Université de Toulouse

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)

Présentée et soutenue par :
Jérémie Bureau

Le mercredi 19 décembre 2012

Titre :

Définition et analyse statistique d'une mesure d'intégrité pour données GPS/EGNOS

ED MITT : Domaine Mathématiques : Mathématiques appliquées

Unité de recherche :

Institut de Mathématiques de Toulouse

Directeur(s) de Thèse :

Jean-Michel Loubes, Université Toulouse III
Fabrice Gamboa, Université Toulouse III

Rapporteurs :

Bernard Bercu, Université Bordeaux I
Anne-Françoise Yao, Université Clermont-Ferrand II

Autre(s) membre(s) du jury :

Jean-Marc Azaïs, Université Toulouse III (Président)
Jean-Marc Bardet, Université Paris I (Examinateur)
Sylvain Tarance, Helileo SA (Invité)

N d'ordre : **** Année 2012

THESE

présentée
pour obtenir le titre de

DOCTEUR DE L'UNIVERSITE DE TOULOUSE

Spécialité : Mathématiques Appliquées

par

Jérémie Bureau

Définition et analyse statistique d'une mesure d'intégrité pour données GPS/EGNOS

Soutenue publiquement le 19 décembre 2012 après autorisations des rapporteurs :

Mme. Anne-Françoise Yao	Université Clermont-Ferrand II
M. Bernard Bercu	Université Bordeaux 1

Devant le jury composé de :

M. Jean-Marc Azaïs	Université Toulouse III	Président du Jury
M. Bernard Bercu	Université Bordeaux 1	Rapporteur
M. Jean-Marc Bardet	Université Paris I	Examinateur
M. Fabrice Gamboa	Université Toulouse III	Directeur de thèse
M. Jean-Michel Loubes	Université Toulouse III	Directeur de thèse
M. Sylvain Tarance	Helileo SA	Invité

Thèse préparée en collaboration entre

L'institut de Mathématiques de Toulouse
Equipe de Statistiques et Probabilités, Bât. 1R1
UniversitÈ Paul Sabatier Toulouse III
31 062 Toulouse Cedex 09

et

HELILEO SA
553, rue Bernard Palissy
Pôle économique d'agglomération
40 990 Saint-Paul-lès-Dax

Remerciements

Tout d'abord, mes remerciements s'adressent aux personnes qui m'ont proposé ce sujet de thèse, qui m'ont fait confiance et qui m'ont encadré pendant ces trois années d'études : Fabrice Gamboa et Jean-Michel Loubes. Merci pour votre disponibilité malgré les 300 kilomètres qui nous séparaient.

Je tiens à exprimer ma profonde gratitude à Bernard Panefieu, qui m'a accueilli au sein de la société HELILEO durant ces trois années.

J'adresse mes sincères remerciements à Bernard Bercu et Anne-Françoise Yao pour l'intérêt qu'ils ont manifesté pour ma thèse en acceptant d'en être rapporteur.

Je remercie Jean-Marc Azaïs qui m'a fait l'honneur de présider mon jury de thèse.

Je tiens également à remercier chaleureusement Jean-Marc Bardet pour avoir accepté de faire partie de mon jury de thèse.

Encore une fois, je remercie tout particulièrement Bernard Bercu qui depuis 2007 m'a suivi et fait confiance. Sans lui, ces travaux n'auraient pas vu le jour.

J'exprime toute mon amitié aux personnes que j'ai croisé durant ces trois années dans la société HELILEO : Sylvain, Jean-Baptiste, Maxence, Fred, Olivier, Eric, Yves, Nicolas, Mélanie, Valérie, Guillaume, Raymond, Yvon, Lise, Maxime et j'en oublie. Merci pour tous les moments passés en votre compagnie.

Merci à mes relecteurs, Marie-Lise et mes parents qui ont eu le courage de lire ce manuscrit du début à la fin malgré les équations mathématiques peu appétissantes pour eux. Mes pensées vont à Marie-Lise pour sa patience et son soutien lors de ces derniers mois éprouvants.
Merci également à Edward, Marie, Cédric, Nathalie et Chloé pour être venu assister à la soutenance.
Je clos enfin ces remerciements en dédiant cette thèse de doctorat à Georges, à mes parents et à mes amis.

Résumé

Parmi les applications GNSS (Global Navigation Satellite System) existantes ou en développement, certaines dont l'aviation, nécessitent de hautes performances en termes de précision de positionnement et de fiabilité.
Ces performances critiques sont évaluées à l'aide d'outils probabilistes et le problème d'appréciation de la précision ou de la fiabilité du système (intégrité) peut être vu comme une estimation de quantile. Ce problème inverse nécessite la connaissance de la fonction de répartition des observations, ce qui n'est pas le cas lorsque l'on travaille sur des données réelles. Il faut alors utiliser des techniques statistiques pour l'estimer.
Les exigences spécifiques à certaines applications, comme par exemple l'atterrissage d'un avion, nécessitent des niveaux de quantiles très élevés atteignant des probabilités de l'ordre de 10^{-7}. Ces probabilités correspondent à des fréquences d'occurrence d'événements rares, situés dans les queues de distribution. Les quantiles associés à de tels niveaux de probabilité sont qualifiés de quantiles extrêmes et se situent le plus souvent au-delà du domaine des observations.
Nous proposons dans cette thèse deux méthodes d'estimation de quantiles extrêmes peu employées dans le domaine du GNSS. La première est une application des modèles issus de la théorie des extrêmes et plus particulièrement du modèle à dépassement de seuil POT (Peak Over Threshold). Cette théorie fournit une classe de modèles permettant l'extrapolation de l'observé vers le non observé et ainsi la caractérisation des événements rares qui peuvent ne jamais avoir été observés. La deuxième méthode fournit une approximation de la décroissance de la queue d'une distribution au moyen de techniques analytiques adaptées à un cadre statistique : il s'agit de la méthode du point selle.
Ces deux techniques de caractérisation des fonctions de répartition sont valables sous certaines hypothèses de stationnarité et d'indépendance des observations ; or les données GPS ne vérifient pas toujours ces conditions. Dans ce travail, nous proposons des méthodes statistiques pour stationnariser les données afin d'utiliser les modèles d'estimation de quantiles extrêmes dans un cadre adéquat.
A partir des outils décrits dans cette thèse, nous fournissons un protocole d'analyse statistique d'intégrité. Les problématiques de calibration de ces outils sont traitées par des processus automatisés dans une plateforme d'analyse de données, support logiciel développé pour cette étude.

Mots-clefs : GNSS, GPS, intégrité, théorie des valeurs extrêmes, quantile extrême, dépendance temporelle, approximation point selle.

Abstract

Among the GNSS (Global Navigation Satellite System) applications currently used or in development, some of them require high performances in terms of precise positioning and reliability for safety of life.
These critical performances are evaluated using statistical tools, and the problem of measuring the position accuracy or the system reliability (integrity) can be modeled as a quantile estimation. This inverse problem requires the knowledge of the cumulative distribution function of the observations. This is not possible when we have to study real data, then it becomes necessary to use statistical techniques to estimate this function.
Specific safety of life applications, such as an airborne precision approach, require very high levels of quantiles which probabilities can reach 10^{-7}. These probabilities correspond to frequencies of rare events occurrence, located in the distribution tails. Quantiles associated to such levels of probability are qualified as extreme quantiles and are generally located beyond the observations domain.
We propose in this work two methods of extreme quantile estimation seldom used in the GNSS field. The first one is a direct application of the models stemming from extreme values theory and more particularly from the model of excesses over a threshold called POT (Peak Over Threshold). This theory provides a class of models allowing an extrapolation from the observed domain to the unobserved domain and then the characterization of rare events which never have been observed. The second method supplies an approximation for the decreasing of a distribution tail by the use of analytical techniques adapted to a statistical framework. This method is called derived from saddle point approximation technics.
These two techniques of tails distributions characterization are valid under certain hypothesis of stationnarity and independency of the observations. GPS data do not always satisfy these conditions. In this work, we propose statistical methods to reach these conditions allowing us to use the models of extreme quantile estimation in an adequate way.
From the tools studied in this thesis, we outline a statistical analysis methodology for integrity measurement. The problems of calibrating these tools are treated by automated processes in a data analysis platform, software developed as a support for this study.

Keywords : GNSS, GPS, integrity, extreme value theory, extreme quantile, temporal dependency, saddle point estimation.

Table des matières

Table des figures

Liste des tableaux

Acronymes

AL	Alarm Limit
CEP	Circular Error Probable
DGPS	Differential GPS
DOP	Dilution Of Precision
ESA	European Space Agency
GEV	Generalized Extreme Value
GNSS	Global Navigation Satellite System
GPD	Generalized Pareto Distribution
GPS	Global Positioning System
OACI	Organisation de l'Aviation Civile Internationale
PE	Position Error
PL	Protection Level
PVT	Position Vitesse Temps *a posteriori*
RAIM	Receiver Autonomous Integrity Monitoring
RMS	Root Mean Square
RTK	Real Time Kinematic
SBAS	Satellite Based Augmentation System
UERE	User Equivalent Range Error

Introduction générale

Contexte de l'étude

22h36 GMT, dans la nuit du 31 août au 1er septembre 1983, le Boeing 747 du vol KAL 007 de la compagnie Korean Airlines reliant New York à Seoul, est abattu par un avion de chasse de l'Union soviétique à l'ouest de l'île de Sakhaline. L'avion de ligne transportait 246 passagers et 23 membres d'équipage. Il n'y eut aucun survivant.
Cet événement majeur, sans doute l'un des plus graves de la Guerre froide, a failli faire basculer le monde dans une guerre nucléaire suite aux déclarations du Président américain Ronald Reagan qualifiant l'acte de "crime contre l'humanité". Les causes techniques de ce drame : une défaillance du système de navigation entraînant une large déviation de la trajectoire programmée et une violation de l'espace aérien soviétique à laquelle l'armée de l'URSS a répondu (Figure 1).
Suite à cet accident, le Président Reagan annonça que la technologie GPS (Global Positioning System), jusque-là destinée à un usage militaire, serait ouverte aux usages civils.

FIGURE 1 – Vol KAL 007 entre Anchorage (après ravitaillement) et Séoul

L'objectif du GPS est de fournir une position à l'utilisateur. Celle-ci doit être la plus précise et la plus fiable possible, tout particulièrement dans l'aviation où la vie des personnes est mise en jeu.
Afin d'accroître les performances du GPS et du futur Galileo (équivalent européen du GPS), un système de satellites géostationnaires européen EGNOS (European Global Navigation Overlay System) a été mis en place dès 2003. EGNOS contribue à améliorer la précision, la fiabilité et la disponibilité du système GPS sur le sol européen. Il fournit un service d'augmentation du GPS spécialement conçu pour les applications critiques liées aux transports. Ce service, connu sous le nom de Safety of Life (SoL) est alors destiné à

l'aviation, les transports maritimes et les transports ferroviaires.
Typiquement dans l'aéronautique, les principaux apports de ce service sont :

- pouvoir réaliser des approches par guidage vertical (approche de précision), jusqu'alors effectuées grâce à des moyens au sol d'aide à la navigation comme l'ILS.
- réaliser des vols de précision sur hélicoptères, en particulier pour les opérations de sauvetage en montagne.
- étendre la disponibilité des procédures d'approche de précision aux aéroports ne pouvant s'équiper de moyens au sol d'aide à la navigation très coûteux de type ILS.
- suivre plus fidèlement les trajectoires de vol et donc réduire l'émission de carbone dans l'atmosphère.

Pour cela, le système doit respecter des normes de sécurité très strictes fixées par l'Organisation de l'Aviation Civile Internationale (OACI) et adaptées à l'Europe par Eurocontrol, l'organisation européenne pour la sécurité de la navigation aérienne.

En pratique, les performances critiques du positionnement par satellites sont évaluées au moyen d'outils statistiques et les niveaux de fiabilité sont exprimés en probabilités.
Le concept de fiabilité des systèmes de navigation par satellites est appelé intégrité, définie comme une mesure de confiance que l'on peut accorder aux informations fournies par le système [DO96]. Les signaux transmis par les satellites peuvent subir des dégradations provenant de divers phénomènes (milieu atmosphérique traversé, interférence, masquage en milieu urbain, etc) ayant pour effet une position inacceptable pour l'utilisateur. Pour garantir un positionnement fiable, des seuils de tolérance ont été fixés afin d'écarter les positions qui les franchissent. On note ces seuils AL pour Alert Limit.
D'un point de vue probabiliste, le modèle de mesure d'intégrité est considéré comme un quantile d'ordre $1-p$ dont on rappelle la définition.

Définition 0.0.1. *Soit X une variable aléatoire réelle suivant une loi F. Pour p un réel de l'intervalle $[0,1]$, le quantile d'ordre $1-p$ est la valeur $q(p)$ vérifiant*

$$\mathbb{P}\big(X \leq q(p)\big) = 1-p$$

A travers ce modèle, on cherche à garantir que la variable X n'excède pas la valeur $q(p)$ pour une probabilité p donnée. Dans le modèle de mesure d'intégrité, on considère X comme le rapport des erreurs de positionnement et des seuils d'alarme AL. Ainsi, on évaluera l'intégrité selon la condition suivante : si $q(p) < 1$ alors l'intégrité est assurée suivant le modèle, sinon, elle ne l'est pas.
Les exigences spécifiques à certaines applications, comme l'atterrissage d'un avion, nécessitent des niveaux de quantiles très élevés (de l'ordre de $1-2.10^{-7}$) qui correspondent à des fréquences d'occurrence très faibles. En effet, on cherche à se protéger d'un événement qui a 2 chances sur 10 millions de se produire, et ce pour chaque approche (dont la durée est fixée à 150 secondes). Ces niveaux sont caractéristiques d'événements dits "rares" situés dans les queues de distribution.
Dans cette étude, nous proposons d'évaluer les performances d'intégrité du système GPS/EGNOS suivant le modèle ci-dessus et ainsi déterminer si les seuils d'alerte sont suffisamment élevés pour satisfaire les niveaux de fiabilité donnés par l'OACI. Le problème se traduit par l'estimation du quantile $q(p)$. De manière générale, l'estimation de quantile nécessite la connaissance de la distribution de la variable observée décrite par sa fonction de répartition.

Les techniques statistiques classiques permettent d'estimer la version empirique de cette fonction à partir de l'échantillon observé. La mesure empirique est efficace en présence d'information, pour des niveaux d'occurrence ni trop élevés ni trop faibles. La principale difficulté du problème de mesure d'intégrité provient des probabilités auxquelles nous sommes confrontés. En effet, ces niveaux très faibles sont associés à des événements très peu, voire jamais observés. Pour espérer obtenir une réalisation d'un événement associé à un niveau tel que $(1-2.10^{-7})/150s$, il faudrait plus de 20 années d'observations. Comment alors garantir que les niveaux de sécurité permettent de se protéger contre l'apparition de ces événements dangereux ?
L'objectif de ce travail de thèse est de fournir une méthodologie permettant de caractériser les queues de distribution, au moyen de méthodes statistiques encore méconnues dans le domaine de la navigation par satellites. Deux approches sont décrites :

- premièrement, la théorie des valeurs extrêmes [Bei04] fondée sur des théorèmes de convergence en loi permettant une extrapolation dans les queues de distribution. Cette théorie fournit une classe d'outils couramment utilisés pour la gestion du risque. Les domaines d'applications utilisant les modèles de la théorie des extrêmes n'ont cessé de se développer ces dernières années touchant des domaines variés comme la finance et l'assurance [EKM97], la climatologie [Mul06], le génie civil et le génie industriel [Gar02].
- la deuxième approche est une approximation de la décroissance d'une queue de distribution à partir de méthodes analytiques de type approximation point selle [But07] appliquées dans un cadre statistique.

A travers ces deux techniques nous formulons plusieurs estimateurs de quantiles dont les propriétés, les hypothèses de validité et les performances sont comparées puis discutées.

Avant que ces travaux de thèse ne soient initiés en 2009 par la société HELILEO, le système EGNOS et le service SoL étaient encore à l'état de validation. Pour leur certification liée aux activités héliportées, la GSA (European GNSS Supervisory Authority) a commandité le projet HEDGE (Helicopters Deploy GNSS in Europe) dans le cadre du 7ème PCRD (Programme Cadre de Recherche et Développement) technologique de l'Union Européenne. HELILEO, acteur européen majeur des essais héliportés [1], a été chargé de réaliser une vaste campagne d'essais sur hélicoptères à la suite de laquelle un rapport de performances a été remis à la GSA.

Le 2 mars 2011, l'ESSP (European Satellite Services Provider) a officiellement déclaré le signal Sol opérationnel et disponible pour l'aviation avec l'autorisation de la commission européenne, offrant ainsi la possibilité aux aéronefs de réaliser des phases d'atterrissage plus sûres quelques soient les conditions météorologiques.

Organisation du manuscrit

Le manuscrit est organisé en 5 chapitres.
Le chapitre 1 présente le contexte de l'étude. Nous introduisons la navigation par satellites ainsi que les concepts d'intégrité associés à des applications aéronautiques. Les

1. HELILEO dispose d'une flotte de 36 hélicoptères EC-120 pouvant réaliser jusqu'à 22 000 heures de vols par an, ce qui lui permet de réaliser des essais en vol pour évaluer les performances des systèmes GNSS

hauts niveaux de fiabilité imposés par l'OACI [DO96] se traduisent par des probabilités très élevées qui mènent le statisticien vers l'étude des queues de distribution.

Le chapitre 2 est consacré à l'étude des valeurs extrêmes d'un échantillon, événements dont les probabilités d'occurrence sont faibles. La théorie des extrêmes vient en complément de la statistique classique où il est usuel d'étudier les variables aléatoires autour de leurs moyennes. Il s'agit dans cette théorie de caractériser le comportement des queues de distribution à l'aide de modèles permettant un bon ajustement au-delà du maximum de l'échantillon. On mesure alors l'épaisseur de la queue de distribution à l'aide d'un unique paramètre appelé indice des valeurs extrêmes. Les fondements de cette théorie reposent sur un équivalent au théorème de la limite centrale appliqué aux queues de distribution : le théorème de Fisher et Tippet (1928) [FT28]. Celui-ci définit la famille de lois limites possibles pour le maximum (renormalisé) d'un échantillon. Le formalisme de la loi des extrêmes généralisée ne tient compte que d'une seule observation, la plus grande. Il n'est pas pensable d'estimer les paramètres de cette loi à partir d'une unique valeur. Pour cette raison, on emploie la méthode dite "block maxima" (Gumbel (1958)[Gum58]) qui consiste à découper l'échantillon initial en plusieurs blocs et à considérer le maximum de chacun d'eux afin de créer artificiellement un échantillon de maxima. Cependant, il peut arriver que plusieurs grandes valeurs apparaissent successivement dans un bloc. Elle seront alors masquées par la plus grande d'entre elles. Cette approche mène à une perte d'information sur les observations extrêmes qui sont par définition très rares.
Une alternative a été proposée suite aux travaux de Balkema, De Haan (1974) [BDH74] et Pickands (1975) [Pic75]. Il n'est plus question d'étudier la loi limite du maximum d'un échantillon mais les valeurs qui dépassent un certain seuil. Cette approche appelée Peak Over Threshold (POT) mène à de meilleurs résultats en termes de caractérisation de loi et d'estimation de quantile.
Nous développons dans ce chapitre les deux formalismes principaux issus de cette théorie dans un cadre univarié et on y compare les performances des estimateurs paramétriques caractérisant les lois limites. Le lecteur y trouvera plusieurs expressions d'estimateurs de quantile extrême ainsi que les intervalles de confiance associés.

Il existe d'autres méthodes statistiques pour caractériser les distributions des variables observées qui s'avèrent être efficaces pour les parties centrales mais aussi pour les queues de distribution. L'approximation point selle, introduite en statistique par Daniels (1954) [Dan54] permet d'estimer la densité de probabilité d'une variable aléatoire en utilisant la transformée de Fourier de cette fonction. A partir de cette approche, Lugannani et Rice [LR80] ont établi l'approximation d'une fonction de répartition menant à la résolution du problème inverse d'estimation de quantile.
On propose dans le chapitre 3 un nouvel estimateur de quantile extrême par inversion de l'approximation de la décroissance de la queue d'une distribution. Par analogie avec la méthode POT, on introduit un modèle à dépassement de seuil.

Afin d'identifier les outils les mieux adaptés au problème de mesure d'intégrité sur les données GPS, les performances des estimateurs, issus de la théorie des extrêmes (chapitre 2) et de l'approximation point selle (chapitre 3), sont évaluées sur des cas théoriques proches des données réelles.

Les données de positionnement GPS auxquelles on s'intéresse dans cette étude présentent

un fort caractère de dépendance temporelle. Cette dépendance temporelle est principalement due à deux facteurs : la récursivité des algorithmes d'estimation de la position implémentés au sein des récepteurs et la présence d'erreurs systématiques sur les signaux satellitaires. Les outils statistiques présentés dans les chapitres 2 et 3 ont été établis pour des variables aléatoires indépendantes et identiquement distribuées. Nous souhaitons casser la structure de dépendance temporelle qui existe au sein des données GPS en appliquant une permutation aléatoire sur les indices temporels du processus observé.
On propose dans un quatrième chapitre une justification théorique de cet heuristique ainsi qu'un théorème d'indépendance asymptotique atteinte après permutation.

Le cinquième et dernier chapitre de ce manuscrit est dédié aux applications et résultats pratiques en lien avec les activités de la société HELILEO.
Il s'agit dans un premier temps de fournir une analyse des données GPS intervenant dans le modèle de mesure d'intégrité afin de s'assurer que les conditions requises par les modèles probabilistes soient satisfaites. Les processus aléatoires décrits par les erreurs de positionnement ne sont pas stationnaires pour des enregistrements de courte durée. Toutefois, on montre que pour de grands échantillons, les données atteignent une stationnarité asymptotique.
Suite à cette analyse, on fournit une méthodologie destinée à évaluer si les positions fournies par les récepteurs GPS remplissent les exigences de sécurité des instances aéronautiques. Parmi les estimateurs de quantiles étudiés, ceux qui sont obtenus à partir du modèle à dépassement de seuil (POT) montrent les meilleures propriétés lorsqu'ils sont confrontés aux données GPS. Ils seront retenus pour l'analyse de performances de ce chapitre.
Les exemples théoriques du chapitre 2 mettent en évidence l'impact important du seuil sur la variabilité des estimateurs et de leurs intervalles de confiance. On propose une procédure automatique de choix de seuil fondée sur deux critères. Le premier est la détection de plages de stabilité des estimateurs. Le second est l'adéquation du modèle aux données.
Dans le cadre du projet HEDGE, HELILEO devait fournir en plus d'un rapport de performances, une justification de la quantité d'heures de vol engagée dans la campagne d'essais héliportés. On justifie ce nombre d'heures enregistrées par les temps de convergence des intervalles de confiance associés aux estimateurs de quantiles extrêmes utilisés pour l'analyse d'intégrité.
Acteur majeur des essais aéroportés pour les systèmes de radionavigation par satellites, la société HELILEO a dû se doter d'un outil de traitement de données performant. Pour répondre à ce besoin, une plateforme d'analyse statistique a été développée ; celle-ci étant capable à la fois d'héberger une base de données et de regrouper divers outils statistiques. En particulier, on y trouve le modèle de mesure d'intégrité décrit dans ce manuscrit. L'architecture ainsi que les fonctionnalités de la plateforme sont détaillées à la fin de ce dernier chapitre.

C H A P I T R E 1

Introduction à la radionavigation par satellites

Résumé :
Le GPS (Global Navigation System) est un système de radionavigation par satellites et a pour objectif de fournir une solution de position à l'utilisateur. Celle-ci doit être la plus fiable possible, en particulier pour les applications aéronautiques civiles comme militaires pour lesquelles les contraintes d'intégrité sont importantes. La position est calculée sur la base de mesures fournies par les satellites de la constellation GPS. Or ces mesures peuvent parfois devenir erronées (dysfonctionnement des satellites, présence de multi-trajets...), entraînant une dégradation de la position calculée inacceptable pour l'utilisateur. Ce chapitre a pour but de présenter les principes de fonctionnement du GPS ainsi que les concepts de fiabilité qui lui sont associés.

Sommaire

1.1 Introduction et historique du GPS

Au siècle dernier, les conflits géopolitiques opposant les idéologies capitalistes au communisme ont suscité une course effrénée à l'armement et à la conquête de l'espace. Ainsi les programmes spatiaux représentaient un enjeu majeur et étaient au centre des attentions du monde de la recherche. L'homme moderne, dans sa course au progrès, a cherché à reproduire un système artificiel permettant une localisation basée sur les mêmes principes que la navigation céleste pratiquée par nos ancêtres marins. Pour cela, il fallait une constellation de satellites en orbite autour de la Terre, diffusant un signal vers l'ensemble des zones visibles sur Terre, et un ensemble d'utilisateurs au sol qui après réception des signaux allaient estimer leur position à partir des informations fournies par les satellites. Les pionniers en matière de positionnement par satellites ont été les Etats-Unis en lançant en 1958, leur programme TRANSIT. Ce système, opérationnel en 1964, était destiné à un usage militaire et malgré un positionnement possible avec une précision de quelques centaines de mètres, il présentait deux inconvénients majeurs : le manque de disponibilité et une précision très médiocre. En effet, avec seulement 6 satellites en orbite autour de la Terre, l'utilisation de ce système n'était pas possible en tout point du globe et 24h sur 24h. C'est principalement ce manque de disponibilité qui a poussé le département de la défense des Etats-Unis à concevoir un système plus efficace qui allait être le futur GPS (Global Navigation System). Ce système devait alors être capable de fournir une position en tout point à la surface de la Terre, une vitesse ainsi qu'une échelle de temps

très précise. Pour cela, les Américains ont mis en oeuvre des moyens considérables et ont ainsi déployé une constellation de 24 satellites, dont le premier fut lancé en 1978. Le système fut déclaré opérationnel en 1995, et compte aujourd'hui plus de 30 satellites en orbite.
Bien conscientes de l'enjeu stratégique (militaire au départ, commercial par la suite) que représentait un tel système de positionnement, d'autres nations ont elles aussi voulu se lancer dans de vastes programmes de recherche visant à développer leur propre système de positionnement par satellites.

1.2 Les autres systèmes de radionavigation par satellites

La première nation à vouloir concurrencer le système américain fut sans nul doute l'URSS, qui durant les années 70, lança le programme GLONASS destiné à fournir un système de positionnement par satellites aux performances similaires à celles du GPS. GLONASS sera déclaré opérationnel en 1996 avec seulement 12 satellites effectifs, la guerre froide et l'explosion du bloc URSS ayant laissé quelques séquelles économiques. La Russie décide de relancer son programme en 2002 avec pour objectif un système complètement opérationnel en 2015 qui comptera à terme une constellation de 24 satellites.
L'Europe a elle aussi lancé son programme de réalisation d'un système de navigation par satellites connu sous le nom de GALILEO. Contrairement aux systèmes américain et russe, GALILEO demeure sous contrôle civil et devrait proposer un large éventail de services aux civils comme aux militaires. Les premières études datent de 1994, et le système est encore à ce jour à l'état de prototype et d'expérimentations avec quatre satellites lancés.
Enfin, la Chine avec le programme BEIDOU a elle aussi initié le développement d'un système de positionnement par satellites. Les premiers satellites ont été lancés en 2000.

1.3 Le GPS

Le système NAVSTAR, mieux connu sous le nom de GPS, système de navigation par satellites américain, est le seul système de positionnement global complètement opérationnel à ce jour. L'objet de ce chapitre est de présenter les principes de fonctionnement du système ainsi que son infrastructure de manière à mieux appréhender les problématiques engendrées par des applications aux niveaux de sécurité et de fiabilité très élevés. Cette description est principalement basée sur les deux ouvrages de référence suivants : [KH06] et [PS96].

1.3.1 Description du système

En introduction, nous avons évoqué uniquement les satellites de la constellation GPS mais ils ne sont pas les seuls éléments de l'architecture du système GPS : on décrit le système par trois segments.

- **un segment spatial**
  Initialement composé d'une constellation de 24 satellites évoluant à une altitude de 20200 km, il en comporte maintenant 32 depuis février 2008, répartis sur 6 plans

orbitaux ayant tous une inclinaison de 55° par rapport au plan équatorial terrestre.

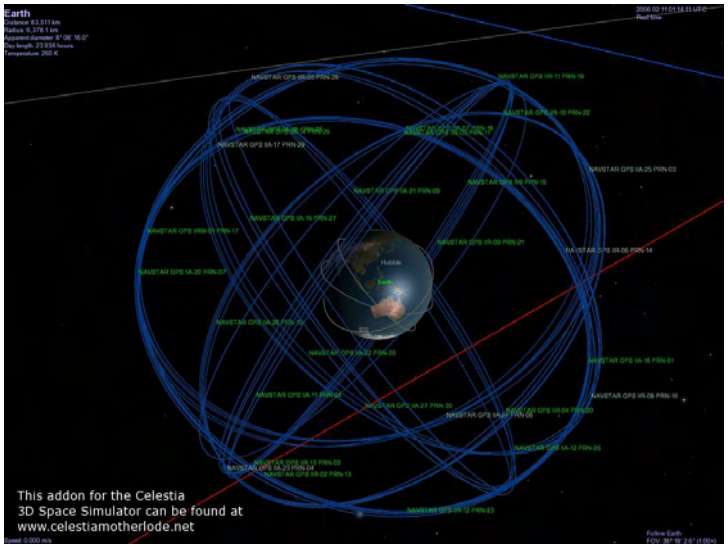


FIGURE 1.1 – Constellation GPS

- **un segment de contrôle**
  Ce segment est composé de 6 stations de contrôle dont une principale située dans l'état du Colorado. Le rôle de ces stations est de suivre, contrôler et piloter les satellites de la constellation. Ces stations auront aussi pour rôle d'enregistrer les signaux de la constellation, de les analyser puis de transmettre des informations de navigation aux satellites qui serviront de relais pour les utilisateurs.
- **un segment utilisateurs**
  L'ensemble des récepteurs civils et militaires à la surface du globe forment ce dernier segment. Les récepteurs reçoivent les signaux satellitaires et estiment leur position, leur vitesse et un temps précis par rapport à l'échelle de temps UTC (Coordinated Universal Time).

