \in \mathbb{R}^2$ tels que $a \leq b$. Si Q_n est une suite de fonctions croissantes de $[a, b]$ dans $\mathbb{R}$ qui converge simplement vers une fonction Q continue, alors Q_n converge uniformément vers Q.

Démonstration:

La fonction Q étant continue sur l'intervalle compact $[a, b]$, le théorème de Heine nous assure de la continuité uniforme de Q. Par conséquent, pour tout $\varepsilon > 0$, il existe alors un certain $h > 0$ tel que :

$$\forall x, y \in [a, b] \ , \ |x - y| \leq h \Rightarrow |Q(x) - Q(y)| \leq \varepsilon.$$

Considérons désormais $x_0, x_1, ..., x_p$ une subdivision de l'intervalle $[a, b]$ telle que $a = x_0 < x_1 < ... < x_p = b$, de pas plus petit que h. Comme Q_n converge vers Q, ceci est donc également vrai pour les points de la subdivision x_i, c'est-à-dire

$$\lim_{n\to\infty} Q_n(x_i) = Q(x_i) \text{ pour tout } 1 \leq i \leq p.$$

Par conséquent, lorsque n est suffisamment grand, on a alors $|Q_n(x_i) - Q(x_i)| \leq \varepsilon$. Soit désormais $i \in \{1, ..., p\}$ fixé et $x \in [x_i, x_{i+1}]$. Ainsi,

$$\begin{aligned} |Q(x) - Q_n(x)| &= |Q(x) - Q(x_i) + Q(x_i) - Q_n(x_i) + Q_n(x_i) - Q_n(x)| \\ &\leq |Q(x) - Q(x_i)| + |Q(x_i) - Q_n(x_i)| + |Q_n(x_i) - Q_n(x)| \\ &\leq \varepsilon + \varepsilon + |Q_n(x_i) - Q_n(x)|. \end{aligned}$$

Les fonctions Q_n étant croissantes, on en déduit alors que $Q_n(x_i) \leq Q_n(x)$, et donc

$$\begin{aligned} |Q_n(x_i) - Q_n(x)| &= Q_n(x) - Q_n(x_i) \\ &\leq |Q_n(x_{i+1}) - Q_n(x_i)| \\ &\leq |Q_n(x_{i+1}) - Q(x_{i+1})| + |Q(x_{i+1}) - Q(x_i)| + |Q(x_i) - Q_n(x_i)| \leq 3\varepsilon. \end{aligned}$$

Finalement, $|Q(x) - Q_n(x)| \leq 5\varepsilon$, nous permettant de conclure de la convergence uniforme. □

Annexe B

Nombres pseudo-aléatoires et fonction quantile

Dans le cadre du travail présenté dans ce manuscrit, de nombreux signaux synthétiques ont été générés à l'aide de simulations numériques aléatoires ayant recours à la bibliothèque Numpy (Harris et coll., 2020) du langage Python (Van Rossum et Drake, 2009) (voir les chapitres 1 et 2). L'utilisation de ces outils numériques permet de simuler le résultat d'une variable aléatoire X, suivant une certaine loi de probabilité (loi normale, loi uniforme,...). Par itérations, il est alors possible de générer un signal aléatoire de longueur n, suivant une loi donnée, comme présenté par exemple dans la figure 1.5.

Il est cependant important de noter que toute simulation numérique repose sur la génération de « nombres pseudo-aléatoires », résultant d'un algorithme déterministe (voir Lagarias (1993); Knuth (1997); Wichmann et Hill (2006) pour plus d'informations). Bien que ces algorithmes permettent d'obtenir une bonne approximation du hasard, il est essentiel de comprendre que ceux-ci ne correspondent pas exactement à la définition mathématique rigoureuse d'une simulation aléatoire. Parmi les algorithmes fréquemment utilisés, on pourra par exemple citer le « Twister de Mersenne » (Matsumoto et Nishimura, 1998), réputé pour la fiabilité de son résultat aléatoire. En effet, celui-ci permet de simuler une variable aléatoire discrète uniforme sur un

ensemble total de $P = 2^{19\,937} - 1$ éléments. Bien que notre utilisation de l'aléatoire numérique se concentre sur la simulation d'un signal synthétique de loi donnée, la génération de nombres aléatoires est également cruciale dans de nombreux autres domaines, comme par exemple en cryptographie, ou encore en mathématiques avec les méthodes de Monte-Carlo (Kroese et Rubinstein, 2012).

Une approche classique de simulation de variable aléatoire, nommée « Méthode de la transformée inverse » (Devroye, 2006), s'appuie justement sur la fonction quantile Q, définie dans le chapitre 1 (1.3). Celle-ci repose sur la proposition suivante :

Proposition B.0.1

Soit $U \sim \mathcal{U}([0,1])$. Alors, pour toute fonction de répartition F (et sa fonction quantile Q associée), la variable $Q(U)$ admet F pour fonction de répartition.

Démonstration:

La fonction F étant croissante et continue à droite, alors, pour tout $u \in]0,1[$ et $x \in \mathbb{R}$, on a

$$Q(u) \leq x \Leftrightarrow u \leq F(x). \text{ Par conséquent,}$$

$$\mathbb{P}\big(Q(U) \leq x\big) = \mathbb{P}\big(U \leq F(x)\big) = F(x).$$

□

Cette proposition rend alors possible de simuler n'importe quelle variable aléatoire à partir de U suivant une loi uniforme sur $[0,1]$. L'obtention de la v.a.r U est toutefois triviale. En effet, les algorithmes classiques permettent de simuler des tirages uniformes discrets d'entiers $(x_n)_{n\geq 1}$, compris entre 0 et un certain $N_{max} \in \mathbb{N}$ (très grand). Par conséquent, la variable $U_n = \dfrac{x_n}{N_{max}}$ est alors une très bonne approximation de v.a.r suivant une loi uniforme sur $[0,1]$.

Par exemple, si on veut simuler $X \sim \mathcal{E}(\lambda)$ $(\lambda > 0)$. Sa densité de probabilité est donnée par (voir section 1.2) :

$$f(t) = \begin{cases} \lambda e^{-\lambda} t \ \text{ si } \ t > 0 \\ 0 \ \text{ si } \ t \leq 0. \end{cases}$$

Par conséquent, sa fonction de répartition F vaut :

$$F(x) = \begin{cases} \displaystyle\int_0^x \lambda e^{-\lambda t} dt = -[e^{-\lambda t}]_0^x = 1 - e^{-\lambda x} \ \text{ si } \ x > 0 \\ 0 \ \text{ si } \ x \leq 0. \end{cases}$$

Par définition de la fonction quantile, on a :

$$Q(u) = \inf\{y \in \mathbb{R} \ ; \ F(y) \geq u\}, \quad \forall u \in [0,1]. \tag{B.1}$$

Si $U \sim \mathcal{U}([0,1])$, on peut supposer que $U \sim \mathcal{U}(]0,1[)$ car $\mathbb{P}(U = 0) = \mathbb{P}(U = 1) = 0$. En reprenant la définition de la fonction quantile dans B.1, on remarque alors que

$$\begin{aligned} F(y) \geq u \ \text{ implique que,} \\ 1 - e^{-\lambda y} &\geq u \\ e^{-\lambda y} &\leq 1 - u \\ -\lambda y &\leq \ln(1-u) \ \text{ car } u < 1 \\ y &\geq \frac{-1}{\lambda} \ln(1-u) \end{aligned}$$

$$\text{Par conséquent, } Q(u) = \frac{-1}{\lambda} \ln(1-u).$$

Finalement, si $U \sim \mathcal{U}([0,1])$, alors la variable $Y = \dfrac{-1}{\lambda} \ln(1-u)$ suit une loi exponentielle de paramètre λ.

