



HAL
open science

Comparaisons de localisations de séismes du fossé rhénan dans des modèles 1D et 3D

Romain Gruber

► **To cite this version:**

Romain Gruber. Comparaisons de localisations de séismes du fossé rhénan dans des modèles 1D et 3D. Méthodologie [stat.ME]. 2012. dumas-00728926

HAL Id: dumas-00728926

<https://dumas.ccsd.cnrs.fr/dumas-00728926>

Submitted on 7 Sep 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Comparaisons de localisations de séismes du fossé rhénan dans des modèles 1D et 3D

Rapport de stage

Romain GRUBER

Master 1 Mathématiques et Applications
Spécialité Statistique

Responsable de stage : Sophie LAMBOTTE

Résumé

La localisation des séismes par des méthodes numériques requiert des modèles de vitesses de propagation des ondes sismiques dans le sous-sol. Bien souvent on utilise des modèles à couches dans lesquels la vitesse des ondes ne varie qu'avec la profondeur. Avec l'augmentation de la puissance de calculs des ordinateurs il devient possible d'utiliser de manière routinière des modèles à trois dimensions tenant compte des spécificités locales du sous-sol, comme l'existence de couches sédimentaires dans certaines régions.

Le but de ce stage est de comparer les localisations des séismes du fossé rhénan (localisation géographique, profondeur, et temps origine) en fonction du modèle de vitesse de propagation des ondes sismiques, et de quantifier l'effet du modèle sur l'erreur de localisation. Trois modèles ont été considérés : un modèle à une dimension simple, noté 1Ds, un modèle à une dimension comportant une couche supplémentaire, noté 1Dc, et un modèle à trois dimensions, noté 3D. Des événements synthétiques, répartis dans tout le fossé rhénan, ont été créés à partir du programme NonLinLoc développé par Anthony Lomax, puis les localisations ont été effectuées avec chaque modèle. Les différences entre les localisations calculées et les localisations exactes ont ensuite été comparées entre elles en fonction du modèle utilisé et d'autres paramètres liés au réseau de stations sismiques. Des analyses graphiques ont permis de mettre en évidence certaines différences de comportements, qui ont ensuite été vérifiées et quantifiées par des méthodes statistiques comme l'analyse de la variance à mesures répétées ou le test de Student apparié.

Table des matières

1. Généralités sur les séismes et les méthodes de localisation	4
Généralités sur la sismologie	4
Objectif du stage	7
Le programme NonLinLoc	8
2. Créations des données et analyse graphique des résultats	11
Création des événements	11
Analyse exploratoire des données	13
Influence du réseau	21
3. Analyse de la variance, comparaisons multiples, tests non paramétriques ..	26
Principes théoriques des méthodes utilisées	26
Vérification des conditions d'application de l'analyse de la variance à mesures répétées	30
Résultats	34
Synthèse	42
Bibliographie	44

1. Généralités sur les séismes et les méthodes de localisation

Généralités sur la sismologie

La planète Terre est secouée par environ cinq cent milles séismes, ou tremblements de terre, chaque année (chiffre variable selon les estimations), dont cent milles suffisamment forts pour être ressentis¹. Un tremblement de terre est dû au déplacement brutal de deux blocs rocheux le long d'une faille. Durant les périodes inter-sismiques les roches subissent une déformation élastique, et sont soumises à des contraintes, accumulant de l'énergie jusqu'au point de rupture, lors de laquelle l'énergie est libérée sous différentes formes : une première partie se dissipe en chaleur lors du glissement, une deuxième partie se transmet sous forme mécanique dans les zones voisines de la rupture, et une troisième partie se propage dans le milieu environnant depuis le point de rupture, appelé foyer du séisme ou hypocentre, sous formes d'ondes élastiques qu'il est possible d'enregistrer à la surface de la Terre. La mise en relation entre l'occurrence d'un séismes et l'enregistrement d'ondes élastiques a été proposée pour la première fois en 1889 suite à un séisme survenu au Japon et dont les ondes sismiques ont été détectées jusqu'à Postdam en Europe.

Les ondes sismiques produites par un séismes sont de différents types selon le mouvement qui affecte une particule traversée par cette onde (figure 1.1). On distingue deux types d'onde : les ondes de volume (ondes P et S) et les ondes de surface (ondes de Love et ondes de Rayleigh).

- Les ondes P provoquent un mouvement de type compression-dilatation suivant la direction de propagation (ondes longitudinales). Elles se propagent dans tous les milieux, qu'ils soient solides ou fluides.
- Les ondes S produisent un mouvement cisailant des particules dans une direction perpendiculaire à la direction de propagation (ondes transversales). Elles ne peuvent se propager dans un fluide, le module de cisaillement d'un fluide étant nul.
- Les ondes de Love provoquent elles-aussi un mouvement de cisaillement perpendiculaire à leur direction de propagation. Elles ne peuvent donc, elles aussi, que se propager dans des milieux solides. Elles se distinguent des ondes S par le fait qu'elle ne se propagent qu'en surface, et à une vitesse moindre que les ondes S.
- Les ondes de Rayleigh provoquent un mouvement des particules de type elliptique, à

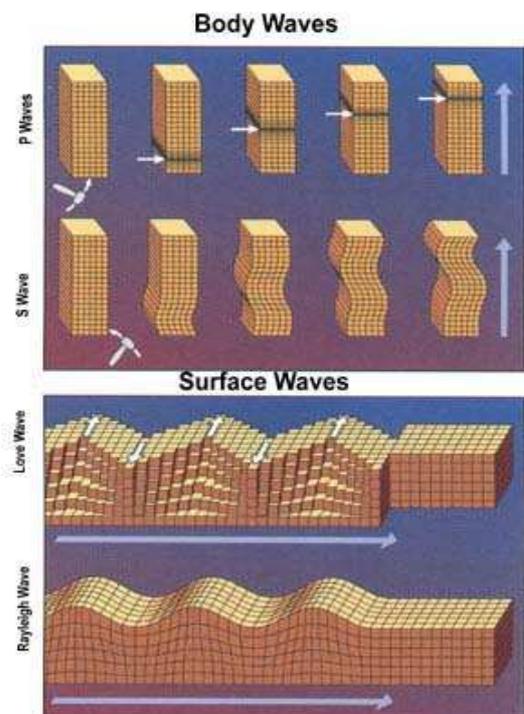


Fig. 1.1 - Les différents types d'ondes sismiques

(Source : United States Geological Survey)

¹ United States Geological Survey, <http://earthquake.usgs.gov/learn/facts.php>

la fois longitudinal et transversal. Elles ne se propagent qu'en surface et ont une vitesse de propagation plus faible que les ondes S.

Ces différents trains d'onde se propagent à des vitesses différentes et les ondes de surface peuvent faire plusieurs fois le tour de la Terre avant d'être totalement amorties (les ondes émises par un séisme de magnitude 8 sont amorties après environ sept jours).

De plus les ondes P et les ondes S voient leur vitesse de propagation varier en fonction du milieu qu'elles traversent et leur direction de propagation se modifier, d'après les lois de Snell-Descartes (que l'on retrouve par exemple en optique). Elles ont une vitesse plus élevée dans un milieu plus dense. La Terre ayant une structure à couches (croûte terrestre, manteau, noyau externe, noyau interne), elle n'est pas un milieu homogène et les ondes sismiques P et S sont donc réfractées, voire réfléchies, lors qu'elles changent de milieu.

Les ondes sismiques provoquées par un séisme sont donc un phénomène complexe à analyser et il est parfois difficile de différencier les différents types d'ondes.

Les enregistrements des ondes sismiques à la surface de la Terre par des stations sismologiques sont appelés des sismogrammes (figure 1.2). On y observe l'amplitude du mouvement du sol selon trois composantes (verticale, est-ouest et nord-sud). Une analyse de ces sismogrammes permet de repérer les temps d'arrivées des différents trains d'onde :

- La première partie d'un sismogramme, où les amplitudes sont extrêmement faibles, correspond à la mesure du bruit de fond, dû notamment aux variations atmosphériques et aux activités humaines.
- L'enregistrement du séisme proprement dit débute par une variation plus ou moins brutale, selon la magnitude du séisme et la distance entre son hypocentre et la station d'enregistrement. Il s'agit de l'arrivée des ondes P, qui sont les ondes sismiques de déplacement le plus rapidement.
- Le second train d'onde détecté correspond à l'arrivée des ondes S.
- Enfin le dernier train d'onde repéré correspond aux ondes de surface (ondes de Love et ondes de Rayleigh).

Il est parfois difficile d'analyser correctement un sismogramme et de repérer de manière précise les temps d'arrivées des différents types d'ondes, d'autant plus qu'on peut aussi détecter les ondes qui ont été réfractées et réfléchies lors de leur trajet.

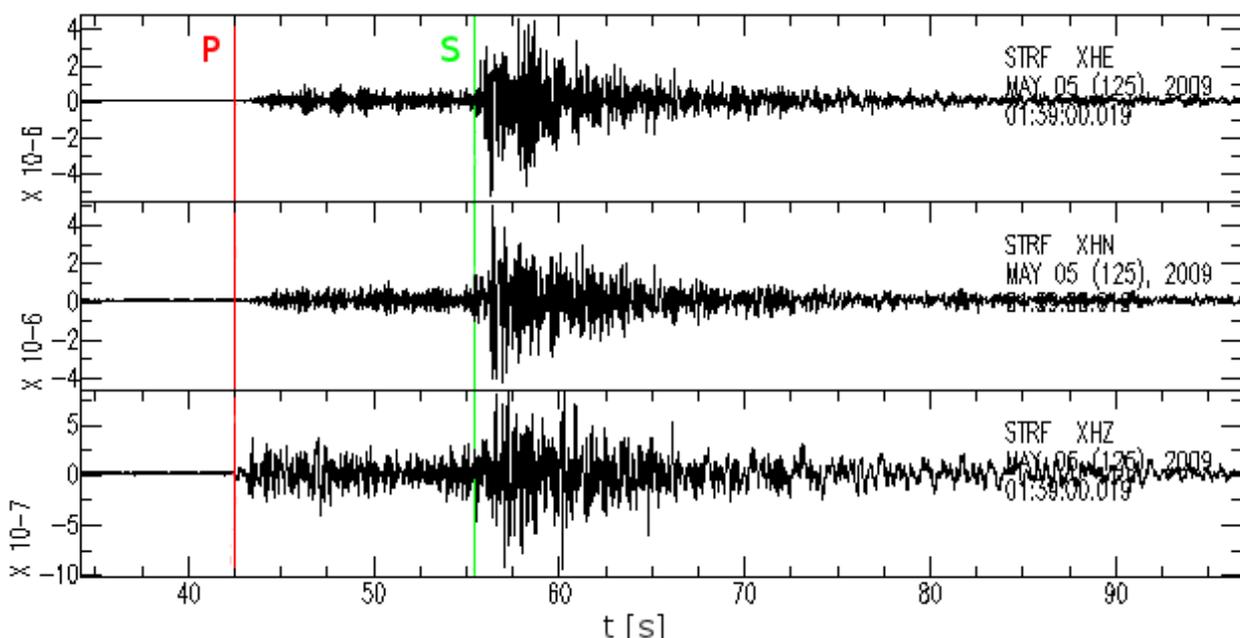


Fig. 1.2 - Les trois composantes d'un sismogramme (Est-Ouest, Nord-Sud, verticale)

La localisation spatio-temporelle d'un tremblement de terre s'effectue à partir de l'analyse des sismogrammes et les temps d'arrivées des ondes P et S à différentes stations sismologiques. Supposons que l'on dispose d'un enregistrement nous fournissant un temps t_p d'arrivée des ondes P et un temps t_s d'arrivée des ondes S. Si on connaît la vitesse V_p des ondes P et la vitesse V_s des ondes S, et si on note d la distance entre la station et le séisme, et t_0 l'instant où se produit le séisme, on a les relations : $t_p = d/V_p$ et $t_s = d/V_s$.

En faisant la différence entre t_p et t_s on a donc : $t_s - t_p = d \times (1/V_s - 1/V_p)$

On peut déterminer d , et donc t_0 .

Si l'on dispose d'au moins trois stations ayant détecté le séisme, on peut tracer un cercle centré en chaque station et de rayon la distance calculée avec la formule précédente, l'intersection des trois cercles donnant ainsi la localisation de l'épicentre : c'est le principe de triangulation. Il ne s'agit que d'une approche très simplifiée, qui ne tient pas compte de la profondeur du séisme.

De nos jours, grâce aux techniques numériques, on procède en ajustant les temps d'arrivées théoriques calculés par rapport aux temps observés, de manière à réduire l'écart entre ces deux temps. On obtient ainsi non seulement la localisation de l'épicentre et le temps origine, mais aussi la profondeur de l'hypocentre. Ici aussi il est nécessaire de disposer d'enregistrements provenant de plusieurs stations, et la précision de la localisation dépendra non seulement du nombre de stations l'ayant détectées, mais aussi de leur distance par rapport à l'épicentre (il est souhaitable d'avoir un enregistrement d'une station proche de l'épicentre) et de leur répartition géographique (il est préférable que les stations soient disposées tout autour du lieu où s'est produit le séisme).

Néanmoins la prédiction d'un séisme étant encore impossible il n'est pas possible de déployer des stations temporaires avant qu'un séisme se produise, et il est dès lors souhaitable de disposer de nombreuses stations réparties sur le territoire de manière à couvrir efficacement les zones où des séismes sont susceptibles de se produire.

Les formules précédentes posent cependant un problème : quelles sont les vitesses V_p et V_s des ondes P et S ? Pour améliorer la détermination de la localisation des séismes il est nécessaire de prendre des modèles de vitesses des ondes P et S réalistes. Une première approximation consiste à définir les vitesses des ondes P et S en fonction de la profondeur, selon un modèle à couches. En effet les vitesses des ondes P et S sont affectées par le milieu dans lequel elles se propagent, or la structure de la Terre présente différentes couches principales. L'un des modèles généraux les plus utilisés est le modèle PREM (Preliminary Reference Earth Model, figure 1.3). Grâce à la puissance des ordinateurs actuels il est possible d'affiner ces modèles à enveloppes sphériques en développant des modèles à trois dimensions, par exemple en modifiant légèrement la profondeur des discontinuités en fonction des régions, ou en ajoutant des zones restreintes où les vitesses de propagation des ondes sont différentes (présence d'une couche de sédiments, ou d'une chambre magmatique par exemple).

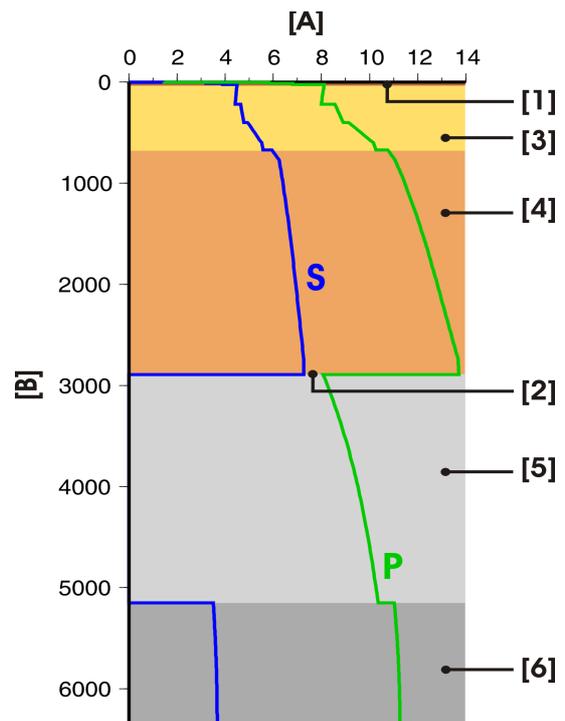


Fig 1.3 - Vitesses des ondes P et S en km/s (A) en fonction de la profondeur en km (B) du modèle PREM.

1 : croûte.

2 : interface noyau-manteau.

3 : manteau supérieur.

4 : manteau inférieur.

5 : noyau externe.

6 : noyau interne.

(Source : Wikipedia)

Objectif du stage

La France métropolitaine n'est pas située sur une frontière de plaques lithosphériques¹ contrairement aux zones à forte sismicité telles que la Californie, la Grèce ou le Japon. Elle est néanmoins située dans une zone de déformation intraplaque et présente une sismicité modérée due principalement au rapprochement de la plaque Afrique et de la plaque Eurasie. La sismicité des deux grands bassins sédimentaires français (bassin aquitaine et bassin parisien) est pratiquement nulle, en revanche quelques zones présentent une sismicité modérée, dont les Alpes, la vallée du Rhône, les Pyrénées, le grand Ouest, ainsi que le fossé rhénan (figure 1.4). Ces séismes ont lieu dans la croûte terrestre et sont donc localisés à faible profondeur (jusqu'à 20 km environ)

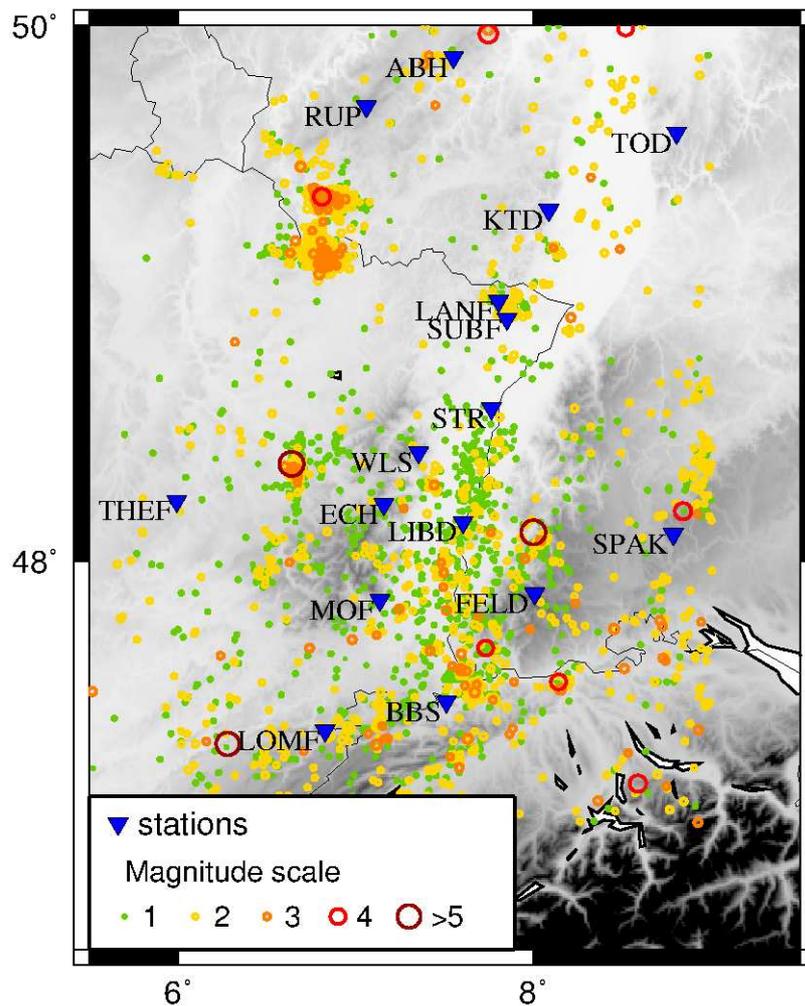


Fig. 1.4 - Sismicité du fossé rhénan

Ce stage a pour but de comparer l'effet de trois modèles de vitesses des ondes P et S dans le fossé rhénan en étudiant leur effet sur la précision des localisations spatio-temporelles des séismes régionaux.

- Le premier modèle est un modèle à une dimension (modèle à enveloppes sphériques) comprenant trois couches : une couche de 0 à 20 km où la vitesse des ondes P est de 5,9 km/s et celle des ondes S de 3,4 km/s, une couche de 20 à 30 km où la vitesse des ondes P

¹ La lithosphère est l'enveloppe rigide de la surface de la Terre comprenant la croûte terrestre et une partie du manteau supérieur. La lithosphère est divisée en plaques tectoniques.

est de 6,5 km/s et celle des ondes S de 3,7 km/s, et enfin une couche à partir de 30 km (qui dans le modèle numérique s'étend jusqu'à l'infini) où la vitesse des ondes P est de 8,2 km/s et celle des ondes S de 4,4 km/s. Par la suite ce modèle sera noté *1Ds*.

- Le deuxième modèle est un modèle à une dimension plus complexe comprenant cette fois quatre couche : une couche de 0 à 2,4 km où la vitesse des ondes P est de 5,63 km/s et celle des ondes S de 3,33 km/s, une couche de 2,4 à 20,1 km où la vitesse des ondes P est de 5,97 km/s et celles des ondes S de 3,4 km/s, une couche de 20,1 à 30,2 km où la vitesse des ondes P est de 6,54 km/s et celle des ondes S de 3,65 km/s, et enfin une couche à partir de 30,2 km où la vitesse des ondes P est de 8,15 km/s et celle des ondes S de 4,4 km/s. Par la suite ce modèle sera noté *1Dc*.
- Le troisième modèle est un modèle à trois dimensions du fossé rhénan. Il s'agit aussi d'un modèle comprenant différentes couches, néanmoins la profondeur de la séparation entre les couches n'est pas la même partout et a été modifiée selon les spécificités géologiques locales. Par la suite ce modèle sera noté *3D*.