1.3.2 Fonctionnement

Le positionnement par satellites est basé sur un principe de triangulation dans lequel interviennent les distances séparant un ensemble de satellites observés et l'antenne du récepteur considéré. On parle aussi de principe de trilatération car on ne regarde pas l'intersection de segments mais de distances. Les satellites émettent des ondes électromagnétiques qui se propagent dans le vide à la vitesse de la lumière. Les distances entre satellites et antenne sont estimées à partir des temps de propagation des signaux. On parlera alors de pseudo-distances. Les satellites GPS sont équipés d'horloges atomiques (césium et rubidium) ultra stables et capables de délivrer la date d'émission d'un signal avec une précision de 10^{-9} secondes. Il n'en est pas de même pour les récepteurs à la surface du globe qui, par soucis de place et de coût, ont des oscillateurs beaucoup moins performants. Il existe donc un décalage entre les horloges des satellites (que l'on considère synchronisées entre elles) et les horloges des récepteurs. En plus, de sa position exprimée dans un repère à trois dimensions, le récepteur doit aussi estimer ce biais d'horloge. Il faut donc prendre en compte au minimum quatre satellites pour résoudre un système de

quatre équations à quatre inconnues.
La structure du signal GPS émis par les satellites permet de transmettre un ensemble d'informations qui sont nécessaires à la résolution du problème d'estimation de la position d'un utilisateur. Ces informations sont contenues dans un message de navigation qui est véhiculé par une onde porteuse.
L'objet de la section suivante est de décrire la méthode usuelle de résolution de ce problème.

1.4 Estimation de la position du récepteur

Toute position se détermine dans un repère donné. Avant de présenter les principes de calcul du positionnement par satellites, il est nécessaire d'introduire les principaux repères qui sont utilisés.

1.4.1 Les repères usuels et les modèles du géoïde

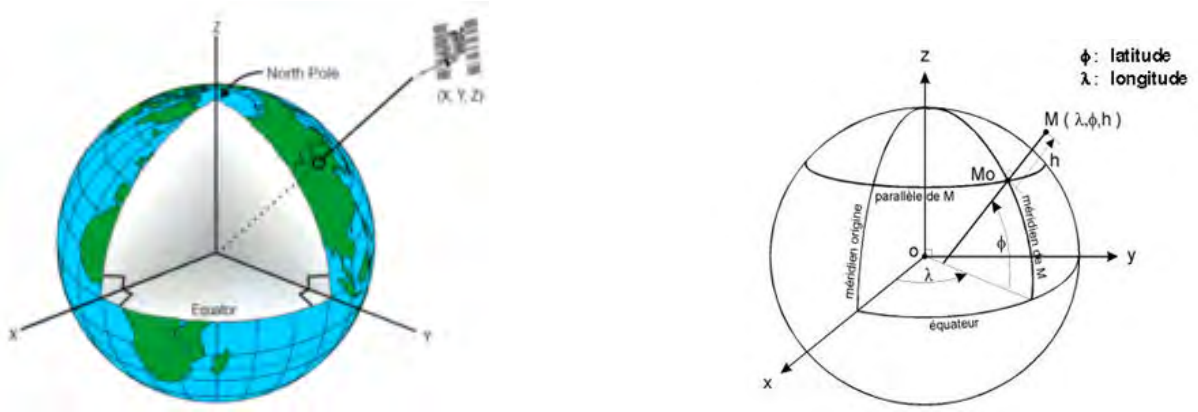


FIGURE 1.2 – Repères usuels

Les coordonnées (longitude, latitude, altitude) sont exprimées en valeurs angulaires (degrés) dans le repère géographique. Ce repère est un repère tridimensionnel dont l'origine est située au centre de la Terre. Les coordonnées d'un utilisateur seront le plus fréquemment données par les récepteurs dans ce repère. Cependant, les grandeurs en degrés présentent quelques inconvénients d'un point de vue calculatoire. Il est plus judicieux d'utiliser le repère cartésien pour faire des calculs de distances et autre.
Le référentiel ECEF (Earth Centered Earth Fixed) est un repère cartésien ayant comme origine le centre de masse de la Terre. L'axe Oz est l'axe de rotation de la Terre, l'axe Ox passe par l'intersection de l'équateur avec le méridien de Greenwich et l'axe Oy est orthogonal à Ox dans le plan équatorial (figure 1.2). Pour les calculs de localisation, les coordonnées des satellites ainsi que celles des utilisateurs seront exprimées dans ce référentiel. Le lecteur voulant connaître les techniques de conversion d'un référentiel à un autre pourra consulter les références suivantes [KH06], [Bor07] ou la notice de changement de système géodésique de l'IGN présentant plusieurs algorithmes de conversion.
Deux autres notions sont importantes lorsque l'on parle de localisation à la surface du

globe : l'ellipsoïde et le modèle de géoïde. En effet, un référentiel géographique est l'ensemble des conventions qui permettent d'associer à tout point de la surface terrestre un point unique sur une carte. Il est donc nécessaire d'utiliser une bonne modélisation du globe terrestre pour obtenir un positionnement précis. Un système de référence de coordonnées est associé à un ellipsoïde de révolution qui est un modèle mathématique simplifié de la Terre débarrassée de ses reliefs. Il s'agit d'un modèle sphérique aplati aux pôles, centré en O et défini par son demi grand axe et son demi petit axe. Afin d'avoir une modélisation plus fine de la surface du globe, on associe à l'ellipsoïde de révolution, un modèle de géoïde. Il décrit les irrégularités gravitationnelles de la Terre et il est défini comme l'équipotentiel du champ de pesanteur correspondant au mieux au niveau des mers. C'est un outil indispensable pour la conversion des hauteurs au dessus de l'ellipsoïde en altitude, coordonnée délivrée par les récepteurs, ainsi que pour le positionnement des satellites. Inventé par les mathématiciens et géodésiens allemands C.F. Gauss et J.B. Listing, il s'exprime sous plusieurs formes mathématiques : développement en harmoniques sphériques, polynômes, développement en ondelettes, grille de valeurs et méthode d'interpolation spatiale. Les coordonnées délivrées par les récepteurs GPS utilisent le modèle global d'ellipsoïde (et de géoïde associé) du DOD (Departement of Defence) Américain World Geodetic system 1984, WGS84. Il en existe d'autres pour des applications locales nécessitant plus de précision, par exemple en France métropolitaine, on pourra utiliser les modèles QGF98, RAF98 ou RAF09.

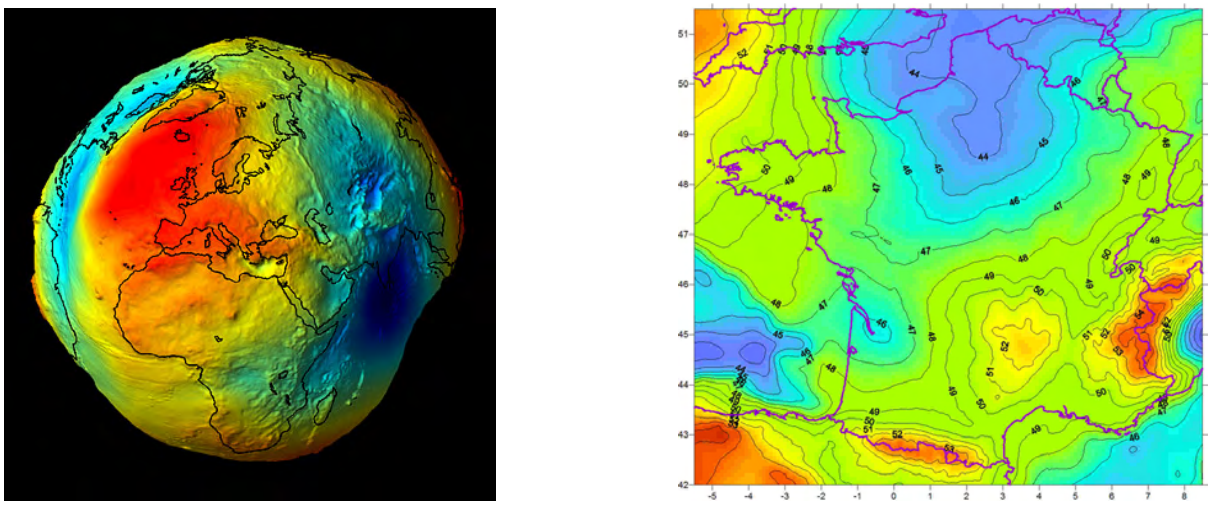


FIGURE 1.3 – Modèles de géoïde

1.4.2 Le modèle de mesure de pseudo-distance

Dans la section précédente, nous avons mentionné l'existence de deux échelles temporelles différentes : l'échelle de temps des horloges embarquées dans les satellites et l'échelle de temps propre au récepteur. Le calcul de la position se fait dans une troisième échelle de temps : le temps GPS. La pseudo distance entre un satellite i et le récepteur est exprimée par :

$$\rho_i = (t_r^{Rx} - t_e^{S_i}).\, c \tag{1.1}$$

où t_r^{Rx} est l'instant de réception du signal par le récepteur dans l'échelle de temps récepteur, $t_e^{S_i}$ est l'instant d'émission dans l'échelle de temps satellite et c la vitesse de propagation de la lumière dans le vide. Si on introduit la troisième échelle de temps GPS, on a

$$t_r^{Rx} = t_r^{GPS} - \Delta t^{Rx} \quad (1.2)$$
$$t_e^{S_i} = t_e^{GPS} - \Delta t^{S_i} \quad (1.3)$$

où Δt^{Rx} et Δt^{S_i} sont les décalages des horloges récepteur et satellite respectivement par rapport au temps GPS. Les horloges satellites étant considérées synchronisées entre elles, Δt^{S_i} sera noté Δt^S.
On note $\|\mathrm{Rx} - \mathrm{Sat}_i\|_2$ la distance géométrique entre le satellite i et le récepteur. Soit (X, Y, Z) et (X_i, Y_i, Z_i) les coordonnées du récepteur et du satellite i respectivement exprimées dans le repère cartésien. Dans l'échelle de temps GPS commune, cette distance s'exprime par

$$\|\mathrm{Rx} - \mathrm{Sat}_i\|_2 = (t_r^{GPS} - t_e^{GPS}).\, c \quad (1.4)$$
$$= \sqrt{(X - X_i)^2 + (Y - Y_i)^2 + (Z - Z_i)^2} \quad (1.5)$$

En combinant les expressions ci-dessus, on forme le modèle de mesure de pseudo-distance suivant :

$$\rho_i = \|\mathrm{Rx} - \mathrm{Sat}_i\|_2 + (\, \Delta t^S - \Delta t^{Rx}).c + \varepsilon_i \quad (1.6)$$

où ε_i est l'erreur de mesure. On note $\Delta t^S - \Delta t^{Rx} = \Delta b_h$ que l'on appelle biais d'horloge. Le terme d'erreur de mesure ε_i requiert une attention particulière et fait l'objet de la section suivante. La résolution du problème de navigation nécessite l'estimation des coordonnées du récepteur (X, Y, Z) ainsi que le biais d'horloge Δb_h. Il faut alors un minimum de quatre équations, et donc quatre satellites en vue pour former quatre mesures de pseudo-distance et résoudre ce système.

1.4.3 Résolution du problème de navigation

La résolution du problème de navigation se fait généralement à l'aide de méthodes statistiques de type moindres carrées. Cependant, on trouve aussi des algorithmes de techniques de filtrage non linéaire [KH06], [GDCT04], [Fau11], principalement lorsqu'il s'agit de fusionner plusieurs capteurs. L'idée est alors de combiner les informations fournies par les différents capteurs de façon à minimiser l'erreur d'estimation de la position. Parmi les hybridations possibles avec les récepteurs GPS, on trouve par exemple des accéléromètres, des centrales inertielles, des altimètres, des caméras vidéo etc... L'objet de cette section est de présenter le principe de la méthode usuelle de résolution du problème par la méthode des moindres carrés.
L'ensemble des n satellites pris en compte pour le positionnement fournissent n mesures de pseudo-distances et forment le système suivant :

$$\rho_1 = \sqrt{(X - X_1)^2 + (Y - Y_1)^2 + (Z - Z_1)^2} + \Delta b_h.c + \varepsilon_1 \quad (1.7)$$
$$\rho_2 = \sqrt{(X - X_2)^2 + (Y - Y_2)^2 + (Z - Z_2)^2} + \Delta b_h.c + \varepsilon_2 \quad (1.8)$$
$$\vdots \quad (1.9)$$
$$\rho_n = \sqrt{(X - X_n)^2 + (Y - Y_n)^2 + (Z - Z_n)^2} + \Delta b_h.c + \varepsilon_n \quad (1.10)$$

On note que les biais d'horloge Δb_h sont identiques quelque soit le satellite considéré car, on rappelle que l'on considère les horloges des différents satellites synchronisées entre elles. Ce n'est pas le cas en réalité, l'erreur de synchronisation sera prise en compte et (partiellement) corrigée par un modèle que l'on évoquera plus tard. Sous forme matricielle, on peut écrire ce système comme :

$$\mathbf{Y} = \mathbf{f}(\mathbf{X}) + \mathbf{E} \tag{1.11}$$

où $\mathbf{X}$ désigne le vecteur $(X, Y, Z, \Delta b_h)$, vecteur position de l'utilisateur concaténé avec le bais d'horloge, $\mathbf{E}$ le vecteur des bruits de mesures $(\varepsilon_1, \cdots, \varepsilon_n)$ supposé stationnaire gaussien centré, et f la fonction :

$$f : (x_1, x_2, x_3, x_4) \longmapsto \sqrt{(x_1 - X_i)^2 + (x_2 - Y_i)^2 + (x_3 - Z_i)^2} + x_4.c \tag{1.12}$$

L'algorithme des moindres carrés est utilisé pour résoudre des problèmes linéaires se présentant sous la forme $\mathbf{Y} = \mathbf{AX} + \mathbf{b}$. La présence de racines carrées dans les expressions des mesures de pseudo-distances implique que le problème n'est pas linéaire. Il existe toutefois des techniques de linéarisation pour contourner ce problème. On effectue un développement de Taylor de la racine carrée à l'ordre 1 au point $(X_0, Y_0, Z_0, \Delta b_{h,0})$. On a alors au voisinage de $\mathbf{X}_0$:

$$f(\mathbf{X}) \approx f(\mathbf{X}_0) + \frac{\partial f}{\partial \mathbf{X}}(\mathbf{X}_0).(\mathbf{X} - \mathbf{X}_0) + r(\mathbf{X}) \tag{1.13}$$

où $r(\mathbf{X})$ est le reste du développement de Taylor. En revenant à l'expression (11), on obtient

$$\mathbf{Y} - f(\mathbf{X}_0) = \frac{\partial f}{\partial \mathbf{X}}(\mathbf{X}_0).(\mathbf{X} - \mathbf{X}_0) + \mathbf{E} \tag{1.14}$$

Le reste $r(\mathbf{X})$ étant maintenant compris dans le vecteur des erreurs de mesures $\mathbf{E}$. Pour une mesure de pseudo-distance provenant du satellite i, les dérivées partielles de la fonction f_i au point $\mathbf{X}_0$ s'écrivent :

$$\frac{\partial f_i}{\partial X}(\mathbf{X}_0) = \frac{X_0 - X_i}{\sqrt{(X_0 - X_i)^2 + (Y_0 - Y_i)^2 + (Z_0 - Z_i)^2}} \tag{1.15}$$

$$\tag{1.16}$$

$$\frac{\partial f_i}{\partial Y}(\mathbf{X}_0) = \frac{Y_0 - Y_i}{\sqrt{(X_0 - X_i)^2 + (Y_0 - Y_i)^2 + (Z_0 - Z_i)^2}} \tag{1.17}$$

$$\tag{1.18}$$

$$\frac{\partial f_i}{\partial Z}(\mathbf{X}_0) = \frac{Z_0 - Z_i}{\sqrt{(X_0 - X_i)^2 + (Y_0 - Y_i)^2 + (Z_0 - Z_i)^2}} \tag{1.19}$$

$$\tag{1.20}$$

$$\frac{\partial f_i}{\partial \Delta b_h}(\mathbf{X}_0) = 1 \tag{1.21}$$

Soit $\mathbf{H}$ la matrice de dimension $n \times 4$ associée au système formé par les n mesures de pseudo-distance provenant de l'ensemble des n satellites en vue pris en compte pour le calcul de la position :

$$\mathbf{H} = \begin{bmatrix} \frac{\partial f_1}{\partial X}(\mathbf{X}_0) & \frac{\partial f_1}{\partial Y}(\mathbf{X}_0) & \frac{\partial f_1}{\partial Z}(\mathbf{X}_0) & \frac{\partial f_1}{\partial \Delta b_h}(\mathbf{X}_0) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_n}{\partial X}(\mathbf{X}_0) & \frac{\partial f_n}{\partial Y}(\mathbf{X}_0) & \frac{\partial f_n}{\partial Z}(\mathbf{X}_0) & \frac{\partial f_n}{\partial \Delta b_h}(\mathbf{X}_0) \end{bmatrix} \tag{1.22}$$

Le problème peut alors s'écrire sous une forme matricielle linéarisée :

$$\mathbf{Y} - f(\mathbf{X}_0) = \mathbf{H}.(\mathbf{X} - \mathbf{X}_0) + \mathbf{E} \tag{1.23}$$

C'est sur ce modèle que l'on va appliquer la méthode des moindres carrés pour estimer le vecteur $(\mathbf{X} - \mathbf{X}_0)$ noté $\Delta\mathbf{X}$. Par souci de lisibilité, on note $\mathbf{Y} - f(\mathbf{X}_0) = \Delta\mathbf{Y}$. Soit σ^2 la variance de l'erreur de mesure de pseudo-distance ε_i fournie par le satellite i. On fait l'hypothèse que pour tout i compris entre 1 et n, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ et $\mathbb{E}[\varepsilon_i] = 0$. Cela signifie que l'on considère que toutes les erreurs de mesures de pseudo-distance seront centrées et auront la même variance. On appelle estimateur des moindres carrés de $\Delta\mathbf{X}$, la valeur qui minimise la fonction :

$$\Delta\mathbf{X} \longmapsto \|\Delta\mathbf{Y} - \mathbf{H}.\Delta\mathbf{X}\| \tag{1.24}$$

L'estimateur des moindres carrés $\widehat{\Delta}\mathbf{X}$ de $\Delta\mathbf{X}$ satisfait :

$$\widehat{\Delta}\mathbf{X} = (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Delta\mathbf{Y} \tag{1.25}$$

A partir de cet estimateur, on peut prédire $\Delta\mathbf{Y}$ par $\widehat{\Delta}\mathbf{Y} = \mathbf{H}.\widehat{\Delta}\mathbf{X}$ et E par $\widehat{E} = \Delta\mathbf{Y} - \mathbf{H}.\widehat{\Delta}\mathbf{X}$. Ces estimateurs seront utiles dans les sections suivantes pour les concepts de contrôle d'intégrité. Lorsque l'on dispose d'au moins quatre mesures de pseudo-distance et d'un point X_0, on est alors capable d'estimer les coordonnées (X, Y, Z) du récepteur ainsi que son biais d'horloge Δb_h. On appelle le point X_0 point d'initialisation. Si on place le problème de navigation précédent dans un contexte temporel, chaque mesure de pseudo-distance sera une fonction du temps. Généralement, la fréquence de mesure des récepteurs est la seconde. On trouve toutefois des matériels pouvant fournir des mesures à une cadence bien plus élevée, de l'ordre de la milliseconde. A chaque instant, le récepteur devra former le système de mesures de pseudo-distance $\Delta\mathbf{Y}(t)$ pour estimer le vecteur $\Delta\mathbf{X}(t)$. Le point d'initialisation $X_0(t_1)$ sera pour la première position à estimer, à l'allumage du récepteur, l'origine du repère cartésien (X_0, Y_0, Z_0). A l'instant suivant, $X_0(t_2)$ sera la position $(\widehat{X}(t_1), \widehat{Y}(t_1), \widehat{Z}(t_1))$ estimée par le récepteur à l'instant précédent t_1 et ainsi de suite. Toutefois, il existe une boucle itérative supplémentaire dans l'algorithme d'estimation de la position : en effet, à chaque instant, le récepteur va estimer successivement la même position tant que la position estimée $(\widehat{X}^j(t_1), \widehat{Y}^j(t_1), \widehat{Z}^j(t_1))$ ne sera pas assez proche de l'estimation de la position suivante $(\widehat{X}^{j+1}(t_1), \widehat{Y}^{j+1}(t_1), \widehat{Z}^{j+1}(t_1))$. La valeur seuil est généralement de 1 à 3 millimètres. On remarque que l'on considère que ces itérations se font au même instant t_k, alors que le calcul n'est pas réellement instantané. La convergence de ces boucles est très rapide [Far08], il faut généralement moins de 6 itérations pour que l'estimation de la position soit définitivement conservée par le récepteur. Ceci explique pourquoi la première position fournie par le récepteur est satisfaisante en ayant pour point d'initialisation X_0, le centre de la Terre.
Il existe plusieurs sources d'erreurs susceptibles d'affecter les mesures de pseudo-distance et le modèle de mesure présenté dans cette section peut être amélioré par des modèles de corrections des différentes sources d'erreurs que l'on a identifiées et qui suscite toujours une grande ferveur dans le monde de la recherche actuelle.

1.4.4 Les différentes sources d'erreurs et leurs impacts

En suivant le trajet parcouru par l'onde électromagnétique émise par le satellite et reçu par le récepteur, on distingue cinq classes de sources d'erreurs.

Les erreurs provenant des satellites :
Bien que très précises, les horloges atomiques embarquées dans les satellites ont une faible dérive au cours du temps. Un des rôles des stations de contrôle au sol est d'évaluer cette dérive de manière à transmettre aux utilisateurs les corrections à appliquer aux mesures de pseudo-distance. Les éphémérides diffusées par les satellites ne sont pas parfaitement exactes. En effet, elles contiennent les positions des satellites de la constellation données par un modèle d'orbite théorique. Les satellites sont en réalité soumis à de faibles mouvements qui les éloignent de leurs orbites théoriques.

Les erreurs provenant du milieu traversé :
L'ionosphère et la troposphère sont deux couches atmosphériques affectant le temps de propagation du signal GPS et ainsi les mesures de pseudo-distances. L'ionosphère est une couche atmosphérique située entre 50 et quelques centaines de kilomètres au dessus de la Terre. Elle a la particularité d'être fortement ionisée par les radiations solaires et ainsi de perturber la propagation des ondes électromagnétiques qui la traversent. Le modèle de Klobuchar [Klo87] permet de corriger partiellement le retard de propagation du signal ; ses paramètres sont transmis via le message de navigation aux utilisateurs. Dans le cycle solaire, dont la période est de 11 ans [Col73], les corrections calculées par le modèle seront moins pertinentes en période de forte activité. C'est une des raisons pour laquelle, l'étude du comportement de l'ionosphère ainsi que son impact sur les mesures GPS sont un des grands centres d'intérêts de la recherche actuelle [NCSS12], [MXWL12].
La troposphère est une couche atmosphérique en contact avec le sol qui s'élève à une altitude comprise entre 10 et 20 kilomètres, des pôles à l'équateur respectivement. La variation de la vitesse de propagation des ondes à travers cette couche atmosphérique est due principalement à des effets résultant du changement de l'indice de réfraction de cette couche. Ce changement est fonction de l'altitude, de la température et de l'humidité. Il existe plusieurs modèles pour corriger ces effets, cependant c'est le modèle de Hopfield [Hop71] qui est implanté dans la plupart des récepteurs.

Les erreurs provenant du mouvement des satellites :
Les mouvements des satellites et des utilisateurs (vitesse de rotation de la Terre) engendrent des effets relativistes agissant sur les horloges embarquées dans les satellites ainsi que des effets Doppler des signaux. Il faudra donc utiliser un modèle de correction des horloges satellites.

Les erreurs provenant de l'environnement de réception :
L'estimation de la position d'un récepteur présentée dans la section précédente suppose que les mesures de pseudo-distance sont faites sur l'onde incidente directe. Toutes réflexions ou diffractions sur les obstacles présents dans le milieu du récepteur induiront des retards de propagations des signaux qui, s'ils sont suffisamment importants, dégraderont les mesures de pseudo-distance et ainsi la précision de positionnement. Ces phénomènes locaux sont appelés multi-trajets et restent une cause d'erreur non modélisée. Il existe toutefois des techniques de filtrages non linéaires [GTC07] qui permettent de les prendre en compte dans l'estimation de la position.
Une autre source d'erreur est la présence d'interférences dans l'environnement de réception. Le principe de fonctionnement des systèmes de localisation par satellites (signaux radio-diffusés) les rendent sensibles à plusieurs types de perturbations provenant d'autres

signaux radio-diffusés. Du fait du niveau très faible de puissance des signaux GPS reçus, ces derniers se voient occultés par de nombreux signaux émis sur des fréquences proches de celles qu'ils utilisent. De plus, il est possible que d'autres émissions, suffisamment puissantes, sur des fréquences non voisines puissent gêner son fonctionnement avec des harmoniques d'ordre relativement élevé. Ce phénomène de perturbation ou de masquage des signaux de radionavigation par d'autres systèmes télécommunicants est ce que l'on appelle interférence. Les erreurs dues aux interférences ne sont pas modélisées dans le cas des systèmes GNSS mais plusieurs techniques agissant à différents étages des récepteurs permettent de s'en protéger [GTD10] [CBV+04].

Les erreurs provenant du récepteur :
Le récepteur lui même induit des erreurs de mesure. En effet, l'antenne qui transforme l'onde électromagnétique bruitée reçue en signal électrique, ainsi que le bruit thermique des composants détériorent la mesure. L'erreur induite par le récepteur sera fonction de la qualité de ce dernier.

Parmi les différentes erreurs que l'on vient de citer, certaines étaient prédictibles par modélisation : les erreurs brutes provenant du milieu traversé (erreurs ionosphériques et troposphériques), les erreurs dues aux effets relativistes et les erreurs dues aux faibles dérives des horloges satellites. Les mesures de pseudo-distance peuvent être partiellement corrigées en injectant ces modèles d'erreurs dans le modèle de mesure de pseudo-distance. Pour une mesure provenant d'un satellite i, on aura alors :

$$\rho_i = \sqrt{(X - X_i)^2 + (Y - Y_i)^2 + (Z - Z_i)^2} + \Delta b_h.c + \Delta_I + \Delta_T + \Delta_{rel} + \Delta_{cl} + \varepsilon_i \quad (1.26)$$

où Δ_I et Δ_T sont les corrections estimées par les modèles d'erreurs ionosphériques et troposphérique (Klobuchar et Hopfield généralement), Δ_{rel} et Δ_{cl} sont les corrections estimées par les modèles associés aux erreurs d'horloges satellites et aux erreurs relativistes. Ces corrections sont exprimées en mètres. On trouvera ces modèles dans [PS96]. L'erreur résiduelle ε_i regroupe alors toutes les erreurs qui n'ont pas pu être corrigées par des modèles. A partir des différentes composantes d'erreurs, on détermine un estimateur appelé UERE (User Equivalent Range Error) de l'erreur globale qui donne la précision de la mesure de pseudo-distance entre l'utilisateur et chaque satellite. Le tableau (1.1) ci-dessous recense les principaux types d'erreurs ainsi que leurs contributions à l'erreur globale UERE.

Segment	Source d'erreur	Erreur (m)
Spacial	Dérive des horloges satellites	3.0
	Effets relativistes	1.0
Contrôle	Éphémérides	4.2
	Autres	0.9
Utilisateur	Erreur ionosphérique	10
	Erreur troposphérique	2.0
	Bruit du récepteur	4.8
UERE		11.3

TABLE 1.1 – Bilan d'erreur pour les mesures de pseudo-distance.

Remarque : Pour construire cet estimateur, chaque composante est supposée suivre une distribution gaussienne centrée (*ie.* d'espérance nulle) et de variance égale à la valeur citée. Ce n'est pas le cas dans la réalité ; les distributions sont rarement centrées. Plusieurs phénomènes peuvent être responsables de l'apparition de petits biais nominaux : multi-trajets, déformation du signal ou mauvaise calibration du centre de phase de l'antenne réceptrice [Mar08].

L'impact de chaque composante est fonction des conditions environnementales rencontrées. Cependant, les performances du système observées sur une longue période sont assez stables (*cf.* chapitre 5).

1.4.5 Performances et limitations du système

Critères de performances :[DO96]
On exprime les performances d'un système de positionnement par satellites selon quatre critères : précision, intégrité, continuité et disponibilité.

- La précision de positionnement est définie comme l'écart entre la valeur mesurée et la valeur vraie de la position de l'utilisateur.
- L'intégrité est une mesure de confiance que l'on peut accorder à l'exactitude des informations fournies par le système. Le concept d'intégrité comprend aussi la capacité du système à avertir correctement l'utilisateur de la présence d'une défaillance du système, dans un temps défini.
  La précision de positionnement et l'intégrité du système sont mesurées par des indicateurs statistiques qui seront présentés dans le paragraphe suivant.
- La continuité du système est sa capacité à fonctionner de manière continue durant une certaine opération, sans interruption non volontaire. La continuité est mesurée par la probabilité que les performances de précision et d'intégrité du système restent remplies durant une opération (ex : durant l'atterrissage d'un avion).
- La disponibilité du système est sa capacité à fournir les services quand les utilisateurs en ont besoin. La disponibilité est mesurée par le pourcentage de temps où, sur une zone géographique, les critères de précision, d'intégrité et de continuité sont remplis.