Annexe C

Figures supplémentaires

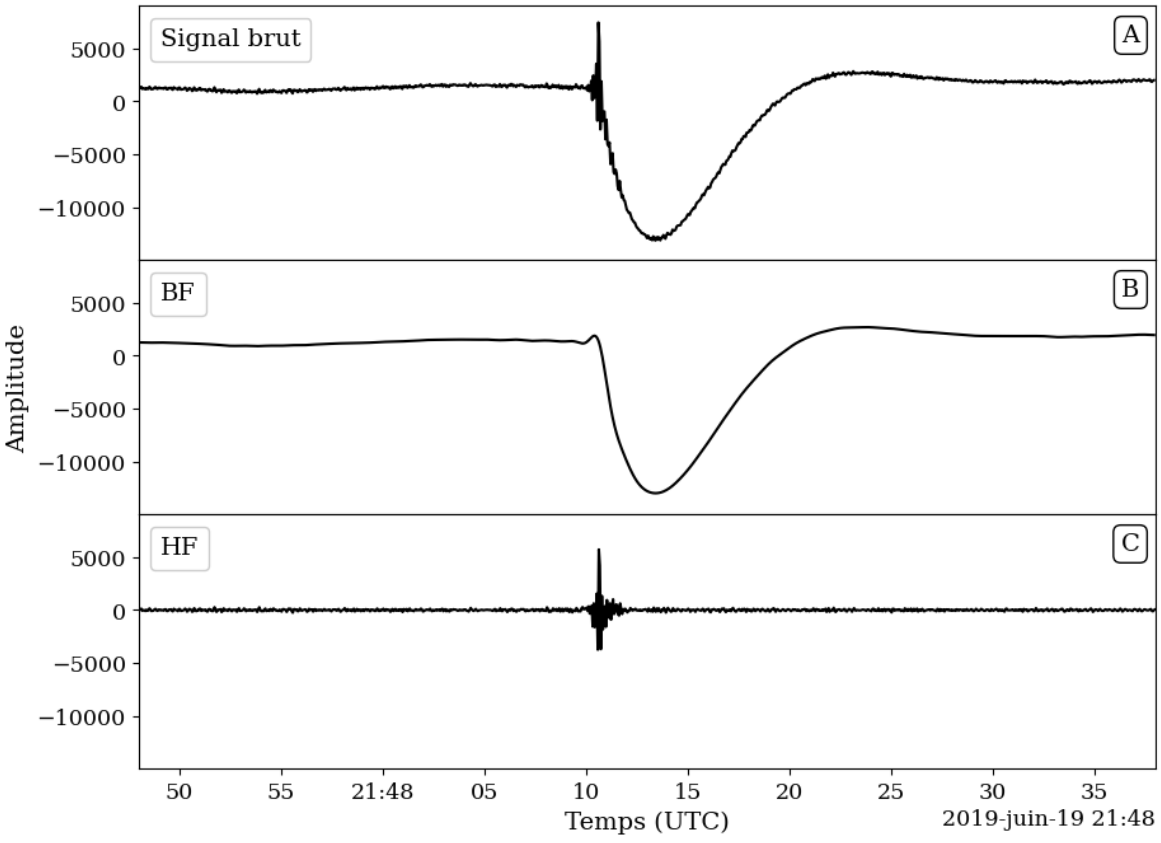

FIGURE C.1 – Forme d'onde d'un *glitch* type, enregistré lors de la mission InSight, pour plusieurs gammes de fréquences. (A) : Signal sismique brut. (B) : Signal basse fréquence BF. (C) : Signal haute fréquence HF.

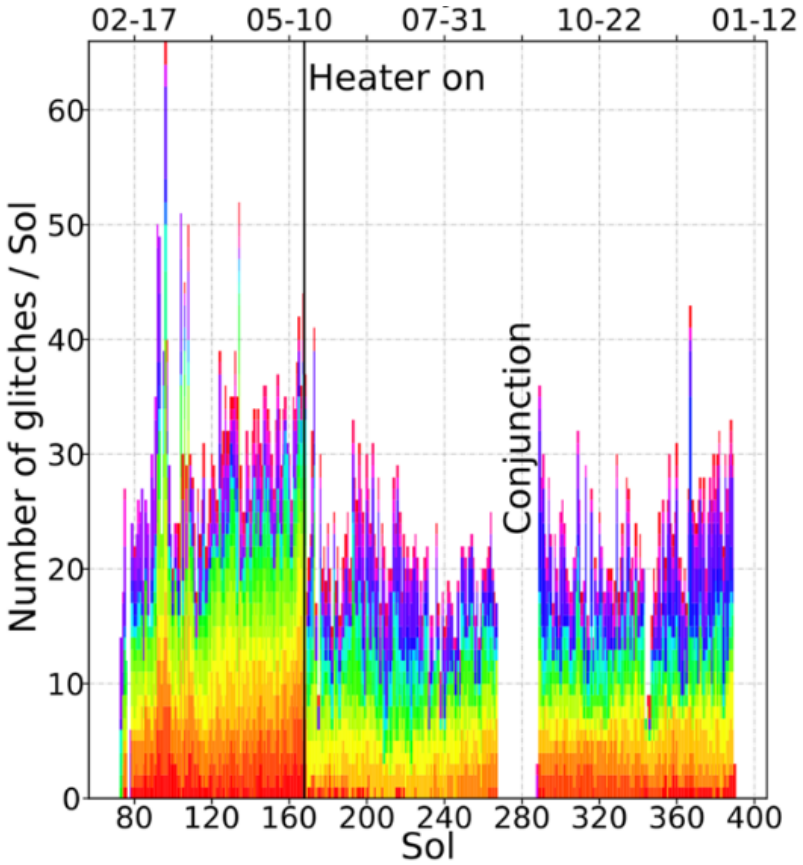

FIGURE C.2 – Nombre de *glitchs* détectés lors des 400 premiers sols de la mission InSight. Une forte diminution est observée suite à la mise en route du chauffage de SEIS au sol 168. Figure extraite de Scholz et coll. (2020).

Table des figures

Bibliographie

Aggarwal, K., Mukhopadhyay, S., et Tangirala, A. K. Statistical characterization and time-series modeling of seismic noise. *arXiv preprint arXiv :2009.01549*, 2020.

Agnew, D., Berger, J., Buland, R., Farrell, W., et Gilbert, F. International deployment of accelerometers : a network for very long period seismology. *EOS, Transactions American Geophysical Union*, 57(4) :180–188, 1976.

Allen, R. Automatic phase pickers : Their present use and future prospects. *Bulletin of the Seismological Society of America*, 72(6B) :S225–S242, 1982.

Allen, R. V. Automatic earthquake recognition and timing from single traces. *Bull. Seismol. Soc. Am.*, 68(5) :1521–1532, 1978. ISSN 0037-1106.