La zone d'intérêt de l'étude est le fossé rhénan. On s'intéressera à la localisation géographique (longitude et latitude), la localisation en profondeur de l'hypocentre, et à la localisation temporelle (temps origine auquel le séisme s'est produit). Il est courant de s'intéresser à une quantité appelée RMS (Root Mean Square) mesurant l'écart entre les temps d'arrivée observés et les temps d'arrivées théoriques, néanmoins ce paramètre a tendance à naturellement augmenter quand le nombre de stations ayant détecté l'événement augmente (il est plus facile d'ajuster le RMS et d'obtenir un RMS faible lorsqu'il y a peu de stations), on n'en tiendra donc pas compte dans cette étude.

On s'intéressera aussi à l'effet d'autres paramètres sur la précision des localisations des séismes : le nombre de phases, égal aux nombres de temps d'arrivées enregistrées, la distance à la station la plus proche, ainsi que le gap azimutal. Le gap azimutal est l'angle azimutal le plus grand entre deux stations ; un gap petit signifie que les stations sont bien réparties autour du séisme, un gap élevé (supérieur à 180°) signifie que le séisme a lieu à l'extérieur du réseau. Il ne faut pas confondre le nombre de phases avec le nombre de stations ayant détecté le séisme : il se peut que certains sismogrammes soient difficiles à analyser et qu'on n'arrive à repérer que le temps d'arrivée des ondes P, mais pas celui des ondes S

Le programme NonLinLoc

Afin de réaliser cette étude on utilisera le programme de localisation NonLinLoc, développé par Anthony Lomax¹. Ce programme permet de localiser des séismes aussi bien avec des modèles 1D qu'avec des modèles 3D. Pour cela on fournit au programme le modèle de vitesses d'ondes P et S souhaité, ainsi que la liste des stations utilisées (avec leur coordonnées géographiques). Le programme crée alors une grille de temps pour chaque station, correspondant, à chaque case de la grille, au temps que mettrait une onde sismique pour aller de la station à cette case (et inversement). La méthode de calcul est basée sur l'algorithme par différences finies de Podvin et Lecomte (1991). Il suffit ensuite d'indiquer au programme le fichier contenant les temps d'arrivées d'onde, et NonLinLoc calcule un estimateur du maximum de vraisemblance de la localisation ainsi que d'autres informations.

La méthode d'inversion utilisée par NonLinLoc consiste à calculer une fonction de densité de

¹ <http://alomax.free.fr/nlloc/>

probabilité à posteriori de la localisation du séisme, selon la méthode développée par Tarantola et Valette (1982). Si p est le vecteur de paramètres que l'on cherche à estimer, d le vecteur des données, $\theta(d,p)$ la relation théorique entre d et p , $\theta(d|p)$ la relation théorique conditionnelle de d sachant p , $\rho_p(p)$ l'information a priori sur p , $\rho_d(d)$ l'information a priori sur d , et $\mu_d(d)$ la fonction de densité d'information nulle, alors on peut montrer que la fonction de densité de probabilité a posteriori de p s'écrit :

$$\sigma_p(p) = \rho_p(p) \int \frac{\rho_d(d)\theta(d|p)}{\mu_d(d)} dd \quad (1)$$

Notons x le vecteur des coordonnées spatiales (abscisse, ordonnée, profondeur), T le temps origine, t le vecteur des temps d'arrivées observés, et h les temps de trajet théoriques des ondes. Si la relation théorique entre (x,T) et t est supposée gaussienne de matrice de covariance C_T , et si le vecteur t suit a priori une loi gaussienne de matrice de covariance C_t , alors il est possible d'intégrer l'équation (1) selon le temps. La fonction de densité de probabilité a posteriori s'écrit alors de la manière suivante (Tarantola et Valette 1982, Moser et al. 1992) :

$$\sigma(x) = K \rho(x) \exp\left[-\frac{1}{2}g(x)\right]$$

$$g(x) = [\hat{t}_0 - \hat{h}(x)]' (C_t + C_T)^{-1} [\hat{t}_0 - \hat{h}(x)]$$

où K est une constante, \hat{t}_0 le vecteur des temps observés moins leur moyenne pondérée, et \hat{h} le vecteur des temps de trajet théoriques moins leur moyenne pondérée, où les poids w_i sont définis par :

$$w_i = \sum_j w_{ij} \quad ; \quad w_{ij} = [(C_t + C_T)^{-1}]_{ij}$$

On obtient donc une solution complète du problème sous la forme d'une fonction de densité de probabilité, qui tient compte de toutes les incertitudes.

NonLinLoc propose ensuite plusieurs algorithmes pour localiser le maximum de vraisemblance et obtenir une représentation de la fonction de densité de probabilité a posteriori : un algorithme de recherche systématique par grille, qui va calculer la fonction de densité de probabilité a posteriori pour chaque point de la grille ; un algorithme de type Metropolis-Gibbs ; et un algorithme appelé Oct-Tree. C'est cet algorithme qui sera utilisé car il donne des résultats proches de la recherche systématique par grille, gère correctement les fonctions de densité de probabilité pluri-modales ou de formes atypiques, tout en étant cent fois plus rapide que la recherche systématique par grille. L'algorithme de type Metropolis-Gibbs est aussi rapide mais a tendance à mal gérer les distributions pluri-modales et à être pris au piège dès qu'il entre dans une zone à forte densité.

Dans un premier temps la zone géographique d'intérêt (en trois dimensions) est divisée en un certain nombre restreint de cellules. L'algorithme calcule $\sigma(x_i)$ au centre de chaque cellule, puis multiplie par le volume de la cellule pour obtenir une probabilité P_i , et classe ensuite les P_i dans l'ordre. La cellule ayant la P_{\max} est ensuite divisée en huit cellules filles. Les P_i sont ensuite calculés pour chaque cellule fille, une nouvelle liste des P_i ordonnées est créée, et l'opération est répétée jusqu'à ce qu'un certain critère soit atteint (taille de cellules inférieure à une limite fixée, nombre de cellules supérieur à une limite fixée, etc). On obtient au final une représentation fidèle de la fonction de densité de probabilité, le nombre de cellules étant plus important dans les zones à forte densité.¹

Le programme renvoie en sortie plusieurs fichiers, dont un fichier d'extension .hyp contenant la

¹ <http://alomal.free.fr/nlloc/octree/OctTree.html>

plupart des informations : coordonnées du maximum de vraisemblance, nombre de phases utilisées, distance à la station la plus proche, RMS, estimateur linéaire de l'hypocentre, ellipsoïde de confiance, etc. Il est aussi possible de tracer dans un fichier image une représentation en nuage de points de la fonction de densité de probabilité estimée (figure 1.5).

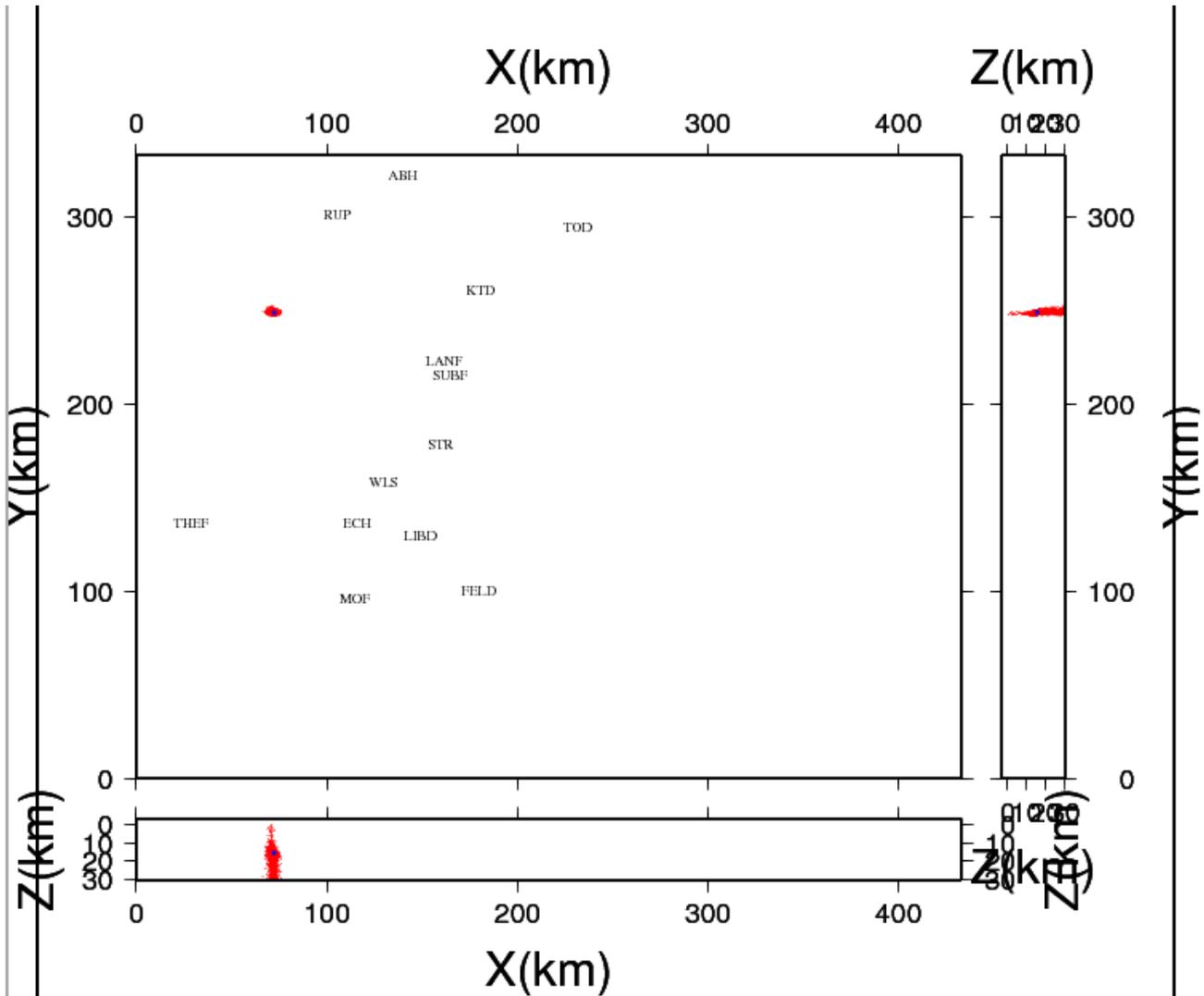


Fig. 1.5 - Exemple d'un nuage de points produit par NonLinLoc. Le point noir correspond à l'hypocentre du maximum de vraisemblance. Le nuage de points rouges est une estimation de la fonction de densité de probabilité.

2. Création des données et analyse graphique des résultats

Création des événements

Le programme NonLinLoc permet de créer des événements synthétiques. Il suffit de fournir un modèle de vitesses des ondes sismiques ainsi qu'un fichier indiquant la localisation du séisme et les stations que l'on souhaite utiliser, et le programme génère un fichier d'observations contenant des temps d'arrivées d'ondes synthétiques. Il est aussi possible de définir les différentes incertitudes (incertitude due au trajet des ondes, incertitude lors du relevé des temps d'arrivées des ondes, etc) afin d'obtenir un fichier d'observation virtuel crédible, proche de ce qu'on observe en réalité. C'est cette méthode qui sera choisie dans notre étude. En effet on cherche à comparer la précision dans la localisation de séismes en fonction du modèle de vitesses d'ondes sismiques utilisé. Pour cela il est nécessaire de connaître la véritable localisation du séisme afin de pouvoir mesurer l'erreur commise, ce qui est impossible dans le cas de séismes réels dont on ignore la véritable localisation¹. On voit néanmoins tout de suite un problème de cette méthode : pour créer les événements virtuels il est nécessaire de fournir à NonLinLoc un modèle de vitesse d'onde que le programme utilisera pour créer des temps d'arrivée d'ondes virtuels, mais ce modèle peut ne pas correspondre tout à fait à la réalité (c'est justement en partie ce que l'on cherche à savoir), on introduit donc peut-être un biais systématique. On n'observera donc pas des imprécisions de localisations absolues, semblables exactement à ce qu'on l'on observerait en réalité, néanmoins cela permet tout de même de faire des comparaisons relatives des modèles entre eux.

Dans notre cas on générera les fichiers d'observations (contenant les temps d'arrivée des ondes virtuelles) en se servant du modèle 3D. D'une certaine manière on suppose donc déjà a priori que ce modèle est plus proche de la réalité que les autres modèles. On dispose des localisations exactes et des stations utilisées, et, pour chaque événement, on effectuera une localisation avec NonLinLoc en utilisant chacun des trois modèles².

Concernant la localisation exacte des séismes synthétiques plusieurs méthodes ont été envisagées. Il a dans un premier temps été envisagé d'utiliser les localisations de séismes réels. Certes ces localisations ne correspondent pas aux localisations exactes des vrais séismes, néanmoins les erreurs ne sont en général que de quelques kilomètres, ces localisations permettent donc d'obtenir une répartition des séismes représentatives de la réalité (on aura beaucoup de séismes virtuels dans les zones à sismicité plus importante). Néanmoins deux problèmes interviennent : D'une part il y aura des zones du fossé rhénan où il n'y aura pas ou très peu de séismes, car peu de vrais séismes y ont eu lieu. Mais rien ne permet d'affirmer qu'il n'y en aura jamais qui se produiront dans ces zones, or on souhaite tester l'efficacité des modèles pour tout le fossé rhénan, et pas seulement pour certaines zones en particulier. D'autre part les séismes étant concentrés dans certaines zones on obtiendra des informations redondantes : en effet imaginons que de nombreux séismes se produisent pratiquement au même endroit (des répliques d'un même séisme par exemple), qui, malheureusement, est mal géré par nos modèles : on aura des informations redondantes et des

1 Il existe quelque séismes dont on connaît la localisation avec un très grande précision grâce au recoupement de différentes méthodes, mais ils ne sont pas en nombre suffisant pour pouvoir en tirer des informations par des méthodes statistiques.

2 Il ne faut donc pas confondre les fichiers d'observations synthétiques, créés en utilisant le modèle 3D, et les localisations, qui sont, elles, effectuées avec le modèle 1D simplifié, le modèle 1D complexe, et le modèle 3D.

localisations mauvaises pour ces séismes qui vont donc biaiser les résultats, puisqu'ils donneront un poids plus important à cette zone. Une meilleure solution serait donc d'avoir une répartition homogène des séismes, de manière à couvrir tout le fossé rhénan, tout en évitant d'avoir des concentrations d'événements à certains endroits.

La méthode qui a été utilisée au final consiste à répartir de manière aléatoire des événements dans le fossé rhénan, en nombre suffisant pour obtenir une couverture homogène de la région, mais pas assez élevé pour obtenir de fortes concentrations dans certaines zones au détriment d'autres. L'avantage de cette méthode est qu'elle permet de s'assurer de l'indépendance entre les événements, tout en obtenant une couverture satisfaisante du fossé rhénan. On peut comparer cette méthode à la méthode de Monte-Carlo (Metropolis et Ulam, 1949), procédé aléatoire utilisé couramment, par exemple pour calculer des intégrales en dimensions élevées. Après plusieurs essais le nombre d'événements créés a été fixé à 2000.

Le nombre de stations pour chaque événement a aussi été choisi de manière aléatoire : pour chaque séisme on trace un cercle de rayon aléatoire et on utilise les stations comprises dans ce cercle. Si le cercle est trop petit on augmente son rayon jusqu'à obtenir un nombre de stations satisfaisant (au minimum trois stations). Pour chaque station le programme NonLinLoc déterminera le temps d'arrivée des ondes P et des ondes S. Le nombre de phases (égal au nombre de temps d'arrivées d'ondes utilisés) est donc égal à deux fois le nombre de stations.

Concernant la profondeur, afin d'obtenir des informations détaillées pour chaque profondeur, il a été décidé de faire varier, pour chaque événement synthétique, la profondeur de 1 km à 25 km avec un pas de 1 km. Le choix de 25 km a été fait car les séismes réels du fossé rhénan sont des séismes d'origine crustale¹ et la profondeur de leur hypocentre ne dépasse que très rarement cette valeur.

On obtient donc un jeu de 2000 événements par profondeur, sur 25 profondeurs, qu'on localise à chaque fois avec les trois modèles différents. On a donc 2000 événement par profondeur par modèle, soit 150 000 localisations différentes au total. En pratique quelques problèmes ont eu lieu qui ont empêché l'utilisation de 2000 événements par profondeur par modèle, on a donc recueilli au final 149 100 localisations.

On considérera trois variables :

- l'erreur sur la localisation géographique, égale à la distance en mètres entre l'épicentre calculé par NonLinLoc et l'épicentre réel. Plusieurs fonctions sous R permettent, à partir des longitudes et latitudes de différents points, de calculer la distance en mètres ou kilomètres entre ces deux points. On a ici utilisé la fonction *geoDist* de la bibliothèque *SoDa*.
- l'erreur sur la profondeur, égale à la distance en kilomètres entre la profondeur exacte et la profondeur calculée par NonLinLoc. Une erreur sur la profondeur positive signifie que la profondeur calculée est plus grande qu'en réalité, une erreur sur la profondeur négative signifie que la profondeur calculée est trop faible qu'en réalité.
- L'erreur sur le temps origine du séisme, égale au temps origine exact moins le temps origine calculé par NonLinLoc. Une erreur sur le temps origine positive signifie que la localisation temporelle calculée est antérieure au temps origine exact, une erreur sur le temps origine négative signifie que la localisation temporelle calculée est postérieure au temps origine exact.

Dans la suite on appellera indifféremment *erreur*, *imprécision* ou *différence* (sur la localisation géographique, le temps origine, la profondeur) ces variables. Lorsqu'on parlera des erreurs sur la localisation il s'agira des trois erreurs confondues (localisation géographique, temps origine, et profondeur).

¹ C'est à dire ayant lieu dans la croûte terrestre.

On effectuera dans un premier temps une analyse exploratoire des données, en traçant des histogrammes, boxplot et autres graphiques, et en calculant les moyennes, médianes, et autres indicateurs afin d'avoir une idée du comportement des erreurs. On s'intéressera aussi aux relations avec les autres paramètres ayant une influence sur l'erreur (nombre de phases, gap azimutal, et distance à la station la plus proche). Dans un second temps on fera intervenir des modèles statistiques et on effectuera des analyses de la variance et des tests statistiques (paramétriques et non-paramétriques) afin d'essayer de confirmer ou d'infirmer d'éventuelles différences entre les trois modèles de vitesse des ondes sismiques considérés.

Analyse exploratoire des données

Il est intéressant dans un premier temps d'étudier graphiquement les répartitions des erreurs afin de dégager certains comportements en fonction des modèles. On rappelle que les événements synthétiques sont créés dans le modèle 3D, puis localisés avec chacun des trois modèles (1Ds, 1Dc, 3D), les erreurs de localisation par le modèle 3D ne s'interprètent donc pas de la même manière que celles par le modèle 1D : les erreurs par le modèle 3D sont dues à la qualité du réseau et aux diverses incertitudes (par exemple l'incertitude sur le pointé des temps d'arrivées des ondes sismiques) alors que les erreurs par les modèles 1D intègrent aussi la part d'erreur due aux différences entre les modèles 1D et le modèle 3D. On s'attend à ce que les erreurs par le modèle 3D soient plus faibles et moins dispersées, puisqu'elles n'intègrent pas la part d'erreur due au modèle. Il faudra donc mettre en relation les graphiques des modèles 1D avec ceux du modèle 3D, regarder les différences relatives, et faire attention aux interprétations.

Les histogrammes des erreurs sur la localisation géographique, toutes profondeurs confondues, pour les trois modèles ont été étudiés (figure 2.1, par souci de place seul le modèle 1Dc est représenté, les histogrammes des autres modèles sont similaires).

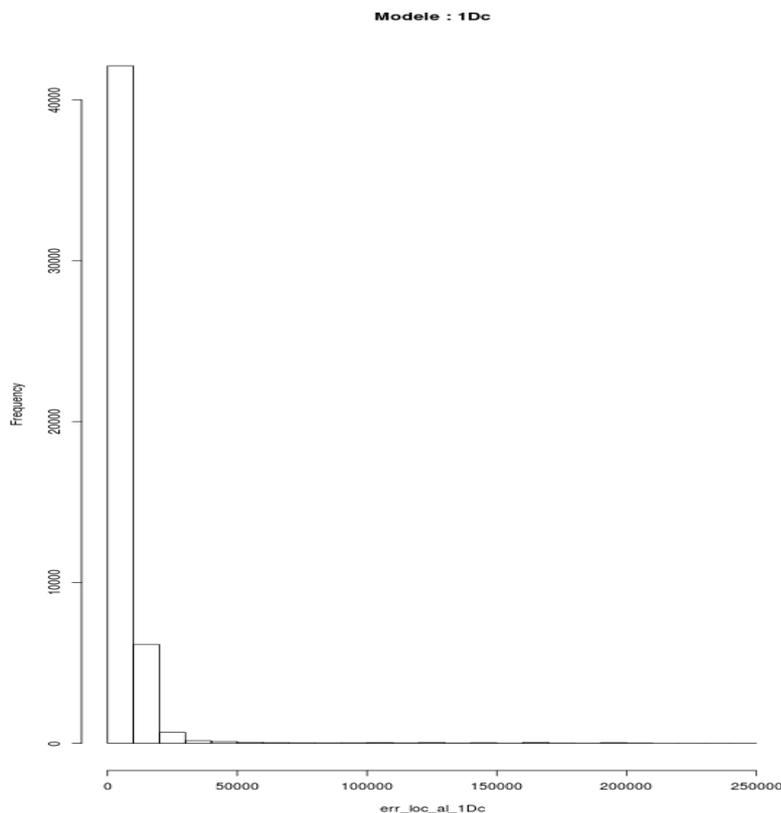


Fig 2.1 - Histogramme des erreurs sur la localisation géographique pour le modèle 1Dc (abscisse en mètres)

Les erreurs sur la localisation géographique sont concentrées en grande partie entre 0 et 10 000 mètres environ, mais les queues de distribution s'étalent sur la droite jusqu'à 150 à 200 km. L'étude des quantiles le confirme :

```
> summary(err_loc_al_1Ds)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 9.73  779.00 1977.00 4985.00 5530.00 228200.00

> summary(err_loc_al_1Dc)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 7.21  823.90 2131.00 5710.00 6841.00 205500.00

> summary(err_loc_al_3D)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 2.92  270.50  491.00  716.00  947.30 84450.00
```

Une transformation logarithmique semble appropriée pour avoir une représentation plus claire de ces erreurs. Le logarithme utilisé est le logarithme en base 10, ainsi la valeur obtenue après transformation pour un événement est égale à l'ordre de grandeur de l'erreur (une valeur de 2 indiquant une erreur de 10^2 mètres, une valeur de 3,5 une erreur de $10^{3,5}$ mètres, etc). Pour tout le reste de l'analyse exploratoire des données on ne s'intéressera plus qu'au logarithme des erreurs sur la localisation géographique. Dans les modèles 1Ds et 1Dc la distribution semble bimodale (figure 2.2, seuls les modèles 1Dc et 3D sont représentés, l'histogramme du modèle 1Ds est similaire à celui du modèle 1Dc) : on observe deux maximums locaux, un premier maximum local à 3 (soit 1 km) et l'autre vers 3,5 pour le modèle 1Ds et vers 4 (soit 10 km) pour le modèle 1Dc, alors que la distribution pour le modèle 3D semble unimodale, avec un maximum un peu avant 3.