Mesure de performances de précision :
Généralement, on étudie la précision d'un système de navigation par satellites sur le plan horizontal, et sur l'axe vertical. On parlera d'erreur horizontale et d'erreur verticale. Il est important de dissocier ces deux composantes de l'erreur globale car les systèmes GNSS sont connus pour être moins précis sur l'axe vertical que sur le plan horizontal.
Les performances de précision sont mesurées à l'aide d'indicateurs statistiques classiques que l'on nomme quantiles, et qui sont définis pour une certaine probabilité p. Par convention [DO96], la précision d'un récepteur est annoncée suivant un quantile de niveau 95%. C'est-à-dire que pour une durée d'enregistrement donnée, si on observe l'erreur de positionnement horizontale, on va retenir le rayon du cercle qui contient 95% des observations.
Il existe d'autres mesures de précision de positionnement. L'indicateur CEP (Circular Error Probable) correspondant à un quantile de niveau 50%, est aussi couramment utilisé par les constructeurs pour annoncer les performances de leurs appareils. Prudence ! Si cet indicateur est plus flatteur d'un point de vue constructeur, il peut s'avérer être traître pour les utilisateurs.
Que signifie la phrase : "mon récepteur GPS est précis à 4 mètres près" ? Tout le monde

comprendra que le récepteur est censé fournir des mesures de position dans un rayon de 4 mètres autour de l'antenne dans le plan horizontal. C'est vrai mais la question à se poser est : selon quel critère statistique ? C'est une nuance à laquelle il faut prêter une attention particulière. Il est clair que pour un enregistrement donné, la valeur du quantile de niveau 95% sera supérieure à la valeur du quantile de niveau 50%.

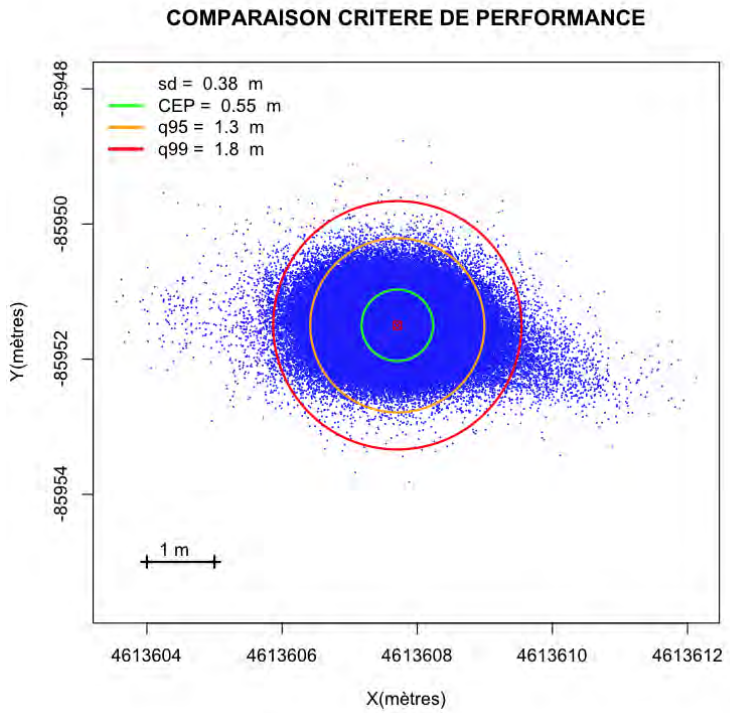
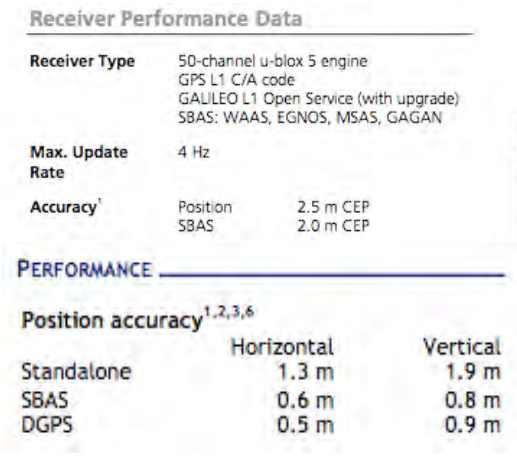


FIGURE 1.4 – Comparaison des différents indicateurs statistiques

On voit sur la figure 1.4 qu'il y a une nette différence entre les rayons des trois cercles. Le nuage de points est toutefois assez concentré puisque 50% des mesures sont comprises dans le cercle vert. Les indicateurs q95 et q99 correspondant aux niveaux de quantile 95% et 99% sont certes plus élevés mais offrent une meilleure garantie pour l'utilisateur. On a peu de chance d'avoir des mesures au delà de ces valeurs. L'idéal pour un constructeur serait de fournir suffisamment d'indicateurs différents de façon à appréhender au mieux le nuage de points et donc le comportement du récepteur : écart-type pour évaluer la dispersion, noté sd sur la figure 1.4, et les quantiles CEP,q95 et q99.
Aussi, l'utilisateur doit être vigilant lorsqu'il s'agit de comparer plusieurs récepteurs. Il faut alors disposer des mêmes indicateurs statistiques de précision pour que la comparaison ait un sens. La figure suivante montre trois extraits de fiche constructeur dont les noms ne seront pas cités. Le premier constructeur (en haut à gauche) annonce les performances du récepteur considéré avec l'indicateur CEP, tandis que le second (à droite) donne à la fois l'indicateur CEP et le quantile à 95%. Quant au troisième, il ne précise pas quel indicateur est utilisé (pas d'information à ce sujet dans les notes 1,2,3,6).
Une dernière remarque concernant la notion de précision : il ne faut pas confondre précision et dispersion. Dans les documents anglo-saxon on trouvera les termes "accuracy" et "precision". "Accuracy" désigne ce que nous appelons précision telle qu'elle est définie au paragraphe précédent. "Precision" désigne le caractère dispersif d'un nuage de



Receiver Performance Data

Receiver Type	50-channel u-blox 5 engine GPS L1 C/A code GALILEO L1 Open Service (with upgrade) SBAS: WAAS, EGNOS, MSAS, GAGAN	
Max. Update Rate	4 Hz	
Accuracy[1]	Position SBAS	2.5 m CEP 2.0 m CEP

PERFORMANCE

Position accuracy[1,2,3,6]

	Horizontal	Vertical
Standalone	1.3 m	1.9 m
SBAS	0.6 m	0.8 m
DGPS	0.5 m	0.9 m

Table 1.2 Performance Specifications

Item	Specification		
Real-time position accuracy		Horizontal CEP	Horizontal 95%
	Autonomous:	3.0m	5.0m
	SBAS:	1.0m	3.0m
	DGPS (local):	0.8m	1.5m

FIGURE 1.5 – Fiches constructeurs

points ; on trouvera cette notion sous le nom d'exactitude dans la littérature francophone.

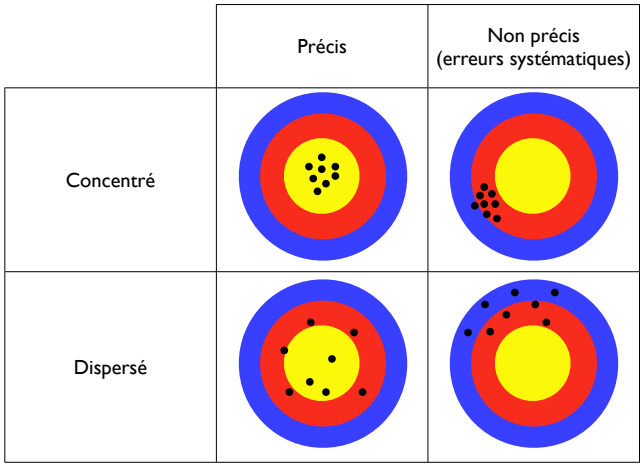


FIGURE 1.6 – Précision et exactitude

Mesure de performances d'intégrité :
L'intégrité d'un système de positionnement par satellites est défini selon trois variables : le risque d'intégrité, les seuils d'alerte et le temps d'alerte.
Le risque d'intégrité est la probabilité que l'erreur de positionnement dépasse un seuil d'alerte sans en informer l'utilisateur, pendant un temps donné.
Les seuils d'alerte HAL et VAL, Horizontal Alert Limit et Vertical Alert Limit, sont les seuils maximaux tolérables pour l'erreur de positionnement. Ces seuils seront variables d'une application à une autre. Par exemple, dans un contexte aéronautique, le HAL pour une phase d'approche guidée de catégorie 1 est de 40 mètres, tandis qu'il est de 7.4 kilomètres pour une phase de vol En-route [ICA06]. Ils sont définis comme les quantiles pour un niveau de probabilité correspondant à une certaine phase de vol, la variable aléatoire observée est l'erreur de positionnement (horizontale et verticale).
Le temps d'alerte est le temps maximal acceptable pour que le système alerte l'utilisateur de la présence d'une défaillance.

Performances de précision du GPS :

	Spécifications	Performances
Horizontale	$\leq$ 17 m (95%)	$\leq$ 7.1 m (95%)
Verticale	$\leq$ 37 m (95%)	$\leq$ 13.2 m (95%)

TABLE 1.2 – Performances du GPS

La précision du GPS s'est fortement améliorée au cours des dernières années notamment grâce aux progrès sur les modèles d'erreurs de pseudo-distance ; de plus en plus d'applications nécessitant une précision sub-métrique ont vu le jour. Concernant l'intégrité, le GPS seul ne permet pas de garantir les très hauts niveaux d'intégrité requis par certaines applications (aéronautique par exemple). On recense trois types d'applications nécessitant de l'intégrité : les applications critiques de transport (mettant en jeu la vie des personnes), les applications commerciales, et les applications juridiquement sensibles. Pour pallier ces limitations, en terme de précision et d'intégrité, des systèmes d'amélioration ont été conçus.

1.5 Les systèmes complémentaires

La grande disponibilité et les bonnes performances du système GPS ont poussé les utilisateurs à utiliser ce système de positionnement pour de nombreuses applications nécessitant des performances de très haut niveau en terme de précision et de fiabilité. On peut citer les procédures d'approche de précision pour l'aéronautique, les systèmes de commande automatique des trains, les applications géodésiques, les manoeuvres portuaires pour les navires de transport de marchandises etc... Cependant, pour que ces applications puissent bénéficier d'un système de positionnement fiable, il a fallu mettre en place des systèmes complémentaires permettant d'améliorer les performances du système et d'en évaluer la fiabilité en temps réel. On appelle ces systèmes complémentaires, système d'augmentation, et on les classe selon trois catégories : les infrastructures spatiales, les infrastructures au sol et les systèmes de contrôle intégrés dans les récepteurs.

1.5.1 Terrestres

Les moyens d'augmentation terrestres sont fondés sur des techniques de positionnement relatif où une position de référence est nécessaire afin de calculer puis transmettre des corrections aux utilisateurs.

GPS différentiel (DGPS) Le GPS différentiel est un moyen de positionnement relatif temps réel qui permet grâce à une station de référence de réduire de manière très significative les erreurs sur les mesures de pseudo-distance et ainsi obtenir un positionnement précis. La station de référence fixe, dont la position est connue avec une grande précision et un haut niveau de confiance, peut évaluer les erreurs commises sur les mesures de pseudo-distance en temps réel. Ces erreurs sont estimées à partir de pseudo-distances

théoriques déterminées grâce aux positions des satellites et de la station de référence. La station va alors émettre en temps réel par voie hertzienne les corrections à appliquer aux pseudo-distances mesurées. Les récepteurs des utilisateurs présents dans la zone de couverture de la station vont intégrer ces informations dans le calcul de leur position et obtenir une position précise. Ce type de système d'augmentation permet de réduire les erreurs de mesure sur les pseudo-distances provenant de la quasi totalité des différentes sources à l'exception des erreurs propres au récepteur et de son environnement (multi-trajets). La précision obtenue est fonction de la distance entre le récepteur et la station de référence.
A plus petite échelle, les systèmes de positionnement utilisés par les géomètres ont été développés selon le même principe : deux récepteurs et un moyen de communication entre les deux (figure 1.7). L'un des deux récepteur sera utilisé comme référence.

FIGURE 1.7 – Système de positionnement géodésique

Ground Based Augmentation System (GBAS) Le GBAS est un système d'augmentation mis en place pour l'OACI (Organisation de l'Aviation Civile Internationale) afin d'augmenter les performances des systèmes de positionnement par satellites pour les phases de vol critiques correspondant au décollage et à l'atterrissage d'un avion. Ce système est composé d'un réseau de stations de référence qui émettent des informations d'intégrité et des corrections différentielles sur les fréquences VHF qui seront prises en compte par les systèmes de bord des avions (figure 1.8).

Real Time Kinematic (RTK) La technique de positionnement RTK est une technique différentielle similaire au DGPS. Cependant, la station de référence ne transmet plus des corrections à appliquer aux mesures de pseudo-distance mais des données brutes. De plus, les différences ne se font plus sur les mesures de code mais sur les mesures de phase. Cette technique nécessite des récepteurs bi-fréquences L1 et L2. Le positionnement RTK permet d'atteindre une précision centimétrique en temps réelle. Les systèmes de positionnement géodésiques utilisent aujourd'hui cette technique de positionnement.

1.5.2 Spatiaux

Les différents systèmes SBAS
Les systèmes d'augmentation spatiaux ont vu le jour à la suite de vastes programmes

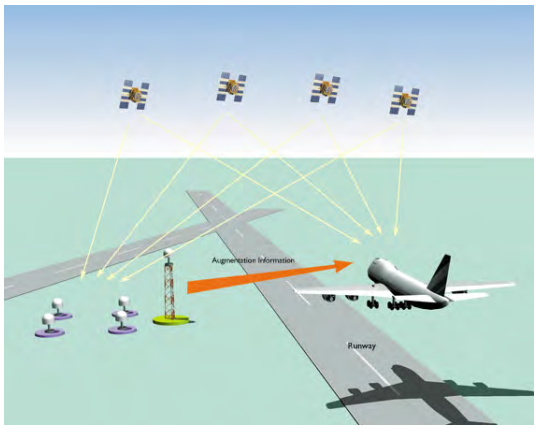


FIGURE 1.8 – Système d'augmentation GBAS

lancés à partir de 1994, suite à l'autorisation du gouvernement américain envers l'OACI pour utiliser le GPS. A l'époque, les signaux des satellites GPS étaient encore dégradés volontairement (Selective Availability, SA, supprimée par les américains en mai 2000) par les autorités américaines et offraientt alors une précision très médiocre par rapport aux performances actuelles (une centaine de mètres). Afin d'utiliser ce moyen de navigation, l'aviation civile internationale a engagé de larges programmes d'études visant à améliorer considérablement les performances du GPS en matière de précision, de disponibilité, de continuité et d'intégrité. Le concept du SBAS (Satellite Based Augmentation System) est apparu. Il s'agit d'un système permettant de transmettre des corrections différentielles aux utilisateurs via un réseau de stations et des satellites géostationnaires. Les stations estiment les corrections à appliquer aux mesures de pseudo-distance par les utilisateurs, les transmettent aux satellites géostationnaires qui eux mêmes les relaient aux utilisateurs. Le principe des satellites géostationnaires permet d'avoir une couverture importante. Plusieurs nations ont mis en place leur propre système SBAS. Ainsi, les américains ont le WAAS (Wide Area Augmentation System), les canadiens le CWAAS (Canadian Wide Area Augmentation System), les européens EGNOS (European Geostationnary Navigation Overlay System), le japon le MSAS (Multi-functional Satellite Augmentation System), la Chine le SNAS (Satellite Navigation Augmentation System), la Russie le SDCM (System of Differential Correction and Monitoring) et enfin l'Inde le GAGAN (GPS and GEO Augmented System), figure (1.9). En dehors de l'aéronautique, les systèmes SBAS sont aujourd'hui utilisés dans tous les domaines nécessitant une grande précision et une forte intégrité.

Le système européen EGNOS

Le système européen EGNOS a été développé par de nombreux acteurs européens réunis autour d'un projet piloté par l'ESA (European Space Agency). Les avantages d'EGNOS sont :

- améliorer la précision de la positionnement fournie par le système GPS seul et le futur Galileo. La précision obtenue est comprise entre 1 et 3 mètres (95%).
- diffuser des informations d'intégrité sur la validité des données transmises par la constellation GPS. Ces informations incluent des intervalles de confiance pour la position estimée ainsi que des informations sur la santé des satellites.
- bénéficier d'un rattachement précis et fiable au temps de référence UTC.

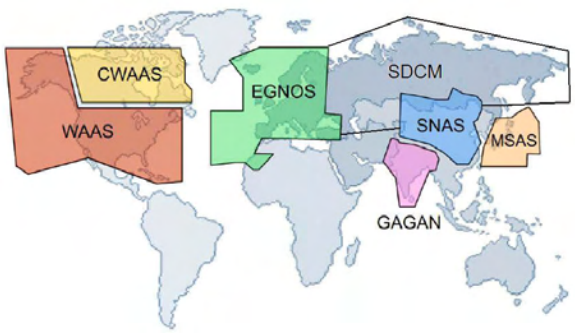


FIGURE 1.9 – Les différents systèmes SBAS

– améliorer la disponibilité du système de positionnement par satellites.

L'architecture d'EGNOS est comme celle du système GPS composée d'un segment spatial, d'un segment sol et d'un segment utilisateur. Le signal EGNOS est comme le signal GPS civil, libre d'accès. La plupart des récepteurs destinés à des usages professionnels sont aujourd'hui compatibles EGNOS.
Le segment spatial est composé de trois satellites géostationnaires positionnés à une altitude d'environ 36 000 kilomètres.
Le segment sol est composé d'un vaste réseau de stations réparties dans toute l'Europe. Le rôle principal de ces stations est de collecter les mesures et données provenant des satellites GPS afin de fournir des corrections différentielles et des informations d'intégrité aux utilisateurs.
L'ouverture du service EGNOS au grand public date d'octobre 2009 et la certification pour la navigation aérienne date de 2010. Des essais pour le service Safety of Life (SoL, désignant l'utilisation de la fonction d'intégrité) sont encore en cours aujourd'hui.

1.6 Applications du GPS

Les applications du GPS sont de plus en plus nombreuses aujourd'hui notamment de part l'expansion massive du marché des smartphones. En dehors de l'aviation civile, le développement des systèmes SBAS a permis d'étendre le champ d'application du GPS dans de nombreux domaines : agriculture de précision pour le guidage des engins agricoles permettant une gestion optimisée des parcelles, applications routières pour le suivi et la gestion de flotte de véhicules, l'exploration pétrolière pour le positionnement des plateformes, les forages en général pour de gros ouvrages en génie civil, les applications scientifiques en sismologie par exemple où le positionnement précis permet de surveiller la tectonique des plaques terrestres, etc ...

1.6.1 Les services autour de la géolocalisation

Les services fondés sur la géolocalisation comprennent tous les services pour lesquels la connaissance de la position de l'utilisateur est nécessaire. Ils reposent sur l'utilisation d'un récepteur GPS et d'un moyen de communication mobile de type GSM. On recense des services d'urgence et d'assistance, avec l'apparition depuis quelques années de la norme

E911, qui impose aux opérateurs de téléphonie mobile de fournir la position à 50 mètres près des utilisateurs qui composent un numéro d'urgence. Des services de dépannage automobile se développent actuellement. Une autre catégorie de service ayant vu le jour ces dernières années est la surveillance et la gestion de flotte. Ces services utilisent la technologie du GPS pour fournir à une station de contrôle la position d'un ensemble de récepteurs. Des services de localisation d'enfants, de personnes âgées atteintes de la maladie d'Alzeihmer par exemple, d'anciens détenus en liberté mais sous surveillance, de véhicules de transport de matières dangereuses etc...sont déjà mis en place. De nouveaux services de facturation routière sont à l'étude. Il sera question de calculer le kilométrage effectué par les transporteurs routiers sur les portions de routes payantes. La majorité des services développés pour le grand public, que l'on trouve sur les smartphones, sont des services d'information en fonction de la position de l'utilisateur.

1.6.2 Les transports

Les techniques de navigation ont toujours été développées initialement pour répondre aux besoins des différents moyens de transport. Un des points forts de la radionavigation par satellites est que l'on peut connaître à tout moment la position d'un utilisateur, à n'importe quel endroit du globe, que ce soit sur terre, dans les airs ou en mer.

Le transport aérien : En vingt ans, le trafic aérien a doublé, les activités de transport aérien sont en constante progression. En Europe, on comptait 7 millions de vols en 1997 contre plus de 11 millions aujourd'hui. L'utilisation des systèmes de radionavigation par satellites offre de nombreux avantages pour le secteur des transports aériens. Par exemple, les contrôleurs aériens cherchent en permanence à éviter une saturation des espaces aériens. Le GPS (augmenté) permet aujourd'hui de diminuer les distances minimales entre les avions tout en conservant la sécurité (grâce aux services d'intégrité fournis par EGNOS en Europe).

Le transport maritime : La navigation par GPS est couramment utilisée aujourd'hui par les navigateurs maritimes. Du fait de l'amélioration des performances du système en terme de précision et d'intégrité, le GPS et le futur Galileo pourront être utilisés dans toutes les phases de la navigation maritime, océanique comme côtière, en approche portuaire ou en manoeuvre d'accostage.

Le transport ferroviaire : La navigation par satellites apporte un support pour moderniser et optimiser ce mode de transport. Le suivi de la position des trains offre de nombreux avantages, notamment l'amélioration du service aux usagers avec l'apport d'informations précises sur les retards éventuels et la réduction de la consommation énergétique des trains en optimisant leurs accélérations et freinages en fonction des conditions de circulation sur le réseau.

Le transport automobile : C'est dans ce domaine que le GPS est le plus connu du grand public. Cependant, de nouvelles applications urbaines voient le jour depuis quelques années. Les taxis des grandes agglomérations peuvent indiquer en temps réel à leurs clients le temps d'attente pour obtenir un véhicule. Les compagnies de bus proposent maintenant des arrêt équipés de panneaux digitaux indiquant l'heure d'arrivée du prochain bus.

1.6.3 L'agriculture

Les activités de production agricole sont soumises à des contraintes de plus en plus fortes, que ce soit en terme de qualité, de respect de l'environnement, de rentabilité économique ou de régulation internationale. L'utilisation de la navigation par satellites apporte une meilleure maîtrise de toutes ces contraintes. Certains agriculteurs utilisent aujourd'hui des moyens très sophistiqués à partir de GPS en complément d'images satellitaires pour, par exemple, réduire l'utilisation de produits chimiques, optimiser l'ensemencement des parcelles, suivre l'évolution des cultures et contrôler les surfaces cultivées.

1.6.4 Géodésie et environnement

Du fait de sa couverture mondiale, la géolocalisation par satellites offre un moyen global de surveillance de la planète et de son environnement. Le suivi des mouvements des plaques terrestres est un élément essentiel pour alimenter les analyses de l'évolution climatique et de l'environnement terrestre. L'océanographie peut elle aussi bénéficier de la géolocalisation par satellites au moyen de balises flottantes permettant d'observer les courants marins. Des études des mouvements des glaces dans les zones polaires ont été menées grâce au positionnement par satellites [MMC02]. L'étude de l'atmosphère bénéficie également des mesures de propagation des signaux satellitaires à travers les différentes couches qui la compose. Des études des mouvements migratoires de la faune ont été menées avec des balises équipées de récepteurs GPS miniaturisés. Il est ainsi possible de suivre certaines espèces protégées dans leur environnement naturel.

1.6.5 Autres exemples d'applications

Parmi les nombreuses autres applications s'appuyant sur la géolocalisation par satellites on peut citer les suivantes : les applications militaires, les transactions financières où l'échelle de temps GPS sert de référence pour la datation et la synchronisation des données informatiques qui circulent dans le monde, les travaux publics, la protection civile en cas de catastrophe naturelle etc...

1.7 Conclusion

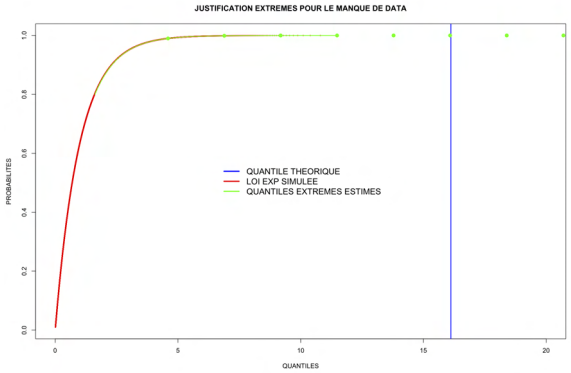
Nous avons évoqué les principes de fonctionnement des systèmes de positionnement par satellites. Ces moyens de navigation, et leurs systèmes d'augmentation, étant de plus en plus évolués à ce jour, il est aujourd'hui possible d'atteindre des degrés de précision que les concepteurs initiaux n'auraient même pas imaginé ; le principe de positionnement RTK fournit une précision centimétrique voire millimétrique pour un enregistrement statique. Ce niveau de performance est d'autant plus impressionnant quand on sait que les positions sont déterminées à partir de distances entre la surface de la Terre et des satellites évoluant à grande vitesse à plus de 20 000 kilomètres de nous.
La qualité de positionnement offerte par de tels systèmes a poussé les instances aéronautiques internationales à vouloir utiliser le GPS et EGNOS comme moyen de navigation à bord des aéronefs. Comme toute technologie mettant en jeu la vie des utilisateurs, cette application critique requiert des niveaux de sécurité maximaux ; des concepts de fiabilité associés au GPS ont alors vu le jour. On parle d'intégrité, qui sera vu comme une mesure de confiance que l'on peut accorder aux informations fournies par le système global.

L'intégrité sera en partie décrite par des modèles probabilistes associés à des niveaux de probabilité de défaillance très faibles, afin de protéger l'utilisateur de l'apparition fortuite d'événements rares (position estimée très éloignée de la position réelle par exemple) qui peuvent avoir des conséquences désastreuses sur la navigation d'un avion.
L'objet du chapitre suivant est de décrire une théorie mathématique dédiée à l'étude des événements rares, que l'on appellera événements extrêmes. On parle de la théorie des extrêmes et de la classe d'outils qu'elle propose.

C H A P I T R E 2

Étude des valeurs extrêmes

Résumé :
Ce chapitre est dédié à l'étude des valeurs extrêmes d'un échantillon, événements dont les probabilités d'occurrence sont faibles. La théorie des extrêmes vient en complément de la statistique classique où il est usuel d'étudier les variables aléatoires autour de leurs moyennes. Il s'agit dans cette théorie de caractériser le comportement des queues de distribution à l'aide de modèles permettant un bon ajustement au-delà du maximum de l'échantillon. Nous présentons dans ce chapitre les deux formalismes principaux de cette théorie dans un cadre univarié et on y compare les performances des estimateurs paramétriques caractérisant les lois limites des queues de distribution. Le lecteur y trouvera plusieurs expressions d'estimateurs de quantile extrême ainsi que les intervalles de confiance associés.



Sommaire

2.1 Introduction

2.1.1 Historique et applications

On s'intéresse dans l'étude des valeurs extrêmes aux grandes valeurs d'un échantillon de variables aléatoires afin de caractériser sa loi. Cette pratique vient en complément de la statistique classique où il est généralement question d'étudier le comportement de variables aléatoires autour de leurs valeurs moyennes. Une théorie dédiée à l'étude de ces valeurs particulières a été développée depuis les années 70 et suscite l'intérêt de nombreux statisticiens, ingénieurs et scientifiques tant le champs d'applications qu'elle touche est vaste. Il s'agit dans cette théorie de caractériser le comportement des queues de distribution à l'aide de modèles spécifiques permettant un bon ajustement au delà du maximum de l'échantillon. On s'intéresse alors aux événements dont les probabilités d'occurrence sont très faibles, on parle d'événements rares. La théorie des extrêmes offre une classe d'outils permettant une extrapolation dans les queues de distribution, à partir des valeurs maximales observées, afin de prédire l'apparition de valeurs non observées.

On donne un exemple illustré de cette approche statistique : on désire estimer un quantile de niveau $1 - 10^{-7}$ à partir d'un échantillon quelconque. On simule un échantillon de 3600 points distribués suivant une loi normale. Pour une loi symétrique, on aura deux quantiles à estimer. Le problème principal est le manque d'observation pour estimer les

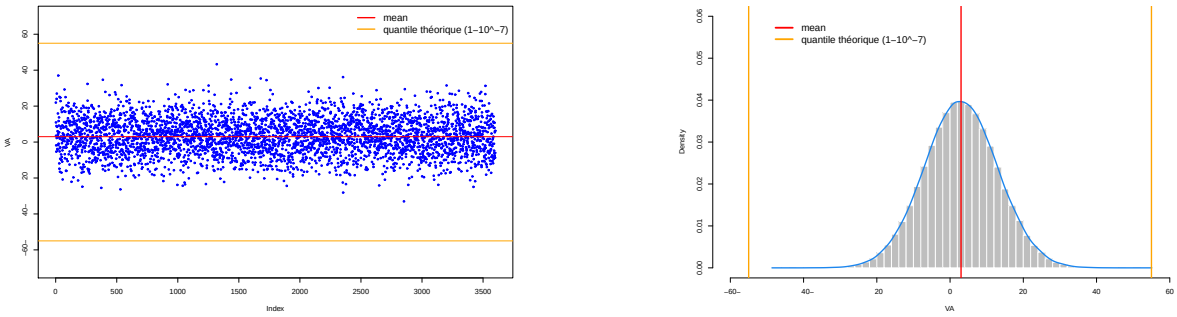


FIGURE 2.1 – Manque de données

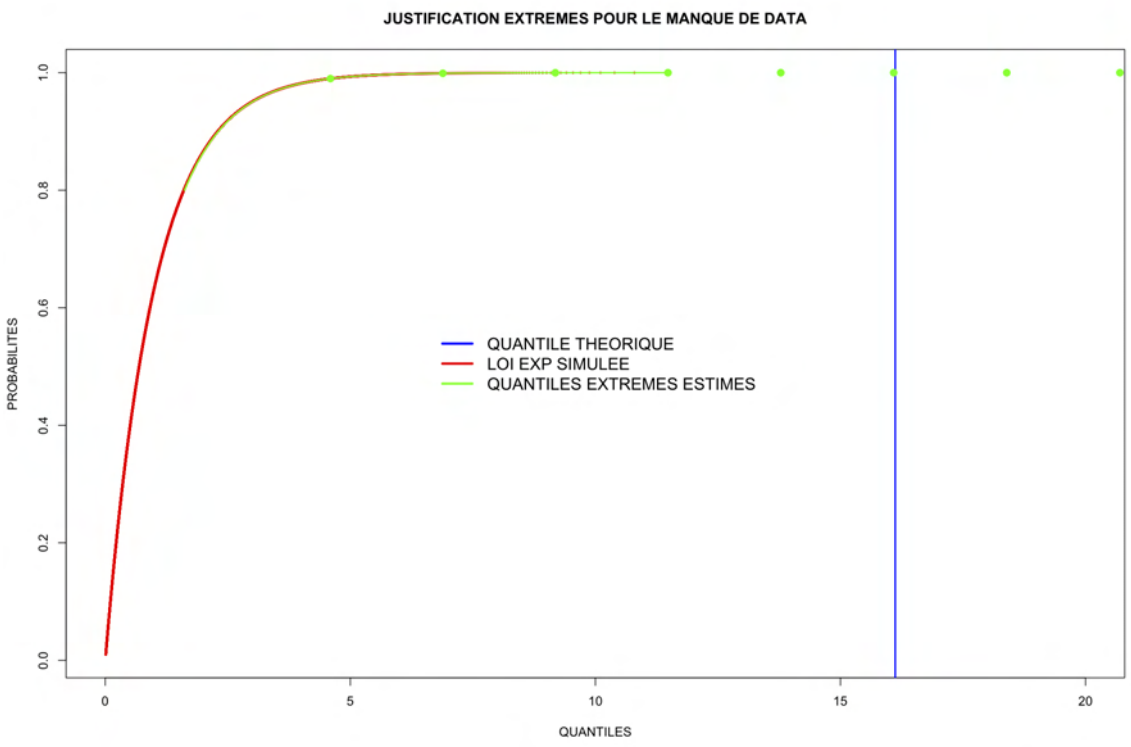


FIGURE 2.2 – Extrapolation dans les queues de distribution

quantiles voulus (ligne continue orange). On voit bien sur la figure 2.1 de gauche, que les observations (points bleus) n'atteignent pas les niveaux recherchés (lignes oranges). On risque de faire une erreur importante si on estime ces quantiles en se basant sur l'échantillon recueilli (quantile empirique) où le quantile recherché sera estimé par la plus grande observation. La théorie des extrêmes permet une extrapolation dans les queues de distribution conduisant à une estimation bien plus fine des quantiles recherchés. La figure 2.2 montre cette extrapolation : on voit qu'il n'y a pas assez d'observation (en

rouge) pour atteindre le quantile voulu (en bleu). La théorie des extrêmes peut alors être employée pour estimer des quantiles à plusieurs niveaux (points verts) y compris au delà des observations.