Allison, M. et McEwen, M. A post-pathfinder evaluation of areocentric solar coordinates with improved timing recipes for mars seasonal/diurnal climate studies. *Planetary and Space Science*, 48(2-3) :215–235, 2000.

Alu, K. I. *Solving the Differential Equation for the Probit Function Using a Variant of the Carleman Embedding Technique*. PhD thesis, East Tennessee State University, 2011.

Anderson, D. L., Duennebier, F. K., Latham, G. V., Toksöz, M. F., Kovach, R. L., Knight, T. C., Lazarewicz, A. R., Miller, W. F., Nakamura, Y., et Sutton, G. The viking seismic experiment. *Science*, 194(4271) :1318–1321, 1976.

Anderson, D. L., Miller, W., Latham, G., Nakamura, Y., Toksöz, M., Dainty, A., Duennebier, F., Lazarewicz, A. R., Kovach, R., et Knight, T. Seismology on mars. *Journal of Geophysical Research*, 82(28) :4524–4546, 1977.

Asgedom, E. G., Gelius, L. J., et Tygel, M. Seismic coherency measures in case of interfering events : A focus on the most promising candidates of higher-resolution algorithms. *IEEE Signal Processing Magazine*, 29(3) :47–56, 2012.

Banerdt, W. B., Smrekar, S. E., Banfield, D., Giardini, D., Golombek, M., Johnson, C. L., Lognonné, P., Spiga, A., Spohn, T., Perrin, C., et al. Initial results from the insight mission on mars. *Nature Geoscience*, 13(3) :183–189, 2020.

Banfield, D., Rodriguez-Manfredi, J., Russell, C., Rowe, K., Leneman, D., Lai, H., Cruce, P., Means, J., Johnson, C., Mittelholz, A., et al. Insight auxiliary payload sensor suite (apss). *Space Science Reviews*, 215 :1–33, 2019.

Banfield, D., Spiga, A., Newman, C., Forget, F., Lemmon, M., Lorenz, R., Murdoch, N., Viudez-Moreiras, D., Pla-Garcia, J., Garcia, R. F., et al. The atmosphere of mars as observed by insight. *Nature Geoscience*, 13(3) :190–198, 2020.

Barkaoui, S., Lognonné, P., Kawamura, T., Stutzmann, É., Seydoux, L., de Hoop, M. V., Balestriero, R., Scholz, J.-R., Sainton, G., Plasman, M., et al. Anatomy of continuous mars seis and pressure data from unsupervised learning. *Bulletin of the Seismological Society of America*, 111(6) :2964–2981, 2021.

Barth, C. A. The atmosphere of mars. *Annual Review of Earth and Planetary Sciences*, 2(1) :333–367, 1974.

Barth, K.-H. The politics of seismology : nuclear testing, arms control, and the transformation of a discipline. *Social Studies of Science*, 33(5) :743–781, 2003.

Bensen, G., Ritzwoller, M., Barmin, M., Levshin, A. L., Lin, F., Moschetti, M., Shapiro, N., et Yang, Y. Processing seismic ambient noise data to obtain reliable broad-band surface wave dispersion measurements. *Geophysical journal international*, 169(3) : 1239–1260, 2007.

Berger, J. et Sax, R. L. Seismic detectors : the state of the art. *Systems, Science and Software. Technical Report, No. SSS*, 1980.

Beucler, É., Mocquet, A., Schimmel, M., Chevrot, S., Quillard, O., Vergne, J., et Sylvander, M. Observation of deep water microseisms in the north atlantic ocean using tide modulations. *Geophysical Research Letters*, 42(2) :316–322, 2015.

Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., et Wassermann, J. Obspy : A python toolbox for seismology. *Seismological Research Letters*, 81(3) : 530–533, 2010.

Blair, J., Edwards, C., et Johnson, J. H. Rational chebyshev approximations for the inverse of the error function. *Mathematics of Computation*, 30(136) :827–830, 1976.

Bliss, C. I. The method of probits. *Science*, 79(2037) :38–39, 1934.

Bormann, P. et Wielandt, E. Seismic signals and noise. In *New manual of seismological observatory practice 2 (NMSOP2)*, pages 1–62. Deutsches GeoForschungsZentrum GFZ, 2013.

Bormann, P., Engdahl, B., et Kind, R. Seismic wave propagation and earth models. In *New manual of seismological observatory practice 2 (NMSOP2)*, pages 1–105. Deutsches GeoForschungsZentrum GFZ, 2012.

Böse, M., Clinton, J. F., Ceylan, S., Euchner, F., van Driel, M., Khan, A., Giardini, D., Lognonne, P., et Banerdt, W. B. A probabilistic framework for single-station location of seismicity on earth and mars. *Physics of the Earth and Planetary Interiors*, 262 : 48–65, 2017.

Carlitz, L. The inverse of the error function. 1963.

Ceylan, S., Clinton, J. F., Giardini, D., Böse, M., Charalambous, C., Van Driel, M., Horleston, A., Kawamura, T., Khan, A., Orhand-Mainsant, G., et al. Companion guide to the marsquake catalog from insight, sols 0–478 : Data content and non-seismic events. *Physics of the Earth and Planetary Interiors*, 310 :106597, 2021.

Ceylan, S., Clinton, J. F., Giardini, D., Stähler, S. C., Horleston, A., Kawamura, T., Böse, M., Charalambous, C., Dahmen, N. L., van Driel, M., et al. The marsquake catalogue from insight, sols 0–1011. *Physics of the Earth and Planetary Interiors*, 333 :106943, 2022.

Ceylan, S., Giardini, D., Clinton, J., Kim, D., Khan, A., Stähler, S. C., Zenhäusern, G., Lognonné, P., et Banerdt, W. B. Mapping the seismicity of mars with insight. *Journal of Geophysical Research : Planets*, page e2023JE007826, 2023.

Charalambous, C., Stott, A. E., Pike, W., McClean, J. B., Warren, T., Spiga, A., Banfield, D., Garcia, R. F., Clinton, J., Stähler, S., et al. A comodulation analysis of atmospheric energy injection into the ground motion at insight, mars. *Journal of Geophysical Research : Planets*, 126(4) :e2020JE006538, 2021.

Chatain, A., Spiga, A., Banfield, D., Forget, F., et Murdoch, N. Seasonal variability of the daytime and nighttime atmospheric turbulence experienced by insight on mars. *Geophysical Research Letters*, 48(22) :e2021GL095453, 2021.

Clinton, J. F., Ceylan, S., van Driel, M., Giardini, D., Stähler, S. C., Böse, M., Charalambous, C., Dahmen, N. L., Horleston, A., Kawamura, T., et al. The marsquake catalogue from insight, sols 0–478. *Physics of the Earth and Planetary Interiors*, 310 :106595, 2021.

Dahmen, N. L., Zenhäusern, G., Clinton, J. F., Giardini, D., Stähler, S. C., Ceylan, S., Charalambous, C., van Driel, M., Hurst, K. J., Kedar, S., et al. Resonances and lander modes observed by insight on mars (1–9 hz). *Bulletin of the Seismological Society of America*, 111(6) :2924–2950, 2021.