L'étude du boxplot du logarithme des erreurs sur la localisation géographique confirme cette différence entre le modèle 3D et les deux autres modèles (figure 2.3).

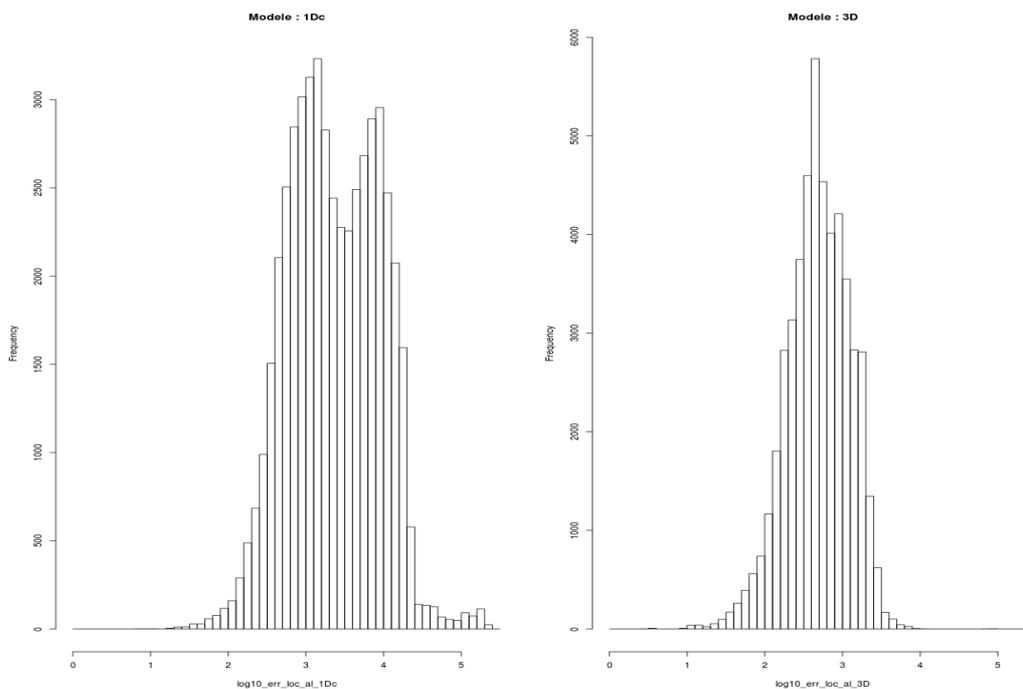


Fig. 2.2 - Histogramme du \log_{10} des erreurs sur la localisation géographique pour les modèles 1Dc et 3D

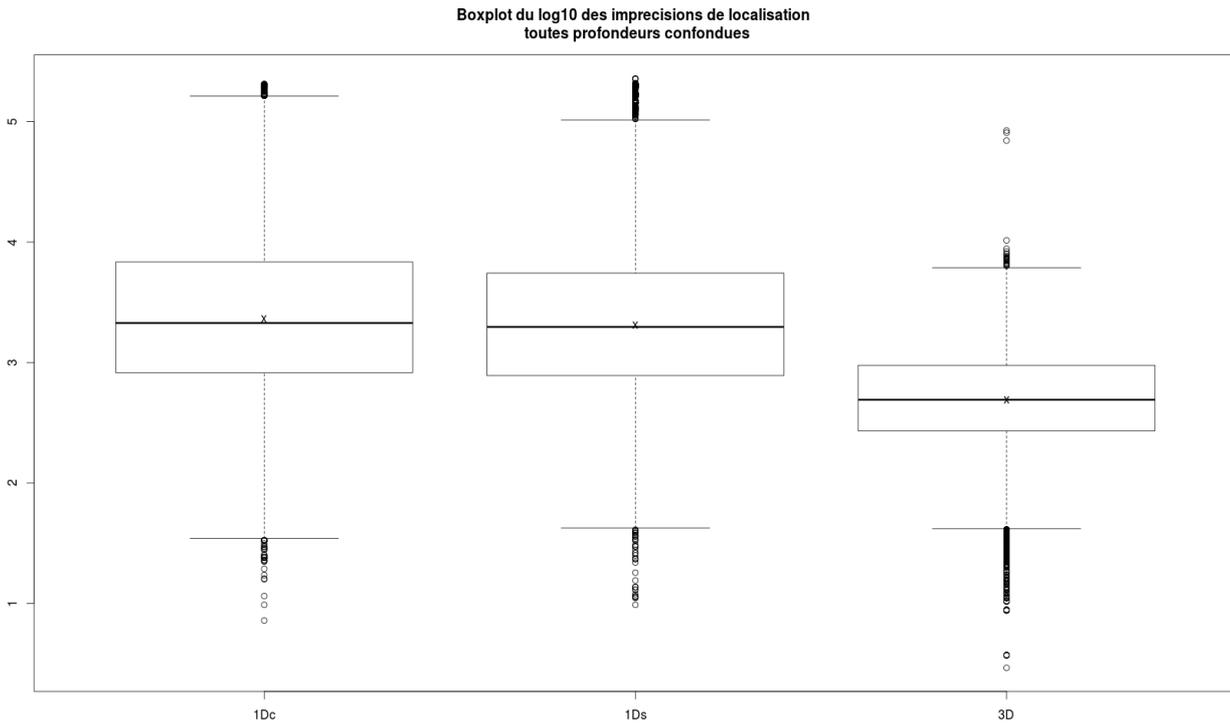


Fig. 2.3 - Boxplot du log10 des erreurs sur la localisation géographique

La bimodalité des erreurs sur la localisation géographique pour les modèles 1Ds et 1Dc s'explique en partie par le fait que l'erreur sur la localisation géographique a tendance à augmenter quand la profondeur augmente, comme le montre la représentation des histogrammes des erreurs sur la localisation géographique pour trois profondeurs (4 km, 13 km et 25 km, figure 2.4, le comportement est le même pour les autres profondeurs). Pour des faibles profondeurs l'erreur sur la localisation géographique est de l'ordre du kilomètre, alors que pour les profondeurs élevées elle est de l'ordre de la dizaine de kilomètres. Pour les profondeurs intermédiaires les erreurs sont réparties entre des erreurs de l'ordre du kilomètre et des erreurs de l'ordre de la dizaine de kilomètres. Cependant l'effet de la profondeur n'explique pas totalement la bimodalité, puisqu'on voit que même à une profondeur faible donnée il subsiste encore un deuxième minimum local, mais moins prononcé qu'auparavant. Le modèle 3D semble peu voire pas affecté par la profondeur.

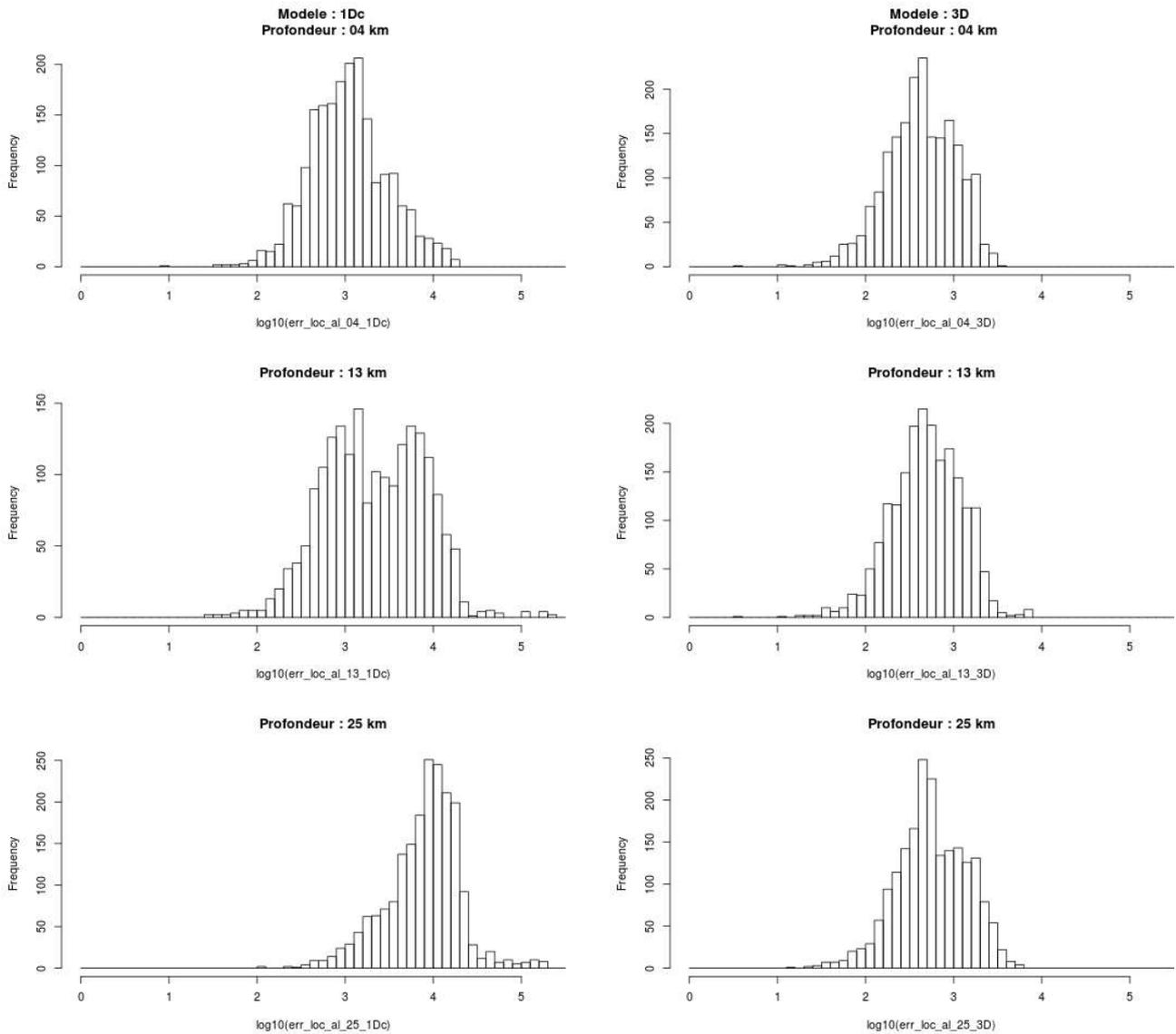


Fig. 2.4 - Histogramme du \log_{10} des erreurs sur la localisation géographique pour trois profondeurs : 4 km, 13 km et 25 km

L'histogramme des erreurs sur le temps origine toutes profondeurs confondues (figure 2.5) montre lui aussi une nette différence entre le modèle 3D et les deux autres modèles 1Ds et 1Dc. Dans le cas du modèle 3D l'erreur est très proche de 0s et est très peu dispersée, alors que l'erreur des modèles 1Ds et 1Dc est plutôt de l'ordre de la seconde, est plus dispersée, et semble même présenter deux maximums locaux. L'étude des quantiles et le tracé des boxplot (figure 2.6) confirment cette différence entre le modèle 3D d'une part et les modèles 1D d'autre part :

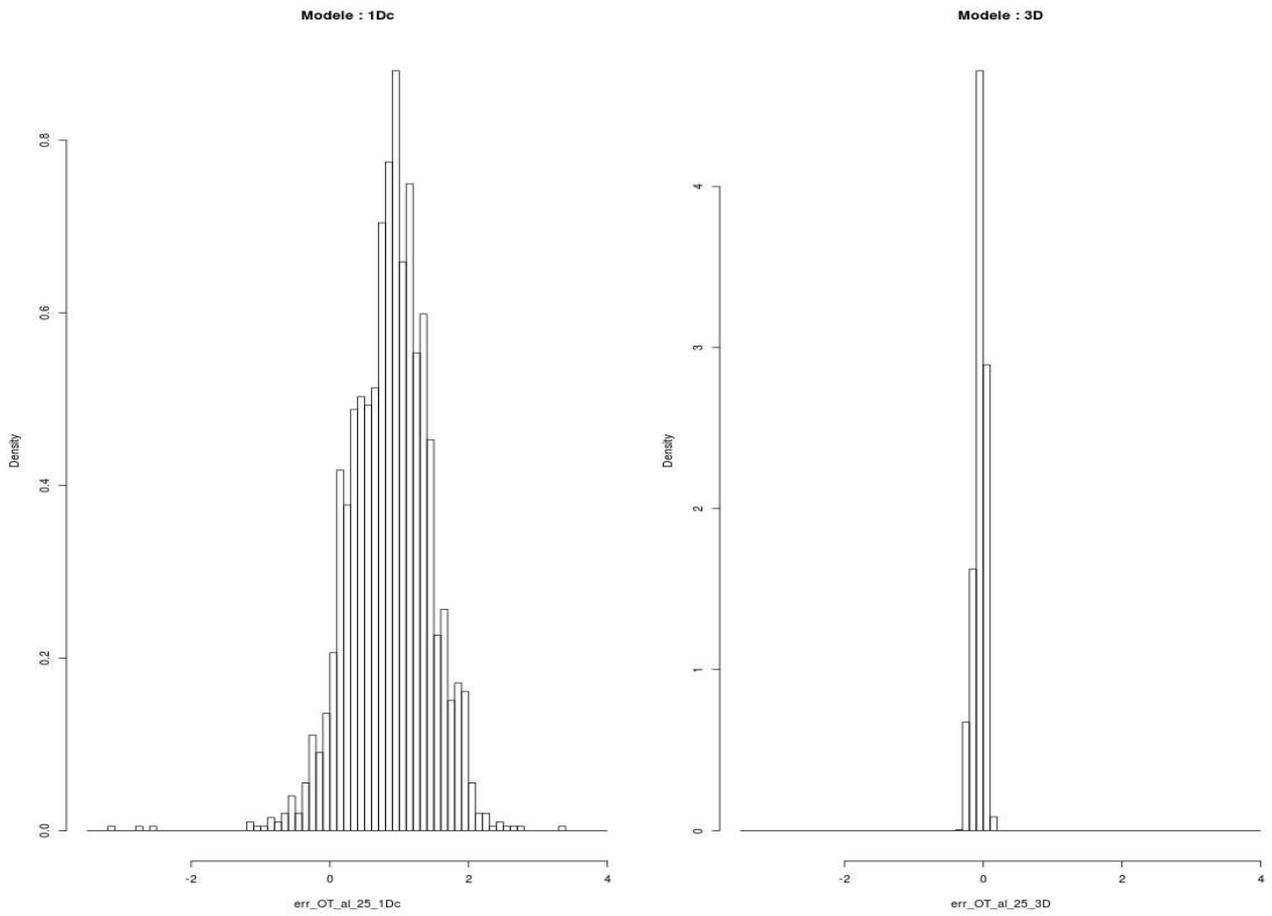


Fig. 2.5 - Histogramme des erreurs sur le temps origine (abscisse en secondes)

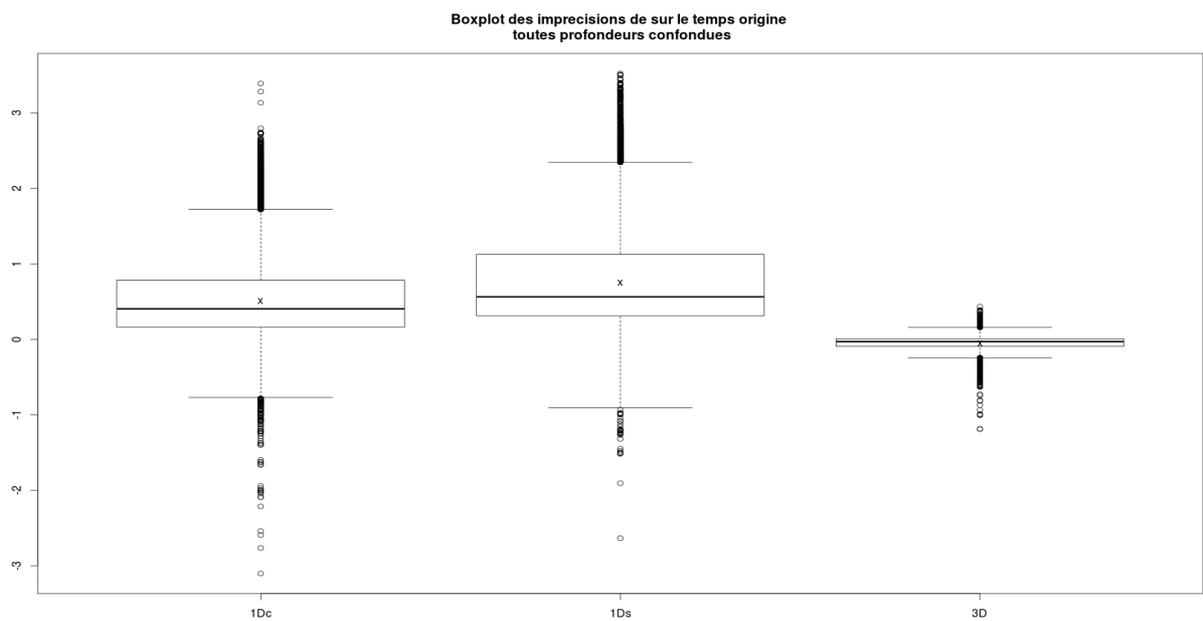


Fig. 2.6 - Boxplot des erreurs sur le temps origine

```

> summary(err_OT_al_1Ds)
Min. 1st Qu. Median Mean 3rd Qu. Max.
-2.6340 0.3149 0.5647 0.7534 1.1270 3.5240
> summary(err_OT_al_1Dc)
Min. 1st Qu. Median Mean 3rd Qu. Max.
-3.1020 0.1620 0.4055 0.5068 0.7867 3.3910
> summary(err_OT_al_3D)
Min. 1st Qu. Median Mean 3rd Qu. Max.
-1.188000 -0.092300 -0.028380 -0.051180 0.008503 0.433100

```

Pour les modèles 1D l'erreur sur le temps origine est le plus souvent positive. Étant donnée la manière dont est calculée l'erreur sur le temps origine (temps origine exact moins temps origine calculé) cela signifie que le temps origine calculé est antérieur au temps origine exact.

Comme précédemment l'erreur sur le temps origine a tendance à augmenter avec la profondeur (figure 2.7).