Le champs d'applications de cette classe d'outils est de plus en plus vaste. On retrouve la théorie des extrêmes dans les domaines suivants :

- finance : gestion de portefeuille, évaluation du risque sur les marchés (prédiction de crises monétaires).
- environnement : évolution du climat, concentration de la pollution, vitesse du vent, surveillance des crues de cours d'eau.
- industrie : fiabilité des structures [Gar02]
- telecom/Spatial : prévision du trafic et de la disponibilité du réseau de télécommunications. Intégrité, disponibilité, continuité GNSS.

2.1.2 Principe de la théorie des extrêmes

Il s'agit dans l'étude des valeurs extrêmes d'analyser l'épaisseur des queues de distributions, ou encore d'étudier les plus grandes observations d'un échantillon pour caractériser sa loi initiale. Ainsi, la théorie des extrêmes vient en complément de la théorie statistique classique où il est plus commun d'étudier le comportement d'une distribution autour de sa moyenne plutôt que dans le domaine des observations extrêmes souvent appelées événements rares. Nous allons voir que toute la théorie des extrêmes est fondée sur un équivalent au théorème central limite mais pour les queues de distribution.
On s'intéresse au comportement du maximum d'un échantillon, variable aléatoire définie par

$$M_n := \max(X_1, \cdots, X_n),$$

où $X_1, \cdots, X_n$ est un échantillon de n variables aléatoires indépendantes et de même loi F.
La distribution de M_n peut s'écrire de la façon suivante :

$$\begin{aligned} \mathbb{P}(M_n \leq x) &= \mathbb{P}(X_1 \leq x, \cdots, X_n \leq x) \\ &= \mathbb{P}(X_1 \leq x) \times \cdots \times \mathbb{P}(X_n \leq x) \\ &= F(x)^n \end{aligned}$$

On ne connaît pas la distribution F en pratique. Les techniques classiques d'estimation de F se basent sur les observations dont on dispose. Cependant de petites erreurs d'estimation sur F peuvent avoir des conséquences désastreuses sur F^n. Une alternative présuppose que l'on conserve F comme inconnue, et que l'on approche cette distribution par des distributions limites estimées dans le domaine des observations extrêmes. Cependant, la loi limite de M_n est dégénérée. Il s'agit alors, tout comme le théorème central limite, d'observer le comportement limite de la variable M_n renormalisée.

2.2 Loi des extrêmes généralisée, approche block maxima

On définit la famille de loi GEV (Generalized Extreme Value) pour $\gamma \in \mathbb{R}$ par :

$$G_\gamma(x) = \begin{cases} \exp\Big(- \Big[1+\gamma x\Big]^{-\frac{1}{\gamma}} \Big) & \text{pour tout } x \text{ tel que } 1+\gamma x \geq 0, \text{ si } \gamma \neq 0 \\ \exp\Big(- \exp\Big(- x\Big)\Big) & \text{pour tout } x \in \mathbb{R}, \text{ si } \gamma = 0 \end{cases} \tag{2.1}$$

2.2.1 Théorème de Fisher et Tippett

Ce théorème a été énoncé pour la première fois par Fisher et Tippett en 1928 [FT28], puis démontré par Gnedenko en 1943 [Gne43].

Théorème 2.2.1. *S'il existe des suites de réels $a_n > 0$ et $b_n \in \mathbb{R}$ telles que :*

$$\lim_{n\to\infty} \mathbb{P}\Big(\frac{M_n - b_n}{a_n} \leq x\Big) = G_\gamma(x) \tag{2.2}$$

ou encore,

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G_\gamma(x) \tag{2.3}$$

pour tout $x \in \mathbb{R}$ et G_γ non dégénérée, alors $G_\gamma(x)$ appartient à la famille de loi GEV (Generalized Extreme Value).

F^n est la fonction de répartition de M_n tandis que $G_\gamma(x)$ est la fonction de répartition limite de M_n correctement renormalisée par a_n et b_n. On dira que F^n est dans le domaine d'attraction D_γ.

2.2.2 Les lois limites possibles

Le comportement de la queue de distribution d'une suite de variables aléatoires sera complètement caractérisé par le paramètre γ appelé indice des valeurs extrêmes. Une partie sera consacrée aux méthodes d'estimation de ce paramètre. Le signe de γ a une forte influence sur la distribution des extrêmes, et on distingue trois cas :

Domaine d'attraction de Gumbel : lorsque $\gamma = 0$, la distribution G_0 est appelée distribution de Gumbel. Le support de cette loi est $\mathbb{R}$ et dans ce cas les queues de distribution sont légères et décroissent de manière exponentielle. G_0 sera parfois notée Λ.

Domaine d'attraction de Fréchet : lorsque $\gamma > 0$, on a : $\Phi(x) := \exp(-x^{-1/\gamma})\mathbb{1}_{x>0}$. De telles distributions possèdent des queues lourdes et la convergence de F^n vers la loi limite se fait très lentement.

Domaine d'attraction de Weibull : lorsque $\gamma < 0$, on pose $\alpha = -1/\gamma > \ 0$ et on note $\Psi_\alpha(x) = \exp(-(-x)^\alpha)$ si x est négatif, et 1 sinon. Dans ce cas la queue de la distribution sera très mince.

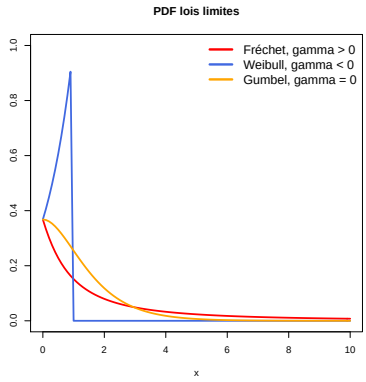


FIGURE 2.3 – Loi des extrêmes généralisée (densité)

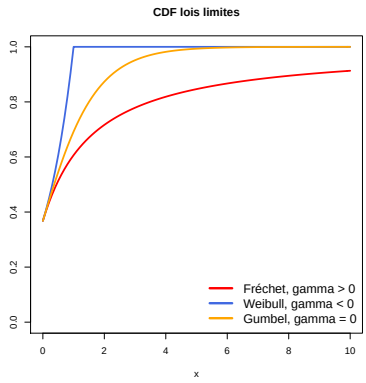


FIGURE 2.4 – Loi des extrêmes généralisée (fonction de répartition)

2.2.3 Estimation des paramètres de la GEV

On donne dans cette partie différents estimateurs de l'indice des valeurs extrêmes proposés dans la littérature. On formulera aussi une expression de l'estimateur du maximum de vraisemblance permettant d'estimer l'ensemble des paramètres de la loi des extrêmes généralisée : l'indice des valeurs extrêmes γ et les deux suites normalisantes a_n et b_n vues dans le théorème de Fisher et Tippet.

Le théorème limite de la loi des extrêmes est fondé sur la convergence renormalisée du maximum d'un échantillon. Cependant, il n'est pas envisageable de construire une estimation pertinente sur une unique réalisation de l'échantillon. Pour pallier cela, Gumbel a introduit en 1958 [Gum58] une méthode baptisée "block maxima". Cette approche consiste à diviser l'échantillon en m blocs de taille l. Sur chacun de ces blocs, on prend en compte le maximum et on obtient donc m valeurs $\max_1, \cdots, \max_m$. Ces valeurs sont donc traitées comme un échantillon de la variable M_m . On notera Z le maximum de l'échantillon $Z_1, \cdots, Z_m$. On pourra alors exprimer la fonction de vraisemblance du maximum de l'échantillon en utilisant les réalisations $Z_1, \cdots, Z_m$ considérées comme indépendantes.
D'autres estimateurs dédiés à l'indice des valeurs extrêmes γ sont construits à partir des k plus grandes observations de l'échantillon réordonné. On parlera de statistiques d'ordre. C'est le cas notamment de l'estimateur de Hill (1975), de l'estimateur des moments (1989), et de l'estimateur de Pickands (1975).

2.2.3.1 Estimation de l'indice des valeurs extrêmes γ

Les deux estimateurs sans doute les plus populaires dans la littérature sont les estimateurs de Hill (1975) [Hil75] et de Pickands (1975) [Pic75].

On note $X_{1,n}, \cdots, X_{n,n}$ les statistiques d'ordre associées à l'échantillon $X_1, \cdots, X_n$. C'est-à-dire que l'on classe $X_1, \cdots, X_n$ par ordre croissant de sorte que :

$$X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n,n}$$

La k-ième plus grande valeur de l'échantillon sera donc notée $X_{n-k,n}$. On a $1 \leq k < n$ avec $k = o(n)$.

L'estimateur de Hill n'est valable que pour $\gamma > 0$, tandis que l'estimateur de Pickands sera valable pour tout $\gamma \in \mathbb{R}$.

Estimateur de Hill
On introduit la notation suivante :

$$M_n^{(j)} := \frac{1}{k} \sum_{i=1}^{k} (\log X_{n-i+1,n} - \log X_{n-k,n})^j$$

L'estimateur de Hill est défini pour $\gamma > 0$ par :

$$\widehat{\gamma}_{X,k,n}^{(H)} = \frac{1}{k} \sum_{i=1}^{k} (\log X_{n-i+1,n} - \log X_{n-k,n}) = M_n^{(1)}$$

avec $X_{1,n}, \cdots, X_{n,n}$ les statistiques d'ordre associées à l'échantillon $X_1, \cdots, X_n$.

Lorsque n et k tendent vers l'infini, on a la normalité asymptotique suivante (Hill,1975 [Hil75]) pour $\gamma > 0$:

$$\sqrt{k} \, (\, \widehat{\gamma}_{X,k,n}^{(H)} - \gamma \,) \longrightarrow \mathcal{N} \, (0, \gamma^2)$$

Les propriétés de consistance forte de cet estimateur ont été montrées par Deheuvels, Haüsler et Mason (1988) [DHM88]. On trouvera l'heuristique de la construction de cet estimateur dans l'article fondateur de Hill [Hil75], ou dans les ouvrages [Bei04] ou [Col01] par exemple.
L'estimateur $\widehat{\gamma}_{X,k,n}^{(H)}$ est très dépendant de la valeur de k. Drees, De Haan et Resnick [DDHR00] fournissent une méthode de choix d'une séquence de k optimaux afin de minimiser asymptotiquement l'erreur MSE (Mean Squared Error).

Estimateur de Pickands
Cet estimateur est défini pour tout $\gamma \in \mathbb{R}$:

$$\widehat{\gamma}_{X,k,n}^{(P)} = \frac{1}{\log 2} \log \left(\frac{X_{n-k+1,n} - X_{n-2k+1,n}}{X_{n-2k+1,n} - X_{n-4k+1,n}} \right)$$

On trouve une étude complète des propriétés de cet estimateur dans Dekkers et De Hann (1989) [DDH89] et principalement la propriété asymptotique suivante pour tout $\gamma \in \mathbb{R}$, lorsque n et k tendent vers l'infini :

$$\sqrt{k}\ (\ \widehat{\gamma}^{(P)}_{X,k,n} - \gamma\ ) \longrightarrow \mathcal{N}\ \left(\,0\,,\, \frac{\gamma^2(2^{\gamma+1}+1)}{(2(2^{\gamma}-1)\log 2)^2}\ \right)$$

Un troisième estimateur populaire est l'estimateur obtenu avec la méthode des moments.

Estimateur des moments
Cet estimateur introduit par Dekkers, Einmahl, de Haan (1989) [DEDH89] est une généralisation de l'estimateur de Hill, valable pour tout $\gamma \in \mathbb{R}$:

$$\widehat{\gamma}^{(M)}_{X,k,n} = \widehat{\gamma}^{(H)}_{X,k,n} + 1 - \frac{1}{2}\Big(1 - \frac{\widehat{\gamma}^{(H)}_{X,k,n}}{M^{(2)}_n}\Big)^{-1}$$

On a la normalité asymptotique suivante lorsque n et k tendent vers l'infini, pour $\gamma \geq 0$:

$$\sqrt{k}\ (\ \widehat{\gamma}^{(M)}_{X,k,n} - \gamma\ ) \longrightarrow \mathcal{N}\ (0, 1+\gamma^2)$$

2.2.3.2 Estimateur du maximum de vraisemblance

L'estimateur du maximum de vraisemblance suivant est directement construit à partir des observations des maxima, et il ne s'agit pas uniquement d'estimer l'indice des valeurs extrêmes : on estimera γ mais aussi les deux suites normalisantes a_n et b_n.
On se place dans le cadre de la méthode "block maxima" et on considère un échantillon de k maxima $Z_1, \cdots, Z_k$ i.i.d. suivant une loi GEV. Pour $\gamma \neq 0$, la fonction de log-vraisemblance est donnée par :

$$\mathcal{L}(Z;\gamma,b_n,a_n) = -k\log a_n - \Big(\frac{1}{\gamma}+1\Big)\sum_{i=1}^{k}\log\Big(1+\gamma\frac{z_i-b_n}{a_n}\Big) - \sum_{i=1}^{k}\Big(1+\gamma\frac{z_i-b_n}{a_n}\Big)^{-\frac{1}{\gamma}}$$

Pour $\gamma = 0$ on aura :

$$\mathcal{L}(Z;0,b_n,a_n) = -k\log a_n - \sum_{i=1}^{k}\exp\Big(\frac{z_i-b_n}{a_n}\Big) - \sum_{i=1}^{k}\Big(\frac{z_i-b_n}{a_n}\Big)$$

On utilisera cet estimateur lorsque $\gamma > -0.5$, cas pour lesquels les propriétés de consistance et la normalité asymptotique sont vérifiées [Smi85]. On remarquera qu'il s'agit en fait d'une pseudo vraisemblance obtenue à partir de la loi limite et non de la loi originale de l'échantillon.
On rappelle que si les paramètres de la distribution ajustée aux observations, ici la loi

GEV, sont estimés par maximum de vraisemblance, on peut utiliser les résultats de Kendall et Stuart (1961) [KA61] pour former un intervalle de confiance asymptotique.
On aura la normalité asymptotique suivante pour $m \to \infty$:

$$\sqrt{m}\,\left((\widehat{a}_n, \widehat{\gamma}, \widehat{b}_n) - (a_n, \gamma\,, b_n)\right) \longrightarrow \mathcal{N}\ (0, I^{-1}) \qquad \gamma > 0.5$$

où I est la matrice d'information de Fisher estimée par sa version empirique donnée par :

$$I(\Theta) = -\mathbb{E}\left(\frac{\partial^2 \mathcal{L}(Z;\Theta\,)}{\partial \Theta^2}\right) \tag{2.4}$$

où $\mathcal{L}(Z;\Theta\,)$ est la fonction de log-vraisemblance associée à la loi de la variable aléatoire Z, paramétrée par un ensemble de paramètres noté Θ. Les calculs des éléments de la matrice I pourront être trouvés dans [CH04] ou [Bei04].

Beirlant (2004) [Bei04] donne en plus de l'estimateur MLE, une estimation suivant la méthode des moments pondérés introduite par Greenwood et al. (1979) [GLMW79] ainsi que les intervalles de confiance associés.

De Haan (2006) [DHF06], présente un comparatif des propriétés asymptotiques des ces estimateurs. L'estimateur de Pickands semble avoir une variance asymptotique plus élevée que les autres. Les estimateurs des moments et MLE ont de bonnes propriétés lorsque γ est autour de 0, c'est à dire quand la loi de l'échantillon observé est proche du domaine d'attraction de Gumbel. L'estimateur de Hill présente la variance asymptotique la plus faible pour $\gamma > 0$, suivi par l'estimateur des moments. Cependant, lorsque γ est négatif, c'est l'estimateur MLE qui a la plus petite variance.
D'autres aperçus de ces estimateurs classiques peuvent être trouvés dans l'article de Csörgo et Viharos (1998) [CV98] ou dans le livre d'Embrechts et al. (1997) [EKM97], par exemple.

De nombreux autres estimateurs ont été proposés : on trouve notamment l'estimateur UH proposé par Beirlant (1996) [BVT96], un estimateur à noyau introduit par Csörgo et Deheuvels (1985) [CDM85a].

2.2.4 Estimation de quantile extrême

2.2.4.1 Estimation par inversion de la fonction de répartition

Une expression d'un quantile extrême de niveau de probabilité p en fonction de la taille de l'échantillon n peut être obtenue en inversant simplement la fonction de répartition GEV des maxima de l'échantillon. On rappelle que $\mathbb{P}(M_n \leq x) = F(x)^n \approx G_\gamma(x)$. On obtient alors [Bei04] :

$$\widehat{q}(p)_{\mathrm{GEV}} = \begin{cases} \widehat{a}_n + \frac{\widehat{b}_n}{\widehat{\gamma}}\left[(-\log(1-p)^n)^{-\gamma} - 1)\right] & , \quad \gamma \neq 0 \\ \\ \widehat{a}_n + \widehat{b}_n \log(-\log(1-p)^n) & , \quad \gamma = 0 \end{cases} \tag{2.5}$$

On construit un intervalle de confiance pour l'estimateur $\widehat{q}(p)_{\mathrm{GEV}}$ à partir de la méthode Delta [VdV00]. Cette méthode analytique repose sur le théorème suivant :

Théorème 2.2.2. *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes distribuées selon une loi paramétrique, de paramètre θ. Soit T_n un estimateur de θ et soit $\phi(\theta)$ une fonction réelle, continûment dérivable, du paramètre θ. On cherche les propriétés asymptotiques de de la fonction $\phi(T_n)$ connaissant celles de l'estimateur T_n.*

Si T_n converge en probabilité vers θ et si ϕ est continue en θ alors $\phi(T_n)$ converge en probabilité vers $\phi(\theta)$.

ϕ étant continûment dérivable, on a

$$\sqrt{n}(\phi(T_n) - \phi(\theta)) \approx \phi'(\theta)\sqrt{n}(T_n - \theta) \tag{2.6}$$

Si $\sqrt{n}(T_n - \theta)$ converge en loi vers une distribution limite T alors

$$\sqrt{n}(\phi(T_n) - \phi(\theta)) \xrightarrow{\mathcal{L}} \phi'(\theta)T \tag{2.7}$$

En particulier, si $\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}\ (0, \sigma^2)$ alors

$$\sqrt{n}(\phi(T_n) - \phi(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}\ (0, \phi'^2\sigma^2) \tag{2.8}$$

Pour le cas où les paramètres de la loi sont estimés par la méthode du maximum de vraisemblance, on a le théorème suivant [Col01] :

Théorème 2.2.3. *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires indépendantes distribuées selon une loi paramétrique, de paramètre θ. Soit T_n un estimateur de θ et soit $\phi(\theta)$ une fonction réelle, continûment dérivable, du paramètre θ. Soit $\nabla\phi = [\frac{\partial\phi}{\partial\theta_1}, \cdots, \frac{\partial\phi}{\partial\theta_d}]^\top$ évalué en T_n, on suppose $\nabla\phi \neq 0$. Alors, d'après le Théorème 2.2.2, $T_n \sim \mathcal{N}(\phi, V)$ où $V = \nabla\phi^\top I(\theta)^{-1}\nabla\phi$ et $I(\theta)$ la matrice d'information de Fisher.*

Si on considère que $\widehat{b}_n$, $\widehat{\gamma}$ et $\widehat{a}_n$ sont estimés à partir de l'estimateur du maximum de vraisemblance, l'intervalle de confiance suivant est obtenu par une application des théorèmes ci-dessus :

$$\sqrt{m}\,(\widehat{q}(p)_{\mathrm{GEV}} - q(p)_{\mathrm{GEV}}) \longrightarrow \mathcal{N}\ (0, \zeta^T\,\Delta\,\zeta) \quad \text{quand} \quad m \to \infty$$

où m est le nombre de blocs qui divisent l'échantillon,

$$\Delta = (1+\gamma)\begin{bmatrix} 1+\gamma & -b_n \\ -b_n & 2b_n^2 \end{bmatrix}$$

et

$$\begin{aligned}
\zeta^T &= \left[\frac{\partial q(p)_{\mathrm{GEV}}}{\partial b_n}, \frac{\partial q(p)_{\mathrm{GEV}}}{\partial \gamma}, \frac{\partial q(p)_{\mathrm{GEV}}}{\partial a_n}\right] \\
&= \left[\frac{1}{\gamma}[(-\log(1-p)^n)^{-\gamma} - 1],\right. \\
&- \left.\frac{b_n}{\gamma^2}[(-\log(1-p)^n)^{-\gamma} - 1] - \frac{b_n}{\gamma}(-\log(1-p)^n)^{-\gamma}\log(-\log(1-p)^n)\ ,\ 1\right]
\end{aligned}$$

2.2.4.2 Autres estimateurs possibles

On peut trouver d'autres estimateurs dans la littérature, construits à partir des estimateurs de l'indice des valeurs extrêmes présentés précédemment. En voici trois.

à partir de l'estimateur de Pickands
Dekkers et al. (1989) [DDH89] donnent l'estimateur de quantile extrême à partir de l'estimateur de Pickands :

$$\widehat{q}(p)^{(P)}_{\mathrm{GEV}} = \frac{(k/pn)^{\widehat{\gamma}^{(P)}_{X,k,n}} - 1}{1 - 2^{-\widehat{\gamma}^{(P)}_{X,k,n}}}(X_{n-k+1,n} - X_{n-2k+1,n}) + X_{n-k+1,n} \tag{2.9}$$

Le théorème de normalité asymptotique de cet estimateur est donné dans [DDH89] (théorème 3.3).

à partir de l'estimateur de Hill
Weismann (1978) [Wei78] a proposé un estimateur de quantile extrême à partir de l'estimateur de Hill :

$$\widehat{q}(p)^{(W)}_{\mathrm{GEV}} = X_{n-k,n}\left(\frac{k}{n.p}\right)^{\widehat{\gamma}^{(H)}_{X,k,n}} \tag{2.10}$$

à partir de l'estimateur des moments
Un estimateur de quantile extrême peut être formulé à partir de l'estimateur des moments de Dekkers et al. [DEDH89] :

$$\widehat{q}(p)^{(M)}_{\mathrm{GEV}} = X_{n-k,n} + \frac{\hat{a}(k/n)}{\widehat{\gamma}^{(M)}_{X,k,n}}\left[(\frac{k}{np})^{\widehat{\gamma}^{(M)}_{X,k,n}} - 1\right] \tag{2.11}$$

avec $\hat{a}(k/n) = X_{n-k,n}\widehat{\gamma}^{(H)}_{X,k,n}\max(1 - \widehat{\gamma}^{(M)}_{X,k,n}, 1)$.

La normalité asymptotique de cet estimateur de quantile est montrée dans [DEDH89].

2.2.5 Discussion sur la méthode block maxima

Le formalisme de la loi des extrêmes généralisée ne tient compte que d'une seule observation, la plus grande. On ne prend pas en compte toutes les autres grandes valeurs de l'échantillon et cela nous conduit à une perte d'information. L'ajustement de la loi limite sera toutefois très influencée par la taille des blocs formés à partir de l'échantillon initial. Cette approche mène aussi à une perte d'information sur les observations extrêmes qui sont par définition très rares. Par exemple, on peut avoir plusieurs observations extrêmes au sein du même bloc, mais seule la plus grande d'entre elles sera prise en compte. Une alternative a été proposée suite aux travaux de Balkema et de Haan [BDH74], ainsi que de Pickands [Pic75]. Contrairement à la loi des extrêmes généralisée, on ne regarde plus le maximum de l'échantillon mais les n_e valeurs qui dépassent un certain seuil u à fixer. On obtient donc un sous échantillon de taille n_e dont on veut caractériser la loi. Cette approche est appelée "pick over threshold" ou POT que nous détaillons maintenant.

2.3 La loi des excès, approche POT

Soit u un réel suffisamment élevé appelé *seuil*. On définit les excès au-delà du seuil u comme l'ensemble des variables aléatoires conditionnelles $\{Y_j\} = \{X_j - u \mid X_j > u\}$. On note que

$$\mathbb{P}(X > u + y | X > u) = \frac{1 - F(u+y)}{1 - F(u)}$$

2.3.1 Théorème de Pickands

Le théorème de Pickands (1975) [Pic75, BDH74] montre que la distribution des excès au delà d'un certain u peut être approchée par une loi de Pareto généralisée. On notera GPD (Generalized Pareto Distribtution) :

Théorème 2.3.1. *Soit $X_1, X_2, \cdots, X_n$ un échantillon de variables aléatoires i.i.d. suivant la loi F. F appartient au domaine d'attraction D_γ si et seulement si il existe une fonction $\sigma(.)$ positive telle que*

$$\lim_{u \to x^*} \sup_{0<x<x^*} \left\{ \left| \mathbb{P}(X - u \leq x | X > u) - H_{\gamma,\sigma}(x) \right| \right\} = 0 \tag{2.12}$$

où $x^* = \sup\{x; F_u < 1\}$ et $H_{\gamma,\sigma}$ est la distribution de Pareto généralisée (GPD) définie par

$$H_{\gamma,\sigma}(x) = \begin{cases} 1 - \left(1 + \frac{\gamma x}{\sigma}\right)^{-\frac{1}{\gamma}} & si\ \gamma \neq 0,\ x \geq 0\ et\ x < -\frac{\sigma}{\gamma}\ si\ \gamma < 0 \\ \\ 1 - \exp\left(-\frac{x}{\sigma}\right) & si\ \gamma = 0\ ,\ x \geq 0 \end{cases}$$

On remarquera que l'on retrouve l'indice des valeurs extrêmes γ identique à celui de la loi GEV des maxima.

Ce théorème signifie que si F vérifie le théorème de Fisher et Tippet 2.2.1, c'est à dire si F appartient au domaine d'attraction D_γ alors il existe une fonction $\sigma(.)$ positive et un réel γ tels que la loi des excès F_u peut être uniformément approchée par une distribution de Pareto généralisée (GPD) notée $H_{\gamma,\sigma}$.

2.3.2 Estimation des paramètres de la loi des excès

Il s'agit dans cette partie de l'estimation des deux paramètres de la loi des excès γ et σ. On rappelle que l'indice des valeurs extrêmes (de la loi limite du maximum de l'échantillon donnée par le théorème de Fisher et Tippet) étant identique au paramètre γ de la loi des excès, les estimateurs présentés plus haut restent valables dans le cadre de la méthode POT.