Dahmen, N. L., Clinton, J. F., Meier, M.-A., Stähler, S. C., Ceylan, S., Kim, D., Stott, A. E., et Giardini, D. A deep catalogue of marsquakes. *Authorea Preprints*, 2022a.

Dahmen, N. L., Clinton, J. F., Meier, M.-A., Stähler, S. C., Ceylan, S., Kim, D., Stott, A. E., et Giardini, D. Marsquakenet : A more complete marsquake catalog obtained by deep learning techniques. *Journal of Geophysical Research : Planets*, 127(11) : e2022JE007503, 2022b.

De Angelis, S. et Bodin, P. Watching the wind : Seismic data contamination at long periods due to atmospheric pressure-field-induced tilting. *Bulletin of the Seismological Society of America*, 102(3) :1255–1265, 2012.

de Laplace, P. S. *Théorie analytique des probabilités*, volume 7. Courcier, 1820.

Derrick, T. et Thomas, J. Time series analysis : the cross-correlation function. 2004.

Devroye, L. Nonuniform random variate generation. *Handbooks in operations research and management science*, 13 :83–121, 2006.

Díaz, J. On the origin of the signals observed across the seismic spectrum. *Earth-Science Reviews*, 161 :224–232, 2016.

Doody, C., Ringler, A. T., Anthony, R. E., Wilson, D. C., Holland, A. A., Hutt, C. R., et Sandoval, L. D. Effects of thermal variability on broadband seismometers : Controlled experiments, observations, and implications. *Bulletin of the Seismological Society of America*, 108(1) :493–502, 2018.

Drezner, Z., Turel, O., et Zerom, D. A modified kolmogorov–smirnov test for normality. *Communications in Statistics—Simulation and Computation®*, 39(4) :693–704, 2010.

Drilleau, M., Beucler, É., Lognonné, P., Panning, M. P., Knapmeyer-Endrun, B., Banerdt, W. B., Beghein, C., Ceylan, S., van Driel, M., Joshi, R., et al. Mss/1 : Single-station and single-event marsquake inversion. *Earth and Space Science*, 7 (12) :e2020EA001118, 2020.

Drilleau, M., Samuel, H., Garcia, R. F., Rivoldini, A., Perrin, C., Michaut, C., Wieczorek, M., Tauzin, B., Connolly, J. A., Meyer, P., et al. Marsquake locations and 1-d seismic models for mars from insight data. *Journal of Geophysical Research : Planets*, 127(9) :e2021JE007067, 2022.

Dybing, S. N., Ringler, A. T., Wilson, D. C., et Anthony, R. E. Characteristics and spatial variability of wind noise on near-surface broadband seismometers. *Bulletin of the Seismological Society of America*, 109(3) :1082–1098, 2019.

Dziewonski, A. M. et Anderson, D. L. Preliminary reference earth model. *Physics of the earth and planetary interiors*, 25(4) :297–356, 1981.

Ebeling, C. W. Inferring ocean storm characteristics from ambient seismic noise : A historical perspective. In *Advances in geophysics*, volume 53, pages 1–33. Elsevier, 2012.

Eibe, F., Hall, M. A., et Witten, I. H. The weka workbench. online appendix for data mining : practical machine learning tools and techniques. In *Morgan Kaufmann*. Morgan Kaufmann Publishers San Francisco, California, 2016.

Ellehoj, M., Gunnlaugsson, H., Taylor, P., Kahanpää, H., Bean, K., Cantor, B., Gheynani, B., Drube, L., Fisher, D., Harri, A.-M., et al. Convective vortices and dust devils at the phoenix mars mission landing site. *Journal of Geophysical Research : Planets*, 115(E4), 2010.

Emmert-Streib, F. et Dehmer, M. Understanding statistical hypothesis testing : The logic of statistical inference. *Machine Learning and Knowledge Extraction*, 1(3) : 945–962, 2019.

Escoffier, J. Probabilités et statistiques pour le capes externe et l'agrégation interne de mathématiques. *Probabilités et statistiques pour le CAPES externe et l'Agrégation interne de Mathématiques*, pages 1–222, 2020.

Fernando, B., Daubar, I. J., Charalambous, C., Grindrod, P. M., Stott, A., Al Ateqi, A., Atri, D., Ceylan, S., Clinton, J., Hauber, E., et al. A tectonic origin for the largest marsquake observed by insight. *Authorea Preprints*, 2023.

Filliben, J. J. The probability plot correlation coefficient test for normality. *Technometrics*, 17(1) :111–117, 1975.

Finney, D. J. Probit analysis, cambridge university press. *Cambridge, UK*, 1971.

Fisher, R. A. Moments and product moments of sampling distributions. *Proceedings of the London Mathematical Society*, 2(1) :199–238, 1930a.

Fisher, R. A. The moments of the distribution for normal samples of measures of departure from normality. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 130(812) :16–28, 1930b.

Folkner, W. M., Dehant, V., Le Maistre, S., Yseboodt, M., Rivoldini, A., Van Hoolst, T., Asmar, S. W., et Golombek, M. P. The rotation and interior structure experiment on the insight mission to mars. *Space Science Reviews*, 214 :1–16, 2018.

Francinou, S., Gianella, H., et Nicolas, S. *Exercices de Mathématiques (oraux X-ENS) : analyse 2*. Cassini, 2013.

Garcia, R. F., Kenda, B., Kawamura, T., Spiga, A., Murdoch, N., Lognonné, P. H., Widmer-Schnidrig, R., Compaire, N., Orhand-Mainsant, G., Banfield, D., et al. Pressure effects on the seis-insight instrument, improvement of seismic records, and characterization of long period atmospheric waves from ground displacements. *Journal of Geophysical Research : Planets*, 125(7) :e2019JE006278, 2020.

Garcia, R. F., Daubar, I. J., Beucler, É., Posiolova, L. V., Collins, G. S., Lognonné, P., Rolland, L., Xu, Z., Wójcicka, N., Spiga, A., et al. Newly formed craters on mars located using seismic and acoustic wave data from insight. *Nature Geoscience*, 15 (10) :774–780, 2022.

Gaudot, I., Beucler, É., Mocquet, A., Schimmel, M., et Le Feuvre, M. Statistical redundancy of instantaneous phases : theory and application to the seismic ambient wavefield. *Geophysical Journal International*, 204(2) :1159–1163, 2016.

Giardini, D., Lognonné, P., Banerdt, W. B., Pike, W. T., Christensen, U., Ceylan, S., Clinton, J. F., van Driel, M., Stähler, S. C., Böse, M., et al. The seismicity of mars. *Nature Geoscience*, 13(3) :205–212, 2020.

Glivenko, V. Sulla determinazione empirica delle leggi di probabilita. *Gion. Ist. Ital. Attauri.*, 4 :92–99, 1933.

Goins, N. R., Dainty, A., et Toksöz, M. Lunar seismology : The internal structure of the moon. *Journal of Geophysical Research : Solid Earth*, 86(B6) :5061–5074, 1981.

Golombek, M., Kipp, D., Warner, N., Daubar, I. J., Fergason, R., Kirk, R. L., Beyer, R., Huertas, A., Piqueux, S., Putzig, N., et al. Selection of the insight landing site. *Space Science Reviews*, 211 :5–95, 2017.