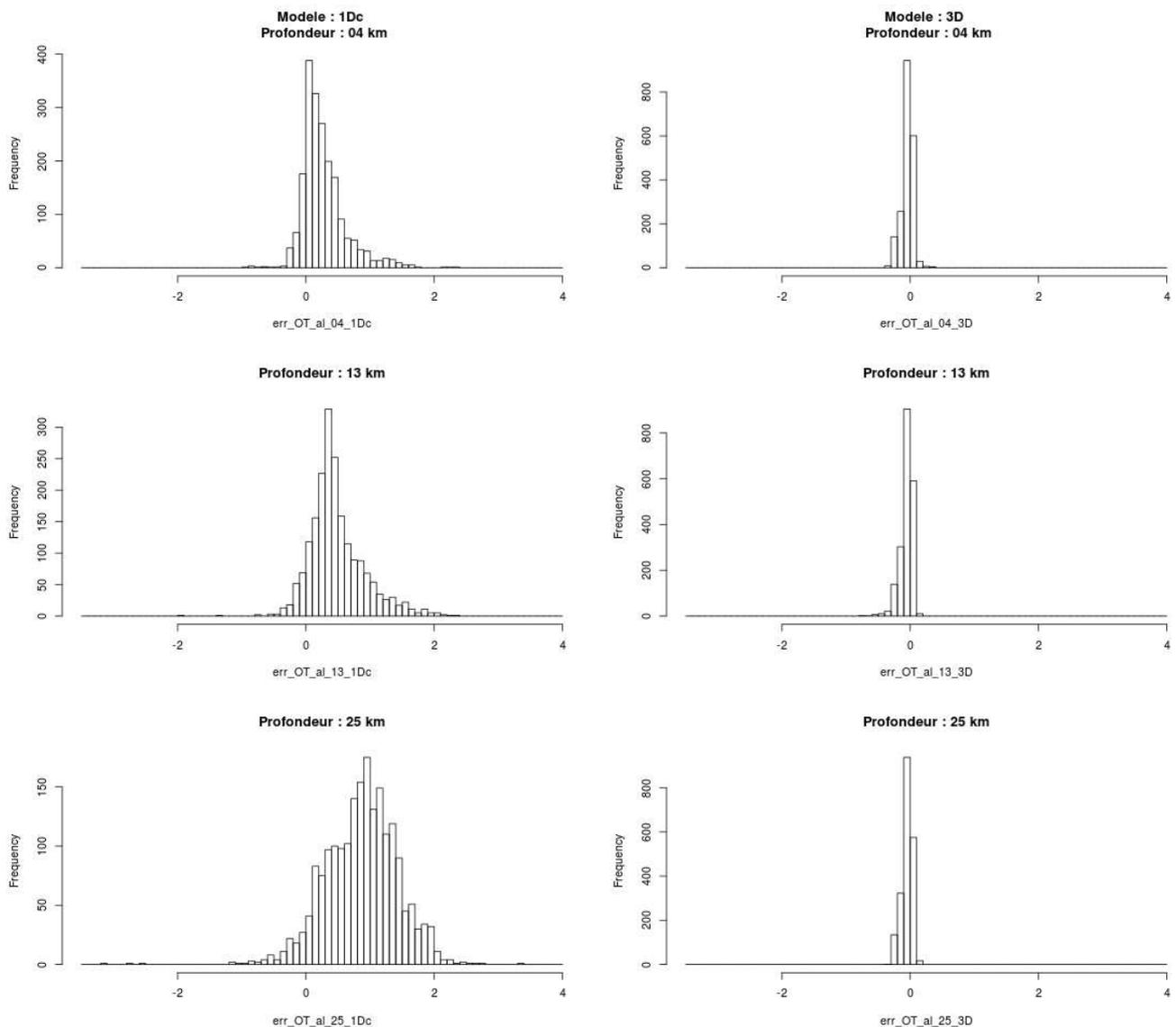


Fig. 2.7 - Histogramme des erreurs sur le temps origine pour trois profondeurs : 4 km, 13 km et 25 km

Concernant l'erreur sur la profondeur il n'y a pas d'intérêt à la regarder toutes profondeurs confondues, puisque l'intervalle des valeurs possibles n'est pas le même en fonction de la profondeur : l'erreur sur la profondeur est bornée, en haut, par la surface et en bas par la limite de la grille spatiale utilisée par le programme¹. Dans notre étude la grille du modèle s'arrête à 31 km. Il faut donc étudier l'erreur sur la profondeur pour chaque profondeur. L'historgramme des erreurs sur la profondeur montre un mauvais comportement des modèles 1D (figure 2.8) : les localisations des événements par ces modèles tendent à être plus profondes que les événements ne le sont réellement, avec une tendance à en placer une partie à la limite inférieure de la grille (c'est à dire à 31 km de profondeur). Ces décalages sont dus aux effets des modèles 1D puisque l'erreur sur la profondeur par le modèle 3D ne présente pas le même comportement. On observe tout de même qu'à de fortes profondeurs les séismes localisés par le modèle 3D sont aussi collés à la limite inférieure de la grille.

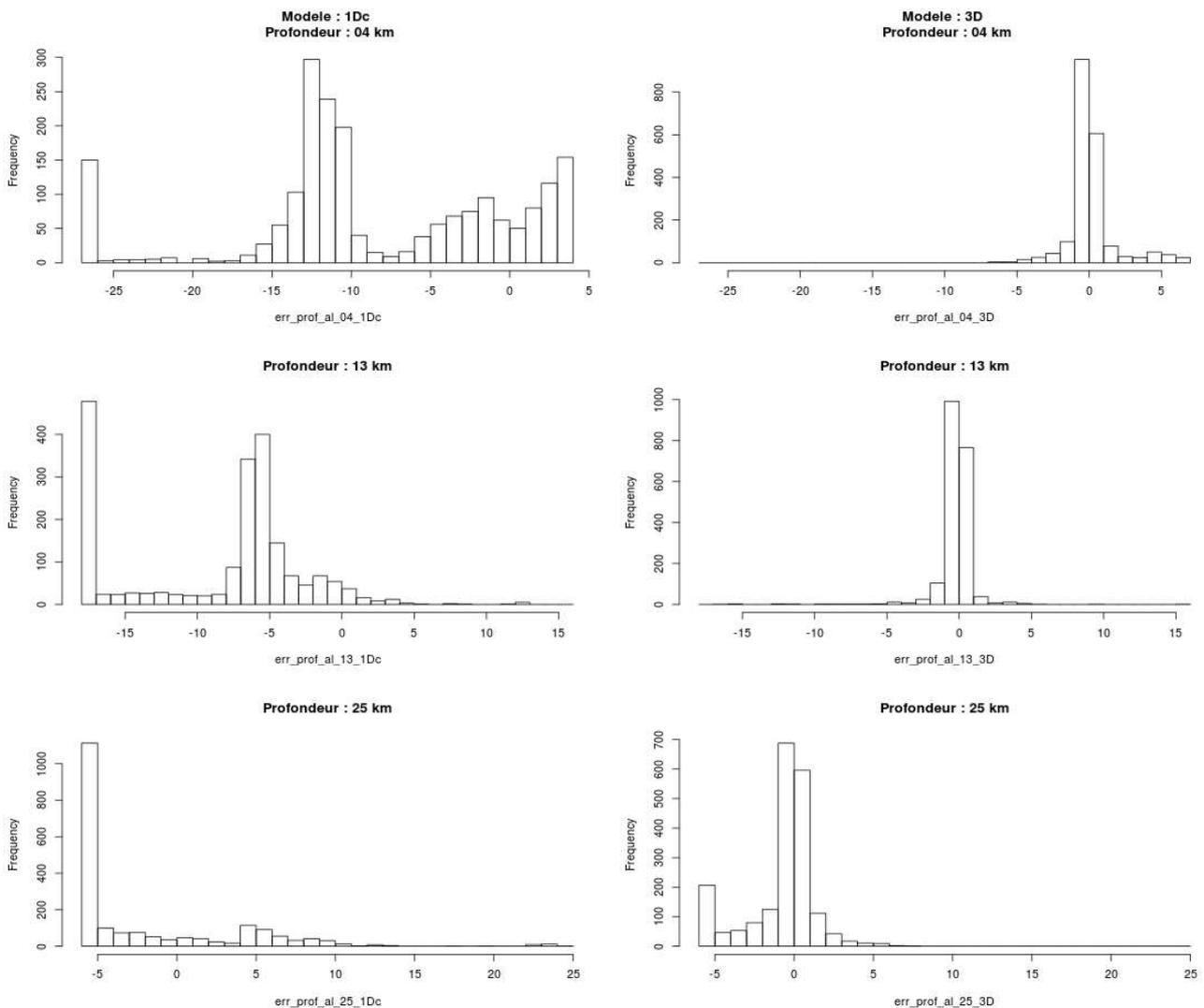


Fig. 2.8 - Histogramme des erreurs sur la profondeur pour trois profondeurs : 4 km, 13 km et 25 km

¹ Lors de la créations des modèles de vitesses par NonLinLoc le programme ne les calcule que jusqu'à une profondeur donnée.

Ces résultats sont cohérents avec ceux trouvés pour l'erreur sur le temps origine, où les localisations calculées sont le plus souvent antérieures aux localisations exacte. Soit deux séismes situés à deux profondeurs différentes (mais à la même longitude et la même latitude) ; pour que les temps d'observations à une station des ondes issues des deux séismes soient les mêmes, il faut que le séisme le plus profond se soit produit avant l'autre, puisque les ondes parcourent une distance plus importante. Ainsi un localisation plus profonde que la localisation exacte, et un temps origine calculé antérieur au temps origine exact vont de paire.

La figure 2.9 résume l'évolution des différentes erreurs par modèle en fonction de la profondeur. On retrouve ce qui a été dit précédemment : les erreurs sur la localisation géographique et le temps origine tendent à augmenter quand la profondeur augmente. Néanmoins les erreurs par le modèle 3D semblent peu voire pas affectées par la profondeur. Il semble donc y avoir des interactions entre le modèle et la profondeur.

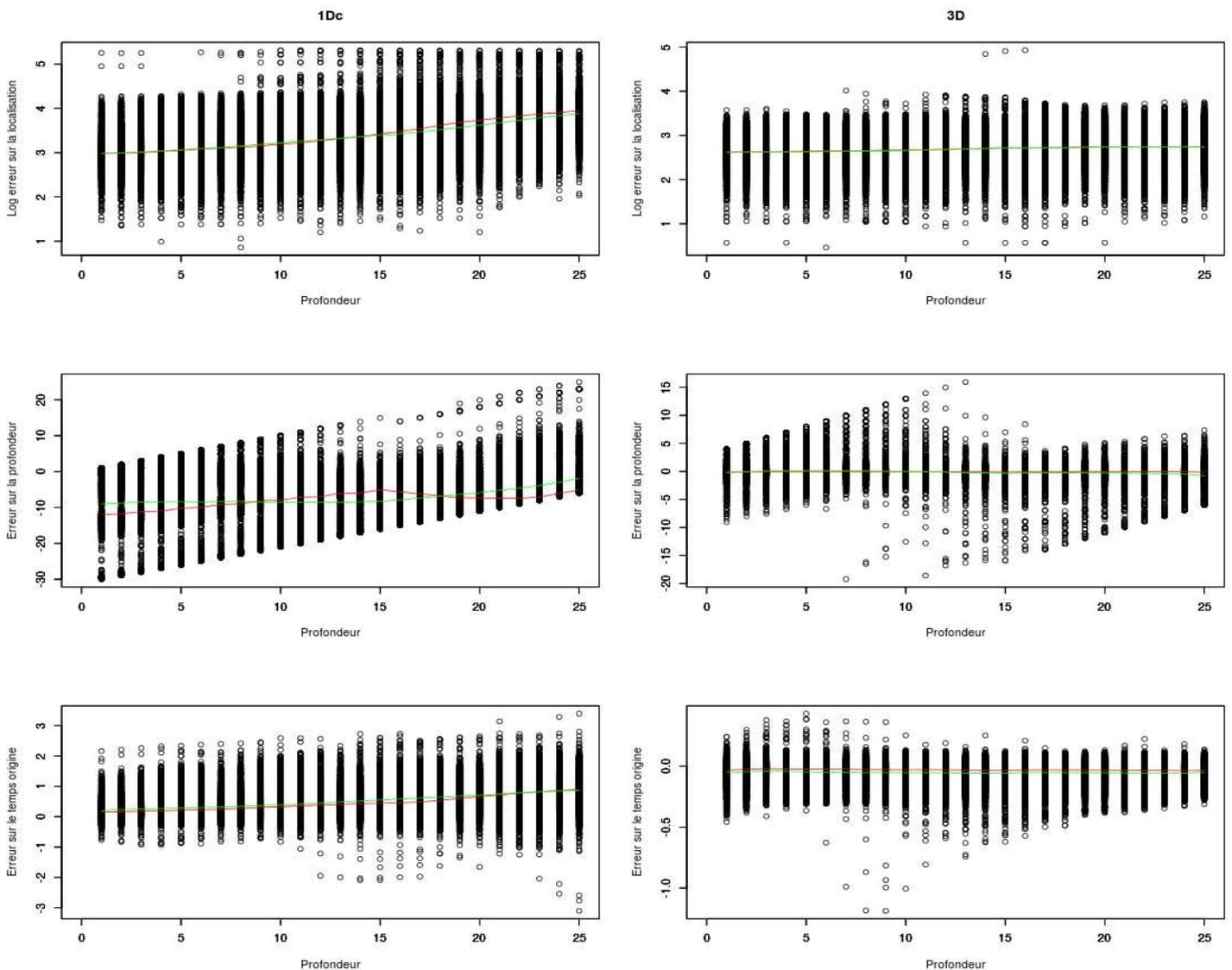


Fig. 2.9 - Les différentes erreurs en fonction de la profondeur. Première ligne : \log_{10} de l'erreur sur la localisation géographique. Deuxième ligne : erreur sur la profondeur. Troisième ligne: erreur sur le temps origine. La ligne rouge correspond à la médiane, la ligne verte à la moyenne.

Influence du Réseau

Il est connu que d'autres paramètres que le modèle de vitesses des ondes ont un effet sur la précision de la localisation des séismes, c'est le cas du nombre de phases, du gap azimuthal et de la distance à la station la plus proche (voir chapitre 1). Les valeurs de ces paramètres étant données par NonLinLoc pour chaque localisation, il est possible de les analyser et d'essayer de dégager des tendances en fonction de ces paramètres.

Il a déjà été dit que les erreurs sur la localisation par le modèle 3D sont dues aux différentes incertitudes ainsi qu'au réseau. Les graphiques des erreurs par le modèle 3D en fonction des paramètres du réseau sont donc plus représentatifs de l'influence du réseau seul. Les graphiques des erreurs par les modèles 1D en fonction des paramètres du réseau permettent d'étudier d'éventuelles interactions entre les modèles et les paramètres du réseau, dans le cas où les comportements seraient différents de ceux par le modèle 3D.

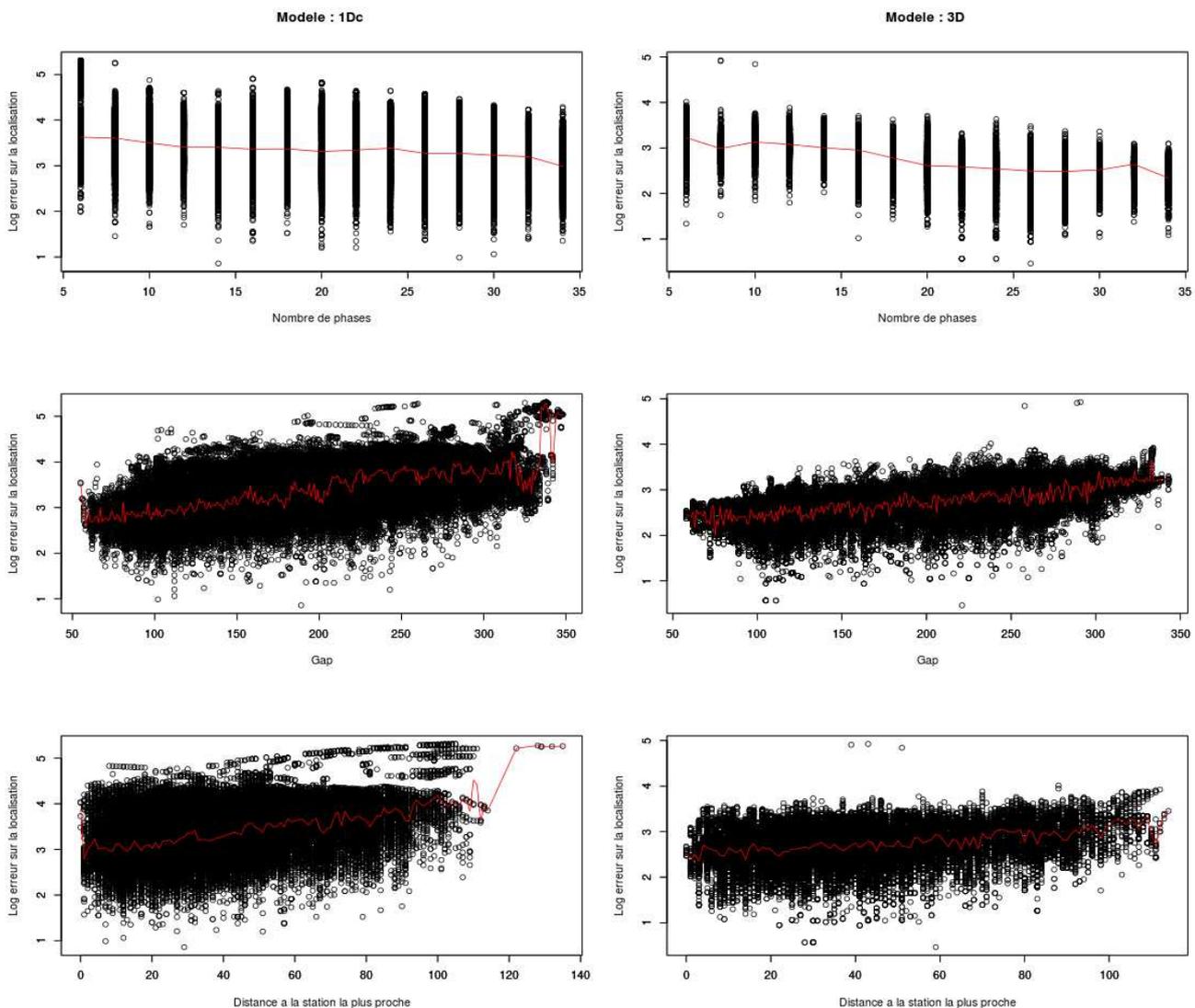


Fig. 2.10 - Log10 de l'erreur sur la localisation géographique en fonction du nombre de phases, du gap et de la distance à la station la plus proche

Une première idée est de tracer les erreurs en fonction de ces paramètres (figure 2.10). On voit que la médiane (trait rouge) semble évoluer de manière linéaire en fonction du nombre de phases, du gap, et de la distance à la station la plus proche. Plus le nombre de phases augmente et moins l'erreur sur la localisation géographique semble grande. Plus le gap azimutal augmente plus l'erreur sur la localisation géographique augmente¹, avec un décrochage pour les valeurs de gap très grande (proche de 360). Cela n'est pas surprenant puisqu'un gap proche de 360 signifie que le séisme est hors réseau. De même l'erreur sur la localisation géographique augmente quand la distance à la station la plus proche augmente. On remarque aussi un certain nombre d'événements répartis approximativement le long d'une droite, sur le graphique du modèle 1Dc, pour lesquels l'erreur sur la localisation géographique est très grande (près de 100 km) même pour une distance à la station la plus proche de l'ordre de la soixantaine de kilomètres. Ces événements sont généralement situés aux extrémités de la zone considérée, ont un nombre de phases faibles, et un gap très élevé². Le dénominateur commun de ces trois paramètres semble être la localisation géographique réelle du séisme : un séisme situé au centre du fossé rhénan aura souvent un gap et une distance à la station la plus proche plus faibles, et un nombre de phases plus grand que les événements situés à l'extérieur du réseau, en bordure du fossé rhénan, où il y a peu de stations sismiques et où celles-ci ne sont réparties que d'un côté du séisme³.

Une représentation géographique des erreurs sur la localisation géographique permet de vérifier cela (figure 2.11 et 2.12). Sur les figures 2.11 et 2.12 l'erreur sur la localisation géographique est représentée par la couleur des points (on considère toujours le logarithme des erreurs sur la localisation géographique), l'échelle va de 0 à 4 (soit 10 km), les erreurs supérieures à 10 km sont plafonnées à 10 km et sont donc représentées par la même couleur. Les points sont placés aux localisations exactes des séismes virtuels. Les événements considérés sont ceux situés à 5 km de profondeur. Le tracé des frontières (bordures noires) permet de se repérer plus facilement.

On voit que plus on s'éloigne du centre du fossé rhénan et plus l'erreur sur la localisation géographique augmente, les événements dont l'erreur sur la localisation géographique est de l'ordre de 10 km ou plus étant situés dans les coins du fossé rhénan, confirmant ainsi l'hypothèse formulée précédemment.

-
- 1 On rappelle que le gap azimutal correspond à l'angle maximal entre deux stations ayant détecté le séisme. Il s'agit donc d'une mesure indiquant si le séisme est bien couvert par les stations ou non.
 - 2 Ce sont les mêmes événements que ceux qui causent le décrochage sur le graphique de l'erreur sur la localisation géographique en fonction du gap.
 - 3 Par exemple un séisme ayant lieu à l'extrémité ouest du fossé rhénan ne sera détecté que par des stations situées à l'est de ce séisme.