2.3.2.1 Estimateur du maximum de vraisemblance

La fonction de log-vraisemblance pour un échantillon de n_e excès $Y_1, \cdots, Y_{n_e}$ *i.i.d.* suivants une loi de Pareto Généralisée est donnée par (Smith, 1985) [Smi85] :

$$\mathcal{L}(Y; \gamma, \sigma) = -n_e \log \sigma - \left(\frac{1}{\gamma} + 1\right) \sum_{i=1}^{n_e} \log\left(1 + \frac{\gamma}{\sigma} y_i\right) \qquad \gamma \neq 0$$

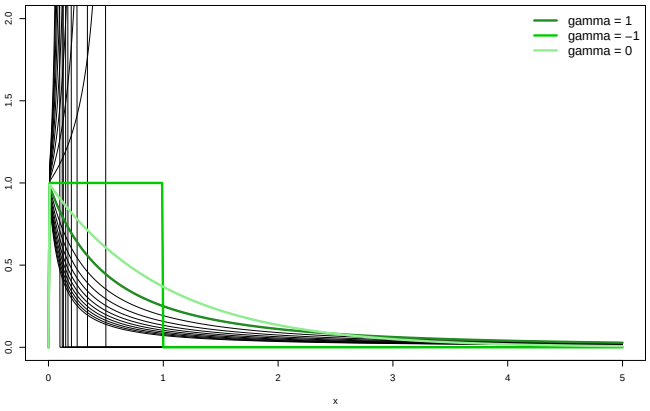


FIGURE 2.5 – Loi de Pareto généralisée (densité), sigma = 1

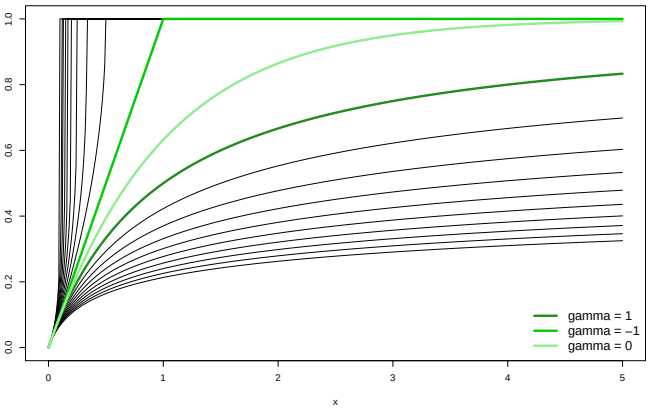


FIGURE 2.6 – Loi de Pareto généralisée (fonction de répartition), sigma = 1

avec $(1 + \frac{\gamma}{\sigma} y_i) > 0$, pour $i = 1, \cdots, n_e$. Dans le cas où $\gamma = 0$ on aura

$$\mathcal{L}(Y; 0, \sigma) = -n_e \log \sigma - \frac{1}{\sigma} \sum_{i=1}^{n_e} y_i \qquad \gamma = 0$$

Le problème de maximisation de la log-vraisemblance nécessite d'avoir recours à des méthodes numériques, type algorithme de Newton-Raphson. Dans le cas de la méthode POT, les conditions de régularité de cet estimateur sont elles aussi atteintes pour $\gamma > -0.5$ (Hosking et Wallis, 1987 [HW87]). La normalité asymptotique de cet estimateur a été établie par Smith (1987) [Smi85] et par Drees, Ferreira et De Haan (2004) [DFDH04] utilisant une autre approche fondée sur des approximations de la queue de la distribution empirique des quantiles, établies par Drees (1998) [Dre98]. Cette méthode d'estimation pour la loi GPD apparaît être la plus performante pour de grands échantillons (Hosking et Wallis, 1987 [HW87]).

On a le résultat suivant pour $\gamma > -1/2$ et $n_e \to \infty$:

$$\sqrt{n_e} \left((\hat{\gamma}^{ML}, \hat{\sigma}^{ML}) - (\gamma^{ML}, \sigma^{ML}) \right) \longrightarrow \mathcal{N}\ (0, \Delta) \ ,$$

avec

$$\Delta = (1+\gamma)\begin{bmatrix} 1+\gamma & -\sigma \\ -\sigma & 2\sigma^2 \end{bmatrix}$$

On peut alors construire les intervalles de confiance de niveau $(1-\alpha)$ suivant pour les paramètres γ et σ :

$$IC_{1-\alpha}(\gamma^{ML}) = \left[\hat{\gamma}^{ML} - \Phi^{-1}(1-\frac{\alpha}{2})\frac{1+\hat{\gamma}^{ML}}{\sqrt{n_e}} \ , \ \hat{\gamma}^{ML} + \Phi^{-1}(1-\frac{\alpha}{2})\frac{1+\hat{\gamma}^{ML}}{\sqrt{n_e}} \right]$$

$$IC_{1-\alpha}(\sigma^{ML}) = \left[\hat{\sigma}^{ML} - \Phi^{-1}(1-\frac{\alpha}{2})\sqrt{\frac{2(1+\hat{\gamma}^{ML})}{n_e}}\hat{\sigma}^{ML} \ , \ \hat{\sigma}^{ML} + \Phi^{-1}(1-\frac{\alpha}{2})\sqrt{\frac{2(1+\hat{\gamma}^{ML})}{n_e}}\hat{\sigma}^{ML} \right]$$

Φ^{-1} désigne la fonction quantile d'une loi normale centrée réduite.

2.3.2.2 Estimateur des moments et des moments pondérés

Les estimateurs des paramètres de la loi de Pareto généralisée issus de la méthode des moments et de la méthode des moments pondérés ont été introduits par Hosking et Wallis (1987) [HW87]. Une description détaillée de la construction de ces estimateurs peut être trouvée dans [Bei04] ou [CH04]. Ces deux méthodes d'estimation sont basées sur la comparaison des moments théoriques d'une distribution et de leurs versions empiriques. Le moment d'ordre r pour la distribution de Pareto Généralisée existe pour $\gamma < 1/r$. Cette contrainte restreint l'utilisation de cette classe d'estimateurs. Cependant nous verrons que pour notre étude, dans le cas de données GPS, γ est souvent compris entre $-1/2$ et $1/2$, ce qui implique que les moments d'ordre 1 et 2 sont bien définis. Soit $Y_1, \cdots, Y_{n_e}$ un échantillon de n_e excès *i.i.d.* suivant une loi de Pareto Généralisée. Les moments d'ordre 1 et 2 sont donnés par

$$\mathbb{E}[Y] = \frac{\sigma}{1-\gamma} \quad \text{et} \quad \mathrm{Var}(Y) = \frac{\sigma^2}{(1-\gamma)^2(1-2\gamma)} \tag{2.13}$$

On considère $\bar{Y}$ et S_Y^2 les estimateurs empiriques des moments d'ordre 1 et 2. On rappelle que l'on utilise l'estimateur sans biais de la variance donné par

$$S_Y^2 = \frac{1}{(n_e-1)}\sum_{i=1}^{n_e}(Y_i - \bar{Y})^2$$

En résolvant les deux équations $\mathbb{E}[Y] = \bar{Y}$ et $\mathrm{Var}(Y) = S_Y^2$, on obtient les estimateurs des moments de γ et σ [CH04] :

$$\widehat{\gamma}^{(M)}_{Y,n_e,n} = \frac{\bar{Y}^2 - S_Y^2}{2S_Y^2} \quad \text{et} \quad \widehat{\sigma}^{(M)}_{Y,n_e,n} = \frac{\bar{Y}(\bar{Y}^2/S_Y^2 + 1)}{2} \tag{2.14}$$

L'estimateur du couple (γ,σ) est donné par [Bei04] :

$$\widehat{\gamma}^{(\mathrm{PWM})}_{Y,n_e,n} = 2 - \frac{\hat{M}_{1,0,0}}{\hat{M}_{1,0,0} - 2\hat{M}_{1,0,1}} \quad \text{et} \quad \widehat{\sigma}^{(\mathrm{PWM})}_{Y,n_e,n} = \frac{2\hat{M}_{1,0,0}\hat{M}_{1,0,1}}{\hat{M}_{1,0,0} - 2\hat{M}_{1,0,1}} \tag{2.15}$$

$$\text{où} \quad M_{1,0,s} = \frac{1}{n_e} \sum_{i=1}^{n_e} \left(1 - \frac{i}{n_e + 1}\right)^s Y_{i,n_e}$$

Le domaine de validité de cet estimateur est $-1/2 < \gamma < 1/2$.

2.3.2.3 Exemple sur un échantillon suivant une loi de Rayleigh

Afin d'illustrer le bon fonctionnement de ces estimateurs à l'usage, on donne un exemple sur une loi de Rayleigh. On génère un échantillon de taille $N = 360000$ suivant une loi de Rayleigh de paramètre $\rho = 2$ (figure 2.7). On choisit d'estimer les paramètres de la loi des excès γ et σ pour une large plage de seuils compris entre 1 et 9. Les trois estimateurs obtenus avec les méthodes présentées dans la section précédente présentent un comportement similaire (figure 2.8). Pour l'estimation de l'indice des valeurs extrêmes γ, on observe une plage de stabilité pour les seuils compris entre 7 et 8, où les valeurs sont proches de 0, valeur théorique pour une loi de Rayleigh appartenant au domaine d'attraction de Gumbel (*cf.* section 2.4.1).

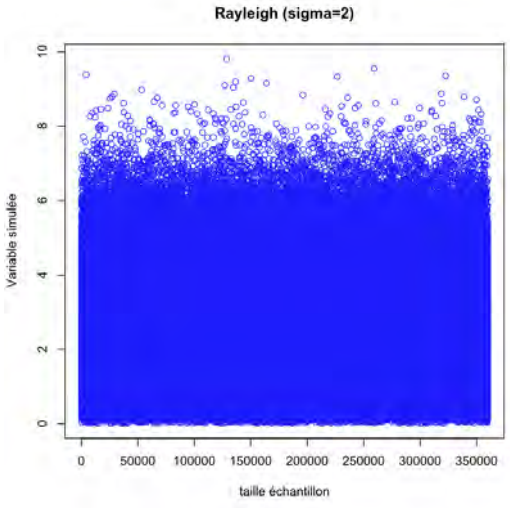


FIGURE 2.7 – Echantillon généré

L'étape qui nous intéresse est l'estimation de quantile extrême ; la section suivante présente la construction de l'estimateur classique POT d'un quantile extrême ainsi que les intervalles de confiance que l'on peut lui associer.

2.3.3 Estimateur de quantile extrême

On souhaite ici estimer un quantile extrême, c'est-à-dire un quantile généralement situé au-delà des observations. On choisit un nombre p positif et inférieur à $1/n$, de sorte que le quantile $q(p)$, d'ordre $1 - p$ soit en général supérieur à x^* l'observation maximale, estimation du quantile $1/n$. On s'intéresse à l'estimateur non paramétrique du quantile basé sur la méthode des excès.

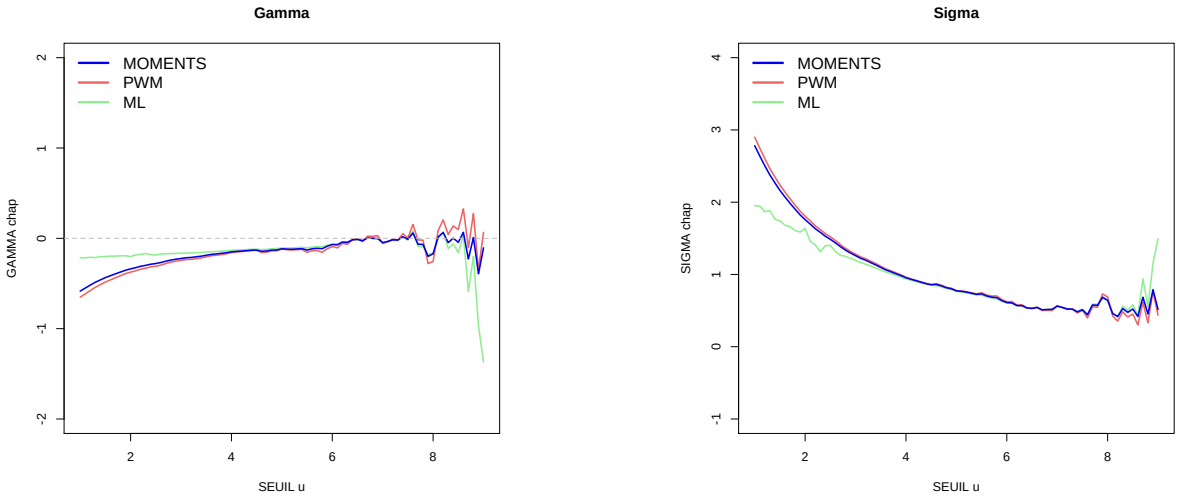


FIGURE 2.8 – Estimation des paramètres de la loi des excès

2.3.3.1 Construction de l'estimateur

On fixe le nombre d'excès, noté n_e, comme un entier tel que $1 < n_e < n$, et étant petit comparé à la taille n de l'échantillon, de sorte que :

$$(1) \quad \lim_{n\to\infty} n_e = +\infty \quad \text{et} \quad (2) \quad \lim_{n\to\infty} \frac{n_e}{n} = 0$$

La condition (1) assure une taille minimale pour n_e alors que la condition (2) impose une valeur maximale de n_e .
On choisit le seuil u comme étant le quantile d'ordre $1 - n_e/n$ de la loi des données. Afin d'obtenir une expression du quantile extrême considéré noté $q(p)_{POT}$, on s'intéresse à la fonction de survie $\mathbb{P}(X > q(p)_{\text{POT}})$.
Pour cela, on cherche tout d'abord la loi asymptotique de la variable aléatoire conditionnelle $\{X > x \,|\, X > u\}$. D'après le théorème de Pickands 2.3.1, pour un seuil u suffisamment élevé et $u > x$, on a :

$$\mathbb{P}(X > x \,|\, X > u) \approx 1 - H_{\gamma,\sigma}(x-u) = \left[1 + \frac{\gamma}{\sigma}(x-u)\right]^{-\frac{1}{\gamma}}$$

or, pour $X > u$, $\mathbb{P}(X > x \cap X > u) = \mathbb{P}(X > u)$, d'après la formule des probabilités conditionnelles on obtient :

$$\mathbb{P}(X > x) = p_u \left[1 + \frac{\gamma}{\sigma}(x-u)\right]^{-\frac{1}{\gamma}} \quad , \quad p_u = \mathbb{P}(X > u) = \frac{n_e}{n}$$

D'après la définition d'un quantile ci-dessus, on cherche alors le quantile $q(p)$ tel que :

$$\mathbb{P}(X > q(p)) = \frac{n_e}{n}\left[1 + \frac{\gamma}{\sigma}\big(q(p) - u\big)\right]^{-\frac{1}{\gamma}} = p$$

enfin, on obtient l'expression connue d'un quantile extrême dans le cadre de la méthode POT :

$$q(p)_{\text{POT}} = u + \frac{\sigma}{\gamma}\left[\left(\frac{n_e}{p\,n}\right)^{\gamma} - 1\right] \tag{2.16}$$

On estime le seuil u par la statistique d'ordre $X_{n-n_e,n}$. On aura n_e observations au dessus du seuil u et finalement on obtient l'estimateur suivant :

$$\widehat{q}(p)_{\text{POT}} = X_{n-n_e,n} + \frac{\widehat{\sigma}}{\widehat{\gamma}}\left[\left(\frac{n_e}{p\,n}\right)^{\widehat{\gamma}} - 1\right] \tag{2.17}$$

Les paramètres de la loi des excès γ et σ seront estimés par l'une des méthodes présentées dans la section précédente (maximum de vraisemblance, moments ou moments pondérés).

La section suivante présente la construction d'intervalles de confiance pour l'estimateur $\widehat{q}(p)_{POT}$. Il existe plusieurs méthodes statistiques pour obtenir de tels intervalles ; nous en présentons trois : les intervalles dérivés de théorèmes asymptotiques, les intervalles obtenus d'après la Méthode Delta et enfin les intervalles obtenus par bootsrap, technique de rééchantillonnage.

2.3.3.2 Intervalle de confiance théorique

Diebolt et al.(2005) [DGR05] démontrent la normalité asymptotique de l'estimateur de quantile extrême ci-dessus dans le cadre de la méthode des excès. Si on estime les paramètres de la loi des excès par la méthode du maximum de vraisemblance, on a la convergence en loi suivante pour $\gamma > -1/2$:
On se donne un niveau n_e élevé tel que $n_e \to \infty$, $n/n_e \to \infty$ et $\sqrt{n_e}A(n/n_e) \to \lambda$, alors

$$\sqrt{n_e}\,\frac{\widehat{q}(p)_{\text{POT}} - q(p)_{\text{POT}}}{\hat{\sigma}^{ML} q_{\hat{\gamma}^{ML}}(n_e/np)} \longrightarrow \mathcal{N}\ (m, v^2) \quad , \quad \text{pour } n_e \to \infty$$

où la fonction $q_\gamma(t)$ est définie par $q_\gamma(t) = \int_1^t s^{\gamma-1}\log(s)ds$, $\forall t > 1$.
De plus on a :

$$m = \begin{cases} \frac{\lambda(1+\gamma)}{(1-\rho)(\gamma-\rho+1)} & \text{si } \gamma \geq 0 \\ \\ \frac{\lambda\rho(1+3\gamma+2\gamma^2)}{(1+\gamma-\rho)(1-\rho)(\gamma+\rho)} & \text{si } \gamma < 0 \end{cases}$$

et

$$v^2 = \begin{cases} (1+\gamma)^2 & \text{si } \gamma \geq 0 \\ 1 + 4\gamma + 5\gamma^2 + 2\gamma^3 + 2\gamma^4 & \text{si } \gamma < 0 \end{cases}$$

ρ est un réel ≥ 0 et A est une fonction telle que $A(t)$ ne change pas de signe pour t grand et $A(t) \to 0$ quand $t \to \infty$. En présence de très grands échantillons, on supposera que $\sqrt{k}A(n/n_e) \to 0$. En particulier on aura $\lambda = 0$ et

$$\sqrt{n_e}\,\frac{\widehat{q}(p)_{\text{POT}} - q(p)_{\text{POT}}}{\hat{\sigma}^{ML} q_{\hat{\gamma}^{ML}}(n_e/np)} \longrightarrow \mathcal{N}\ (0, v^2) \quad , \quad \text{pour } n_e \to \infty$$

Reste à calculer $q_\gamma(t)$. Une intégration par parties donne :

$$q_\gamma(t) = \frac{1}{\gamma}\,t^\gamma\,\left(\log(t) - \frac{1}{\gamma}\right) + \frac{1}{\gamma^2}$$

Finalement, on aura l'intervalle de confiance suivant :

$$IC_{1-\alpha}\big(q(p)_{\mathrm{POT}}\big)_{th} = \left[\widehat{q}(p)_{\mathrm{POT}} - \frac{1}{\sqrt{n_e}}\Phi^{-1}(1-\frac{\alpha}{2})\sigma q_\gamma(\frac{n_e}{np})v \ , \ \widehat{q}(p)_{\mathrm{POT}} + \frac{1}{\sqrt{n_e}}\Phi^{-1}(1-\frac{\alpha}{2})\sigma q_\gamma(\frac{n_e}{np})v\right]$$

Φ^{-1} désigne la fonction quantile d'une loi normale centrée réduite.

2.3.3.3 Intervalle de confiance par la Méthode Delta

A partir des propriétés asymptotiques du couple (γ,σ), on peut établir une convergence en loi pour l'estimateur du quantile POT et calculer un intervalle de confiance à l'aide de la Méthode Delta [Bei04]. j sera égal à 1,2 ou 3 selon que l'on estime le couple (γ,σ) avec la méthode du maximum de vraisemblance, la méthode des moments, ou la méthode des moments pondérés respectivement.

$$\sqrt{n_e}(\widehat{q}(p)_{\mathrm{POT}} - q(p)_{\mathrm{POT}}) \longrightarrow \mathcal{N}\ (0, \zeta^T\, \Delta_j\, \zeta) \quad , \quad \text{pour } n_e \to \infty$$

avec

$$\begin{aligned}\zeta^T &= \left[\frac{\partial q(p)_{\mathrm{POT}}}{\partial\gamma} \ , \ \frac{\partial q(p)_{\mathrm{POT}}}{\partial\sigma}\right] \\ &= \left[-\frac{\sigma}{\gamma^2}\,(p^{-\gamma}-1) - \frac{\sigma}{\gamma}\,p^{-\gamma}\log p \ , \ \frac{1}{\gamma}\,(p^{-\gamma}-1)\right]\end{aligned}$$

et

$$\Delta_1 = (1+\gamma)\begin{bmatrix} 1+\gamma & -\sigma \\ -\sigma & 2\sigma^2 \end{bmatrix} \quad \text{et } \gamma > -1/2$$

$$\Delta_2 = C\begin{bmatrix} (1-2\gamma)^2(1-\gamma+6\gamma^2) & -\sigma(1-2\gamma)(1-4\gamma+12\gamma^2) \\ -\sigma(1-2\gamma)(1-4\gamma+12\gamma^2) & 2\sigma^2(1-6\gamma+12\gamma^2) \end{bmatrix} \quad \text{et } \gamma < 1/4$$

$$\text{où} \quad C = \frac{(1-\gamma)^2}{(1-2\gamma)(1-3\gamma)(1-4\gamma)}$$

$$\Delta_3 = C\begin{bmatrix} (1-\gamma)(1-2\gamma)^2(1-\gamma+2\gamma^2) & -\sigma(2-\gamma)(2-6\gamma+7\gamma^2-2\gamma^3) \\ -\sigma(2-\gamma)(2-6\gamma+7\gamma^2-2\gamma^3) & \sigma^2(7-18\gamma+11\gamma^2-2\gamma^3) \end{bmatrix} \quad \text{et } \gamma < 1/2$$

$$\text{où} \quad C = \frac{1}{(1-2\gamma)(3-2\gamma)}$$

On construit alors l'intervalle de confiance suivant :

$$IC_{1-\alpha}\big(q(p)_{\mathrm{POT}}\big)_{DM} = \left[\widehat{q}(p)_{\mathrm{POT}} - \Phi^{-1}(1-\frac{\alpha}{2})\sqrt{\frac{\zeta^T\,\Delta_j\,\zeta}{n_e}} \ , \ \widehat{q}(p)_{\mathrm{POT}} + \Phi^{-1}(1-\frac{\alpha}{2})\sqrt{\frac{\zeta^T\,\Delta_j\,\zeta}{n_e}}\right]$$

Φ^{-1} désigne la fonction quantile d'une loi normale centrée réduite.

2.3.3.4 Intervalle de confiance par bootstrap

Le bootsrap est une méthode de rééchantillonnage statistique non paramétrique introduite par Efron (1979) [Efr79].
Soit n observations $X_1, \cdots, X_n$, on choisit un entier fixe n_{boot} assez grand (au moins 1000 en général) et on procède à n_{boot} tirages avec remise de n observations parmi l'échantillon $X_1, \cdots, X_n$ initial. On obtient alors n_{boot} nouveaux échantillons issus du jeu de données original. L'idée est alors de s'inspirer d'un théorème central limite pour décrire le comportement des n_{boot} estimateurs qu'on peut former à partir des n_{boot} nouveaux jeux de données.
Sur chacun de ces jeux de données, on calcule alors $\widehat{q}(p)^j_{POT}$ pour $j = 1, \cdots, n_{boot}$. On estime empiriquement la moyenne $\bar{q}(p)_{boot}$ et l'écart-type $sd(p)_{boot}$ de notre estimateur. La méthode consiste à définir un intervalle de confiance boostrapé par

$$IC_{1-\alpha}\big(q(p)_{\text{POT}}\big)_{boot} = \left[\bar{q}(p)_{boot} - sd(p)_{boot}\Phi^{-1}(1-\frac{\alpha}{2}) \;\;,\;\; \bar{q}(p)_{boot} + sd(p)_{boot}\Phi^{-1}(1-\frac{\alpha}{2}) \right]$$

Φ^{-1} étant la fonction quantile d'une loi normale centrée réduite.

2.3.3.5 Exemple sur un échantillon suivant une loi de Rayleigh

On reprend l'exemple de l'échantillon généré suivant une loi de Rayleigh de paramètre 2. On fixe le niveau de probabilité du quantile que l'on veut estimer à $p = 1-10^{-7}$. La valeur théorique de ce quantile vaut $q_{\text{th}} = 11.35538$. On estime le quantile par l'estimateur POT décrit précédemment ainsi que les trois intervalles de confiance associés. Afin d'observer la variabilité de notre estimateur en fonction des valeurs du seuil u et après examen de l'échantillon considéré (figure 2.7), on estime le quantile pour une plage de seuils compris en 2 et 8. De plus, on fixe le niveau des intervalles de confiance à $\alpha = 0.05$.

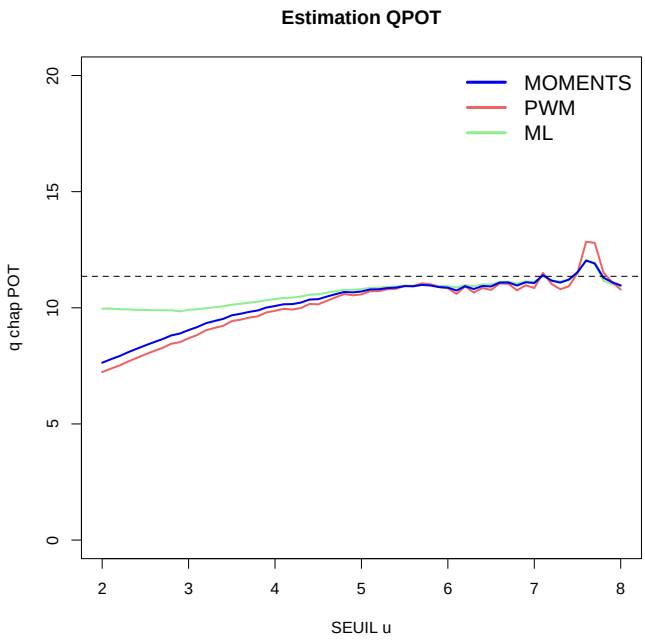


FIGURE 2.9 – Estimation du quantile de niveau $p = 1 - 10^{-7}$

On compare trois estimateurs $\widehat{q}(p)_{\mathrm{POT}}$ obtenus en injectant dans l'expression (5.12) les valeurs des paramètres estimés avec les trois méthodes présentées. Ces trois estimateurs présentent une plage de stabilité autour du seuil de valeur 7. Cependant l'estimateur du maximum de vraisemblance pour les paramètres de la loi des excès, est plus stable que les deux autres. La méthode du maximum de vraisemblance a l'avantage d'avoir de bonnes propriétés asymptotiques mais du fait de la nécessité de résoudre un problème d'optimisation, elle est plus coûteuse d'un point de calculatoire et n'aboutit pas forcément à un résultat explicite. Lorsque les ressources de calculs le permettent, cet estimateur sera toutefois privilégié.

Les courbes ci-dessous (figure 2.10) illustrent les intervalles de confiance obtenus par la Méthode Delta, pour les trois estimateurs du couple (γ,σ).

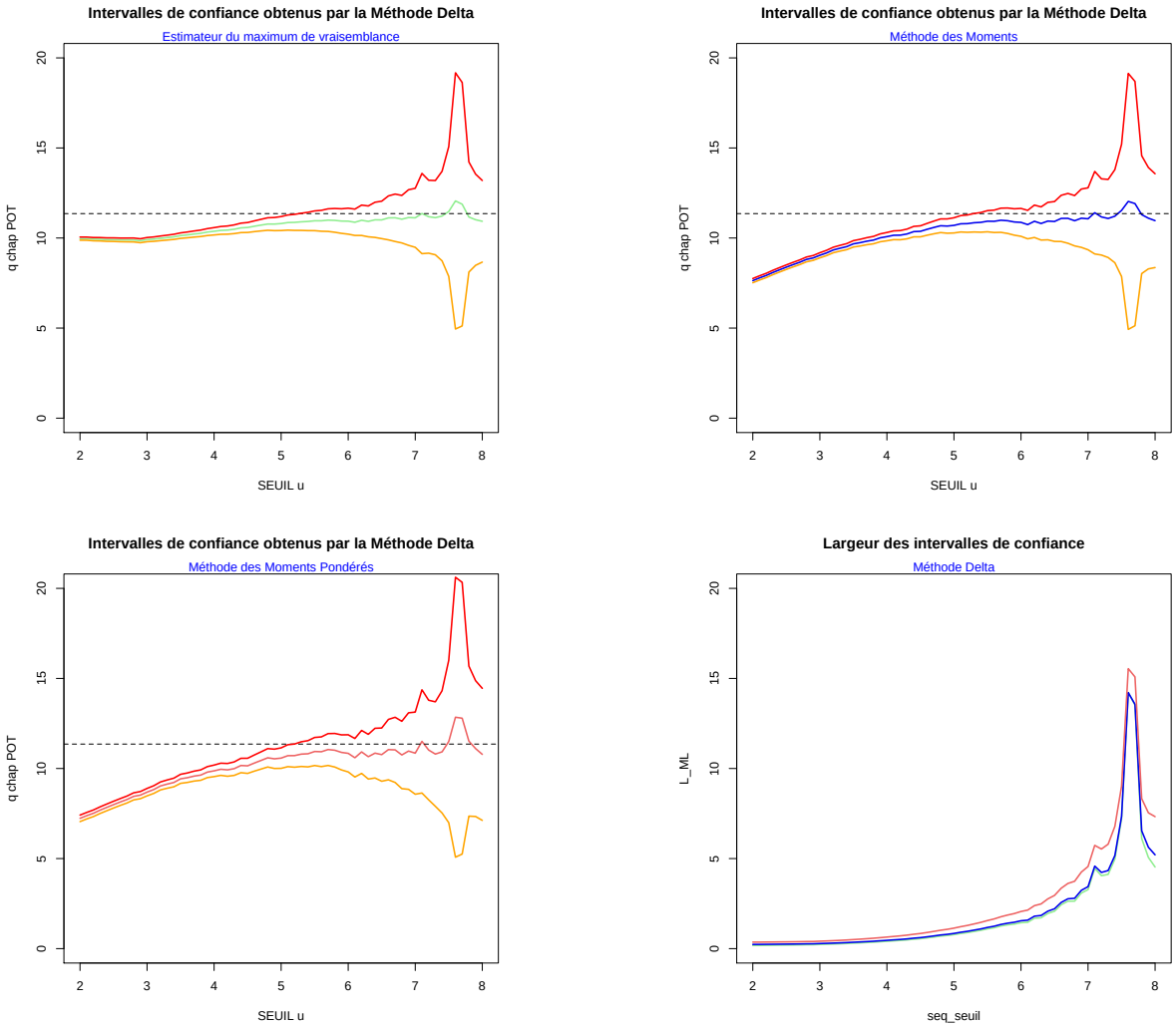


FIGURE 2.10 – Intervalles de confiance par Méthode Delta

Les trois intervalles de confiance présentent un comportement ainsi qu'une largeur en fonction du seuil similaires. La valeur théorique du quantile que l'on cherche à estimer est comprise dans les intervalles de confiance pour une plage de seuils assez large comprise entre 5 et 8.

Les intervalles de confiance issus de théorèmes asymptotiques [DGR05] sont beaucoup plus étroits (figure 2.11). Lorsqu'on souhaite déterminer un u parmi la plage de seuils testés, l'avantage de tels intervalles est de contenir la valeur théorique du quantile que l'on cherche sur une plage assez étroite. Cependant, cette finesse d'intervalle peut mener à de mauvais résultats lorsque la valeur théorique du quantile est inconnue. On observe que la valeur théorique du quantile est comprise dans l'intervalle pour trois valeurs du seuil comprises entre 7 et 8.