Groos, J. C. et Ritter, J. R. R. Time domain classification and quantification of seismic noise in an urban environment. *Geophys. J. Int.*, 179(2) :1213–1231, 2009. ISSN 0956-540X. doi : 10.1111/j.1365-246X.2009.04343.x.

Gualtieri, L., Stutzmann, É., Capdeville, Y., Farra, V., Mangeney, A., et Morelli, A. On the shaping factors of the secondary microseismic wavefield. *Journal of Geophysical Research : Solid Earth*, 120(9) :6241–6262, 2015.

Gutenberg, B. *Physics of the Earth's Interior*. Elsevier, 2016.

Halmos, P. R. *Naive set theory*. van Nostrand, 1960.

Hamamoto, A. H., Carvalho, L. F., Sampaio, L. D. H., Abrão, T., et Proença Jr, M. L. Network anomaly detection system using genetic algorithm and fuzzy logic. *Expert Systems with Applications*, 92 :390–402, 2018.

Hanasoge, S. M. et Branicki, M. Interpreting cross-correlations of one-bit filtered seismic noise. *Geophysical Journal International*, 195(3) :1811–1830, 2013.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., et Oliphant, T. E. Array programming with NumPy. *Nature*, 585(7825) : 357–362, 2020. doi : 10.1038/s41586-020-2649-2.

Heine, E. Die elemente der functionenlehre. 1872.

Hess, S., Henry, R., Leovy, C. B., Ryan, J., et Tillman, J. E. Meteorological results from the surface of mars : Viking 1 and 2. *Journal of Geophysical Research*, 82(28) : 4559–4574, 1977.

Hoffman, D. L. et Low, S. A. An application of the probit transformation to tourism survey data. *Journal of Travel Research*, 20(2) :35–38, 1981.

Hourcade, C., Bonnin, M., et Beucler, É. New cnn-based tool to discriminate anthropogenic from natural low magnitude seismic events. *Geophysical Journal International*, 232(3) :2119–2132, 2023.

InSight Marsquake Service. Mars seismic catalogue, insight mission ; v12 2022-10-01, 2022.

InSight Marsquake Service. Mars seismic catalogue, insight mission ; v13 2023-01-01, 2023.

Jackson, B. Vortices and dust devils as observed by the mars environmental dynamics analyzer instruments on board the mars 2020 perseverance rover. *The Planetary Science Journal*, 3(1) :20, 2022.

Jeffreys, H. et Bullen, K. Seismological tables, brit. assoc. *Adv. Sci., London*, 1940.

Jordan, M. I. et Mitchell, T. M. Machine learning : Trends, perspectives, and prospects. *Science*, 349(6245) :255–260, 2015.

Kahanpää, H., Newman, C., Moores, J., Zorzano, M.-P., Martín-Torres, J., Navarro, S., Lepinette, A., Cantor, B., Lemmon, M. T., Valentín-Serrano, P., et al. Convective vortices and dust devils at the msl landing site : Annual variability. *Journal of Geophysical Research : Planets*, 121(8) :1514–1549, 2016.

Kawamura, T., Clinton, J. F., Zenhäusern, G., Ceylan, S., Horleston, A. C., Dahmen, N. L., Duran, C., Kim, D., Plasman, M., Stähler, S. C., et al. S1222a—the largest marsquake detected by insight. *Geophysical Research Letters*, 50(5) :e2022GL101543, 2023.

Kenda, B., Drilleau, M., Garcia, R. F., Kawamura, T., Murdoch, N., Compaire, N., Lognonné, P., Spiga, A., Widmer-Schnidrig, R., Delage, P., et al. Subsurface structure at the insight landing site from compliance measurements by seismic and meteorological experiments. *Journal of Geophysical Research : Planets*, 125(6) :e2020JE006387, 2020.

Khan, A., van Driel, M., Böse, M., Giardini, D., Ceylan, S., Yan, J., Clinton, J., Euchner, F., Lognonné, P., Murdoch, N., et al. Single-station and single-event marsquake location and inversion for structure using synthetic martian waveforms. *Physics of the Earth and Planetary Interiors*, 258 :28–42, 2016.

Kim, D., Davis, P., Lekić, V., Maguire, R., Compaire, N., Schimmel, M., Stutzmann, E., CE Irving, J., Lognonné, P., Scholz, J.-R., et al. Potential pitfalls in the analysis and structural interpretation of seismic data from the mars insight mission. *Bulletin of the Seismological Society of America*, 111(6) :2982–3002, 2021.

Kim, D., Banerdt, W., Ceylan, S., Giardini, D., Lekić, V., Lognonné, P., Beghein, C., Beucler, É., Carrasco, S., Charalambous, C., et al. Surface waves and crustal structure on mars. *Science*, 378(6618) :417–421, 2022.

Knapmeyer-Endrun, B., Panning, M. P., Bissig, F., Joshi, R., Khan, A., Kim, D., Lekić, V., Tauzin, B., Tharimena, S., Plasman, M., et al. Thickness and structure of the martian crust from insight seismic data. *Science*, 373(6553) :438–443, 2021.

Knuth, D. E. *The art of computer programming*, volume 3. Pearson Education, 1997.

Kockelman, K. M. et Kweon, Y.-J. Driver injury severity : an application of ordered probit models. *Accident Analysis & Prevention*, 34(3) :313–321, 2002.

Kolmogorov. Sulla determinazione empirica di una legge didistribuzione. *Giorn Dell'inst Ital Degli Att*, 4 :89–91, 1933.

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

Krischer, L., Megies, T., Barsch, R., Beyreuther, M., Lecocq, T., Caudron, C., et Wassermann, J. Obspy : A bridge for seismology into the scientific python ecosystem. *Computational Science & Discovery*, 8(1) :014003, 2015.

Kroese, D. P. et Rubinstein, R. Y. Monte carlo methods. *Wiley Interdisciplinary Reviews : Computational Statistics*, 4(1) :48–58, 2012.

Ksanfomaliti, L., Zubkova, V., Morozov, N., et Petrova, E. Microseisms at the venera-13 and venera-14 landing sites. *Soviet Astronomy Letters, vol. 8, July-Aug. 1982, p. 241, 242. Translation Pisma v Astronomicheskii Zhurnal, vol. 8, July 1982, p. 444-447*, 8 :241, 1982.

Lagarias, J. C. Pseudorandom numbers. *Statistical Science*, 8(1) :31–39, 1993.

Lam, S. K., Pitrou, A., et Seibert, S. Numba : A llvm-based python jit compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, pages 1–6, 2015.

Latham, G., Ewing, M., Press, F., et Sutton, G. The apollo passive seismic experiment : The first lunar seismic experiment is described. *Science*, 165(3890) :241–250, 1969.

Latham, G., Ewing, M., Dorman, J., Press, F., Toksoz, N., Sutton, G., Meissner, R., Duennebier, F., Nakamura, Y., Kovach, R., et al. Seismic data from man-made impacts on the moon. *Science*, 170(3958) :620–626, 1970a.

Latham, G., Ewing, M., Dorman, J., Lammlein, D., Press, F., Toksoz, N., Sutton, G., Duennebier, F., et Nakamura, Y. Moonquakes. *Science*, 174(4010) :687–692, 1971.

Latham, G. V., Ewing, M., Press, F., Sutton, G., Dorman, J., Nakamura, Y., Toksöz, N., Wiggins, R., Derr, J., et Duennebier, F. Passive seismic experiment. *Science*, 167(3918) :455–457, 1970b.