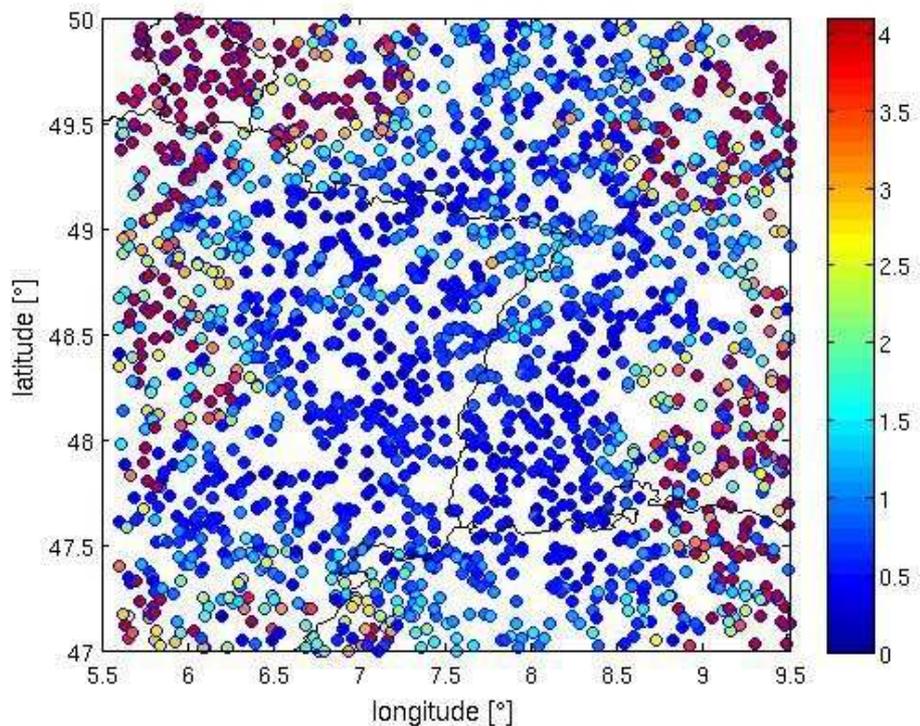


Fig. 2.11 - Représentation géographique des erreurs sur la localisation géographique par le modèle 1Ds pour les événements à 5 km de profondeur

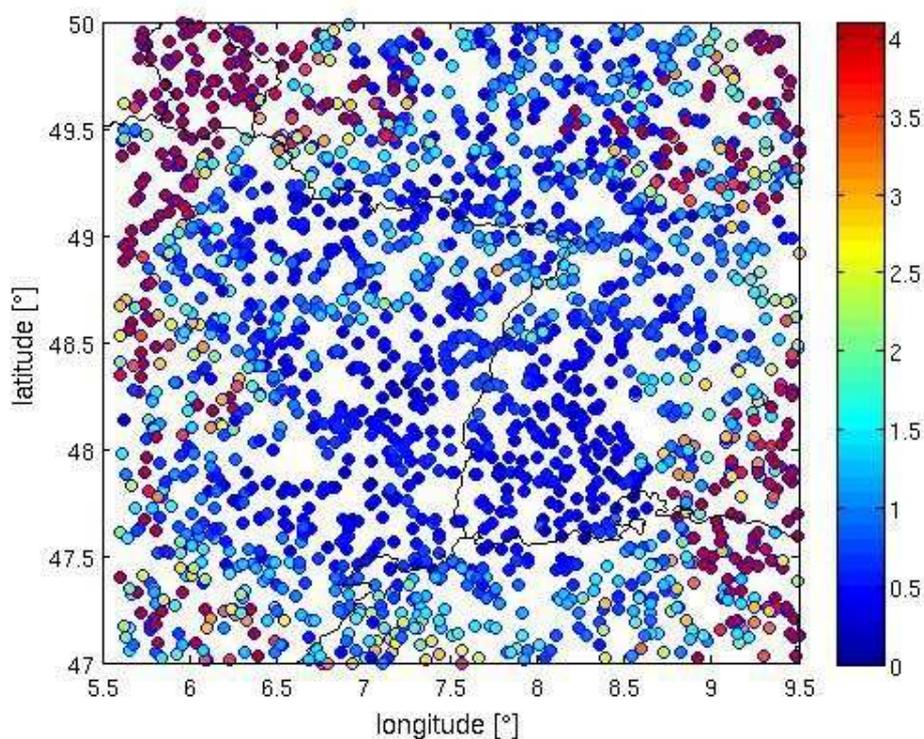


Fig. 2.11 - Représentation géographique des erreurs sur la localisation géographique par le modèle 1Dc pour les événements à 5 km de profondeur

De plus, en comparant les deux modèles 1Ds et 1Dc on observe que pour certaines zones les erreurs sur la localisation géographique semblent légèrement plus faibles pour le modèle 1Ds que pour le modèle 1Dc. Cela se voit bien pour les séismes ayant une latitude entre 8 et 9E et une longitude entre 47 et 47,5N. Cela s'explique par le fait que la couche supplémentaire du modèle 1Dc correspond à une couche sédimentaire située au centre du fossé rhénan mais a priori inexistante aux bords. La généralisation de cette couche à tout le fossé rhénan dans le modèle 1Dc peut donc entraîner une erreur sur la localisation géographique plus importante (le modèle 3D utilisé pour créer les événements limitant la couche sédimentaire au centre du fossé rhénan).

Une autre approche a ensuite été considérée pour tenir compte des paramètres du réseau (nombre de phases, gap, distance à la station la plus proche), en essayant de trouver des seuils, ou des intervalles de valeurs, pour ces paramètres, pour lesquels les comportements des erreurs sont facilement différenciables. Après plusieurs essais une différence apparaît pour l'erreur sur la localisation géographique entre les événements ayant un nombre de phases inférieur ou égal à douze, et ceux ayant un nombre de phases strictement supérieur à douze (figure 2.13). Concernant l'erreur sur le temps origine une différence très nette existe entre les événements ayant un nombre de phases égal à six, et ceux ayant un nombre de phases strictement supérieur à six. Aucun résultat aussi clair n'a pu être mis en évidence pour l'erreur sur la profondeur.

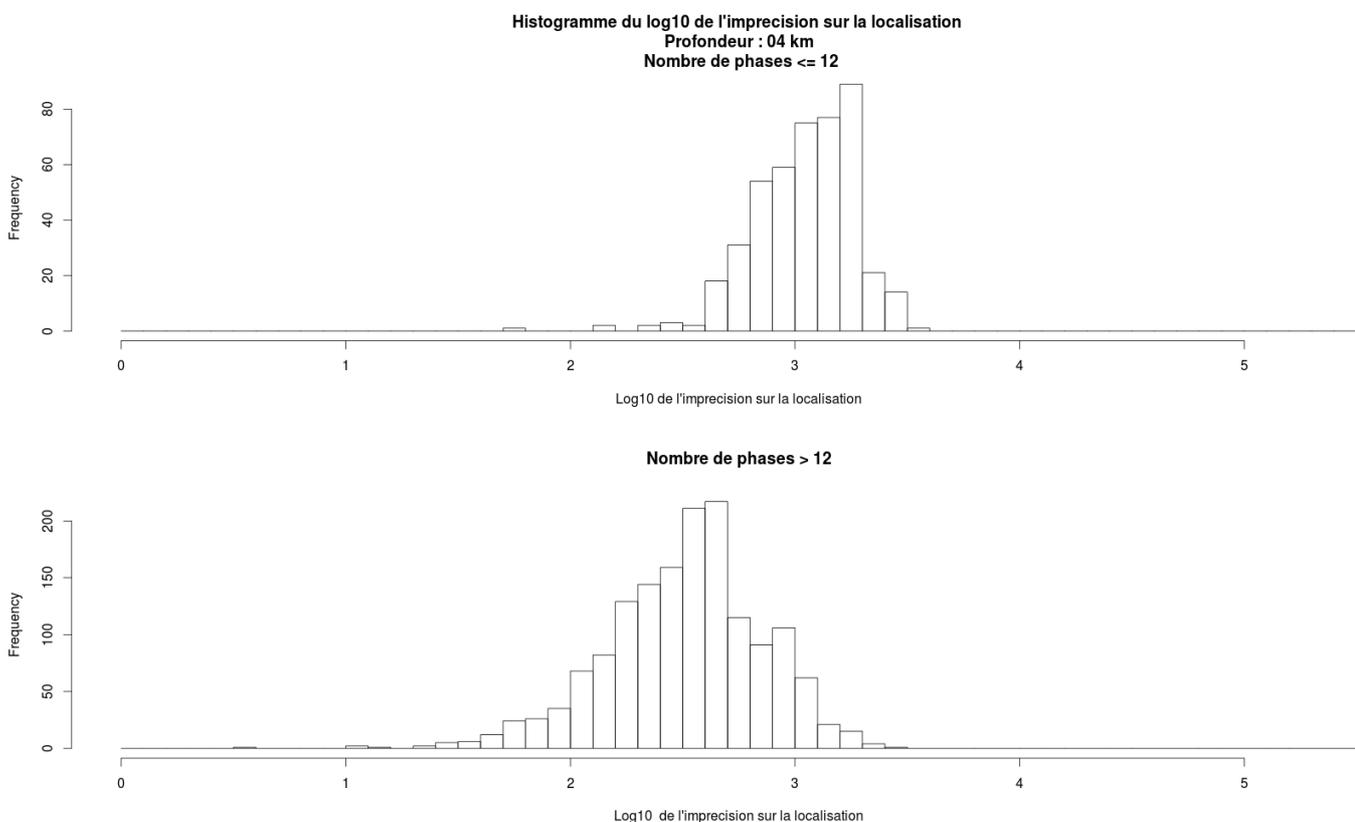


Fig. 2.13 - Histogrammes du log10 des erreurs sur la localisation géographique par le modèle 3D selon le nombre de phases

Les paramètres liés au réseau semblent donc influencer l'erreur sur la localisation géographique. Les erreurs semblent évoluer de manière linéaire en fonction de ces paramètres, néanmoins les différents paramètres étant liés entre eux, et étant liés à la localisation géographique des événements, il est possible de simplifier ces tendances en séparant les événements principalement en fonction du nombre de phases. Il sera intéressant de tenir compte de l'influence de ces paramètres lors d'analyses statistiques plus poussées (voir chapitre 3).

3. Analyse de la variance, comparaisons multiples et tests non paramétriques

Principes théoriques des méthodes utilisées

L'objectif de l'étude est de déterminer l'effet du modèle sur les erreurs dans la localisation des séismes. On dispose de 49700 séismes, 1988 par profondeur pour une profondeur allant de 1 km à 25 km, et pour chacun de ces séismes on effectue la localisation en utilisant les trois modèles (1Ds, 1Dc et 3D). Une analyse de la variance à mesures répétées semble donc appropriée pour traiter notre problème. L'analyse de la variance à mesures répétées se distingue de l'analyse de la variance classique par le fait qu'on ne suppose plus l'indépendance entre tous les événements, mais qu'on tient compte d'un nouveau facteur « sujet », correspondant dans notre cas aux différents séismes, sur lesquels on effectue plusieurs mesures. Les hypothèses sous-jacentes à l'analyse de la variance à mesures répétées sont similaires à celles de l'analyse de la variance classique :

- L'indépendance entre les différents sujets (ici entre les différents séismes), qui remplace l'indépendance entre les observations : puisqu'on effectue plusieurs mesures sur un même sujet cette indépendance entre toutes les observations est impossible, on s'intéresse alors à l'indépendance entre les différents sujets.
- La normalité des résidus, qui peut être vérifiée en utilisant un test de Shapiro-Wilk (ou l'une de ses variantes). Néanmoins le test de Shapiro-Wilk a tendance à systématiquement détecter des écarts à la normalité lorsque l'échantillon est grand. Plutôt que de regarder la normalité des résidus il est aussi possible d'étudier la normalité des observations pour chaque niveau du facteur.
- L'hypothèse de sphéricité, qui remplace l'hypothèse d'homogénéité des variances des observations des différents groupes. L'hypothèse de sphéricité signifie que la variance des différences entre deux mesures doit être la même pour toutes les paires de mesures. Par exemple si l'on mesure une certaine quantité A à trois temps différents $temps1$, $temps2$ et $temps3$, l'hypothèse de sphéricité signifie que la variance des différences des mesures de A entre $temps1$ et $temps2$ est égale à celle entre $temps1$ et $temps3$, et à celle entre $temps2$ et $temps3$. Cette hypothèse peut être vérifiée en utilisant le test de Mauchly (Mauchly, 1940). Néanmoins ce test a tendance à ne pas détecter les écarts à la sphéricité pour les échantillons trop petits et à systématiquement détecter un écart à la sphéricité lorsque l'échantillon est trop grand, son intérêt est donc limité.

Lorsque l'hypothèse de sphéricité n'est pas vérifiée les degrés de liberté classiques des statistiques utilisées dans l'analyse de la variance ne sont plus corrects et peuvent entraîner un mauvais calcul de la p-valeur réelle. Plusieurs corrections existent pour tenir compte des écarts à la sphéricité, comme la correction de Greenhouse-Geiser et la correction de Huynh-Feldt. Ces deux corrections se basent sur l'estimation d'un paramètre ϵ représentant l'écart à la sphéricité. Un paramètre ϵ égal à 1 signifie que la sphéricité est vérifiée, et plus l'écart à la sphéricité est grand plus le paramètre ϵ décroît. Les degrés de liberté sont ensuite multipliés par la valeur de ce paramètre. Ainsi la statistique F n'est pas modifiée, puisqu'elle est le rapport des sommes des carrés divisés par les degrés de libertés, donc les ϵ s'annulent, mais les degrés de liberté de la loi de Fisher utilisée le sont.

Une autre possibilité en cas de violation de l'hypothèse de sphéricité est d'utiliser une analyse de la variance multivariée (ou MANOVA), qui ne requiert pas cette hypothèse. Les recommandations quant à la méthode à utiliser varient d'un auteur à l'autre. La correction de Huynh-Feldt est souvent

considérée comme trop libérale (tendance à repérer des différences significatives là où il n'y en a pas) alors que la correction de Greenhouse-Geiser et l'approche multivariée sont considérées comme trop conservatives (des différences significatives peuvent ne pas être détectées). De manière générale il est recommandé d'utiliser la correction de Huynh-Feldt lorsque l'écart à la sphéricité est faible (*epsilon* supérieur à 0,75), et d'utiliser la correction de Greenhouse-Geiser ou l'approche multivariée lorsque l'écart à la sphéricité est important (*epsilon* inférieur à 0,75).

L'analyse de la variance à mesures répétées est aussi affectée, comme l'analyse de la variance classique, par les valeurs manquantes ou des tailles d'échantillons selon chaque facteur inégales (plans déséquilibrés).

Dans notre cas il serait possible d'effectuer une analyse de la variance à mesures répétées en ajoutant, en plus du facteur à mesures répétées « Modèle » à trois niveaux (« 1Ds », « 1Dc » et « 3D »), un deuxième facteur « Profondeur » à 25 modalités (qui ne serait donc pas à mesures répétées). Néanmoins comme on suspecte une interaction entre le modèle et la profondeur (voir chapitre 2), et qu'on accorde de l'importance aux différences entre modèles en fonction de la profondeur¹, on préférera effectuer une analyse de la variance par profondeur. Il faudra dans ce cas faire attention au seuil de signification.

Plusieurs fonctions existent sous R pour effectuer une analyse de la variance à mesures répétées. Ainsi il suffit d'ajouter « + *Error*(...) » dans la fonction *aov* pour ajouter l'effet aléatoire du sujet, la syntaxe étant alors du type :

```
> aov_loc_01 <- aov(err_loc_al_01 ~ Modele + Error(Seismes/Modele))
```

On obtient ensuite le tableau classique de l'analyse de la variance grâce à la fonction *summary*.

Une autre fonction utilisable est la fonction *lme* de la bibliothèque *nlme*. Cette fonction permet de créer des modèles linéaires à effets mixtes (*linear mixed-effects models*), les effets aléatoires (ici ceux du sujet) étant introduits par l'option *random*. La syntaxe est alors du type :

```
> lme_loc_01 <- lme(err_loc_al_01 ~ Modele, random = ~1|Seismes)
```

Les résultats d'une régression linéaire par cette fonction s'obtiennent grâce à la fonction *summary*, et le tableau classique de l'analyse de la variance grâce à la fonction *anova* ou bien grâce à la fonction *Anova* de la bibliothèque *car*.

Une troisième méthode consiste à utiliser la fonction *lmer* de la bibliothèque *lme4* écrite par Douglas Bates. Cette fonction permet notamment de gérer les modèles linéaires à effets mixtes, ainsi que les plans déséquilibrés. L'introduction d'un facteur à effet aléatoire se fait par une formule de type « + (1|Sujets) », la syntaxe étant alors du type :

```
> lmer_loc_01 <- lmer(err_loc_al_01 ~ Modele + (1|Seismes))
```

Si on utilise ensuite la fonction *summary* on se rend compte que les p-valeurs ne sont pas affichées. L'auteur de la fonction s'en justifie par le fait que la fonction *lmer* permet de gérer les plans déséquilibrés et les modèles à effets mixtes, mais que dans ce cas les degrés de liberté des statistiques peuvent être modifiés, les p-valeurs données classiquement par d'autres logiciels étant

¹ Les séismes du fossé rhénan se produisant le plus souvent à des profondeurs faibles il est plus intéressant de savoir pour quelles profondeurs un modèle est meilleur qu'un autre, plutôt que de savoir quel est le meilleur modèle toutes profondeurs confondues.

alors fausses¹. Il est néanmoins possible de récupérer ces p-valeurs grâce à la fonction *Anova* de la bibliothèque *car* écrite par John Fox. Cette fonction fournit par défaut les sommes des carrés de type II, et non les sommes des carrés de type III comme le font d'autres fonctions et d'autres logiciels. Un article intéressant de John Fox sur le sujet peut être trouvé en ligne².

Enfin une quatrième méthode consiste à nouveau à utiliser la fonction *Anova* de la bibliothèque *car*, en l'appliquant cette fois-ci à un objet de type modèle linéaire multivarié. Pour cela on commence par créer un facteur avec un niveau par modèle de vitesses des ondes sismiques, puis on crée une matrice de nos variables grâce à la fonction *cbind*, on crée ensuite le modèle linéaire multivarié, et enfin on applique la fonction *Anova*, en spécifiant le facteur sujet grâce aux options *idata* et *idesign*:

```
> rmodele <- factor(c("1Ds", "1Dc", "3D"))
> err_loc_al_01m <- cbind(err_loc_al_01_1Ds, err_loc_al_01_1Dc, err_loc_al_01_3D)
> mlm_loc_01 <- lm(err_loc_al_01m ~ 1)
> mlm_loc_01.aov <- Anova(mlm_loc_01, idata = data.frame(rmodele), idesign = ~rmodele)
```

Ici puisqu'il n'y a qu'un facteur dans le modèle les sommes des carrés de type II et les sommes des carrés de type III sont les mêmes. Elles seraient différentes si on introduisait un deuxième facteur (par exemple si on regardait l'erreur toutes profondeurs confondues et qu'on rajoutait le facteur « profondeur »).

On récupère ensuite les informations grâce à la fonction *summary*. En donnant la valeur *TRUE* à l'option *multivariate* on obtient les résultats de l'analyse de la variance multivariée. Cette méthode présente l'avantage de fournir les résultats selon les différentes corrections possibles : On obtient les statistiques de test et les p-valeur en supposant la sphéricité, avec la correction de Greenhouse-Geiser, avec la correction de Huynh-Feldt, et par analyse multivariée. On obtient aussi les résultats du test de sphéricité de Mauchly.

Si un effet significatif du facteur est trouvé on peut ensuite effectuer des comparaisons multiples pour déterminer entre quels niveaux du facteurs se trouvent les différences. Dans le cas de l'analyse de la variance à mesure répétées il est déconseillé d'utiliser la méthode de Tukey HSD (Honestly Significant Difference) car celle-ci est sensible aux écarts à la sphéricité. Il est généralement recommandé d'utiliser des tests de Student appariés, pour lesquels la statistique utilisée est la différence des moyennes qui est supposée suivre une loi normale, avec une correction de Bonferroni pour ajuster le seuil de signification.

Pour tenir compte des paramètres liés au réseau (nombre de phases, gap azimutal, distance à la station la plus proche) plusieurs approches sont possibles.

La première consiste à effectuer une analyse de la covariance (ou ANCOVA). L'analyse de la covariance est un mélange de l'analyse de la variance et de la régression linéaire : dans le cas d'une analyse de la variance les variables explicatives sont qualitatives alors que dans une régression linéaire elles sont quantitatives, mais dans une analyse de la covariance on introduit à la fois des variables qualitatives et des variables quantitatives. L'idée de l'analyse de la covariance est d'introduire des variables quantitatives qu'on pense être liées à la variable dépendante mais dont l'estimation de l'effet n'est pas l'intérêt premier. En introduisant ces variables quantitatives on diminue la variance résiduelle (puisque la variance due aux variables quantitatives est désormais

1 <https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html>

2 <https://stat.ethz.ch/pipermail/r-help/2010-March/230280.html>

identifiée et est isolée), de ce fait on diminue le dénominateur de la statistique F, augmentant ainsi la valeur de F et donc la puissance du test.

La seconde approche consiste à grouper les variables liées au réseau par intervalles, afin de les transformer en facteur au lieu de les voir comme des variables quantitatives. Cette approche peut-être intéressante dans le cas où l'on pense que certains paliers existent tels que la variable dépendante aura un comportement différent si le paramètre est en-dessous du palier plutôt que s'il est au-dessus, et qu'au contraire il n'y aura pas de différence de comportement importante au sein d'un même intervalle de ce paramètre. Par exemple on pourrait grouper le gap en plusieurs intervalles, comme $[0,110[$, $[110,180[$ et $[180,360[$, si on pense que le comportement des erreurs de localisation sera différente selon la plage de valeurs dans laquelle on se trouve, mais qu'il ne sera pas différent au sein de ces intervalles.

Du fait des corrélations entre le nombre de phases, le gap et la distance à la station la plus proche, et du fait de mon manque d'expérience concernant l'ANCOVA, en particulier dans le cas de mesures répétées, la seconde méthode a été préférée. On ne s'intéressera qu'au nombre de phases, qu'on séparera en groupes selon les résultats du chapitre 2 : concernant l'erreur sur la localisation géographique on regardera les événements ayant un nombre de phases strictement supérieur à douze et ceux ayant un nombre de phases inférieur ou égal à douze ; concernant l'erreur sur le temps origine on séparera les événements ayant six phases des autres.