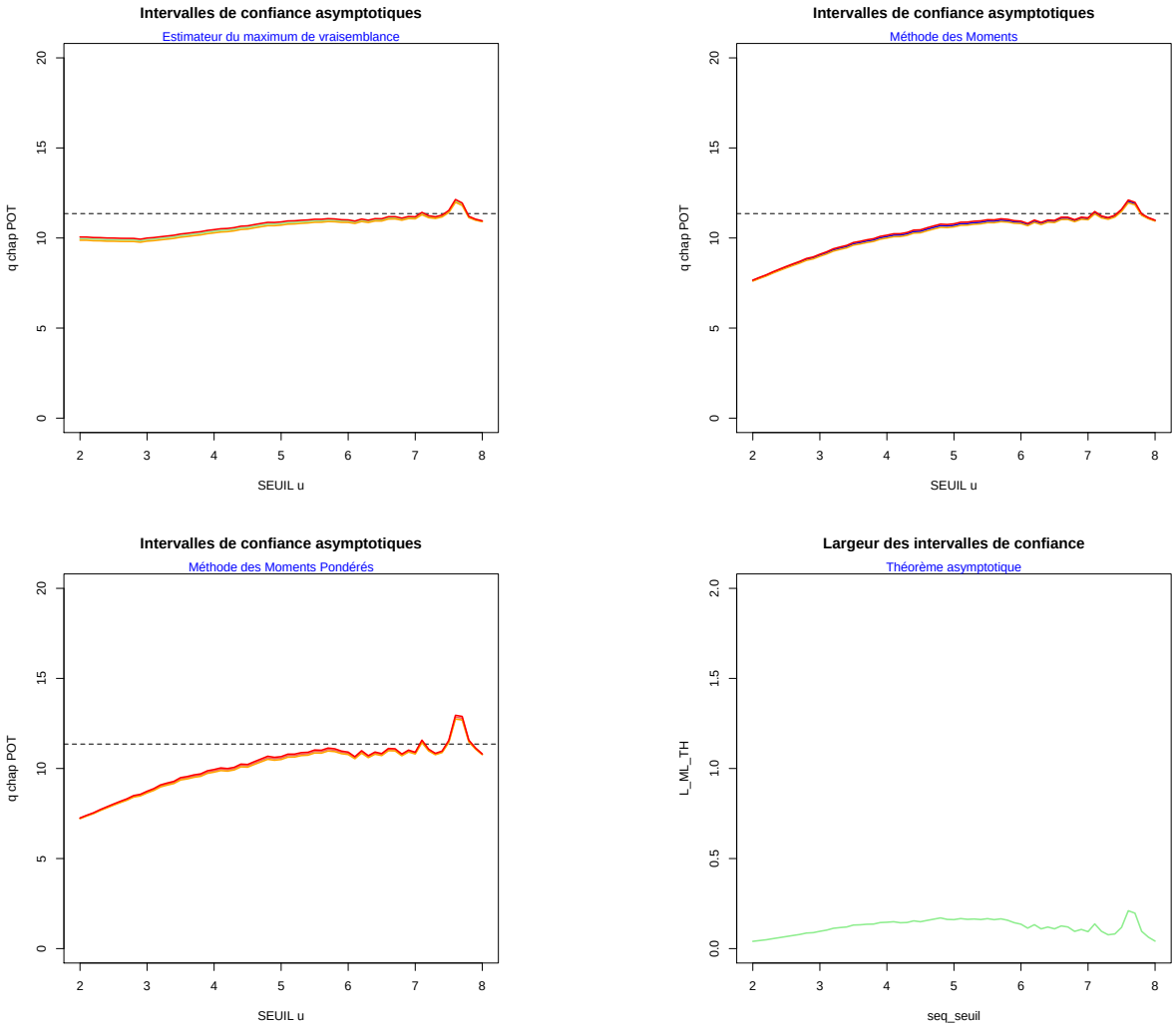


FIGURE 2.11 – Intervalles de confiance asymptotiques

Enfin, les figures suivantes représentent les intervalles de confiance obtenus par rééchantillonnage. A titre d'exemple, on illustre le cas où le couple (γ,σ) est estimé avec la méthode des moments. Afin d'obtenir des intervalles représentatifs, on fixe le nombre de tirages à $n_{boot} = 1000$.
La largeur de ces derniers intervalles de confiance (figure 2.12) est moindre que ceux obtenus par la Méthode Delta. De plus, la plage de seuils comprenant la valeur théorique du quantile à estimer est plus réduite. Ces deux remarques font que ces intervalles de confiance sont un critère plus efficace pour le choix du seuil u. Cependant, pour des es-

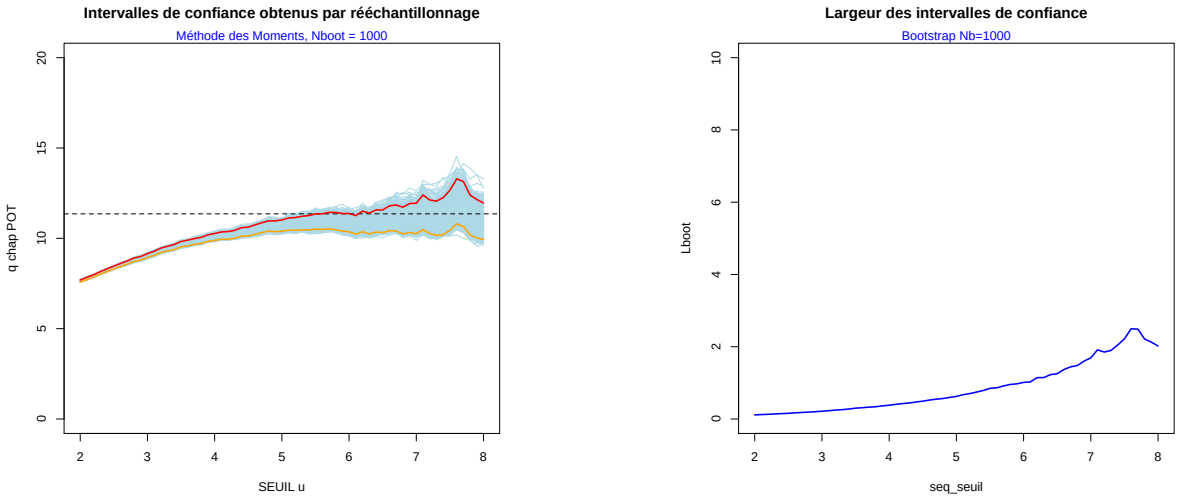


FIGURE 2.12 – Intervalles de confiance bootstrap

timations online, ou pour le traitement d'échantillon de grande taille, le temps de calcul important pour obtenir ce type d'intervalle est à prendre en compte.

L'estimation d'un quantile extrême par la méthode POT présentée jusqu'ici ainsi que l'estimation des paramètres de la loi des excès ont été faites en fonction d'une plage de seuils u. Or on s'intéresse à l'estimation d'une valeur unique. Il existe de nombreux critères de choix de seuil, graphiques ou calculatoires, menant à une estimation correcte du quantile recherché.

2.3.4 Choix du seuil u

Dans un cadre statistique, le choix du seuil u est très important car il induit une grande variabilité dans l'estimation des quantiles extrêmes et des paramètres de la loi des excès.

Il existe différentes approches pour le choix du seuil u de la méthode POT. En effet, le seuil doit être suffisamment grand pour satisfaire la caractère asymptotique du modèle, mais pas trop élevé non plus, afin de garder un nombre d'excès suffisant pour estimer convenablement les paramètres du modèle. Une technique peu coûteuse consiste à estimer les paramètres d'échelle et de forme ainsi que le quantile pour un niveau voulu, en fonction de u et retenir le seuil u pour lequel ces estimations sont stables. Sur les exemples précédents, on a retenu les valeurs proches de 7. Il faut toutefois s'assurer que le modèle choisi est bien ajustée aux données. On peut le faire graphiquement en estimant le quantile $q(p)_{\text{POT}}$ pour des niveaux de probabilité compris entre 0 et 1, en conservant la valeur de u retenue précédemment.

On observe sur la figure 2.13 que le modèle obtenu en utilisant l'estimateur du maximum de vraisemblance pour les paramètres de la loi GEV (courbe de gauche) ainsi que pour la loi des excès GPD (courbe de droite) est correctement ajusté aux données. De plus, cette adéquation reste valable quelque soit la méthode d'estimation choisie (figure 2.14). Cette méthode nous mène à la conclusion suivante pour le quantile recherché de niveau

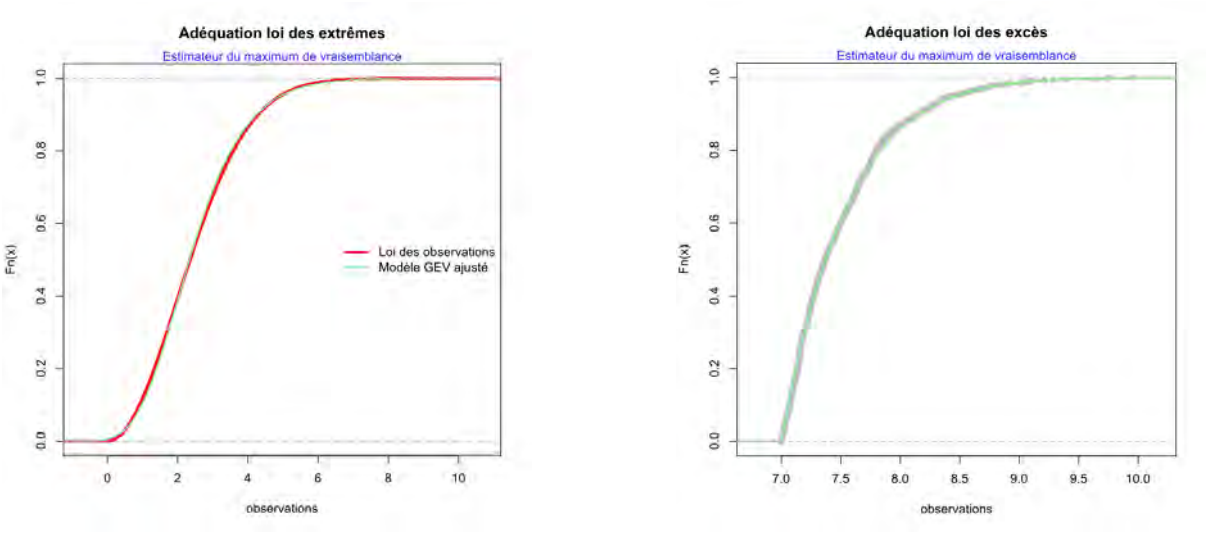


FIGURE 2.13 – Adéquation du modèle ajusté par maximum de vraisemblance

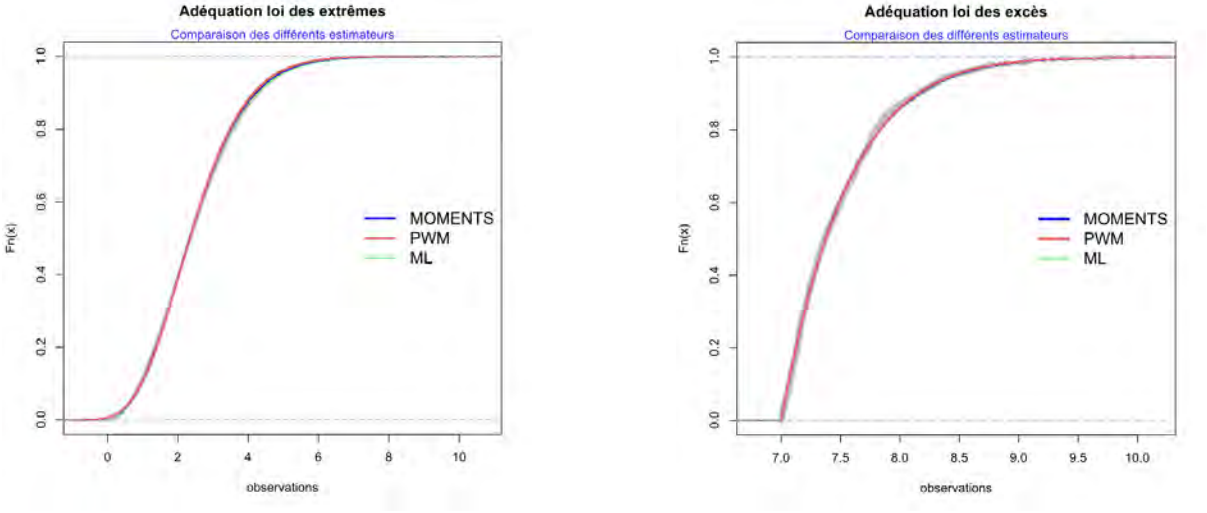


FIGURE 2.14 – Adéquation pour les trois estimateurs

$p = 1 - 10^{-7}$, et pour un niveau d'intervalle de confiance $\alpha = 0.05$. On rappelle la valeur théorique du quantile recherché : $q_{th} = 11.35$.

Les trois estimations sont satisfaisantes dans cet exemple car, dans les trois cas, la valeur théorique du quantile que l'on cherche est bien contenue dans les intervalles de confiance (Table 2.1).

Une autre approche présentée par Coles (2001) [Col01], appelée "Mean Residual Plot", est fondée sur la linéarité de l'estimation de la moyenne de la loi des excès. Cette méthode est peu employée car le graphe en u de l'estimateur de $\mathbb{E}[X - u \,|\, X > u]$ est souvent d'allure très saccadée, et donc difficile à interpréter.

Il est possible de choisir un seuil grâce à des méthodes calculatoires. Toulemonde (2008) [Tou08] propose un test d'ajustement pour la distribution de Pareto qui, après une repa-

Estimateurs	$\widehat{\gamma}$	$\widehat{\sigma}$	$\widehat{q}(p)_{\mathrm{POT}}$	$IC\big(q(p)_{\mathrm{POT}}\big)_{th}$	$IC\big(q(p)_{\mathrm{POT}}\big)_{DM}$	$IC\big(q(p)_{\mathrm{POT}}\big)_{boot}$
ML	-0.08	0.55	11.32	$[10.98, 11.47]_{th}$	$[9.48, 12.76]_{DM}$	$[10.24, 12.48]_{boot}$
MOM	-0.08	0.55	11.26	$[11.01, 11.52]_{th}$	$[9.34, 12.78]_{DM}$	$[10.48, 12.11]_{boot}$
WPM	-0.10	0.56	10.85	$[10.20, 11.89]_{th}$	$[8.57, 13.13]_{DM}$	$[9.99, 13.01]_{boot}$

TABLE 2.1 – Bilan de l'estimation du quantile de niveau $1 - 10^{-7}$ d'une loi de Rayleigh de paramètre (2)

ramétrisation adéquate, peut-être considérée comme un cas particulier de la loi de Pareto Généralisée. Davison et Smith (1990) [DS90] mettent en évidence que ce type de test peut être utile pour l'élaboration d'une méthode de choix de seuil, cependant, l'optimalité du seuil retenu n'est pas prouvée.
Beirlant (2004) [Bei04] propose une méthode de choix de seuil optimal basée sur la minimisation des erreurs quadratiques moyennes asymptotiques (Asymptotic Mean Squared Error) pour différents estimateurs de l'indice des valeurs extrêmes γ. Dupuis et al. (2006) [DVF06] développent un critère robuste de prévision pour la sélection du nombre k de statistiques d'ordres et de l'indice des valeurs extrêmes (pour γ exclusivement positif, ce qui n'est pas notre cas sur des données GPS). Ces AMSE ainsi que leurs estimateurs sont donnés dans Dekkers et al. (1989) [DEDH89] pour l'estimateur des moments et dans Drees (2004) [DFDH04] pour l'estimateur du maximum de vraisemblance. Cette méthode de minimisation des AMSE nécessite l'estimation de nouveaux paramètres (intervenant dans les théorèmes asymptotiques) ainsi que de nouvelles hypothèses sur ces derniers. Cette contrainte nous a menés à ne pas utiliser ces méthodes de choix de seuil qui s'avéraient être peu pratiques à l'usage pour les hommes de métier du GNSS.
Une procédure automatique de sélection de seuil est présentée dans le dernier chapitre de ce manuscrit. Cette procédure utilise les critères suivants :

- la détection de zones de stabilité sur les estimateurs des paramètres de la loi des excès et du quantile recherché.
- l'optimisation de l'adéquation des modèles générés par les seuils retenus.

2.4 Cas du domaine d'attraction de Gumbel

On s'intéresse dans cette partie au domaine d'attraction de Gumbel, c'est à dire aux modèles pour lesquels la loi des excès tend vers une loi exponentielle. En effet, nous verrons que les lois des erreurs de positionnement qui nous intéressent sont proches des lois de Rayleigh (dans le cas horizontal) et gaussienne (dans le cas vertical). On rappelle la définition de la loi de Pareto Généralisée pour le cas où l'indice des valeurs extrêmes γ est nul.

Soit $X_1, \cdots, X_n$ une suite de variables aléatoires *i.i.d.* de fonction de répartition F et soit Y un échantillon de n_e excès au dessus d'un seuil u, supposés indépendants et de fonction de répartition inconnue $F_u(y)$. Alors le théorème de Pickands 2.3.1 assure que

Théorème 2.4.1. *Si F appartient au domaine d'attraction de Gumbel, il existe une fonction $\sigma(.)$ positive telle que*

$$\lim_{u \to x^*} \sup_{0<y<x^*-u} \left| F_u(y) - \left(1 - \exp(-\frac{x}{\sigma})\right) \right| = 0 \tag{2.18}$$

où x^* est le point terminal de la fonction de répartition F.

Propriété du domaine d'attraction de Gumbel : Pour tout échantillon issu d'une loi appartenant au domaine d'attraction de Gumbel, et pour un seuil convenablement choisi, la loi des excès au-dessus de ce seuil suit approximativement une loi exponentielle, lorsque la taille de l'échantillon et le seuils tendent vers l'infini.
Beaucoup de lois classiques se trouvent dans le domaine d'attraction de Gumbel. C'est le cas des lois normales, exponentielles, Gamma, Weibull, Rayleigh, log-normale, Gumbel, Logistic.
En pratique, il existe un diagnostic graphique [Col01] pour vérifier cette propriété à partir d'un Quantile-Quantile plot (QQplot) indiquant l'appartenance probable à l'un des trois domaines d'attraction des valeurs extrêmes. Il s'agit de tracer les quantiles de la loi exponentielle standard $(-\ln(i/n_e)_{i=1,\cdots,n_e})$ contre les excès ordonnés donnés par $(X_{n-n_e+i,n} - X_{n-n_e,n})_{i=1,\cdots,n_e}$. On suppose que le nombre d'excès utilisé est adéquat. Sous cette hypothèse, si les données appartiennent au DA(Gumbel), alors les excès suivent approximativement la loi exponentielle. Les points du graphe du QQplot seront donc approximativement alignés (figure 2.15, centre). En revanche, si les données n'appartiennent pas au domaine d'attraction de Gumbel, on peut observer deux phénomènes : soit une incurvation du QQplot vers le haut (convexité) témoignant de l'appartenance des données au domaine d'attraction de Fréchet (figure 2.15, droite), ou au contraire une incurvation vers le bas ou la droite (concavité) si les données sont dans le domaine d'attraction de Weibull (figure 2.15, gauche).

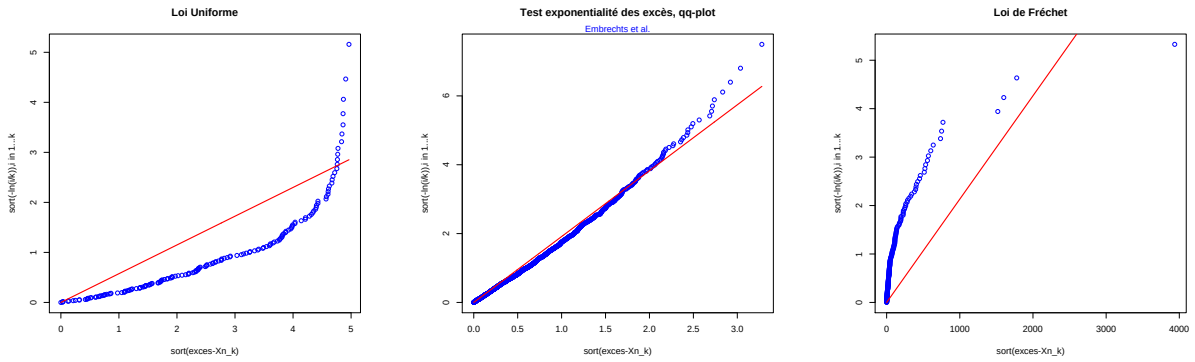


FIGURE 2.15 – Diagnostic graphique du domaine d'attraction

2.4.1 Condition suffisante d'appartenance au domaine d'attraction de Gumbel

Gnedenko (1943) [Gne43] donne une condition suffisante d'appartenance au domaine d'attraction de Gumbel :

Théorème 2.4.2. *Si F est de classe $\mathcal{C}^2$, F appartient au domaine d'attraction de Gumbel si*

$$\lim_{x\to\infty} \frac{(1-F(x))\partial^2 F(x)}{(\partial F(x))^2} = -1 \tag{2.19}$$

A titre d'exemple, on étudie le cas d'une loi de Rayleigh de paramètre ρ dont la fonction de répartition est définie par

$$F(x) = 1 - \exp(-\frac{x^2}{2\rho^2}) \tag{2.20}$$

Afin de vérifier l'appartenance de la loi de Rayleigh au domaine d'attraction de Gumbel on calcule le rapport énoncé dans le théorème 5.2.1. On a

$$\partial F(x) = \frac{1}{\rho^2} x \exp(-\frac{x^2}{2\rho^2})$$

et,

$$\partial^2 F(x) = \frac{1}{\rho^2} \exp(-\frac{x^2}{2\rho^2}) - \frac{x^2}{\rho^4} \exp(-\frac{x^2}{2\rho^2})$$

donc,

$$\frac{(1-F(x))\partial^2 F(x)}{(\partial F(x))^2} = \frac{\exp(-\frac{x^2}{2\rho^2})\left[\frac{1}{\rho^2}\exp(-\frac{x^2}{2\rho^2}) - \frac{x^2}{\rho^4}\exp(-\frac{x^2}{2\rho^2})\right]}{\left(\frac{1}{\rho^2} x \exp(-\frac{x^2}{2\rho^2})\right)^2} = \frac{\rho^2}{x^2} - 1$$

On a bien la limite qui vaut -1 quand x tend vers l'infini ; la condition est donc vérifiée.

La loi limite, comme mentionné plus haut, est une loi exponentielle de paramètre σ. Elle ne comporte plus qu'un seul paramètre, là où elle en avait deux dans le cas général. L'unique paramètre σ de la loi des excès pourra être estimé par l'estimateur du maximum de vraisemblance du paramètre d'une loi exponentielle (Breiman, 1990) [BSK90] :

$$\widehat{\sigma} = \frac{1}{n_e - 1} \sum_{i=1}^{n_e} \left(Y_i - Y_{n_e}\right) \tag{2.21}$$

Le même estimateur exprimé sous forme de statistiques d'ordre de l'échantillon original des observations dont on dispose s'écrit

$$\widehat{\sigma} = \frac{1}{n_e - 1} \sum_{i=1}^{n_e} \left(X_{n-i+1,n} - X_{n-n_e+1,n}\right) \tag{2.22}$$

2.4.2 Estimateur de quantile extrême dans le cas $\gamma = 0$

De façon analogue au cas général, on construit l'estimateur de quantile suivant à partir de la fonction de répartition de la loi des excès :

$$\widehat{q}(p)_{\gamma=0} = u + \widehat{\sigma} \log\big(\frac{n_e}{np}\big)$$

le seuil u sera estimé par la statistique d'ordre $X_{n-n_e,n}$ et on aura :

$$\widehat{q}(p)_{\gamma=0} = X_{n-n_e,n} + \widehat{\sigma} \log\big(\frac{n_e}{np}\big) \tag{2.23}$$

2.4.3 Intervalles de confiance dans le cas $\gamma = 0$

Intervalle de confiance asymptotique :
Garrido (2002) [Gar02] donne un intervalle de confiance pour cet estimateur. Cet intervalle de confiance prend en compte l'erreur d'approximation du quantile théorique q_{th} par $q_{\gamma=0}$. En effet, on suppose que les excès suivent une loi exponentielle. Le théorème de Pickands étant un résultat asymptotique, cette approximation n'est vraie que pour un très grand nombre d'excès. L'intervalle de confiance prend aussi en compte l'erreur d'estimation de $q_{\gamma=0}$ par l'estimateur $\widehat{q}_{\gamma=0}$ dont la loi asymptotique est donnée dans [dHR93] et [DR84]. Cet intervalle de confiance s'exprime par

$$IC_{1-\alpha}\big(q(p)_{\gamma=0}\big)_{DM} = \left[\widehat{q}_{\gamma=0} - \Phi^{-1}(1-\frac{\alpha}{2})\,\widehat{\sigma}\frac{\log(n_e/np)}{\sqrt{n_e}}, \widehat{q}_{\gamma=0} + \Phi^{-1}(1-\frac{\alpha}{2})\,\widehat{\sigma}\frac{\log(n_e/np)}{\sqrt{n_e}} \right]$$

2.4.4 Exemple

On utilise le même tirage de données que dans les exemples précédents. On rappelle (*cf.* 2.3.2.3) qu'il s'agit d'un échantillon de taille $N = 360000$ généré suivant une loi de Rayleigh de paramètre $\rho = 2$. On désire estimer le quantile de niveau $p = 1 - 10^{-7}$ dont la valeur théorique est $q_{th} = 11.35$. On parcourt une plage de seuils allant de 2 à 8 avec un pas de 0.1.
On remarque sur la courbe de gauche (figure 2.16) que la valeur théorique est cette fois approchée par le haut. Elle est contenue dans l'intervalle de confiance ($\alpha = 0.05$) sur une plage de seuils assez courte. L'estimateur se stabilise pour des valeurs de seuils autour de 7 comme les estimateurs précédents et l'intervalle de confiance est étroit (courbe de droite).
On peut dans ce cas aussi trouver des intervalles de confiances aléatoires par rééchantillonnage. On fixe le nombre de tirages bootstrap à $n_{boot} = 1000$ et on obtient un intervalle de confiance (figure 2.17) dont le comportement et la largeur sont proches de l'intervalle asymptotique précédent (figure 2.16).

Le tableau 2.2 suivant donne les résultats obtenus pour un seuil fixé à titre d'exemple à 7.

2.5 Cas de données dépendantes

Les outils issus de la théorie des extrêmes présentés jusqu'ici, étaient valables pour des variables aléatoires **indépendantes** et identiquement distribuées. Cependant lorsque les

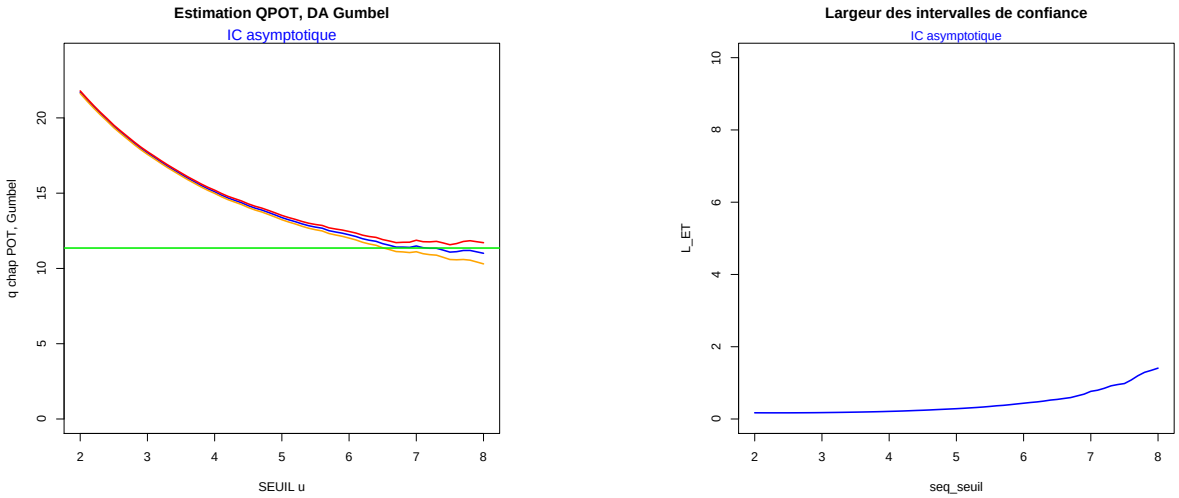


FIGURE 2.16 – Intervalle de confiance asymptotique dans le cas Gumbel

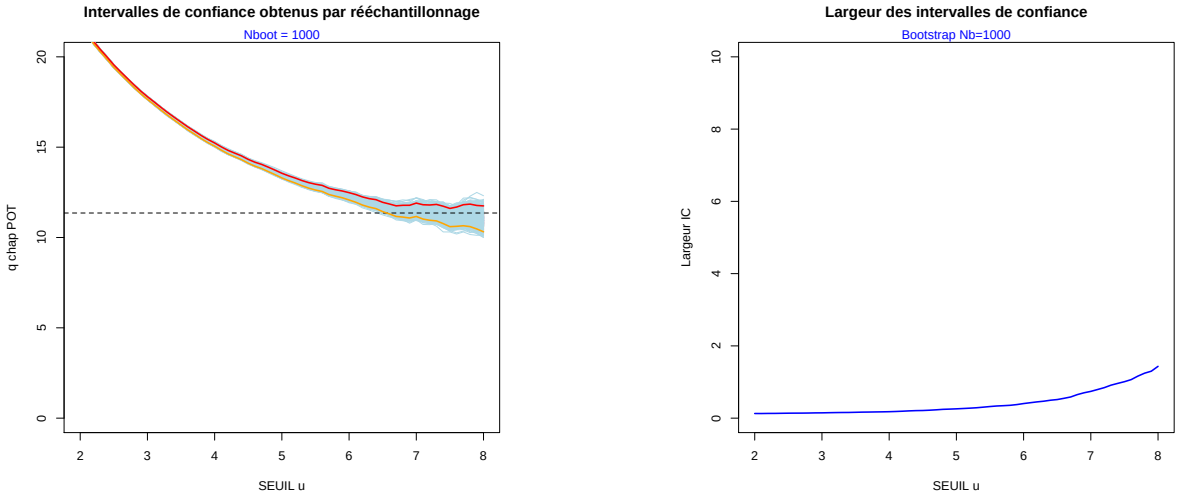


FIGURE 2.17 – Intervalle de confiance bootstrap dans le cas Gumbel

γ	$\widehat{\sigma}$	$\widehat{q}(p)_{\gamma=0}$	$IC\big(q(p)_{\gamma=0}\big)_{DM}$	$IC\big(q(p)_{\gamma=0}\big)_{boot}$
0	0.50	11.36	$[10.87, 11.86]_{DM}$	$[10.62, 11.63]_{boot}$

TABLE 2.2 – Bilan de l'estimation avec l'estimateur de type Gumbel

données à analyser présentent un caractère important de dépendance, l'étude du comportement des valeurs extrêmes est différente et implique la manipulation d'outils supplémentaires que nous allons présenter succinctement.
Certaines séries de données et notamment les mesures GPS, sont dépendantes temporellement. On parlera de corrélation temporelle. Cette dépendance est due à la récursivité des

algorithmes d'estimation de la position, présents au sein des récepteurs GNSS. L'étude de la corrélation temporelle existant au sein des données GPS recueillies sera discutée en détail au chapitre 4.
La dépendance forte des données implique un comportement particulier des valeurs extrêmes d'un échantillon : ces événements rares ont tendance à apparaître regroupés par paquets que l'on nommera clusters. Les phénomènes extrêmes peuvent persister sur plusieurs observations consécutives. Si on prend pour exemple des relevés journaliers de températures, lors d'une période de canicule, on observera des valeurs très élevées de température pendant plusieurs jours. Ceci formera un cluster.