Lee, W. H. K., Bennett, R., et Meagher, K. *A method of estimating magnitude of local earthquakes from signal duration*. US Department of the Interior, Geological Survey, 1972.

Lehmann, I. P'. *Bureau Central Séismologique International Strasbourg : Publications du Bureau Central Scientifiques*, 14 :87–115, 1936.

Leovy, C. Weather and climate on mars. *Nature*, 412(6843) :245–249, 2001.

Li, J., Beghein, C., Davis, P., Wieczorek, M. A., McLennan, S. M., Kim, D., Lekić, V., Golombek, M., Schimmel, M., Stutzmann, E., et al. Crustal structure constraints

from the detection of the sspp phase on mars. *Earth and Space Science*, 10(3) : e2022EA002416, 2023.

Lognonné, P., Banerdt, W. B., Giardini, D., Pike, W. T., Christensen, U., Laudet, P., De Raucourt, S., Zweifel, P., Calcutt, S., Bierwirth, M., et al. Seis : Insight's seismic experiment for internal structure of mars. *Space Science Reviews*, 215 :1–170, 2019.

Lognonné, P., Banerdt, W. B., Pike, W. T., Giardini, D., Christensen, U., Garcia, R. F., Kawamura, T., Kedar, S., Knapmeyer-Endrun, B., Margerin, L., et al. Constraints on the shallow elastic and anelastic structure of mars from insight seismic data. *Nature Geoscience*, 13(3) :213–220, 2020.

Lognonné, P., Banerdt, W., Clinton, J., Garcia, R., Giardini, D., Knapmeyer-Endrun, B., Panning, M., et Pike, W. Mars seismology. *Annual Review of Earth and Planetary Sciences*, 51, 2023.

Lorenz, R. D., Kedar, S., Murdoch, N., Lognonné, P., Kawamura, T., Mimoun, D., et Bruce Banerdt, W. Seismometer detection of dust devil vortices by ground tilt. *Bulletin of the Seismological Society of America*, 105(6) :3015–3023, 2015.

Lorenz, R. D., Spiga, A., Lognonné, P., Plasman, M., Newman, C. E., et Charalambous, C. The whirlwinds of elysium : A catalog and meteorological characteristics of "dust devil" vortices observed by insight on mars. *Icarus*, 355 :114119, 2021.

Lu, W. et Ghorbani, A. A. Network anomaly detection based on wavelet analysis. *EURASIP Journal on Advances in Signal Processing*, 2009 :1–16, 2008.

Maki, J., Golombek, M., Deen, R., Abarca, H., Sorice, C., Goodsall, T., Schwochert, M., Lemmon, M., Trebi-Ollennu, A., et Banerdt, W. The color cameras on the insight lander. *Space Science Reviews*, 214 :1–34, 2018.

Matsumoto, M. et Nishimura, T. Mersenne twister : a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 8(1) :3–30, 1998.

Megies, T., Beyreuther, M., Barsch, R., Krischer, L., et Wassermann, J. Obspy–what can it do for data centers and observatories ? *Annals of Geophysics*, 54(1) :47–58, 2011.

Meier, M.-A., Ross, Z. E., Ramachandran, A., Balakrishna, A., Nair, S., Kundzicz, P., Li, Z., Andrews, J., Hauksson, E., et Yue, Y. Reliable real-time seismic signal/noise discrimination with machine learning. *Journal of Geophysical Research : Solid Earth*, 124(1) :788–800, 2019.

Mohorovicic, A. Jahrbuch des meteorologischen observatoriums in zagreb (agram) für das jahr 1909. 1910.

Mousavi, S. M. et Beroza, G. C. Deep-learning seismology. *Science*, 377(6607) : eabm4470, 2022.

Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., et Beroza, G. C. Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature communications*, 11(1) :3952, 2020.

Murdoch, N., Spiga, A., Lorenz, R., Garcia, R. F., Perrin, C., Widmer-Schnidrig, R., Rodriguez, S., Compaire, N., Warner, N., Mimoun, D., et al. Constraining martian regolith and vortex parameters from combined seismic and meteorological measurements. *Journal of Geophysical Research : Planets*, 126(2) :e2020JE006410, 2021.

Nakamura, Y., Duennebier, F. K., Latham, G. V., et Dorman, H. J. Structure of the lunar mantle. *Journal of Geophysical Research*, 81(26) :4818–4824, 1976.

Nakamura, Y., Latham, G. V., et Dorman, H. J. Apollo lunar seismic experiment—final summary. *Journal of Geophysical Research : Solid Earth*, 87(S01) :A117–A123, 1982.

Onodera, K., Nishida, K., Kawamura, T., Murdoch, N., Drilleau, M., Otsuka, R., Lorenz, R. D., Horleston, A. C., Widmer-Schnidrig, R., Schimmel, M., et al. Systematic catalog of martian convective vortices observed by insight. *Authorea Preprints*, 2023.

Ordonez-Etxeberria, I., Hueso, R., et Sánchez-Lavega, A. A systematic search of sudden pressure drops on gale crater during two martian years derived from msl/rems data. *Icarus*, 299 :308–330, 2018.

Ouvrard, J.-Y. *Probabilités : Masteur-Agrégation*. Cassini, 2004.

Page, E. On problems in which a change in a parameter occurs at an unknown point. *Biometrika*, 44(1/2) :248–252, 1957.

Pearson, E. S. A further development of tests for normality. *Biometrika*, pages 239–249, 1930.

Pearson, E. S. I. note on tests for normality. *Biometrika*, 22(3-4) :423–424, 1931.

Pearson, E. S. Some problems arising in approximating to probability distributions, using moments. *Biometrika*, 50(1/2) :95–112, 1963.

Pearson, E. S. Tables of percentage points of b_1 and b_2 in normal samples ; a rounding off. *Biometrika*, 52(1/2) :282–285, 1965.

Pearson, K. X. contributions to the mathematical theory of evolution.—ii. skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London.(A.)*, (186) :343–414, 1895.

Pearson, K. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302) :157–175, 1900.

Pedersen, H. et Colombi, A. Body waves from a single source area observed in noise correlations at arrival times of reflections from the 410 discontinuity. *Geophysical Journal International*, 214(2) :1125–1135, 2018.

Pedersen, H. A. et Krüger, F. Influence of the seismic noise characteristics on noise correlations in the baltic shield. *Geophysical Journal International*, 168(1) :197–210, 2007.

Pedersen, H. A., Leroy, N., Zigone, D., Vallée, M., Ringler, A. T., et Wilson, D. C. Using component ratios to detect metadata and instrument problems of seismic stations : Examples from 18 yr of geoscope data. *Seismological Research Letters*, 91 (1) :272–286, 2020.

Perrin, C., Rodriguez, S., Jacob, A., Lucas, A., Spiga, A., Murdoch, N., Lorenz, R., Daubar, I., Pan, L., Kawamura, T., et al. Monitoring of dust devil tracks around the insight landing site, mars, and comparison with in situ atmospheric data. *Geophysical Research Letters*, 47(10) :e2020GL087234, 2020.