Dans un dernier temps on effectuera des tests non-paramétriques, afin de voir si les résultats vont dans le même sens que ceux trouvés par l'analyse de la variance. L'équivalent non-paramétrique de l'analyse de la variance à mesures répétées est le test de Friedman (Friedman, 1937). Tout comme d'autres tests non-paramétriques, comme le test de Kruskal-Wallis ou le test de Mann-Whitney, le test de Friedman se base sur les rangs des observations. Pour chaque observation on remplace la valeur mesurée par son rang (si on a effectué trois mesures on attribuera donc les rangs 1, 2 et 3). S'il n'y a pas de différence significatives entre les mesures les rangs devraient être distribués équitablement. D'après le théorème central limite la moyenne des rangs pour chaque mesure suit approximativement une loi normale, on construit donc une statistique de test suivant une loi du χ^2 à partir de ces moyennes des rangs pour déterminer s'il y a une différence significative entre les moyennes des rangs. Le test de Friedman ne teste donc pas la même chose que l'analyse de la variance, et présente l'inconvénient de ne pas fournir de mesure des différences, puisqu'il remplace les valeurs mesurées par les rangs. En revanche il présente l'avantage de détecter des différences qu'une analyse de la variance ne remarquerait pas. Imaginons par exemple que toutes les mesures selon un niveau du facteur sont presque systématiquement décalées d'un côté (par exemple qu'elles sont systématiquement supérieures aux mesures selon les autres niveaux du facteur). Si cet écart systématique est compensé dans la moyenne des différences par une mesure fortement inférieure aux mesures selon les autres niveaux, l'analyse de la variance classique ne détectera pas de différence significative, en revanche le test de Friedman étant basé uniquement sur les rangs, il détectera que les rangs sont systématiquement plus grands pour cette mesure que pour les autres, et détectera une différence significative. On voit donc que le test de Friedman et l'analyse de la variance ne testent pas tout à fait la même chose, et en ce sens ils peuvent permettre d'obtenir des

résultats complémentaires. Dans notre cas il peut être intéressant d'effectuer un test de Friedman pour savoir si un modèle de vitesse des ondes entraîne des erreurs sur la localisation géographique systématiquement plus grandes ou plus faibles.

Pour effectuer un test de Friedman on peut utiliser simplement la fonction par défaut *friedman.test*, mais on lui préférera la fonction *friedman* de la bibliothèque *agricolae* qui effectue en plus les comparaisons post-hoc.

Vérification des conditions d'application de l'analyse de la variance à mesures répétées

D'après le protocole expérimental, et étant donné que les séismes virtuels sont issus d'une répartition aléatoire et ont un nombre de phases aléatoire, l'indépendance des séismes entre eux est crédible. Même dans le cas de deux séismes situés pratiquement au même endroit, cette coïncidence serait due au hasard, et le nombre de phases étant lui aussi aléatoire et ayant une influence sur les erreurs de localisation, l'indépendance demeure crédible.

L'hypothèse de normalité ne semble pas être vérifiée, au vu des histogrammes vus précédemment. Pour l'erreur sur la localisation géographique on a vu que les distributions s'étaient énormément sur la droite et qu'il fallait une transformation logarithmique pour obtenir une distribution plus proche de la normalité, comme le confirme l'étude des QQ-plot (ou droite de Henry) des résidus (figure 3.1 et 3.2). Pour l'erreur sur le temps origine, même si les histogrammes semblent plus proches d'une loi normale, on s'en écarte encore sensiblement. De même pour l'erreur sur la profondeur, l'hypothèse de normalité semble peu probable, surtout à des profondeurs élevées.

On pourrait être tenté, dans le cas de l'erreur sur la localisation géographique, d'utiliser le logarithme, puisque la distribution des résidus ne semble pas beaucoup s'écarter d'une loi normale (figure 3.2). Néanmoins cela pose des problèmes d'interprétation des résultats : si un effet significatif du modèle de vitesses des ondes est trouvé et que des comparaisons multiples sont faites, les résultats porteront sur la moyenne des différences des logarithmes des erreurs sur la localisation géographique, autrement dit sur la moyenne du logarithme des rapports des erreurs. Une interprétation concrète de ces résultats serait donc difficile, la question que l'on se pose étant « Y a-t-il un effet des modèles de propagation des ondes sur les erreurs de localisation, et si oui de quel ordre ? ».

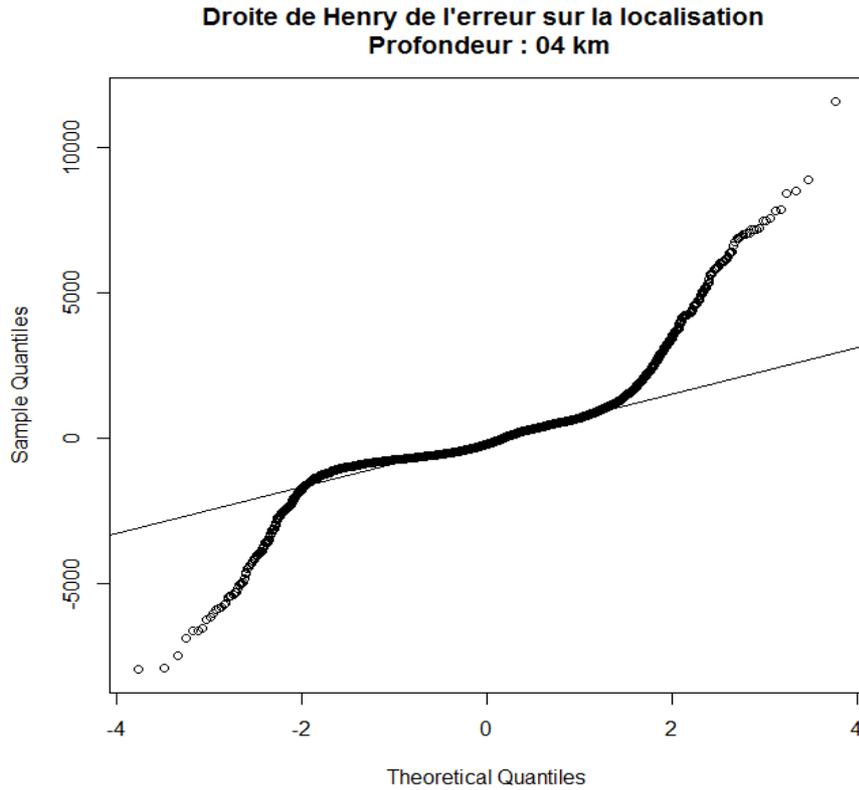


Fig. 3.1 - QQ-plot des résidus de l'erreur sur la localisation géographique

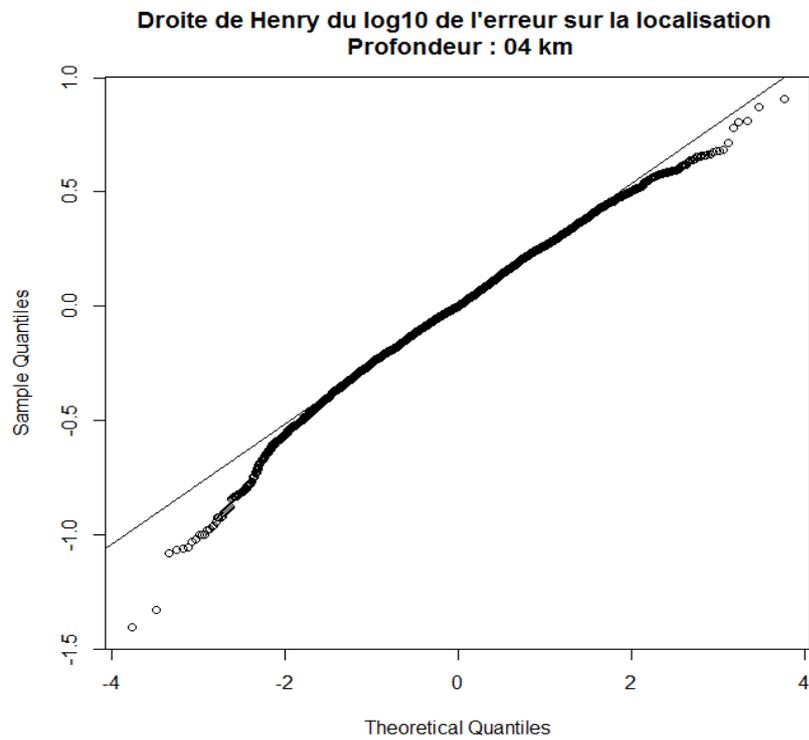


Fig. 3.2 - QQ-plot des résidus du log10 de l'erreur sur la localisation géographique

Il a finalement été décidé de conserver les erreurs sous leur forme originale, afin d'avoir une interprétation immédiate des résultats, quitte à s'écarter de manière importante de l'hypothèse de normalité (surtout pour l'erreur sur la localisation géographique). Cette décision s'appuie sur le fait que l'on dispose du même nombre d'observations par niveau du facteur (plan équilibré) et que le nombre d'observations est très grand : il a été montré que dans ce cas les écarts à la normalité, même importants, n'invalident pas les tests de l'analyse de la variance et les tests de Student. On citera, entre autres, les travaux sur le sujet de Pearson (1931), Box et Anderson (1955), Glass (1972), Harwell (1992), Berkovits, Hancock et Nevitt (2000), et Lumley, Diehr, Emerson et Chen (2003). Des distributions encore plus éloignées de la normalité que les nôtres ont été utilisées dans certains de ces articles, mais les effets sur les risques de première et seconde espèce se sont avérés faibles voire négligeables dans le cas de plans équilibrés et d'une taille suffisamment grande des échantillons. On se rapportera notamment à Lumley, Diehr, Emerson et Chen (2003), où une distribution ayant un coefficient de dissymétrie égale à 8,8 et un kurtosis de 131 a été utilisée et où les résultats étaient valides pour un échantillon de cinq cent observations. Dans notre cas le kurtosis à par exemple 4 km de profondeur est au pire de 12,4815 (pour le modèle 1Dc) et le coefficient de dissymétrie de 3,20 (pour le modèle 1Dc).

L'hypothèse de sphéricité est rarement vérifiée dans l'analyse de la variance à mesures répétées. Le test de Mauchly n'est pas intéressant dans notre cas puisque ce test est connu pour détecter systématiquement des écarts à la sphéricité lorsque la taille des observations est trop importante. On peut cependant regarder les variances des différences afin d'avoir une idée de la plausibilité de l'hypothèse de sphéricité ou non. Ainsi à 4 km de profondeur, on obtient les résultats suivant :

```
> sd(err_loc_al_04_1Ds - err_loc_al_04_1Dc)
[1] 1127.889
> sd(err_loc_al_04_1Ds - err_loc_al_04_3D)
[1] 2371.346
> sd(err_loc_al_04_1Dc - err_loc_al_04_3D)
[1] 2356.484
```

On regarde ici l'écart-type plutôt que la variance, mais cela revient pratiquement au même. On voit que l'écart-type des différences entre le modèle 1Ds et 1Dc est environ deux fois plus petit que l'écart-type entre le modèle 1Ds et 3D, et l'écart-type entre le modèle 1Dc et 3D, la variance est donc environ 4 fois plus petite. Les résultats sont similaires pour les autres profondeurs.

Pour l'erreur sur le temps origine on obtient, à 5 km de profondeur, les résultats suivants :

```
> var(err_OT_al_05_1Ds - err_OT_al_05_1Dc)
[1] 0.09274526
> var(err_OT_al_05_1Ds - err_OT_al_05_3D)
[1] 0.2526627
> var(err_OT_al_05_1Dc - err_OT_al_05_3D)
[1] 0.1260627
```

De même pour l'erreur sur la profondeur, à 5 km de profondeur :

```
> var(err_prof_al_05_1Ds - err_prof_al_05_1Dc)
[1] 15.84332
> var(err_prof_al_05_1Ds - err_prof_al_05_3D)
[1] 47.67053
> var(err_prof_al_05_1Dc - err_prof_al_05_3D)
[1] 60.0717
```

Là encore on a des différences importantes entre les variances des différences, avec des résultats similaires pour les autres profondeurs.

Il est donc peu probable que l'hypothèse de sphéricité soit vérifiée. On s'intéressera aux résultats des différentes méthodes et corrections afin de vérifier si elles sont cohérentes entre elles et si on obtient des résultats similaires et des p-valeurs proches avec toutes les méthodes. Pour cela on utilisera, d'après ce qui a été dit précédemment, la fonction *Anova* de la bibliothèque *car* appliquée à un modèle linéaire multivarié, et l'option *multivariate=TRUE* de la fonction *summary*.

L'étude des valeurs aberrantes ou non-représentatives a aussi été faite. Aucune valeur aberrante ne semble présente : mêmes les quelques erreurs très élevées (notamment les erreurs de localisation géographique de 200 km) ne sont pas dues à un dysfonctionnement du programme NonLinLoc où à une mauvaise saisie des données, mais sont un résultat possible, d'autant plus qu'il a été vu précédemment que ces événements sont en général situés à l'extérieur du réseau et ont un nombre de phases faibles. Le nombre de valeurs extrêmes est conforme à ce qui est attendu : si on s'intéresse aux boîtes à moustaches des erreurs et que l'on regarde la proportion de valeurs extrêmes¹ on obtient le résultat suivant :

```
> length(boxplot(err_loc_al_1Ds)$out)/49700
[1] 0.06822938
> length(boxplot(err_loc_al_1Dc)$out)/49700
[1] 0.06108652
> length(boxplot(err_loc_al_3D)$out)/49700
[1] 0.04865191
```

(49700 étant le nombre d'observations total par modèle)

Ce résultat est cohérent avec la théorie qui dit que cette proportion tend vers 0,7% quand l'effectif de l'échantillon tend vers l'infini. Les proportions sont encore plus faibles pour l'erreur sur le temps origine et l'erreur sur la profondeur.

1 C'est à dire des observations en dehors de l'intervalle $[Q1 - 3/2*IQR, Q3 + 3/2*IQR]$, où Q1 et Q3 sont respectivement le premier et le troisième quartile de l'échantillon, et IQR l'étendue interquartile égale à Q3-Q1

Résultats

Les résultats des analyses de la variance ont été consignés dans des tableaux. La première colonne donne la profondeur à laquelle l'analyse de la variance a été effectuée, la deuxième donne la valeur de la statistique F en supposant la sphéricité, la troisième donne la p-valeur en supposant la sphéricité, la quatrième donne l'estimation de *epsilon* pour la correction de Greenhouse-Geiser, la cinquième donne la p-valeur pour la correction de Greenhouse-Geiser, la sixième donne l'estimation de *epsilon* pour la correction de Huynh-Feldt, la septième donne la p-valeur pour la correction de Huynh-Feldt, la huitième donne la statistique de Wilk pour l'analyse de la variance multivariée, la neuvième donne la p-valeur associée à la statistique de Wilk¹.

Le seuil de signification est fixé à 0,2% (soit 0,002). Cela peut sembler très restrictif, néanmoins étant donné qu'on effectue un grand nombre d'analyse de la variance (une par profondeur, soit vingt-cinq en tout), on cherche à ce que le risque de commettre au moins une erreur soit assez faible. On applique la correction de Bonferroni, qui consiste à diviser le seuil de signification que l'on souhaite avoir par le nombre de tests effectués. Cette méthode est généralement utilisé pour les comparaisons multiples mais peut tout aussi bien être utilisé dans notre cas. Si l'on se fixe un risque d'erreur de premier espèce général de 5%, soit 0,05, on obtient bien un seuil de signification par profondeur de 0,2%, soit 0,002. On peut se demander si un risque d'erreur de premier espèce général de 5% n'est pas non plus trop restrictif : comme on effectue une analyse de la variance pour chaque profondeur allant d'un kilomètre à vingt cinq kilomètres, commettre une erreur de première espèce parmi les vingt-cinq tests n'est pas nécessairement très pénalisant. Un risque de premier espèce général de 10% aurait peut-être été satisfaisant. On verra cependant, au vu des p-valeurs obtenues, que cela n'a pas beaucoup d'importance. La correction de Bonferroni est parfois considérée comme trop conservatrice, c'est à dire que des différences significatives peuvent ne pas être détectées, réduisant ainsi la puissance des tests. Cependant dans notre cas on dispose d'observations en nombre très important et on s'attend donc à ce que les tests soient suffisamment puissants et que les effets significatifs soient bien détectés.

Les résultats obtenus par analyses de la variance sont les suivants (figure 3.3, 3.4 et 3.5) :

¹ Les valeurs des autres statistiques pour l'analyse de la variance multivariée, comme la statistique de Pillai par exemple, sont ici systématiquement égales à la statistique de Wilk.

Tableau récapitulatif des statistiques de l'effet du modèle sur l'erreur de localisation								
Profondeur	En supposant la sphéricité		Correction de Greenhouse-Geiser		Correction de Huynh-Feldt		MANOVA : Statistique de Wilk	
	Valeur de F	P-valeur	Epsilon	P-valeur	Epsilon	P-valeur	Valeur de F	P-valeur
1	94,929	< 2,2e-16	0,63829	< 2,2e-16	0,63852	< 2,2e-16	290,9901	< 2,22e-16
2	108,74	< 2,2e-16	0,65816	< 2,2e-16	0,65843	< 2,2e-16	287,4048	< 2,22e-16
3	121,74	< 2,2e-16	0,66966	< 2,2e-16	0,66994	< 2,2e-16	291,7247	< 2,22e-16
4	531,97	< 2,2e-16	0,67525	< 2,2e-16	0,67555	< 2,2e-16	315,1102	< 2,22e-16
5	156,37	< 2,2e-16	0,6928	< 2,2e-16	0,69313	< 2,2e-16	312,098	< 2,22e-16
6	122,53	< 2,2e-16	0,81339	< 2,2e-16	0,81396	< 2,2e-16	118,5126	< 2,22e-16
7	119,15	< 2,2e-16	0,76445	< 2,2e-16	0,76492	< 2,2e-16	98,3577	< 2,22e-16
8	100,58	< 2,2e-16	0,96996	< 2,2e-16	0,97089	< 2,2e-16	88,50843	< 2,22e-16
9	144,91	< 2,2e-16	0,91409	< 2,2e-16	0,91489	< 2,2e-16	127,074	< 2,22e-16
10	120	< 2,2e-16	0,93775	< 2,2e-16	0,93861	< 2,2e-16	97,70478	< 2,22e-16
11	157,34	< 2,2e-16	0,62768	< 2,2e-16	0,62789	< 2,2e-16	103,1649	< 2,22e-16
12	187,92	< 2,2e-16	0,62558	< 2,2e-16	0,62578	< 2,2e-16	117,0825	< 2,22e-16
13	178,46	< 2,2e-16	0,78853	< 2,2e-16	0,78905	< 2,2e-16	123,527	< 2,22e-16
14	201,9	< 2,2e-16	0,74143	< 2,2e-16	0,74185	< 2,2e-16	127,0877	< 2,22e-16
15	208,06	< 2,2e-16	0,70902	< 2,2e-16	0,70838	< 2,2e-16	129,8269	< 2,22e-16
16	219,89	< 2,2e-16	0,70988	< 2,2e-16	0,71024	< 2,2e-16	136,802	< 2,22e-16
17	264,63	< 2,2e-16	0,65516	< 2,2e-16	0,65542	< 2,2e-16	164,0027	< 2,22e-16
18	261,53	< 2,2e-16	0,80577	< 2,2e-16	0,80633	< 2,2e-16	175,5274	< 2,22e-16
19	299,98	< 2,2e-16	0,72289	< 2,2e-16	0,72328	< 2,2e-16	194,7553	< 2,22e-16
20	320,95	< 2,2e-16	0,72723	< 2,2e-16	0,72762	< 2,2e-16	218,872	< 2,22e-16
21	387,92	< 2,2e-16	0,69018	< 2,2e-16	0,6905	< 2,2e-16	248,6657	< 2,22e-16
22	429,42	< 2,2e-16	0,64502	< 2,2e-16	0,64526	< 2,2e-16	273,8199	< 2,22e-16
23	455,71	< 2,2e-16	0,71743	< 2,2e-16	0,7178	< 2,2e-16	310,6653	< 2,22e-16
24	508,73	< 2,2e-16	0,69146	< 2,2e-16	0,69178	< 2,2e-16	358,2329	< 2,22e-16
25	626,76	< 2,2e-16	0,64529	< 2,2e-16	0,64529	< 2,2e-16	409,1621	< 2,22e-16

Fig. 3.3 - Tableau récapitulatif des analyses de la variance pour l'erreur sur la localisation géographique