2.5.1 Condition de mélange et Extremal Index

Il existe une généralisation de l'étude des valeurs extrêmes dans le cas *i.i.d.* pour le cas des séries stationnaires. En effet, si les données répondent à la condition de mélange $D(u_n)$ (Leadbetter, 1983 [Lea83]), c'est à dire si la dépendance n'est pas trop présente dans le temps, alors la seule loi limite non dégénérée du maximum (correctement renormalisé) d'un échantillon stationnaire appartient à la famille paramétrique des lois Generalized Extreme Value [Lea83]. Toutefois, cette loi limite ne sera pas identique à celle de l'échantillon *i.i.d.* correspondant (avec la même loi F) du fait de l'apparition d'un nouveau paramètre : l'extremal index, noté θ. θ caractérise la dépendance à court terme au sein de l'échantillon. Il existe plusieurs méthodes pour estimer le paramètre θ que l'on trouvera dans [Bei04], [Col01] ou [EKM97].

2.5.2 Estimation de quantile extrême avec des données dépendantes

Coles [Col01] donne un estimateur de quantile extrême lorsque les observations ne sont plus distribuées de façon indépendante.

$$\widehat{q}(p)^{\text{IND}}_{\text{POT}} = X_{n-k,n} + \frac{\widehat{\sigma}}{\widehat{\gamma}}\left[\left(\frac{n_e}{p\,n}\,.\,\hat{\theta}\right)^{\widehat{\gamma}} - 1\right] \tag{2.24}$$

2.6 Conclusion

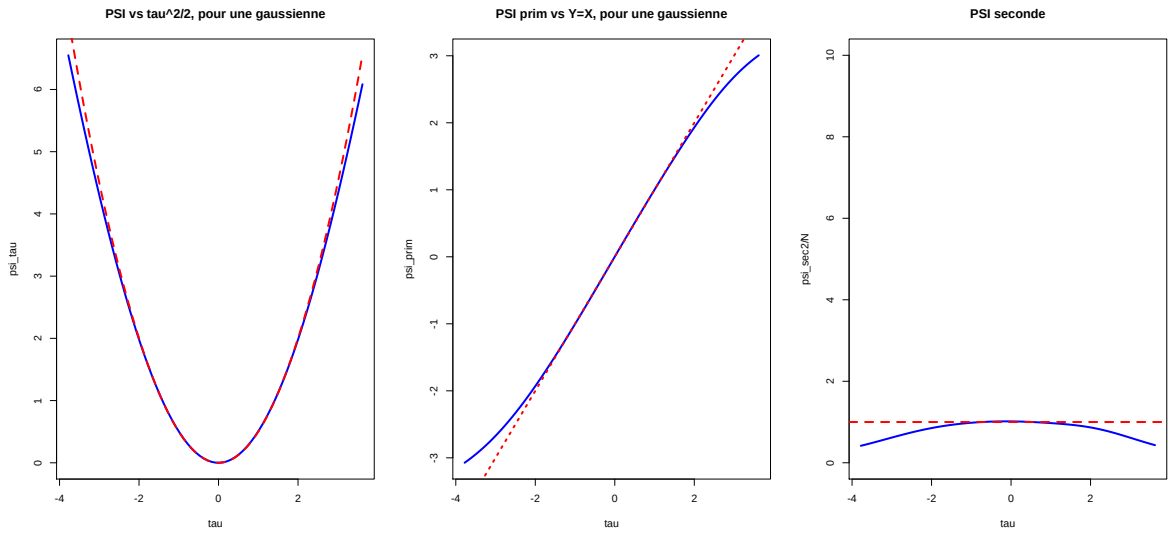
L'estimation de quantile extrême par la méthodologie issue de la théorie des extrêmes présentée dans ce chapitre, s'est avérée être très efficace sur des cas théoriques. En particuliers, nous avons vu que le modèle des observations au dessus d'un seuil, méthode POT, était plus adapté à notre future étude que la modélisation par blocs pour le modèle suivant une loi d'extrême généralisée. Concernant les problématiques d'inférence liées au modèle POT, plusieurs estimateurs ont été testés sur un ensemble de lois théoriques dont l'une d'entre elle est exposée dans ce chapitre. Les estimateurs des moments et du maximum de vraisemblance ont présenté les meilleures propriétés et ont donc été retenus pour l'étude sur les données GPS. Les erreurs de positionnement à observer pour l'analyse qui sera présentée par la suite, suivent des lois proches de lois théoriques appartenant au domaine d'attraction de Gumbel. Un estimateur de quantile extrême exprimé pour le cas où la loi des observations appartient à ce domaine d'attraction a été présenté, et sera lui aussi retenu pour l'étude sur les données GPS. Enfin, un fort accent a été mis sur les

intervalles de confiance associés aux différents estimateurs vu jusqu'ici afin d'insister sur la notion de garantie et de qualité d'estimation, indispensable en statistique.

C H A P I T R E 3

Estimation de quantile extrême par approximation point selle

Résumé :
Un nouvel estimateur de quantile extrême est étudié dans ce chapitre. Il est construit à partir d'un théorème que nous démontrons. Ce théorème décrit le comportement de la décroissance d'une queue de distribution. Nous construisons un estimateur par inversion de cette approximation. Le théorème sera démontré dans ce chapitre et des simulations y sont réalisées afin d'évaluer les performances de ce nouvel estimateur. Cette approche diffère de la théorie des extrêmes dans la façon de caractériser la distribution de la variable observée. En effet, nous donnons une approximation de type point selle de la transformée de Laplace (ou fonction génératrice des cumulants) de la variable considérée.



3.1 Introduction

Le problème inverse d'estimation d'un quantile implique la connaissance de la distribution de la variable observée. Bien souvent, il n'est pas possible d'en connaître une expression exacte et on doit employer des méthodes d'estimation statistiques pour en obtenir une version la plus précise possible. Les niveaux de probabilité très élevés des quantiles présents dans le modèle d'intégrité auquel on s'intéresse, nous mènent à l'étude des queues de distribution.
Les méthodes statistiques dédiées à l'étude comportementale des queues de distribution sont fondées sur des théorèmes asymptotiques. C'est le cas dans la théorie des valeurs extrêmes avec le théorème de Fisher-Tippet (1928) [FT28] qui est un équivalent du théorème limite central pour la valeur maximale d'un échantillon. Il existe d'autres moyens à partir de méthodes analytiques pour caractériser les distributions des variables observées qui s'avèrent être performants pour les parties centrales mais aussi pour les queues de distribution.

Lorsque la fonction de distribution, ou de manière équivalente la densité de probabilité, d'une variable aléatoire sont inconnues, on peut utiliser les méthodes d'approximation point selle introduite dans un cadre statistique par Daniels en 1954 [Dan54]. Cet auteur considère le problème d'estimation de la densité comme une inversion de sa transformée de Fourier. Le coût calculatoire de ces méthodes est assez réduit comparé à d'autres méthodes de simulation stochastique comme les méthodes de Monte-Carlo.
L'approximation point selle d'une fonction de répartition a été introduite par Lugannani-Rice (1980) [LR80]. Les méthodes de type approximation point selle sont comparées dans cet article au développement d'Edgeworth et montrent de meilleures propriétés l'évaluation des queues de distribution, pour des échantillons de petites tailles.

On propose dans ce chapitre de construire un nouvel estimateur d'un quantile extrême à partir d'une inversion de l'approximation point selle de la queue d'une distribution. Cette approximation nécessite que l'on évalue au préalable la transformée de Fenchel-Legendre de la fonction de répartition. Pour cela, on introduit les fonctions suivantes :

On suppose que l'on travaille sur des variables aléatoires réelles X dont le support des lois est tel que

$$\exists a \in \mathbb{R} \ / \ \operatorname{supp} \mathcal{L}(X) \supset]a, +\infty[\tag{3.1}$$

Définition 3.1.1. *Soit X une variable aléatoire à valeurs dans $\mathbb{R} \cup \{+\infty\}$. La* ***fonction génératrice des cumulants*** *ou cumulant generating function (CGF) est définie par :*

$$\Lambda(\tau) = \log \mathbb{E}[e^{\tau X}] \ , \ \tau \in \mathbb{R}.$$

Remarque : lorsque τ est tel que $\mathbb{E}[e^{\tau X}] = +\infty$, on pose $\Lambda(\tau) = +\infty$. De plus, si τ est tel que $\Lambda(\tau) < +\infty$ on pose pour $r \in \mathbb{R}$, $\Lambda(\tau + ir) = \log \mathbb{E}[e^{(\tau+ir)X}]$.

Définition 3.1.2. *On définit la* ***transformée de Fenchel-Legendre*** *de Λ par :*

$$\Lambda^*(x) = \sup_t \{t\, x - \Lambda(t) \ ; \ t \in \mathbb{R}\}$$

Par analogie avec la méthode POT (Peak Over Threshold), issue de l'étude des valeurs extrêmes, on introduit un modèle à dépassement de seuil.
Le théorème suivant fournit une approximation de la queue d'une distribution à partir de ce modèle. La suite du chapitre sera consacrée à la preuve de ce résultat et à des simulations.

Théorème 3.1.1. *Soit X une variable aléatoire à valeur dans $\mathbb{R} \cup \{+\infty\}$ et soit Λ la fonction génératrice des cumulants de X. On suppose que les hypothèses suivantes sont satisfaites :*
(i) X admet une densité de probabilité dans $L^2(\mathbb{R})$.
(ii) hypothèse H_1 : on suppose que pour $\tilde{\tau} \in]\tau; \tau + a[$, $\tau \in \mathbb{R}$ et $a \in \mathbb{R}$

$$\frac{\Lambda''(\tilde{\tau})}{\Lambda''(\tau)} \longrightarrow 1 \quad \textit{quand } \tau \to \infty \tag{3.2}$$

(iii) hypothèse H_2 : pour $\tau \in \mathbb{R}$ grand, on suppose que $\forall r > 0$, $Re(\Lambda^{(3)}(\tau + ir)) \geq 0$.
(iv) hypothèse H_3 : pour $R \in \mathbb{R}$ on suppose que

$$\frac{1}{R} \log \Lambda''(R) \to 0 \quad \textit{quand } R \to \infty \tag{3.3}$$

(v) pour un réel u assez grand, on suppose qu'il existe un τ_u tel que $\Lambda'(\tau_u) = u$.
Alors pour tous $u, t > 0$ on a

$$\log \frac{\mathbb{P}(X > u + t)}{\mathbb{P}(X > u)} \sim -\tau_u t \quad \textit{quand } u \to \infty \tag{3.4}$$

On note que ce théorème s'applique aux lois à queue légère appartenant au domaine d'attraction de Gumbel défini au chapitre 2. On rappelle la définition d'une distribution à queue légère :
La loi de X est une distribution à queue légère, s'il existe un $\theta > 0$ tel que

$$\limsup_{x \to +\infty} \frac{\mathbb{P}(X > x)}{e^{-\theta x}} < \infty \tag{3.5}$$

3.2 Preuve, exemple et contre exemple

3.2.1 Preuve du théorème

Soit X une variable aléatoire à valeur dans $\mathbb{R}$ de loi F_X. On rappelle que l'on suppose dans les hypothèses du théorème que la loi de X est de support infini et admet une densité de probabilité dans $L^2(\mathbb{R})$.

On cherche un quantile u tel que $\mathbb{P}(X > u) = \alpha$ avec α fixé et $u > 0$.

Soit $\Lambda(\tau) = \log \mathbb{E}[e^{\tau X}]$ la fonction génératrice des cumulants associée à X. On suppose qu'il existe un τ_u^* tel que :

$$\Lambda'(\tau_u^*) = u \tag{3.6}$$

L'existence et l'unicité d'un tel τ_u^* sont assurés par la convexité de la fonction $\Lambda(.)$ [Dan54] [Dan87]. En effet, on peut montrer que la transformée de Laplace a pour propriété d'être log convexe [BN78].

Comme dans la méthodologie POT présentée au chapitre 2, on cherche une expression du quotient $\frac{\mathbb{P}(X>u+t)}{\mathbb{P}(X>u)}$.

$$\mathbb{P}(X > u) = \int_{\{x>u\}} dF_X(x) = \int_{\{x>u\}} \exp(\tau_u^* x - \tau_u^* x + \Lambda(\tau_u^*) - \Lambda(\tau_u^*)) dF_X(x)$$

On effectue le changement de mesure de probabilité suivant :

$$dF_{\tau_u^*}(x) = \exp(\tau_u^* x - \Lambda(\tau_u^*)) dF_X(x)$$

On a alors en réinjectant dans l'expression de $\mathbb{P}(X > u)$ précédente :

$$\begin{aligned} \mathbb{P}(X > u) &= \exp\big(\Lambda(\tau_u^*)\big) \int_{\{x>u\}} \exp(-\tau_u^* x) dF_{\tau_u^*}(x) \\ &= \exp\big(\Lambda(\tau_u^*)\big) \int_{\{u>0\}} \exp(-\tau_u^*(u+t)) dG(t) \end{aligned}$$

où G est la loi de la variable aléatoire recentrée $Y_{\tau_u^*} = X_{\tau_u^*} - u$.

$$\mathbb{P}(X > u) = \exp(\Lambda(\tau_u^*) - \tau_u^* u) \int_{\{t>0\}} \exp(-\tau_u^* t) dG(t)$$

En résumé, si on pose $\Lambda^*(u) = \sup_\tau\{u\tau - \Lambda(\tau)\}$ alors $\Lambda^*(u) = u\tau_u^* - \Lambda(\tau_u^*)$ car la dérivée par rapport à τ_u^* est nulle en utilisant l'équation (3.6). On a :

$$\mathbb{P}(X > u) = \exp(-\Lambda^*(u)) \mathbb{E}\Big[\mathbb{1}_{[Y_{\tau_u^*}>0]} \exp(-Y_{\tau_u^*}\tau_u^*)\Big] \tag{3.7}$$

où $Y_{\tau_u^*} = X_{\tau_u^*} - u$ (recentrage de la loi comme précédemment).

On doit donc étudier le comportement de la partie exponentielle puis de l'espérance.

En effectuant les mêmes étapes menant à l'expression (3.7), on a pour $\mathbb{P}(X > u + t)$:

$$\mathbb{P}(X > u + t) = \exp(-\Lambda^*(u+t)) \mathbb{E}\Big[\mathbb{1}_{[Y_{\tau_{u+t}^*}>0]} \exp(-Y_{\tau_{u+t}^*}\tau_{u+t}^*)\Big] \tag{3.8}$$

Le quotient vaut alors,

$$\frac{\mathbb{P}(X > u+t)}{\mathbb{P}(X > u)} = \exp\big(-(\Lambda^*(u+t) - \Lambda^*(u))\big) \frac{\mathbb{E}\Big[\mathbb{1}_{[Y_{\tau_{u+t}^*}>0]} \exp(-Y_{\tau_{u+t}^*}\tau_{u+t}^*)\Big]}{\mathbb{E}\Big[\mathbb{1}_{[Y_{\tau_t^*}>0]} \exp(-Y_{\tau_t^*}\tau_t^*)\Big]} \tag{3.9}$$

Afin de démontrer le Théorème 3.1.1, on s'intéresse au log du rapport $\frac{\mathbb{P}(X>u+t)}{\mathbb{P}(X>u)}$, et on va étudier le comportement de chacun des termes de l'égalité de droite dans (3.9).

$$\log \frac{\mathbb{P}(X > u+t)}{\mathbb{P}(X > u)} = -(\Lambda^*(u+t) - \Lambda^*(u)) + \log \frac{\mathbb{E}\Big[\mathbb{1}_{[Y_{\tau^*_{u+t}}>0]} \exp(-Y_{\tau^*_{u+t}}\tau^*_{u+t})\Big]}{\mathbb{E}\Big[\mathbb{1}_{[Y_{\tau^*_u}>0]} \exp(-Y_{\tau^*_u}\tau^*_u)\Big]} \quad (3.10)$$

En utilisant les deux lemmes suivants et l'expression (3.10), on retrouve l'approximation du Théorème 3.1.1. On donne une preuve pour ces deux lemmes.

Lemme 3.2.1.

$$\Lambda^*(u+t) - \Lambda^*(u) \sim \tau^*_u\, t \quad \textit{quand } u \to \infty. \quad (3.11)$$

Lemme 3.2.2.

$$\log \frac{\mathbb{E}\Big[\mathbb{1}_{[Y_{\tau^*_{u+t}}>0]} \exp(-Y_{\tau^*_{u+t}}\tau^*_{u+t})\Big]}{\mathbb{E}\Big[\mathbb{1}_{[Y_{\tau^*_u}>0]} \exp(-Y_{\tau^*_u}\tau^*_u)\Big]} = o(\tau^*_u) \quad (3.12)$$

On donne une preuve pour chacun de ces lemmes.

Preuve du Lemme 3.2.1. On cherche à connaître $\Lambda^*(u+t) - \Lambda^*(u)$. Le Théorème des accroissements finis donne

$$\Lambda^*(u+t) - \Lambda^*(u) = (\Lambda^*)'(\tilde{u})t \quad \text{, pour } \tilde{u} \in]u; u+t[\text{ et } t \text{ fixé.}$$

Or d'après la proposition B.2.4 (p.173) dans [Pha07], la transformée de Fenchel-Legendre a la propriété suivante :

$$\big(\Lambda^*\big)'(x) = \big(\Lambda'\big)^{-1}(x) \quad \text{, } \forall x \in \mathbb{R}.$$

Donc

$$\Lambda^*(u+t) - \Lambda^*(u) = (\Lambda')^{-1}(\tilde{u})t. \quad (3.13)$$

Montrons que $(\Lambda')^{-1}(\tilde{u})t \sim \tau^*_u t$ lorsque $u \to \infty$.

D'après l'équation point selle définie dans le Théorème (3.1.1), on a

$$(\Lambda')^{-1}(\tilde{u}) = \tau^*_{\tilde{u}} \quad (3.14)$$

or $\tilde{u} \in]u; u+t[$ donc

$$\begin{aligned} u &< \tilde{u} < u+t \\ 1 &< \frac{\tilde{u}}{u} < \frac{t}{u} + 1 \end{aligned}$$

le quotient $\tilde{u}/u$ tend alors vers 1 quand u tend vers l'infini. Donc $\tilde{u}/u = 1 + o(u)$, d'où $\tilde{u} \sim u$ lorsque $u \to \infty$. On a $(\Lambda')^{-1}(\tilde{u}) = \tau^*_{\tilde{u}}$, or on a montré que $\tilde{u} \sim u$ quand $u \to \infty$. $(\Lambda')^{-1}$ étant continue, il existe un $\varepsilon > 0$ tel que $\tau^*_{\tilde{u}} = \tau^*_u + \varepsilon$. D'où $\tau^*_{\tilde{u}}/\tau^*_u = 1 + o(\tau^*_u)$ et donc $\tau^*_{\tilde{u}} \sim \tau^*_u$ quand $u \to \infty$.

Finalement, on a montré que $\Lambda^*(u+t) - \Lambda^*(u) \sim \tau^*_u\, t$ quand $u \to \infty$. □

Preuve du lemme 3.2.2. On s'intéresse pour l'instant au dénominateur du quotient des espérances. On effectue le recentrage et la renormalisation suivants :

$$\mathbb{E}\Big[\mathbb{1}_{[Z_{\tau_u^*}>0]}\exp\Big(\sqrt{\Lambda''(\tau_u^*)}\left(-\frac{X_{\tau_u^*}-u}{\sqrt{\Lambda''(\tau_u^*)}}\right)\tau_u^*\Big)\Big] \tag{3.15}$$

où $Z_{\tau_u^*}$ est la variable aléatoire recentrée et renormalisée telle que :

$$Z_{\tau_u^*} = \frac{X_{\tau_u^*}-u}{\sqrt{\Lambda''(\tau_u^*)}} \tag{3.16}$$

On remarque que la fonction $\Lambda''(\tau_u^*)$, dérivée seconde de la fonction génératrice des cumulants, est une variance.

On veut montrer que $Z_{\tau_u^*}$ tend vers une variable aléatoire gaussienne centrée réduite. Ainsi, il restera à contrôler le terme $\tau_u^*\sqrt{\Lambda''(\tau_u^*)}$ dans l'expression de l'espérance. On cherche tout d'abord à calculer la log Laplace de $Z_{\tau_u^*}$.

On calcule la log Laplace de $Z_{\tau_u^*}$. Tout d'abord, la transformée de Laplace de la variable $Z_{\tau_u^*}$ s'exprime par :

$$\begin{aligned}\mathbb{E}\big[\exp(Z_{\tau_u^*}.\xi)\big] &= \mathbb{E}\big[\exp(\xi.\frac{X_{\tau_u^*}-u}{\sqrt{\Lambda''(\tau_u^*)}})\big] &(3.17)\\ &= \exp(-\frac{u.\xi}{\sqrt{\Lambda''(\tau_u^*)}}).\mathbb{E}\big[\exp(X_{\tau_u^*}\ \frac{\xi}{\sqrt{\Lambda''(\tau_u^*)}})\big] &(3.18)\end{aligned}$$

or l'espérance restante peut s'écrire :

$$\begin{aligned}\mathbb{E}\big[\exp(X_{\tau_u^*}\ \frac{\xi}{\sqrt{\Lambda''(\tau_u^*)}})\big] &= \int_{\mathbb{R}}\exp(\frac{x.\xi}{\sqrt{\Lambda''(\tau_u^*)}})\exp(\tau_u^*x-\Lambda(\tau_u^*))dF_X(x) &(3.19)\\ &= \exp(-\Lambda(\tau_u^*))\exp(\Lambda(\frac{\xi}{\sqrt{\Lambda''(\tau_u^*)}}+\tau_u^*)) &(3.20)\end{aligned}$$

En revenant à l'équation (3.17), et en passant au logarithme on obtient :

$$\log\mathbb{E}\big[\exp(Z_{\tau_u^*}.\xi)\big] = -\frac{u.\xi}{\sqrt{\Lambda''(\tau_u^*)}} - \Lambda(\tau_u^*) + \Lambda\,(\frac{\xi}{\sqrt{\Lambda''(\tau_u^*)}}+\tau_u^*) \tag{3.21}$$

On fait maintenant un développement limité à l'ordre 2 de l'accroissement $\Lambda(\frac{\xi}{\sqrt{\Lambda''(\tau_u^*)}}+\tau_u^*)$

$$\Lambda(\frac{\xi}{\sqrt{\Lambda''(\tau_u^*)}}+\tau_u^*) = \Lambda\,(\tau_u^*) + \Lambda'(\tau_u^*)\frac{\xi}{\sqrt{\Lambda''(\tau_u^*)}} + \Lambda''(\eta_u)\frac{\xi^2}{2\Lambda''(\tau_u^*)} + o(\frac{\xi^2}{\Lambda''(\tau_u^*)}) \tag{3.22}$$

où $\eta_u \in]\tau_u^*, \tau_u^* + \frac{\xi}{\sqrt{\Lambda_n''(\tau_u^*)}}[$.

$$\text{Donc,}\quad \Lambda(\frac{\xi}{\sqrt{\Lambda''(\tau_u^*)}}+\tau_u^*) \sim \Lambda(\tau_u^*) + \Lambda'(\tau_u^*)\frac{\xi}{\sqrt{\Lambda''(\tau_u^*)}} + \Lambda''(\eta_u)\frac{\xi^2}{2\Lambda''(\tau_u^*)} \tag{3.23}$$

D'après l'hypothèse de départ (équation point selle) selon laquelle il existe un τ_u^* tel que $\Lambda'(\tau_u^*) = u$, on a

$$\Lambda(\frac{\xi}{\sqrt{\Lambda''(\tau_u^*)}} + \tau_u^*) \sim \Lambda(\tau_u^*) + \frac{u\,\xi}{\sqrt{\Lambda''(\tau_u^*)}} + \Lambda''(\eta_u)\frac{\xi^2}{2\Lambda''(\tau_u^*)} \tag{3.24}$$

On revient à l'expression de la fonction log Laplace et il reste

$$\log \mathbb{E}\big[\exp(Z_{\tau_u^*}\,.\,\xi)\big] = \frac{\Lambda''(\eta_u)}{\Lambda''(\tau_u^*)}.\frac{\xi^2}{2} \tag{3.25}$$

Grâce à l'hypothèse (H_1) émise dans l'énoncé du théorème, on a la limite suivante :

$$\frac{\Lambda_n''(\eta_u)}{\Lambda_n''(\tau_u^*)} \longrightarrow 1 \quad \text{quand} \quad \tau_u^* \to \infty \tag{3.26}$$

Donc $\log \mathbb{E}\big[\exp(Z_{\tau_u^*}\,.\,\xi)\big] \to \frac{\xi^2}{2}$, c'est-à-dire que la variable $Z_{\tau_u^*}$ tend bien vers une gaussienne centrée réduite.