Perrin, C., Jacob, A., Lucas, A., Myhill, R., Hauber, E., Batov, A., Gudkova, T., Rodriguez, S., Lognonné, P., Stevanović, J., et al. Geometry and segmentation of cerberus fossae, mars : Implications for marsquake properties. *Journal of Geophysical Research : Planets*, 127(1) :e2021JE007118, 2022.

Peterson, J. et al. Observations and modeling of seismic background noise, open-file report 93–322. *US Geological Survey, Albuquerque, NM*, 1993.

Plackett, R. L. Karl pearson and the chi-squared test. *International statistical review/-revue internationale de statistique*, pages 59–72, 1983.

Posiolova, L., Lognonné, P., Banerdt, W. B., Clinton, J., Collins, G. S., Kawamura, T., Ceylan, S., Daubar, I. J., Fernando, B., Froment, M., et al. Largest recent impact craters on mars : Orbital imaging and surface seismic co-investigation. *Science*, 378 (6618) :412–417, 2022.

Pou, L., Nimmo, F., Lognonné, P., Mimoun, D., Garcia, R. F., Pinot, B., Rivoldini, A., Banfield, D., et Banerdt, W. B. Forward modeling of the phobos tides and applications to the first martian year of the insight mission. *Earth and Space Science*, 8(7) :e2021EA001669, 2021.

Pourhoseingholi, A., Pourhoseingholi, M. A., Vahedi, M., Safaee, A., Moghimi-Dehkordi, B., Ghafarnejad, F., et Zali, M. R. Relation between demographic factors and type of gastrointestinal cancer using probit and logit regression. *Asian Pac J Cancer Prev*, 9(4) :753–5, 2008.

Queffélec, H. et Zuily, C. *Eléments d'analyse : agrégation de mathématiques*. Dunod, 2002.

Ramis, J.-P., Warusfel, A., Buff, X., Garnier, J., Halberstadt, E., Lachand-Robert, T.,

Moulin, F., et Sauloy, J. *Mathématiques Tout-en-un pour la Licence-Niveau L1-2e édition : Cours complet, exemples et exercices corrigés*. Dunod, 2013.

Romanowicz, B. et Dziewonski, A. Toward a federation of broadband seismic networks. *EOS, Trans. Am. geophys. Un.*, 67 :541, 1987.

Romanowicz, B. Seismic tomography of the earth's mantle. *Annual Review of Earth and Planetary Sciences*, 19(1) :77–99, 1991.

Romanowicz, B., Cara, M., Fel, J. F., et Rouland, D. Geoscope : A french initiative in long-period three-component global seismic networks. *Eos, Transactions American Geophysical Union*, 65(42) :753–753, 1984.

Rombaldi, J.-É. et Darracq, M. C. *Analyse pour la licence*. De Boeck Superieur, 2020.

Ryan, J. et Lucich, R. Possible dust devils, vortices on mars. *Journal of Geophysical Research : Oceans*, 88(C15) :11005–11011, 1983.

Samuel, H., Ballmer, M. D., Padovan, S., Tosi, N., Rivoldini, A., et Plesa, A.-C. The thermo-chemical evolution of mars with a strongly stratified mantle. *Journal of Geophysical Research : Planets*, 126(4) :e2020JE006613, 2021.

Scales, J. A. et Snieder, R. What is noise ? *Geophysics*, 63(4) :1122–1124, 1998.

Schimmel, M., Stutzmann, E., et Ventosa, S. Measuring group velocity in seismic noise correlation studies based on phase coherence and resampling strategies. *IEEE Transactions on Geoscience and Remote Sensing*, 55(4) :1928–1935, 2017.

Scholz, J.-R., Widmer-Schnidrig, R., Davis, P., Lognonné, P., Pinot, B., Garcia, R. F., Hurst, K., Pou, L., Nimmo, F., Barkaoui, S., et al. Detection, analysis, and removal of glitches from insight's seismic data from mars. *Earth and Space Science*, 7(11) : e2020EA001317, 2020.

Seed Reference Manual. Standard for the exchange of earthquake data. 2012.

Seydoux, L., Balestriero, R., Poli, P., Hoop, M. d., Campillo, M., et Baraniuk, R. Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning. *Nature communications*, 11(1) :3972, 2020.

Shapiro, N. M. et Campillo, M. Emergence of broadband rayleigh waves from correlations of the ambient seismic noise. *Geophys. Res. Lett.*, 31(7), 2004. ISSN 1944-8007. doi : 10.1029/2004GL019491.

Shapiro, S. S. et Wilk, M. B. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4) :591–611, 1965.

Shapiro, S. et Wilk, M. B. Approximations for the null distribution of the w statistic. *Technometrics*, 10(4) :861–866, 1968.

Silverman, R. A. et al. *Special functions and their applications*. Courier Corporation, 1972.

Smith, S. W. Iris—a university consortium for seismology. *Reviews of Geophysics*, 25 (6) :1203–1207, 1987.

Sorrells, G., McDonald, J. A., Der, Z., et Herrin, E. Earth motion caused by local atmospheric pressure changes. *Geophysical Journal International*, 26(1-4) :83–98, 1971.

Spiga, A., Banfield, D., Teanby, N. A., Forget, F., Lucas, A., Kenda, B., Rodriguez Manfredi, J. A., Widmer-Schnidrig, R., Murdoch, N., Lemmon, M. T., et al. Atmospheric science with insight. *Space Science Reviews*, 214 :1–64, 2018.

Spiga, A., Murdoch, N., Lorenz, R., Forget, F., Newman, C., Rodriguez, S., Pla-Garcia, J., Moreiras, D. V., Banfield, D., Perrin, C., et al. A study of daytime convective vortices and turbulence in the martian planetary boundary layer based on half-a-year of insight atmospheric measurements and large-eddy simulations. *Journal of Geophysical Research : Planets*, 126(1) :e2020JE006511, 2021.

Spohn, T., Grott, M., Smrekar, S., Knollenberg, J., Hudson, T., Krause, C., Müller, N., Jänchen, J., Börner, A., Wippermann, T., et al. The heat flow and physical properties package (hp 3) for the insight mission. *Space Science Reviews*, 214 :1–33, 2018.

Stähler, S. C., Khan, A., Banerdt, W. B., Lognonné, P., Giardini, D., Ceylan, S., Drilleau, M., Duran, A. C., Garcia, R. F., Huang, Q., et al. Seismic detection of the martian core. *Science*, 373(6553) :443–448, 2021.

Steakley, K. et Murphy, J. A year of convective vortex activity at gale crater. *Icarus*, 278 :180–193, 2016.

Stott, A., Garcia, R., Chédozeau, A., Spiga, A., Murdoch, N., Pinot, B., Mimoun, D., Charalambous, C., Horleston, A., King, S., et al. Machine learning and marsquakes : a tool to predict atmospheric-seismic noise for the nasa insight mission. *Geophysical Journal International*, 233(2) :978–998, 2023.

Stutzmann, E., Ardhuin, F., Schimmel, M., Mangeney, A., et Patau, G. Modelling long-term seismic noise in various environments. *Geophysical Journal International*, 191(2) :707–722, 2012.

Stutzmann, É., Schimmel, M., Lognonné, P., Horleston, A., Ceylan, S., van Driel, M., Stahler, S., Banerdt, B., Calvet, M., Charalambous, C., et al. The polarization of ambient noise on mars. *Journal of Geophysical Research : Planets*, 126(1) : e2020JE006545, 2021.