Tableau récapitulatif des statistiques de l'effet du modèle sur l'erreur sur le temps origine								
Profondeur	En supposant la sphéricité		Correction de Greenhouse-Geiser		Correction de Huynh-Feldt		Manova : Statistique de Wilk	
	Valeur de F	P-valeur	Epsilon	P-valeur	Epsilon	P-valeur	Valeur de F	P-valeur
1	1457,7	< 2,2e-16	0,71534	< 2,2e-16	0,71572	< 2,2e-16	915,105	< 2,22e-16
2	1639,4	< 2,2e-16	0,70914	< 2,2e-16	0,7095	< 2,2e-16	1018,311	< 2,22e-16
3	1735,6	< 2,2e-16	0,70794	< 2,2e-16	0,7083	< 2,2e-16	1068,748	< 2,22e-16
4	1763,4	< 2,2e-16	0,7134	< 2,2e-16	0,71378	< 2,2e-16	1091,603	< 2,22e-16
5	1786,7	< 2,2e-16	0,72239	< 2,2e-16	0,72278	< 2,2e-16	1111,656	< 2,22e-16
6	1850,5	< 2,2e-16	0,72753	< 2,2e-16	0,72793	< 2,2e-16	1167,808	< 2,22e-16
7	1934,9	< 2,2e-16	0,71462	< 2,2e-16	0,71499	< 2,2e-16	1214,736	< 2,22e-16
8	2016,7	< 2,2e-16	0,7102	< 2,2e-16	0,71056	< 2,2e-16	1258,769	< 2,22e-16
9	2136,5	< 2,2e-16	0,71611	< 2,2e-16	0,71648	< 2,2e-16	1341,245	< 2,22e-16
10	2208,2	< 2,2e-16	0,72471	< 2,2e-16	0,7251	< 2,2e-16	1441,667	< 2,22e-16
11	2315,1	< 2,2e-16	0,73631	< 2,2e-16	0,73672	< 2,2e-16	1534,802	< 2,22e-16
12	2480,1	< 2,2e-16	0,7257	< 2,2e-16	0,72608	< 2,2e-16	1631,481	< 2,22e-16
13	2693,3	< 2,2e-16	0,72302	< 2,2e-16	0,72341	< 2,2e-16	1805,216	< 2,22e-16
14	2919,6	< 2,2e-16	0,75187	< 2,2e-16	0,75232	< 2,2e-16	1992,325	< 2,22e-16
15	3183	< 2,2e-16	0,7846	< 2,2e-16	0,78511	< 2,2e-16	2237,8	< 2,22e-16
16	3427,5	< 2,2e-16	0,77169	< 2,2e-16	0,77218	< 2,2e-16	2414,324	< 2,22e-16
17	3568,4	< 2,2e-16	0,77665	< 2,2e-16	0,77715	< 2,2e-16	2515,097	< 2,22e-16
18	3836,9	< 2,2e-16	0,78972	< 2,2e-16	0,79024	< 2,2e-16	2690,686	< 2,22e-16
19	4138	< 2,2e-16	0,80104	< 2,2e-16	0,80158	< 2,2e-16	2871,992	< 2,22e-16
20	4453,3	< 2,2e-16	0,79102	< 2,2e-16	0,79155	< 2,2e-16	3014,996	< 2,22e-16
21	4751,2	< 2,2e-16	0,82804	< 2,2e-16	0,82864	< 2,2e-16	3294,178	< 2,22e-16
22	5087,9	< 2,2e-16	0,8497	< 2,2e-16	0,85035	< 2,2e-16	3586,885	< 2,22e-16
23	5367,6	< 2,2e-16	0,85028	< 2,2e-16	0,85064	< 2,2e-16	3780,647	< 2,22e-16
24	5420,7	< 2,2e-16	0,90129	< 2,2e-16	0,90129	< 2,2e-16	4071,427	< 2,22e-16
25	5406,8	< 2,2e-16	0,87456	< 2,2e-16	0,87527	< 2,2e-16	3921,595	< 2,22e-16

Fig. 3.4 - Tableau récapitulatif des analyses de la variance pour l'erreur sur le temps origine

Tableau récapitulatif des statistiques de l'effet du modèle sur l'erreur sur la profondeur								
Profondeur	En supposant la sphéricité		Correction de Greenhouse-Geiser		Correction de Huynh-Feldt		Manova : Statistique de Wilk	
	Valeur de F	P-valeur	Epsilon	P-valeur	Epsilon	P-valeur	Valeur de F	P-valeur
1	2889,9	< 2,2e-16	0,74141	< 2,2e-16	0,74184	< 2,2e-16	2248,854	< 2,22e-16
2	2878,8	< 2,2e-16	0,73636	< 2,2e-16	0,73678	< 2,2e-16	2275,235	< 2,22e-16
3	2832,2	< 2,2e-16	0,72748	< 2,2e-16	0,72788	< 2,2e-16	2199,865	< 2,22e-16
4	2824,5	< 2,2e-16	0,71909	< 2,2e-16	0,71947	< 2,2e-16	2191,305	< 2,22e-16
5	2851	< 2,2e-16	0,70976	< 2,2e-16	0,71012	< 2,2e-16	2188,395	< 2,22e-16
6	2890,2	< 2,2e-16	0,70316	< 2,2e-16	0,70316	< 2,2e-16	2198,014	< 2,22e-16
7	2891,7	< 2,2e-16	0,71126	< 2,2e-16	0,71162	< 2,2e-16	2138,86	< 2,22e-16
8	2934,7	< 2,2e-16	0,70848	< 2,2e-16	0,70884	< 2,2e-16	2104,994	< 2,22e-16
9	3025,9	< 2,2e-16	0,7141	< 2,2e-16	0,71447	< 2,2e-16	2075,004	< 2,22e-16
10	3252,7	< 2,2e-16	0,71343	< 2,2e-16	0,7138	< 2,2e-16	2131,118	< 2,22e-16
11	3308,8	< 2,2e-16	0,72365	< 2,2e-16	0,72404	< 2,2e-16	2122,709	< 2,22e-16
12	3473,6	< 2,2e-16	0,72972	< 2,2e-16	0,73012	< 2,2e-16	2197,783	< 2,22e-16
13	3550,5	< 2,2e-16	0,73555	< 2,2e-16	0,73596	< 2,2e-16	2241,845	< 2,22e-16
14	3655,6	< 2,2e-16	0,75745	< 2,2e-16	0,7579	< 2,2e-16	2353,516	< 2,22e-16
15	3608,5	< 2,2e-16	0,77736	< 2,2e-16	0,77786	< 2,2e-16	2365,951	< 2,22e-16
16	3460,3	< 2,2e-16	0,81026	< 2,2e-16	0,81082	< 2,2e-16	2366,046	< 2,22e-16
17	3064,7	< 2,2e-16	0,81883	< 2,2e-16	0,81941	< 2,2e-16	2119,229	< 2,22e-16
18	2689,3	< 2,2e-16	0,84499	< 2,2e-16	0,84563	< 2,2e-16	1925,532	< 2,22e-16
19	2335,8	< 2,2e-16	0,83417	< 2,2e-16	0,83479	< 2,2e-16	1685,578	< 2,22e-16
20	2056,7	< 2,2e-16	0,83977	< 2,2e-16	0,8404	< 2,2e-16	1505,832	< 2,22e-16
21	1677,6	< 2,2e-16	0,88747	< 2,2e-16	0,88821	< 2,2e-16	1337,912	< 2,22e-16
22	1276,9	< 2,2e-16	0,87884	< 2,2e-16	0,87956	< 2,2e-16	1025,282	< 2,22e-16
23	867,07	< 2,2e-16	0,89385	< 2,2e-16	0,8946	< 2,2e-16	709,9863	< 2,22e-16
24	466,46	< 2,2e-16	0,90572	< 2,2e-16	0,9065	< 2,2e-16	418,0076	< 2,22e-16
25	161,65	< 2,2e-16	0,91022	< 2,2e-16	0,91101	< 2,2e-16	172,0398	< 2,22e-16

Fig. 3.5 - Tableau récapitulatif des analyses de la variance pour l'erreur sur la profondeur

On a une p-valeur extrêmement faible pour toutes les profondeurs et pour toutes les méthodes et corrections. Cela n'est pas surprenant au vu de l'analyse des graphiques faite précédemment. De plus le nombre d'observations étant très grand on s'attendait à ce que la présence d'un effet du facteur soit nettement détectée.

Si on tient compte du nombre de phases les résultats sont différents. Concernant l'erreur sur le temps origine on observe toujours un effet significatif quelle que soit la profondeur et la correction utilisée, et ce pour un nombre de phases égal à six ou pour un nombre de phases strictement supérieur à six. Pour l'erreur sur la localisation géographique dans le cas d'un nombre de phases strictement supérieur à douze on observe aussi toujours un effet significatif quelle que soit la profondeur et la correction utilisée. En revanche pour un nombre de phases inférieur ou égal à douze il est plus délicat de conclure (figure 3.6).

Tableau récapitulatif des statistiques de l'effet du modèle sur l'erreur sur la localisation avec un nombre de phases <= 12								
Profondeur	En supposant la sphéricité		Correction de Greenhouse-Geiser		Correction de Huynh-Feldt		Manova : Statistique de Wilk	
	Valeur de F	P-valeur	Epsilon	P-valeur	Epsilon	P-valeur	Valeur de F	P-valeur
1	6,3448	0,001836	0,53595	0,01058	0,5362	0,01057	16,09039	1,79E-007
5	8,2062	0,000294	0,55864	0,003179	0,55905	0,003172	46,73546	< 2,22e-16
10	18,528	1,31E-008	0,94739	2,82E-008	0,9513	2,66E-008	15,53399	3,00E-007
15	42,096	< 2,2e-16	0,71528	9,60E-014	0,71693	9,05E-014	27,08302	7,90E-012
20	72,844	< 2,2e-16	0,73528	< 2,2e-16	0,73711	< 2,2e-16	48,51015	< 2,22e-16
25	130,74	< 2,2e-16	0,65104	< 2,2e-16	0,65215	< 2,2e-16	80,38025	< 2,22e-16

Fig. 3.6 - Tableau récapitulatif des analyses de la variance pour l'erreur sur la localisation géographique dans le cas d'un nombre de phases inférieur ou égal à douze (seules les résultats pour cinq profondeurs sont affichés)

A partir de dix kilomètres de profondeur on peut aussi conclure quant à un effet significatif du modèle de vitesse des ondes au seuil de risque fixé précédemment. En revanche les p-valeurs obtenues avec les corrections de Greenhouse-Geiser et Huynh-Feldt pour les faibles profondeurs ne permettent pas de conclure quant à un effet significatif du modèle de vitesse des ondes au seuil de risque fixé précédemment, contrairement aux p-valeurs obtenues en supposant la sphéricité et par analyse de la variance multivariée. La p-valeur obtenue en supposant la sphéricité n'est sans doute pas intéressante puisqu'on a vu que l'hypothèse de sphéricité était peu crédible, en revanche il n'est pas clair quelle méthode privilégier entre l'analyse de la variance multivariée et les corrections de Greenhouse-Geiser et Huynh-Feldt. Une hypothèse expliquant ces résultats est que pour les profondeurs faibles et un nombre de phases faibles l'effet du réseau est plus important, diminuant la part due au modèle dans l'erreur. La différence entre les différents modèles serait donc plus faible. On effectuera des comparaisons multiples afin d'étudier ces cas plus en détail.

On effectue ensuite des comparaisons multiples afin de savoir où résident les différences. On utilisera des tests de Student, en appliquant à nouveau une correction de Bonferroni. Cette fois-ci on effectue soixante-quinze comparaisons (trois par profondeur). Afin de ne pas avoir un seuil de signification trop faible, et des intervalles de confiance trop larges, on se limitera à un risque de premier espèce général de 10%, soit 0,1. Le seuil de signification pour chaque comparaison sera alors de 0,001, et les intervalles de confiance seront données à 99,9%. On rappelle que le nombre d'observations étant très grand on s'attend à pouvoir détecter nettement des différences et à ne pas avoir des intervalles de confiance trop larges, même à 99,9%. Dans le cas de l'erreur sur la localisation géographique avec un nombre de phases inférieur ou égal à douze, on prendra des intervalles de confiance à 97,5%, soit 0,975, car le nombre d'observations est plus petit (il y a 449 observations ayant un nombre de phases inférieur ou égal à 12 par profondeur et par modèle). Pour chaque comparaison on donnera la p-valeur, la borne inférieure de l'intervalle de confiance (désignée par *IC inf*), la borne supérieure de l'intervalle de confiance (désignée par *IC sup*), et la moyenne des différences des erreurs (désignée par *moyenne*).

Les résultats obtenus pour les comparaisons multiples sont les suivants (figure 3.7, 3.8 et 3.9) :

Tableau récapitulatif des comparaisons multiples pour l'erreur sur la localisation. IC à 0,999												
Profondeur	Comparaison 1Ds-1Dc				Comparaison 1Ds-3D				Comparaison 1Dc-3D			
	P-valeur	IC inf	IC sup	moyenne	P-valeur	IC inf	IC sup	moyenne	P-valeur	IC inf	IC sup	moyenne
1	8,86E-001	-326	356	15	< 2,2e-16	936	1234	1085	< 2,2e-16	713	1427	1070
2	9,38E-001	-344	329	-8	< 2,2e-16	990	1308	1149	< 2,2e-16	801	1514	1157
3	7,07E-001	-377	300	-39	< 2,2e-16	1055	1390	1222	< 2,2e-16	897	1625	1261
4	3,70E-002	-31	136	53	< 2,2e-16	1141	1491	1316	< 2,2e-16	1089	1438	1263
5	2,81E-001	-233	460	113	< 2,2e-16	1143	1906	1524	< 2,2e-16	1224	1598	1411
6	4,02E-001	-277	466	95	< 2,2e-16	1261	2280	1770	< 2,2e-16	1316	2036	1676
7	3,84E-001	-276	473	99	< 2,2e-16	1405	2599	2002	< 2,2e-16	1456	2351	1904
8	3,48E-001	-709	394	-157	< 2,2e-16	1514	2801	2158	< 2,2e-16	1711	2919	2315
9	2,06E-001	-660	294	-183	< 2,2e-16	1720	2956	2338	< 2,2e-16	1993	3050	2521
10	1,06E-001	-949	324	-312	< 2,2e-16	2034	3628	2831	< 2,2e-16	2377	3910	3143
11	2,76E-003	-672	32	-320	< 2,2e-16	2314	4068	3191	< 2,2e-16	2694	4329	3511
12	4,17E-008	-903	-227	-565	< 2,2e-16	2473	4138	3306	< 2,2e-16	3029	4712	3871
13	7,72E-004	-1130	-12	-571	< 2,2e-16	2765	4589	3677	< 2,2e-16	3356	5141	4248
14	5,08E-011	-1681	-562	-1121	< 2,2e-16	2847	4564	3706	< 2,2e-16	3822	5832	4827
15	8,26E-013	-1814	-675	-1245	< 2,2e-16	3198	5175	4186	< 2,2e-16	4321	6542	5431
16	1,18E-013	-1948	-755	-1352	< 2,2e-16	3440	5484	4462	< 2,2e-16	4655	6971	5813
17	< 2,2e-16	-1881	-859	-1370	< 2,2e-16	3900	6029	4965	< 2,2e-16	5176	7493	6334
18	1,88E-012	-2456	-897	-1677	< 2,2e-16	4228	6384	5306	< 2,2e-16	5741	8223	6983
19	2,43E-014	-2242	-896	-1569	< 2,2e-16	4844	7246	6045	< 2,2e-16	6342	8885	7614
20	8,44E-010	-2034	-617	-1326	< 2,2e-16	5505	8130	6818	< 2,2e-16	6858	9428	8143
21	6,33E-015	-2135	-873	-1504	< 2,2e-16	5915	8413	7164	< 2,2e-16	7385	9951	8668
22	1,62E-015	-1996	-835	-1415	< 2,2e-16	6670	9347	8008	< 2,2e-16	8086	10762	9424
23	8,86E-008	-1958	-469	-1214	< 2,2e-16	7510	10362	8936	< 2,2e-16	8806	11493	10150
24	1,80E-006	-1759	-325	-1042	< 2,2e-16	8115	11041	9578	< 2,2e-16	9303	11937	10620
25	< 2,2e-16	-2108	-923	-1515	< 2,2e-16	8592	11355	9973	< 2,2e-16	10150	12827	11489

Fig. 3.7 - Tableau récapitulatif des comparaisons multiples pour les erreurs sur la localisation géographique

Tableau récapitulatif des comparaisons multiples pour l'erreur sur le temps origine. IC à 0,999												
Profondeur	Comparaison 1Ds-1Dc				Comparaison 1Ds-3D				Comparaison 1Dc-3D			
	P-valeur	IC inf	IC sup	moyenne	P-valeur	IC inf	IC sup	moyenne	P-valeur	IC inf	IC sup	moyenne
1	< 2,2e-16	0,124	0,162	0,143	< 2,2e-16	0,36	0,421	0,39	< 2,2e-16	0,226	0,268	0,247
2	< 2,2e-16	0,138	0,177	0,157	< 2,2e-16	0,4	0,465	0,433	< 2,2e-16	0,253	0,298	0,276
3	< 2,2e-16	0,151	0,191	0,171	< 2,2e-16	0,421	0,499	0,466	< 2,2e-16	0,271	0,318	0,294
4	< 2,2e-16	0,162	0,205	0,184	< 2,2e-16	0,461	0,532	0,496	< 2,2e-16	0,288	0,337	0,313
5	< 2,2e-16	0,171	0,216	0,194	< 2,2e-16	0,488	0,563	0,526	< 2,2e-16	0,305	0,358	0,332
6	< 2,2e-16	0,181	0,229	0,205	< 2,2e-16	0,518	0,596	0,557	< 2,2e-16	0,325	0,378	0,351
7	< 2,2e-16	0,199	0,25	0,224	< 2,2e-16	0,554	0,635	0,594	< 2,2e-16	0,342	0,397	0,37
8	< 2,2e-16	0,206	0,257	0,232	< 2,2e-16	0,582	0,666	0,624	< 2,2e-16	0,364	0,421	0,393
9	< 2,2e-16	0,219	0,272	0,245	< 2,2e-16	0,621	0,707	0,664	< 2,2e-16	0,389	0,448	0,419
10	< 2,2e-16	0,234	0,292	0,263	< 2,2e-16	0,661	0,751	0,706	< 2,2e-16	0,413	0,473	0,443
11	< 2,2e-16	0,235	0,294	0,265	< 2,2e-16	0,687	0,778	0,733	< 2,2e-16	0,438	0,499	0,468
12	< 2,2e-16	0,233	0,292	0,263	< 2,2e-16	0,72	0,813	0,767	< 2,2e-16	0,473	0,535	0,504
13	< 2,2e-16	0,229	0,289	0,26	< 2,2e-16	0,752	0,845	0,798	< 2,2e-16	0,508	0,57	0,539
14	< 2,2e-16	0,236	0,297	0,267	< 2,2e-16	0,792	0,885	0,838	< 2,2e-16	0,54	0,604	0,572
15	< 2,2e-16	0,235	0,297	0,266	< 2,2e-16	0,827	0,918	0,872	< 2,2e-16	0,574	0,639	0,607
16	< 2,2e-16	0,238	0,299	0,269	< 2,2e-16	0,862	0,955	0,908	< 2,2e-16	0,607	0,672	0,64
17	< 2,2e-16	0,247	0,31	0,279	< 2,2e-16	0,895	0,989	0,942	< 2,2e-16	0,63	0,696	0,663
18	< 2,2e-16	0,251	0,314	0,283	< 2,2e-16	0,931	1,024	0,978	< 2,2e-16	0,661	0,729	0,695
19	< 2,2e-16	0,252	0,314	0,283	< 2,2e-16	0,968	1,06	1,014	< 2,2e-16	0,697	0,766	0,731
20	< 2,2e-16	0,249	0,309	0,279	< 2,2e-16	0,997	1,089	1,043	< 2,2e-16	0,728	0,8	0,764
21	< 2,2e-16	0,26	0,321	0,291	< 2,2e-16	1,045	1,137	1,091	< 2,2e-16	0,762	0,837	0,8
22	< 2,2e-16	0,25	0,311	0,281	< 2,2e-16	1,078	1,167	1,123	< 2,2e-16	0,804	0,88	0,842
23	< 2,2e-16	0,261	0,322	0,292	< 2,2e-16	1,115	1,205	1,16	< 2,2e-16	0,83	0,907	0,868
24	< 2,2e-16	0,258	0,324	0,291	< 2,2e-16	1,148	1,238	1,193	< 2,2e-16	0,863	0,942	0,902
25	< 2,2e-16	0,252	0,316	0,284	< 2,2e-16	1,155	1,246	1,2	< 2,2e-16	0,876	0,957	0,916

Fig. 3.8 - Tableau récapitulatif des comparaisons multiples pour les erreurs sur le temps origine

Tableau récapitulatif des comparaisons multiples pour l'erreur sur la profondeur. IC à 0,999