Taner, M. T., Koehler, F., et Sheriff, R. Complex seismic trace analysis. *Geophysics*, 44(6) :1041–1063, 1979.

Tang, G. et Ma, J. Application of total-variation-based curvelet shrinkage for three-dimensional seismic data denoising. *IEEE geoscience and remote sensing letters*, 8 (1) :103–107, 2010.

Tanimoto, T. et Wang, J. Low-frequency seismic noise characteristics from the analysis of co-located seismic and pressure data. *Journal of Geophysical Research : Solid Earth*, 123(7) :5853–5885, 2018.

Taylor, W. A. Change-point analysis : a powerful new tool for detecting changes, 2006.

Trebi-Ollennu, A., Kim, W., Ali, K., Khan, O., Sorice, C., Bailey, P., Umland, J., Bonitz, R., Ciarleglio, C., Knight, J., et al. Insight mars lander robotics instrument deployment system. *Space Science Reviews*, 214 :1–18, 2018.

Tytell, J., Vernon, F., Hedlin, M., de Groot Hedlin, C., Reyes, J., Busby, B., Hafner, K., et Eakins, J. The usarray transportable array as a platform for weather observation and research. *Bulletin of the American Meteorological Society*, 97(4) :603–619, 2016.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Van Rossum, G. et Drake, F. L. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., et SciPy 1.0 Contributors. SciPy 1.0 : Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17 :261–272, 2020. doi : 10.1038/s41592-019-0686-2.

Wang, D., Li, Y., et Nie, P. A study on the gaussianity and stationarity of the random noise in the seismic exploration. *Journal of Applied Geophysics*, 109 :210–217, 2014.

Weaver, R. L. Information from seismic noise. *Science*, 307(5715) :1568–1569, 2005.

Webb, S. C. The earth's 'hum'is driven by ocean waves over the continental shelves. *Nature*, 445(7129) :754–756, 2007.

Wichmann, B. A. et Hill, I. Generating good pseudo-random numbers. *Computational Statistics & Data Analysis*, 51(3) :1614–1622, 2006.

Wieczorek, M. A., Broquet, A., McLennan, S. M., Rivoldini, A., Golombek, M., Antonangeli, D., Beghein, C., Giardini, D., Gudkova, T., Gyalay, S., et al. Insight constraints on the global character of the martian crust. *Journal of Geophysical Research : Planets*, 127(5) :e2022JE007298, 2022.

Williams, P. Note on the sampling distribution of $\sqrt{\beta_1}$, where the population is normal. *Biometrika*, 27(1/2) :269–271, 1935.

Yazici, B. et Yolacan, S. A comparison of various tests of normality. *Journal of Statistical Computation and Simulation*, 77(2) :175–183, 2007.

Zhong, T., Li, Y., Wu, N., Nie, P., et Yang, B. Statistical properties of the random noise in seismic data. *Journal of applied Geophysics*, 118 :84–91, 2015a.

Zhong, T., Li, Y., Wu, N., Nie, P., et Yang, B. Statistical analysis of background noise in seismic prospecting. *Geophysical Prospecting*, 63(5) :1161–1174, 2015b. doi : 10.1111/1365-2478.12237.

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

COLLEGE DOCTORAL PAYS DE LA LOIRE
ED MATIERE,MOLECULES MATERIAUX ET GEOSCIENCES

Nantes Université

Titre : Nouvelle méthode de localisation d'échantillons non Gaussiens : applications au signal sismique terrestre et martien

Mots clés : Sismologie, statistiques, traitement du signal, loi normale

Résumé : Au cours des dernières décennies, un intérêt particulier fut apporté à l'étude du signal sismique continu, se démarquant de l'enjeu primaire et historique de la sismologie, principalement centré autour de la détection d'évènements impulsifs associés aux tremblements de terres. L'étude d'un tel signal, communément appelé « bruit sismique » se révèle toutefois riche en informations et témoigne de la présence de phénomènes atmosphériques, de perturbations anthropiques, ou encore d'altérations mécanique locales (glitchs).

Dans ce contexte, nous proposons de procéder à l'analyse du signal sismique continu via l'une de ses propriétés remarquables : sa distribution Gaussienne. Afin de répondre à ce besoin, nous présentons dans ce manuscrit la méthode **NG-loc**, une approche statistique originale permettant de localiser les échantillons altérant la distribution Gaussienne d'une série de données. L'utilisation de NG-loc apporte des résultats prometteurs, permettant d'estimer la qualité de n'importe quel signal sismique mais aussi de mettre en lumière les possibles dégradations de ce dernier, affectant certaines stations sur des périodes temporelles, composantes et bandes de fréquences spécifiques. De plus, NG-loc permet de détecter tout type de perturbation, dès lors que celle-ci altère la distribution statistique du signal, et s'avère donc sensible à un large panel d'altérations d'origines naturelles ou anthropiques (événements sismiques, passage de machines agricoles,...). Une seconde application de NG-loc est également proposée par l'analyse du signal sismique enregistré sur la planète Mars, lors de la mission spatiale **InSight** (2018-2022), nous permettant de mettre en évidence de fortes fluctuations de sa qualité au cours de la mission. De plus, l'analyse des perturbations détectées par NG-loc via une approche de discrimination par *machine learning* nous permet également de proposer une classification d'un certain type d'altérations, causées par des des tornades de poussière.

Finalement, bien que NG-loc puisse apporter d'intéressantes contributions dans le domaine de la sismologie, nous pensons que cette nouvelle approche peut également trouver son intérêt dans un champ d'application plus large, de par l'analyse de n'importe quel signal, à priori Gaussien.

Title : New method for locating non-Gaussian samples: applications to the terrestrial and martian seismic signal

Keywords : Seismology, statistics, data processing, Gaussian law

Abstract : In recent decades, particular interest has been shown in the study of the continuous seismic signals, moving away from the primary and historical focus of seismology, which was mainly centred on the detection of impulsive events associated with earthquakes. However, the study of such a signal, commonly known as 'seismic noise', is rich in information and reveals the presence of atmospheric phenomena, human disturbances and local mechanical alterations (glitches).

In this context, we propose to analyse the continuous seismic signal via one of its remarkable properties: its Gaussian distribution. To meet this need, we present in this manuscript the NG-loc method, an original statistical approach for locating samples that alter the Gaussian distribution of a series of data. The use of NG-loc yields promising results, making it possible to estimate the quality of any seismic signal but also to highlight possible degradations of the latter, affecting certain stations over specific time periods, components and frequency bands. In addition, NG-loc can detect any type of disturbance that alters the statistical distribution of the signal, and is therefore sensitive to a wide range of alterations of natural or man-made origin (seismic events, passage of agricultural machinery, etc.). A second application of NG-loc is the analysis of the seismic signal recorded on Mars during the InSight space mission (2018-2022), enabling us to highlight significant fluctuations in its quality over the course of the mission. In addition, analysis of the disturbances detected by NG-loc using a machine learning discrimination approach has also enabled us to propose a classification of a certain type of alteration caused by dust devils.

Finally, although NG-loc can provide interesting contributions in the field of seismology, we believe that this new approach can also be of interest in a wider field of application, through the analysis of any signal, a priori Gaussian.