Profondeur	Comparaison 1Ds-1Dc				Comparaison 1Ds-3D				Comparaison 1Dc-3D			
	P-valeur	IC inf	IC sup	moyenne	P-valeur	IC inf	IC sup	moyenne	P-valeur	IC inf	IC sup	moyenne
1	< 2.2e-16	-1,9563	-1,314	-1,6351	< 2.2e-16	-10,9693	-9,9305	-10,4499	< 2.2e-16	-9,3974	-8,2321	-8,8147
2	< 2.2e-16	-1,9082	-1,2763	-1,5922	< 2.2e-16	-10,8007	-9,7808	-10,2908	< 2.2e-16	-9,277	-8,1201	-8,6985
3	< 2.2e-16	-1,9212	-1,3036	-1,6124	< 2.2e-16	-10,7374	-9,7062	-10,2218	< 2.2e-16	-9,187	-8,0318	-8,6094
4	< 2.2e-16	-1,8484	-1,2444	-1,5464	< 2.2e-16	-10,6231	-9,5991	-10,1111	< 2.2e-16	-9,1401	-7,9894	-8,5647
5	< 2.2e-16	-1,7681	-1,1797	-1,4739	< 2.2e-16	-10,574	-9,5534	-10,0637	< 2.2e-16	-9,1627	-8,017	-8,5898
6	< 2.2e-16	-1,5836	-1,0087	-1,2962	< 2.2e-16	-10,4305	-9,4269	-9,9287	< 2.2e-16	-9,2002	-8,0648	-8,6325
7	< 2.2e-16	-1,5779	-1,0071	-1,2925	< 2.2e-16	-10,292	-9,2942	-9,7931	< 2.2e-16	-9,0551	-7,9461	-8,5006
8	< 2.2e-16	-1,465	-0,9139	-1,1894	< 2.2e-16	-10,0947	-9,1119	-9,6033	< 2.2e-16	-8,9534	-7,8743	-8,4139
9	< 2.2e-16	-1,272	-0,7287	-1,0004	< 2.2e-16	-9,9783	-9,0058	-9,492	< 2.2e-16	-9,0166	-7,9668	-8,4917
10	< 2.2e-16	-1,1945	-0,6769	-0,9357	< 2.2e-16	-9,928	-8,9741	-9,451	< 2.2e-16	-9,0141	-8,0165	-8,5153
11	< 2.2e-16	-1,0881	-0,5714	-0,8297	< 2.2e-16	-9,7802	-8,8366	-9,3084	< 2.2e-16	-8,9616	-7,9957	-8,4787
12	< 2.2e-16	-1,103	-0,5934	-0,8482	< 2.2e-16	-9,7818	-8,8493	-9,3155	< 2.2e-16	-8,9314	-8,0032	-8,4673
13	< 2.2e-16	-1,0146	-0,5133	-0,764	< 2.2e-16	-9,5802	-8,6716	-9,1259	< 2.2e-16	-8,8104	-7,9135	-8,3619
14	< 2.2e-16	-1,1695	-0,662	-0,9158	< 2.2e-16	-9,5032	-8,6209	-9,062	< 2.2e-16	-8,576	-7,7166	-8,1463
15	< 2.2e-16	-1,1105	-0,5962	-0,8534	< 2.2e-16	-9,232	-8,374	-8,803	< 2.2e-16	-8,3685	-7,5308	-7,9496
16	< 2.2e-16	-1,1931	-0,6593	-0,9262	< 2.2e-16	-8,9602	-8,1297	-8,545	< 2.2e-16	-8,0315	-7,206	-7,6187
17	< 2.2e-16	-1,1916	-0,6537	-0,9226	< 2.2e-16	-8,4344	-7,6112	-8,0228	< 2.2e-16	-7,5101	-6,6904	-7,1002
18	< 2.2e-16	-1,2419	-0,6826	-0,9623	< 2.2e-16	-7,9663	-7,1532	-7,5598	< 2.2e-16	-7,0059	-6,1891	-6,5975
19	< 2.2e-16	-1,3085	-0,7564	-1,0324	< 2.2e-16	-7,5168	-6,7055	-7,1111	< 2.2e-16	-6,4924	-5,665	-6,0787
20	< 2.2e-16	-1,3525	-0,8009	-1,0767	< 2.2e-16	-7,0568	-6,2547	-6,6559	< 2.2e-16	-5,9886	-5,1697	-5,5792
21	< 2.2e-16	-1,4981	-0,9038	-1,201	< 2.2e-16	-6,5059	-5,7254	-6,1156	< 2.2e-16	-5,3236	-4,5057	-4,9146
22	< 2.2e-16	-1,4203	-0,8282	-1,1242	< 2.2e-16	-5,7874	-5,0021	-5,3947	< 2.2e-16	-4,6847	-3,8562	-4,2705
23	< 2.2e-16	-1,3339	-0,7175	-1,0257	< 2.2e-16	-4,9601	-4,1622	-4,5612	< 2.2e-16	-9,9552	-3,1157	-3,5354
24	< 2.2e-16	-1,3855	-0,7292	-1,0573	< 2.2e-16	-3,9594	-3,1475	-3,5535	< 2.2e-16	-2,935	-2,0573	-2,4961
25	< 2.2e-16	-1,4461	-0,7506	-1,0984	< 2.2e-16	-2,6646	-1,8279	-2,2463	< 2.2e-16	-1,609	-0,6867	-1,1479

Fig. 3.9 - Tableau récapitulatif des comparaisons multiples pour les erreurs sur la profondeur

Pour des faibles profondeurs (un kilomètre à onze kilomètres) la différence entre l'erreur de localisation pour le modèle 1Ds et le modèle 1Dc n'est pas significative au seuil fixé, mais l'est pour des profondeurs plus importantes, et est de l'ordre du kilomètre. La moyenne des différences étant négatives, cela signifie que l'erreur sur la localisation par le modèle 1Dc est plus grande que par le modèle 1Ds. Une raison possible de cette moins bonne efficacité du modèle 1Dc a déjà été donnée précédemment : la couche supplémentaire du modèle 1Dc correspond à une couche sédimentaire présente seulement au centre du fossé rhénan mais qui est généralisée à tout le fossé rhénan dans le modèle 1Dc (voir chapitre 2). On observe par contre une différence significative au seuil fixé pour toutes les profondeurs entre le modèle 1Ds et le modèle 3D, et entre le modèle 1Dc et le modèle 3D. Cette différence est de l'ordre du kilomètre pour les faibles profondeurs et augmente avec la profondeur jusqu'à atteindre la dizaine de kilomètres à vingt-cinq kilomètres de profondeur.

Concernant l'erreur sur le temps origine les différences sont significatives au seuil fixé pour toutes les profondeurs et pour toutes les paires de modèle et ont tendance à augmenter avec la profondeur. En revanche la différence entre le modèle 1Dc et le modèle 1Ds est relativement faible (de 0,1 seconde à 0,3 seconde) alors que la différence la plus faible entre le modèle 1Ds et 3D est de 0,39 seconde à un kilomètre de profondeur (intervalle de confiance de 0,36 à 0,421 seconde) et augmente jusqu'à atteindre 1,2 seconde à vingt-cinq kilomètres de profondeur (intervalle de confiance de 1,155 à 1,246 seconde), et la différence la plus faible entre le modèle 1Dc et 3D est de 0,247 seconde à un kilomètre de profondeur (intervalle de confiance de 0,226 à 0,268 seconde) et augmente jusqu'à atteindre 0,916 seconde à vingt-cinq kilomètres (intervalle de confiance de 0,876 à 0,957 seconde). On remarque que cette fois-ci c'est le modèle 1Dc qui est meilleur que le modèle

1Ds : la différence entre les deux étant positive, et l'erreur étant positive pour les deux modèles (voir page 18), même si la différence entre les deux modèles est faible.

Pour l'erreur sur la profondeur les différences sont significatives au seuil fixé pour toutes les profondeurs et pour toutes les paires de modèle. La différence entre les modèles 1Ds et 1Dc est de l'ordre du kilomètre pour toutes les profondeurs. La différence entre les modèles 1Ds et 3D, et la différence entre les modèles 1Dc et 3D, sont de l'ordre de la dizaine de kilomètres pour des profondeurs faibles, et tendent à décroître quand la profondeur augmente. Ce résultat peut paraître surprenant, néanmoins elle s'explique par l'étude des sismogrammes : avec les modèles 1Ds et 1Dc les séismes sont localisés à des profondeurs supérieures à la profondeur réelle, qu'un nombre important de séismes est placé à la limite inférieure de la grille spatiale du logiciel (à 31 km de profondeur), et qu'à de fortes profondeurs la plupart des événements sont situés sur cette limite inférieure (voir figure 2.8). Le modèle 3D produisant généralement de faibles erreurs sur la profondeur¹, les séismes sont presque localisés à la bonne profondeur, et la différence entre les erreurs sur la profondeur avec les modèles 1D (1Ds et 1Dc) et le modèle 3D pour les fortes profondeurs ne peut pas être très grande². Il est possible que si la grille spatiale du logiciel dépassait les 31 km une partie des événements localisés par les modèles 1D serait encore collée à la limite inférieure, et que les différences entre les modèles 1D et 3D pour les profondeurs élevées (environ 18 à 25 km) serait aussi d'une dizaine de kilomètres, comme pour les faibles profondeurs. Néanmoins il ne s'agit que de suppositions qu'il faudrait vérifier par de nouvelles simulations.

Les résultats des comparaisons multiples pour l'erreur sur la localisation géographique dans le cas d'un nombre de phases inférieur ou égal à douze sont donnés ci-dessous (figure 3.10) :

Tableau récapitulatif des comparaisons multiples pour l'erreur sur la localisation, nombre de phases <= 12. IC à 0,975												
Profondeur	Comparaison 1Ds-1Dc				Comparaison 1Ds-3D				Comparaison 1Dc-3D			
	P-valeur	IC inf	IC sup	moyenne	P-valeur	IC inf	IC sup	moyenne	P-valeur	IC inf	IC sup	moyenne
1	7.49E-002	-1809	208	-801	2.56E-007	293	734	513	3.79E-003	298	2329	1314
5	6.75E-001	-833	1215	191	2.33E-003	381	2486	1433	< 2.2e-16	953	1532	1242
10	4.37E-001	-2519	1225	-647	7.64E-006	2333	6936	4635	9.39E-008	3093	7470	5282
15	3.67E-004	-4203	-965	-2584	2.52E-009	4798	10427	7613	9.08E-013	7079	13314	10197
20	6.57E-002	-3677	363	-1657	4.53E-016	10182	17583	13882	< 2.2e-16	11994	19084	15539
25	9.50E-002	-2918	428	-1245	< 2.2e-16	15461	23140	19300	< 2.2e-16	16904	24186	20545

Fig. 3.10 - Tableau récapitulatif des comparaisons multiples pour les erreurs sur la localisation géographique dans le cas d'un nombre de phases inférieur ou égal à douze

Les résultats de l'analyse de la variance dans cette situation étaient contradictoires pour les faibles profondeurs selon la correction utilisée. Avec des tests de Student appariés on obtient des différences significatives même pour les faibles profondeurs : à un kilomètre de profondeur il y a une différence significative au seuil fixé entre le modèle 1Ds et le modèle 3D et entre le modèle 1Dc et le modèle 3D. On rappelle qu'on s'est cette fois-ci fixé un risque de premier espèce de 0,25%. Si on s'était fixé un risque de premier espèce de 0,1% la différence entre le modèle 1Dc et le modèle 3D n'aurait pas

¹ La moyenne des erreurs sur la profondeur avec le modèle 3D, toutes profondeurs confondues, est de 161 mètres.

² Si par exemple un séisme situé à 25 km de profondeur est localisé à exactement 25 kilomètres de profondeur par le modèle 3D et à 31 km par les modèles 1D (limite de la grille du programme) la différence est de seulement 6 km.

été significative. A cinq kilomètres de profondeur il y a une différence significative au seuil fixé entre le modèle 1Ds et le modèle 3D, et entre le modèle 1Dc et le modèle 3D. Avec un risque de première espèce de 0,1% la différence entre le modèle 1Ds et le modèle 3D n'aurait pas été significative.

On note que pour des profondeurs modérées à élevées la moyenne des différences entre les modèles 1D et le modèle 3D sont supérieures aux valeurs obtenues précédemment, quand on ne tenait pas compte du nombre de phases. Ainsi à quinze kilomètres de profondeur la moyenne de la différence entre le modèle 1Ds et le modèle 3D était de 4186 mètres (intervalle de confiance à 99,9% de 3198 mètres à 5175 mètres) alors qu'elle est ici de 7613 mètres (intervalle de confiance à 97,5% de 4798 mètres à 10427 mètres), et à vingt-cinq kilomètres de profondeur la moyenne de la différence entre le modèle 1Ds et le modèle 3D était de 9973 mètres (intervalle de confiance à 99,9% de 8592 mètres à 11355 mètres) alors qu'elle est ici de 19300 mètres (intervalle de confiance à 97,5% de 15461 mètres à 23140 mètres).

Les tests de Friedman ne seront donnés que pour l'erreur sur la localisation géographique à six profondeurs différentes, par souci de place, car ils sont donnés essentiellement à titres indicatifs. Les résultats sont les suivants (figure 3.11) :

Profondeur	Test de Friedman		Somme des rangs			Tests post-hoc : Différence significative entre		
	Valeur de F	P-valeur	Modele 1Ds	Modele 1Dc	Modele 3D	1Ds – 1Dc	1Ds – 3D	1Dc – 3D
1	9,26E+002	< 2,2e-16	4602,5	4452,5	2873	oui	oui	oui
5	1,21E+003	< 2,2e-16	4600,5	4613,5	2714	non	oui	oui
10	1,64E+003	< 2,2e-16	4848,5	4567,5	2512	oui	oui	oui
15	2,04E+003	< 2,2e-16	5181	4343	2404	oui	oui	oui
20	2,76E+003	< 2,2e-16	5429,5	4325,5	2173	oui	oui	oui
25	3268,991	< 2,2e-16	5582	4319	2027	oui	oui	oui

Fig. 3.11 - Tableau récapitulatif des tests de Friedman pour l'erreur sur la localisation géographique

Les tests de Friedman sont significatifs au seuil fixé pour toutes les profondeurs étudiées. D'après les tests post-hoc il n'y a pas de différence significative entre le modèle 1Ds et le modèle 1Dc à cinq kilomètres de profondeur.

La significativité des comparaisons post-hoc est moins intéressante à regarder que la somme des rangs : on voit que les modèles 1Ds et 1Dc ont une somme des rangs proches pour les profondeurs faibles (1 à 10 kilomètres), que le modèle 1Dc a une somme des rangs plus petite que celle du modèle 1Ds pour les profondeurs modérées à élevées (15 à 25 kilomètres), mais, surtout, les sommes des rangs pour le modèle 3D est systématiquement beaucoup plus faible que pour les modèles 1D (environ deux fois plus faibles). On rappelle qu'on a 1988 événements par modèle et par profondeur, le minimum de la somme des rangs pour un modèle est donc 1988. La somme des rangs du modèle 3D est donc très petite, ce qui signifie que l'erreur sur la localisation géographique par le modèle 3D est presque systématiquement plus petite que par les modèles 1D. Les tests de Friedman confirment donc ce que laissaient penser les histogrammes et boîtes à moustaches vus précédemment.

Synthèse

Le but de cette étude était d'étudier l'effet de trois modèles de propagations des ondes sismiques dans le fossé rhéno-germanique, deux modèles 1D à couches et un modèle 3D, sur les erreurs de localisation spatio-temporelle (localisation géographique, profondeur et temps origine), de mettre en évidence la présence ou l'absence de différences des erreurs en fonction des modèles, et de quantifier cette différence si elle existe. Afin de faire cela le programme NonLinLoc développé par Anthony Lomax a été utilisé. Ce programme permet la localisation des séismes à partir des temps d'arrivées des ondes sismiques à différentes stations, mais permet aussi de créer des séismes synthétiques en introduisant différentes incertitudes afin d'obtenir des observations crédibles de ces séismes synthétiques. On a utilisé cette fonctionnalité en créant 1988 événements répartis aléatoirement dans le fossé rhéno-germanique, et en faisant varier la profondeur d'un kilomètre à vingt-cinq kilomètres avec un pas d'un kilomètre. Chaque événement virtuel a été localisé en utilisant chaque modèle, et les erreurs sur la localisation entre les différents modèles ont ensuite été comparées.

Une première étude graphique a permis de mettre en évidence des comportements, qui ont ensuite été confirmés par des méthodes statistiques comme l'analyse de la variance à mesures répétées, le test de Student apparié, et le test non-paramétrique de Friedman. Globalement les erreurs sur la localisation géographique sont de l'ordre du kilomètre à la dizaine de kilomètres, en revanche pour certains événements elles peuvent atteindre cent à deux cents kilomètres. Pour des faibles profondeurs (un à dix kilomètres) les différences entre les deux modèles 1D à couches pour l'erreur sur la localisation géographique n'étaient pas significatives au seuil fixé d'après les tests de Student effectués, en revanche les différences entre les modèles 1D et le modèle 3D étaient de l'ordre du kilomètre. Pour des profondeurs intermédiaires les différences entre les deux modèles 1D à couches étaient significatives mais relativement faibles (de l'ordre du kilomètre) comparé aux différences entre les modèles 1D et le modèle 3D (la moyenne des différences allant jusqu'à dix kilomètres). On a aussi pu montrer que cette moyenne des différences a tendance à augmenter avec la profondeur, passant d'environ un kilomètre pour les événements à un kilomètre de profondeur, à dix kilomètres pour les événements à vingt-cinq kilomètres. Les erreurs sur le temps origine des séismes ont un comportement similaire, mais demeurent globalement faibles (inférieures à une seconde pour la plupart des événements). Les erreurs sur la profondeur sont globalement très grandes pour les modèles 1D, et les moyennes des différences avec le modèle 3D sont de l'ordre de sept à dix kilomètres (on note une baisse de cette différence pour les profondeurs élevées due à la limite inférieure de la grille spatiale utilisée par le logiciel).

On a aussi essayé de tenir compte d'autres paramètres, liés au réseau, connus pour avoir une influence sur les erreurs de localisation. Ces paramètres sont le nombre de phases (c'est à dire le nombre de temps d'arrivées d'ondes), le gap azimutal (qui permet de mesurer la couverture de l'événement par le réseau) et la distance à la station la plus proche. Graphiquement une augmentation linéaire de l'erreur sur la localisation en fonction de ces paramètres a pu être observée. On a montré de manière plus générale que les erreurs de localisation spatio-temporelle étaient affectés par la localisation exacte du séisme, les séismes situés aux bords du fossé rhéno-germanique

étant mal localisés, contrairement à ceux situés au centre du fossé rhénan, qui sont bien localisés car de nombreuses stations réparties tout autour du séisme peuvent les détecter. Ainsi un nombre plus important de stations et une répartition de celles-ci dans tout le fossé rhénan permettrait d'améliorer la localisation des séismes. Lors de l'analyse statistique on n'a tenu compte que du nombre de phases en séparant les événements selon certaines plages de valeurs. Pour un nombre de phases faible les erreurs ont effectivement tendance à être plus grandes. Il serait intéressant de tenir compte plus finement de ces paramètres, par exemple grâce à l'utilisation d'une analyse de la covariance, puisque les erreurs semblent évoluer linéairement en fonction de ces paramètres.

Il faut faire attention à l'interprétation des résultats trouvés. Les erreurs obtenues ne correspondent pas aux erreurs observées pour des séismes réels. En effet pour créer les événements virtuels avec le programme NonLinLoc il a fallu utiliser un modèle de propagation des ondes sismique de référence, qui a servi à créer les temps d'arrivées des ondes virtuels. C'est le modèle 3D qui a été utilisé pour cela, introduisant alors comme a priori que ce modèle est plus représentatif de la véritable structure du fossé rhénan. Il est donc normal que les erreurs de localisation obtenues en localisant les événements par le modèle 3D soient plus faibles que celles obtenues par les modèles 1D, puisqu'elles correspondent à la part d'erreur due à la qualité du réseau et aux différentes incertitudes, comme l'incertitude sur le temps d'arrivée des ondes sismiques. Ainsi les différences observées dans les erreurs de localisations selon les modèles ne permettent d'avoir que des estimations des différences relatives entre ces modèles, de mesurer l'effet dû aux différences entre les modèles, et de comparer ces modèles entre eux, mais ne permettent pas d'avoir une estimation des erreurs absolues que l'on commet lors de la localisation de séismes réels. Il s'agit là d'un problème inhérent à la modélisation de la Terre et des modèles de vitesses de propagations des ondes sismiques, dont il est impossible de se défaire. Il est nécessaire de procéder ainsi, par des comparaisons relatives des nouveaux modèles avec les anciens, et des améliorations successives. Il serait intéressant de disposer d'un nombre suffisant d'événements réels pour lesquels une localisation quasiment exacte est connue (en recoupant différentes méthodes par exemple) et d'effectuer des comparaisons des modèles et des analyses statistiques sur ces événements.

Bibliographie

Berkovits, I., G.R. Hancock et J. Nevitt (2000), Bootstrap Resampling Approaches for Repeated Measure Designs : Relative Robustness to Sphericity and Normality Violations, *Educational and Psychological Measurement*, **60**, 877-892.

Friedman, M. (1937), The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance, *Journal of the American Statistical Association*, **32**, 675-701.

Glass, G. V., P. D. Pechman et J. R. Sanders (1972), Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analysis of Variances and Covariance, *Rev. Educ. Res.*, **42**, 237-288.

Harwell, M. R., E. N. Rubinstein, W. S. Hayes et C. C. Olds (1992), Summarizing Monte Carlo Results in Methodological Research : the One- and Two-Factor Fixed Effects ANOVA Cases, *J. Educ. Stat.*, **17**, 315-339.

Husen, S., E. Kissiling, N. Deichmann, S. Wiemer, D. Giardini et M. Baer (2003), Probabilistic Earthquake Location in Complex Three-Dimensional Velocity Models : Application to Switzerland, *J. Geophys. Res.*, **108**, doi : 10.1029/2002JB001778.

Husen, S. et J. L. Hardebeck (2010), Earthquake Location Accuracy, *Community Online Resource for Statistical Seismicity Analysis*, doi : 10.2078/corssa-55815573.

Lomax, A., J. Virieux, P. Volant et C. Berge-Thierry (2000), Probabilistic Earthquake Location in 3-D and Layered Models, *Advances in Seismic Event Location*, pp. 101-134, N. Kluwer Academic Publishers.

Lumley, T., P. Diehr, S. Emerson et L. Chen (2002), The Importance of the Normality Assumption in Large Public Health Data Sets, *Annu. Rev. Public Health* 2002, **23**, 151-169.

Moser, T. J., T. van Eck et G. Nolet (1992), Hypocenter Determination in Strongly Heterogeneous Earth Models Using the Shortest Path Method, *J. Geophys. Res.*, **97**, 6563-6572.

Pearson, E. S. (1931), The Analysis of Variance in Cases of Non-Normal Variation, *Biometrika*, **23**, 114-133.

Tarantola, A., et B. Valette (1982), Inverse Problems = Quest for Information, *J. Geophys.*, **50**, 159-170.