On revient à l'espérance (3.15) qui nous intéresse :

$$\mathbb{E}\Big[\mathbb{1}_{[Z_{\tau_u^*}>0]} \exp\Big(-Z_{\tau_u^*}\,\tau_u^*\sqrt{\Lambda''(\tau_u^*)}\Big)\Big] \tag{3.27}$$

On note $\alpha_u = \sqrt{\Lambda''(\tau_u^*)}\,\tau_u^*$. Soit f_Z la densité de probabilité de la variable aléatoire $Z_{\tau_u^*}$. On rappelle que l'on suppose que $f_Z \in L^2(\mathbb{R})$.

$$\begin{aligned}
\mathbb{E}\Big[\mathbb{1}_{[Z_{\tau_u^*}>0]} \exp\Big(-\alpha_u\, Z_{\tau_u^*}\Big)\Big] &= \int_{-\infty}^{+\infty} \exp(-\alpha_u Z_{\tau_u^*}) f_Z\, dF_{Z_{\tau_u^*}} \\
&= \frac{1}{\alpha_u}\int_{-\infty}^{+\infty} \alpha_u \exp(-\alpha_u Z_{\tau_u^*}) f_Z\, dF_{Z_{\tau_u^*}}
\end{aligned}$$

On a fait apparaître un produit d'une densité de loi exponentielle $\mathcal{E}(\alpha_u)$ et de la densité f_Z. Le Théorème de Parseval permet d'écrire

$$\begin{aligned}
\mathbb{E}\Big[\mathbb{1}_{[Z_{\tau_u^*}>0]} \exp\Big(-\alpha_u\, Z_{\tau_u^*}\Big)\Big] &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} \frac{1}{\alpha_u - i\omega}\,\varphi_Z(\omega)d\omega \\
&= \frac{1}{2\pi\alpha_u}\int_{-\infty}^{+\infty} \frac{1}{1 - i\omega/\alpha_u}\,\varphi_Z(\omega)d\omega
\end{aligned}$$

où $\varphi_Z(\omega)$ est la fonction caractéristique de la variable $Z_{\tau_u^*}$ définie par (3.16) et $1/(\alpha_u - i\omega)$ est la fonction caractéristique de la loi exponentielle de paramètre α_u. On calcule $\varphi_Z(\omega)$ définie pour tout $\omega \in \mathbb{R}$ par :

$$\begin{aligned}
\varphi_Z(\omega) &= \mathbb{E}[\exp(Z_{\tau_u^*}\, i\omega)] \\
&= \mathbb{E}\Big[\exp(\frac{X_{\tau_u^*} - u}{\sqrt{\Lambda''(\tau_u^*)}}\, i\omega)\Big] \\
&= \exp(\frac{-u}{\sqrt{\Lambda''(\tau_u^*)}}\, i\omega)\mathbb{E}\Big[X_{\tau_u^*}\,\frac{i\omega}{\sqrt{\Lambda''(\tau_u^*)}}\Big]
\end{aligned}$$

or $\mathbb{E}\Big[X_{\tau_u^*}\ \frac{i\omega}{\sqrt{\Lambda''(\tau_u^*)}}\Big]$ peut s'écrire

$$\begin{aligned}\mathbb{E}\Big[X_{\tau_u^*}\ \frac{i\omega}{\sqrt{\Lambda''(\tau_u^*)}}\Big] &= \int_{\mathbb{R}} \exp(\frac{x\ i\omega}{\sqrt{\Lambda''(\tau_u^*)}}) \exp(\tau_u^* x - \Lambda(\tau_u^*))\ dF_X(x) \\ &= \int_{\mathbb{R}} \exp\Big(x(\frac{i\omega}{\sqrt{\Lambda''(\tau_u^*)}} + \tau_u^*)\Big) \exp(-\Lambda(\tau_u^*))\ dF_X(x) \\ &= \exp(-\Lambda(\tau_u^*)) \exp\Big(\Lambda(\frac{i\omega}{\sqrt{\Lambda''(\tau_u^*)}} + \tau_u^*)\Big)\end{aligned}$$

En revenant à $\varphi_Z(\omega)$ on a :

$$\varphi_Z(\omega) = \exp\Big(\frac{-u}{\sqrt{\Lambda''(\tau_u^*)}}\ i\omega\Big) \exp\Big(-\Lambda(\tau_u^*) + \Lambda\,(\frac{i\omega}{\sqrt{\Lambda''(\tau_u^*)}} + \tau_u^*)\Big) \tag{3.28}$$

On fait un développement limité avec reste intégral, à l'ordre 2, de l'accroissement $(\frac{i\omega}{\sqrt{\Lambda''(\tau_u^*)}} + \tau_u^*)$. On va chercher à obtenir un résultat d'uniformité sur le reste. Ce développement s'écrit :

$$\Lambda(\frac{i\omega}{\sqrt{\Lambda''(\tau_u^*)}} + \tau_u^*) = \Lambda\,(\tau_u^*) + \Lambda'(\tau_u^*)\frac{i\omega}{\sqrt{\Lambda''(\tau_u^*)}} - \Lambda''(\tau_u^*)\frac{\omega^2}{2\Lambda''(\tau_u^*)} + R(\frac{i\omega}{\sqrt{\Lambda''(\tau_u^*)}} + \tau_u^*)$$

où le reste intégral $R(\frac{i\omega}{\sqrt{\Lambda''(\tau_u^*)}} + \tau_u^*)$ s'exprime par

$$R(\frac{i\omega}{\sqrt{\Lambda''(\tau_u^*)}} + \tau_u^*) = \int_{\tau_u^*}^{\tau_u^* + \frac{i\omega}{\sqrt{\Lambda''(\tau_u^*)}}} \frac{(\tau_u^* + \frac{i\omega}{\sqrt{\Lambda''(\tau_u^*)}} - t)^2}{2} \Lambda^{(3)}(t)dt$$

On réécrit cette intégrale en $t = \tau_u^* + i\theta$, $\theta > 0$.

$$\begin{aligned}\int_0^{\frac{\omega}{\sqrt{\Lambda''(\tau_u^*)}}} (\frac{\omega}{\sqrt{\Lambda''(\tau_u^*)}} - i\theta)^2 \Lambda^{(3)}(\tau_u^* + i\theta)d\theta &= -\int_0^{\frac{\omega}{\sqrt{\Lambda''(\tau_u^*)}}} (\frac{\omega}{\sqrt{\Lambda''(\tau_u^*)}} - \theta)^2 \Lambda^{(3)}(\tau_u^* + i\theta)d\theta \\ &= -\int_0^{\frac{\omega}{\sqrt{\Lambda''(\tau_u^*)}}} \xi^2 \Lambda^{(3)}(\tau_u^* + i(\frac{\omega}{\sqrt{\Lambda''(\tau_u^*)}} - \xi))d\xi\end{aligned}$$

avec le changement de variable $\xi = \frac{\omega}{\sqrt{\Lambda''(\tau_u^*)}} - \theta$. En revenant à l'expression de la fonction caractéristique (3.28) et en considérant l'hypothèse initiale selon laquelle il existe un τ_u^* tel que $\Lambda'(\tau_u^*) = u$ on obtient :

$$\varphi_Z(\omega) = \exp\Big(\frac{-u}{\sqrt{\Lambda''(\tau_u^*)}}\ i\omega\Big) \exp\Big(\frac{-u}{\sqrt{\Lambda''(\tau_u^*)}}\ i\omega - \frac{\omega^2}{2} - \int_0^{\frac{\omega}{\sqrt{\Lambda''(\tau_u^*)}}} \xi^2 \Lambda^{(3)}(\tau_u^* + i(\frac{\omega}{\sqrt{\Lambda''(\tau_u^*)}} - \xi))d\xi\Big)$$

En regroupant les exponentielles et en passant au module, on a

$$|\varphi_Z(\omega)| \leq |\exp\Big(-\frac{\omega^2}{2} - \int_0^{\frac{\omega}{\sqrt{\Lambda''(\tau_u^*)}}} \xi^2 Re(\Lambda^{(3)}(\tau_u^* + i(\frac{\omega}{\sqrt{\Lambda''(\tau_u^*)}} - \xi)))d\xi\Big)|$$

On rappelle l'hypothèse H_2 :
pour $\tau \in \mathbb{R}$ grand, on suppose que $\forall r > 0,\ Re(\Lambda^{(3)}(\tau + ir)) \geq 0$.

En utilisant cette hypothèse dans l'inégalité précédente, on a pour τ_u^* grand,

$$|\varphi_Z(\omega)| \leq \exp(-\frac{\omega^2}{2}) \quad \forall \omega.$$

Le membre de droite est intégrable et ne dépent pas de u. On avait avec le Théorème de Parseval,

$$\mathbb{E}\Big[\mathbb{1}_{[Z_{\tau_u^*}>0]} \exp\Big(-\alpha_u Z_{\tau_u^*}\Big)\Big] = \frac{1}{2\pi\alpha_u}\int_{-\infty}^{+\infty} \frac{1}{1 - i\omega/\alpha_u}\varphi_Z(\omega)d\omega \tag{3.29}$$

En appliquant le théorème de convergence dominé on a :

$$\begin{aligned}
\lim_{u\to\infty} \mathbb{E}\Big[\mathbb{1}_{[Z_{\tau_u^*}>0]} \exp\Big(-\alpha_u Z_{\tau_u^*}\Big)\Big] &= \lim_{u\to\infty} \frac{1}{2\pi\alpha_u}\int_{-\infty}^{+\infty} \frac{1}{1 - i\omega/\alpha_u}\varphi_Z(\omega)d\omega \\
&= \frac{1}{2\pi\alpha_u}\int_{-\infty}^{+\infty} 1 \lim_{u\to\infty} \varphi_Z(\omega)d\omega \\
&= \frac{1}{2\pi\alpha_u}\int_{-\infty}^{+\infty} \exp(-\frac{\omega^2}{2})d\omega = \frac{1}{2\pi\alpha_u}.\sqrt{2\pi} = \frac{1}{\sqrt{2\pi}\alpha_u}
\end{aligned}$$

d'où en passant au log,

$$\log \mathbb{E}\Big[\mathbb{1}_{[Z_{\tau_u^*}>0]} \exp\Big(-\alpha_u Z_{\tau_u^*}\Big)\Big] \longrightarrow \log(1/\sqrt{2\pi}\alpha_u) \quad , \quad \text{quand } u \to \infty. \tag{3.30}$$

On cherche à montrer que l'espérance considérée est un $o(\tau_u^*)$. On regarde donc la limite du quotient suivant :

$$\frac{1}{\tau_u^*}\frac{\log \mathbb{E}\Big[\mathbb{1}_{[Z_{\tau_u^*}>0]} \exp\Big(-\alpha_u Z_{\tau_u^*}\Big)\Big]}{\log(1/\sqrt{2}\pi\alpha_u)} = \frac{1}{\tau_u^*}\frac{\log \mathbb{E}\Big[\mathbb{1}_{[Z_{\tau_u^*}>0]} \exp\Big(-\sqrt{\Lambda''(\tau_u^*)}\,\tau_u^* Z_{\tau_u^*}\Big)\Big]}{\log(1/\sqrt{2\pi}\alpha_u)} \tag{3.31}$$

$$\text{A-t-on} \quad \lim_{u\to\infty} \frac{1}{\tau_u^*} \quad \frac{\log \mathbb{E}\Big[\mathbb{1}_{[Z_{\tau_u^*}>0]} \exp\Big(-\sqrt{\Lambda''(\tau_u^*)}\,\tau_u^* Z_{\tau_u^*}\Big)\Big]}{\log(1/\sqrt{2\pi}\alpha_u)} = 0 \qquad ? \tag{3.32}$$

$$\text{Oui si} \quad \lim_{u\to\infty} \frac{1}{\tau_u^*} \log(\sqrt{\Lambda''(\tau_u^*)}\,\tau_u^*) = 0 \tag{3.33}$$

L'hypothèse (H_3) selon laquelle pour $R \in \mathbb{R}$ on suppose que $\frac{1}{R}\log\Lambda''(R) \to 0$ quand $R \to \infty$, permet de conclure. Cette hypothèse n'est pas trop restrictive de par la présence du log. Cependant il est nécessaire de la prendre en compte car certaines lois peuvent ne pas la satisfaire. Un exemple particulier sera donné à la fin de cette partie.

Ainsi, on a montré que

$$\log \mathbb{E}\Big[\mathbb{1}_{[Y_{\tau_u^*}>0]} \exp(-Y_{\tau_u^*}\tau_u^*)\Big] = o(\tau_u^*) \tag{3.34}$$

Un calcul analogue permet d'obtenir le même résultat pour l'espérance du numérateur, à t fixé et ainsi obtenir l'approximation du Lemme 3.2.2 :

$$\log \frac{\mathbb{E}\left[\mathbb{1}_{[Y_{\tau^*_{u+t}}>0]} \exp(-Y_{\tau^*_{u+t}}\tau^*_{u+t})\right]}{\mathbb{E}\left[\mathbb{1}_{[Y_{\tau^*_u}>0]} \exp(-Y_{\tau^*_u}\tau^*_u)\right]} = o(\tau^*_u) \tag{3.35}$$

□

3.2.2 Contre exemple pour l'hypothèse (H_3)

On donne ici un contre exemple à l'hypothèse (H_3).
Soit W une variable aléatoire suivant une loi de Poisson de paramètre 1 et soit $N_1, \cdots, N_j$ une suite de variables aléatoires indépendantes et identiquement distribuées suivant une loi de Poisson de paramètre 1. On suppose que W est indépendante de la suite (N_j). Soit N la variable aléatoire définie par

$$N = \sum_{i=1}^{W} N_i \tag{3.36}$$

Sa transformée de Laplace est donnée par

$$\Lambda_N(t) = \exp(\exp(t) - 1) - 1\ , \quad t \in \mathbb{R}. \tag{3.37}$$

En effet,

$$\begin{aligned} \mathbb{E}[\exp(tW)] &= \mathbb{E}[\exp(t\sum_{i=1}^{W} N_i)] \\ &= \mathbb{E}\{\mathbb{E}[\exp(t\sum_{i=1}^{W} N_i)/W]\} \end{aligned}$$

de plus,

$$\begin{aligned} \mathbb{E}[\exp(t\sum_{i=1}^{W} N_i)/W] &= \mathbb{E}[\exp(tN_1)]^W \\ &= (\exp(e^t - 1))^W \end{aligned}$$

donc,

$$\begin{aligned} \mathbb{E}[\exp(tW)] &= \mathbb{E}[\exp(W(e^t - 1))] \\ &= \sum_{k\geq 0} \exp[k(e^t - 1)]\frac{e^{-1}}{k!} \\ &= e^{-1}\exp(\exp(e^t - 1)) \end{aligned}$$

En passant au log on trouve l'expression :

$$\Lambda_N(t) = \exp(\exp(t) - 1) - 1 \tag{3.38}$$

On a ensuite,

$$\begin{aligned}\Lambda'_N(t) &= \exp(t)\exp(\exp(t)-1)\\ \Lambda''_N(t) &= (\exp(t)+\exp(2t))\exp(\exp(t)-1)\end{aligned}$$

Si on regarde la limite de $\frac{1}{t}\log\Lambda''_N(t)$, on a pas la limite supposée dans l'hypothèse (H_3) : $\frac{1}{R}\log\Lambda''(R) \to 0$ quand $R \to \infty$.

3.2.3 Exemple du cas gaussien

Dans le cas gaussien les hypothèses sont vérifiées. Soit X et X^* deux variables aléatoires réelles distribuées respectivement suivant $F_X \sim N(0,1)$ et $F_{X^*} \sim N(u,1)$. on s'intéresse au rapport $\frac{\mathbb{P}(X>u+t)}{\mathbb{P}(X>u)}$. Premièrement, on a

$$\begin{aligned}\mathbb{P}(X>u) &= \int_{\{x>u\}} dF_X(x)\\ &= \int_{\{x>u\}} \exp(tx-\frac{t^2}{2})\exp(-tx+\frac{t^2}{2})dF_X(x)\\ &= \int_0^{+\infty} \exp(-tx+\frac{t^2}{2})dF_{X^*}(x)\\ &= \int_0^{+\infty} \exp(-tx+\frac{t^2}{2})dF_{X^*}(x)\\ &= \exp(\frac{t^2}{2})\int_0^{+\infty}\exp(-tx)dF_{X^*}(x) \quad \text{, on pose} \quad t=x-u,\end{aligned}$$

$$\begin{aligned}\mathbb{P}(X>u) &= \exp(\frac{t^2}{2})\int_0^{+\infty}\exp(-t(u+t))dF_X(x)\\ &= \exp(-\frac{t^2}{2})\int_0^{+\infty}\exp(-tu)dF_X(x)\end{aligned}$$

On fait le même calcul pour $\mathbb{P}(X>u+t)$ et on obtient le rapport suivant :

$$\frac{\mathbb{P}(X>u+t)}{\mathbb{P}(X>u)} = \exp(-\frac{1}{2}(t+u)^2-\frac{u^2}{2})\ \frac{\int_0^{+\infty}\exp(-t(u+t))dF_X(x)}{\int_0^{+\infty}\exp(-tu)dF_X(x)} \tag{3.39}$$

et ainsi,

$$\log\frac{\mathbb{P}(X>u+t)}{\mathbb{P}(X>u)} = (-\frac{1}{2}(t+u)^2-\frac{u^2}{2}) + \log\ \frac{\int_0^{+\infty}\exp(-t(u+t))dF_X(x)}{\int_0^{+\infty}\exp(-tu)dF_X(x)} \tag{3.40}$$

Pout t fixé, on a bien l'approximation

$$(-\frac{1}{2}(t+u)^2-\frac{u^2}{2}) \sim -ut, \quad \text{quand } u \longrightarrow \infty \tag{3.41}$$

car $(-\frac{1}{2}(t+u)^2 - \frac{u^2}{2}) = -\frac{1}{2}t^2 - ut$ et en divisant par ut on a bien une limite qui vaut -1. Reste à montrer que le log du quotient des espérances tend vers 0, ou que le quotient des espérances tend vers 1.

$$\begin{aligned} E_1 &= \int_0^{+\infty} \exp(-t(u+t)) dF_X(x) \\ &= \exp(-t^2) \int_0^{+\infty} \exp(-tu) \frac{\exp(-u^2/2)}{\sqrt{2\pi}} du \end{aligned}$$

On souhaite se ramener à une seule gaussienne :

$$E_1 = \exp(-t^2) \int_0^{+\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(u+t)^2) \exp(\frac{1}{2}t^2) du$$

Soit $V \sim N(0,1)$ telle que $v = u + t$. On a

$$E_1 = \exp(-\frac{t^2}{2}) \mathbb{P}(v > t) \tag{3.42}$$

De même,

$$\begin{aligned} E_2 &= \int_0^{+\infty} \exp(-tu) \frac{\exp(-u^2/2)}{\sqrt{2\pi}} du \\ &= \exp(\frac{t^2}{2}) \mathbb{P}(v > t) \end{aligned}$$

$\log \frac{E_1}{E_2} = -t^2$ à t fixé, ne dépend pas de u et apparaît comme borné en terme de u. Dans le cas gaussien, on a

$$\Lambda(x) = \frac{x^2}{2} \quad \text{et} \quad \Lambda'(x) = u$$

Donc avec l'équation point selle, on a $\tau_u^* = u$. On cherche à montrer que le log du quotient des espérance est un $o(\tau_u^*)$ ou de manière équivalente un $o(u)$. Grâce aux résultats précédents, on a bien

$$\frac{\log(E_1/E_2)}{u} \to 0, \quad \text{quand } u \to \infty. \tag{3.43}$$

Ainsi, l'approximation donnée par le Théorème 3.1.1 est vérifiée dans le cas gaussien.

3.3 Construction de l'estimateur d'un point de vue statistique

On construit un estimateur empirique d'un quantile à partir du Théorème 3.1.1 .

$$\log \frac{\mathbb{P}(X > u + t)}{\mathbb{P}(X > u)} \sim -\tau_u^* . t \quad \text{lorsque } u \to +\infty \text{ et } t \text{ fixé.}$$

On cherche un T_α (valant $u + t$) tel que $\mathbb{P}(X > T_\alpha) = \alpha$.
On a :

$$\log \alpha = -\tau_u^*(T_\alpha - u) + \log \mathbb{P}(X > u)$$

On choisit d'estimer $\mathbb{P}(X > u)$ par la loi empirique, comme dans l'approche POT, et on va travailler sur les statistiques d'ordre associées à l'échantillon de taille n.
On rappelle que l'on note $X_{1,n}, \cdots, X_{n,n}$ les statistiques d'ordre associées à l'échantillon $X_1, \cdots, X_n$. C'est-à-dire que l'on classe $X_1, \cdots, X_n$ par ordre croissant de sorte que :

$$X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n,n}$$

La k-ième plus grande valeur de l'échantillon est notée $X_{n-k,n}$.

Soit $u = X_{j,n}$ pour $j > J_0$, un entier à déterminer. On a alors

$$\log \alpha = -\tau^*_{X_{j,n}}(T_\alpha - X_{j,n}) + \log(1 - \frac{j}{n})$$

On peut voir cette équation comme un modèle de régression où T_α serait le régresseur :

$$-(\log \alpha - \log(1 - \frac{j}{n})) + \tau^*_{X_{j,n}} X_{j,n} = \tau^*_{X_{j,n}} T_\alpha$$

L'estimateur des moindres carrés $\widehat{T}_\alpha$ de T_α s'écrit alors :

$$\widehat{T}_\alpha = \frac{\sum_{j=J_0}^{n-1} \tau^*_{X_{j,n}} \left[\tau^*_{X_{j,n}} . X_{j,n} + \log(1 - \frac{1}{n}) - \log \alpha\right]}{\sum_{j=J_0}^{n-1} (\tau^*_{X_{j,n}})^2} \tag{3.44}$$

Le paramètre $\tau^*_{X_{j,n}}$ (ou u) permet de faire le bon changement de loi de la section précédente, cependant il est inconnu. On le détermine de façon empirique.

$$\text{on a} \quad \Lambda'(\tau^*_{X_{j,n}}) = X_{j,n} \quad J_0 < j < n - 1 \tag{3.45}$$

avec J_0 rang de statistique d'ordre correspondant à une valeur seuil à déterminer. On utilise la version empirique de la fonction génératrice des cumulants :

$$\widehat{\Lambda}_n(\tau) = \log \frac{1}{n} \sum_{i=1}^{n} \exp(\tau X_{j,n}) \tag{3.46}$$

Remarque :
1) résoudre l'équation (3.45) revient à maximiser la fonction

$$\sup_\tau \left(\tau X_{j,n} - \log \frac{1}{n} \sum_{i=1}^{n} \exp(\tau X_{j,n})\right) \tag{3.47}$$

2) la fonction Λ' est croissante.

On veut résoudre l'équation $\Lambda'(\tau^*_{X_{j,n}}) - X_{j,n} = 0$. On utilise la méthode de Newton-Raphson pour sa résolution numérique. A partir de la version empirique Λ_n de Λ, on exprime Λ'_n et Λ''_n :

$$\widehat{\Lambda}'_n(\tau) = \frac{\sum_{i=1}^{n-1} X_{i,n} \exp(\tau X_{i,n})}{\sum_{i=1}^{n-1} \exp(\tau X_{i,n})} \tag{3.48}$$

$$\widehat{\Lambda}''_n(\tau) = \frac{\sum_{i=1}^{n-1} X_i^2 \exp(\tau X_{i,n})}{\sum_{i=1}^{n-1} \exp(\tau X_{i,n})} - (\widehat{\Lambda}'_n(\tau))^2 \tag{3.49}$$

3.4 Simulations et résultats

On désire évaluer les performances de l'estimateur construit dans la section précédente. Pour cela, on décide d'estimer des quantiles dont les valeurs sont connues pour plusieurs niveaux de probabilité que l'on fixe à $p = 1 - 10^{-3}$, $p = 1 - 10^{-5}$ et $p = 1 - 10^{-7}$.

On s'intéresse dans un premier temps à une loi normale centrée réduite pour laquelle on connaît les fonctions $\Lambda(\tau)$, $\Lambda'(\tau)$ et $\Lambda''(\tau)$. En effet, dans ce cas, $\Lambda(\tau) = \tau^2/2$, $\Lambda'(\tau) = \tau$ et $\Lambda''(\tau) = 1$. On génère un échantillon de 10 000 points selon une $\mathcal{N}(0, 1)$.

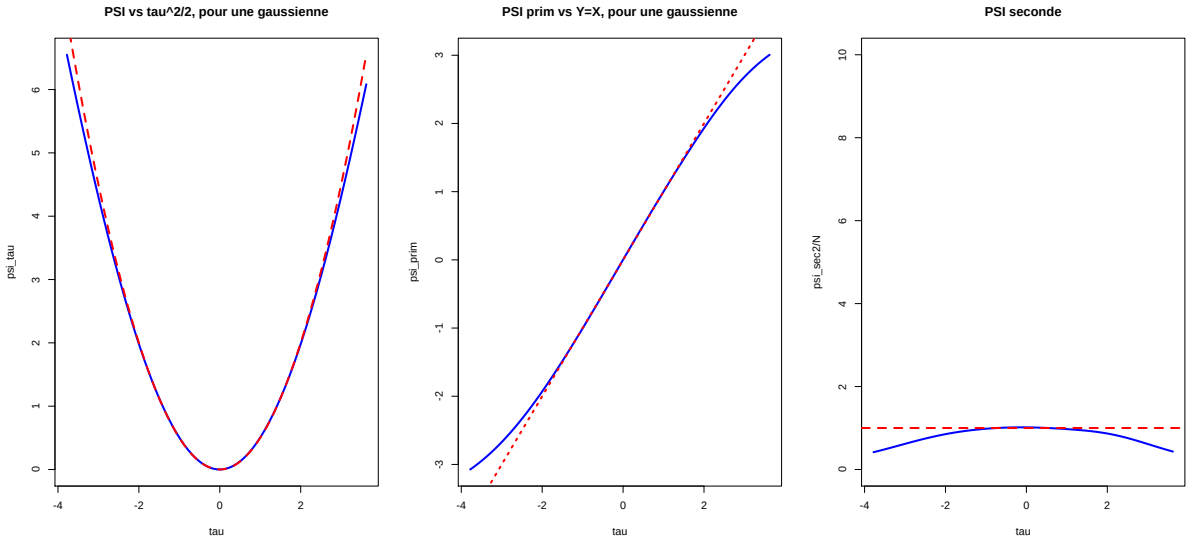


FIGURE 3.1 – Estimation des fonctions Λ, Λ' et Λ''

La figure 3.1 montre que les estimations sont bonnes dans les parties centrales mais pas sur les bords. La mauvaise estimation sur les bords est plus flagrante sur des échantillons de plus petite taille.
On applique un algorithme de Newton-Raphson pour déterminer le τ_t^* solution de l'équation point selle pour chaque $X_{j,n}$. La figure 3.2 montre les fonctions $\widehat{\Lambda}_n(\tau_t^*)$, $\widehat{\Lambda}'_n(\tau_t^*)$ et $\widehat{\Lambda}''_n(\tau_t^*)$ estimées avec les τ_t^* obtenus numériquement. On observe que la sensibilité aux bords s'est accentuée, et que la fonction $\Lambda''(\tau_t^*)$ n'est pas bien estimée.
Une fois la suite des τ_t^* calculée, on peut alors estimer le quantile T_α. Au vu des résultats des estimations de Λ, Λ' et Λ'', il apparaît raisonnable de conserver uniquement les valeurs centrales des τ_t^*. Cependant, afin d'observer le comportement global de l'estimateur, on évalue T_α pour une suite d'entiers J_0 balayant la quasi totalité de l'échantillon. J_0 correspond à un rang de statistique d'ordre pour laquelle on a une proportion p de valeurs au-dessus de $X_{J_0,n}$. On balaye une plage de proportions allant de 0.1 à 0.95 avec un pas de 0.5, correspondant à autant de valeurs de J_0. Les valeurs théoriques des quantiles de niveaux $p = 1 - 10^{-3}$, $p = 1 - 10^{-5}$ et $p = 1 - 10^{-7}$ sont respectivement 3.09, 4.26 et 5.19 pour une loi gaussienne centrée réduite.

Pour les trois estimations représentées sur la figure 3.3, on observe le même phénomène. L'estimation n'est pas bonne lorsqu'il y a soit, trop d'observations au dessus de J_0, soit pas assez. Contrairement à la méthode POT où l'on doit trouver un seuil assez grand

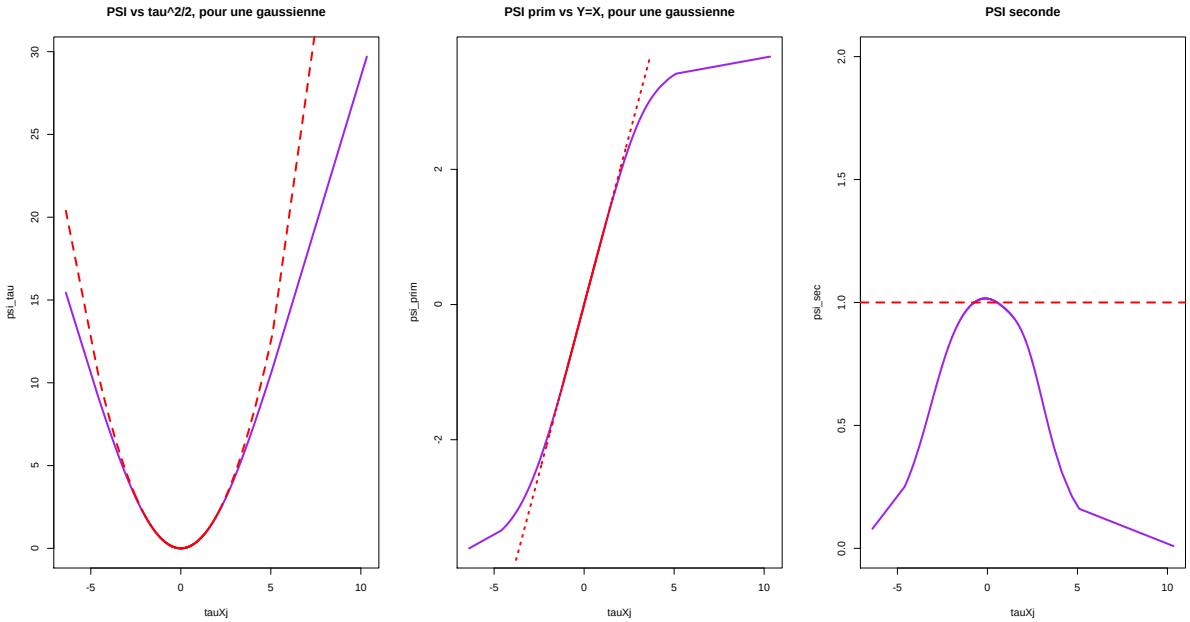


FIGURE 3.2 – Estimation des fonctions Λ_n, Λ'_n et Λ''_n avec les τ_t^*

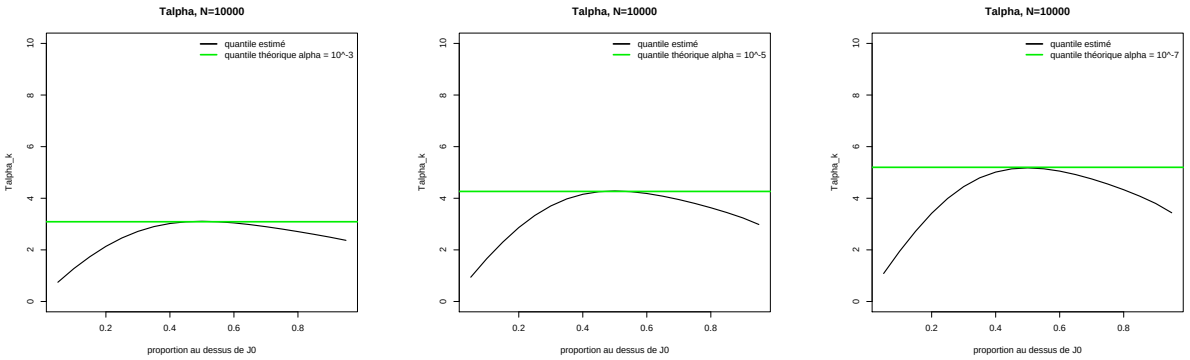


FIGURE 3.3 – Estimation pour une loi $\mathcal{N}(0,1)$

pour satisfaire les conditions asymptotiques du modèle, il semble ici que les proportions correspondant aux seuils pour lesquels l'estimation est la meilleure se situent autour de la médiane de l'échantillon.

On évalue maintenant l'estimateur sur un échantillon distribué suivant une loi de Rayleigh de paramètre $\rho = 2$ pour les trois niveaux de probabilité précédents (figure 3.4). Les valeurs théoriques des quantiles à estimer sont respectivement 7.43, 9.59 et 11.35.
Quel que soit le niveau de probabilité testé, l'estimateur présente le même comportement que pour une loi symétrique.

Le tableau (3.1) regroupe les résultats obtenus pour trois tirages de 10 000 observations distribuées selon trois lois paramétriques : la loi normale centrée réduite, la loi exponentielle de paramètre $\lambda = 1$, et la loi de Rayleigh de paramètre $\rho = 2$. On a déterminé visuellement la proportion adéquate menant à une estimation correcte.

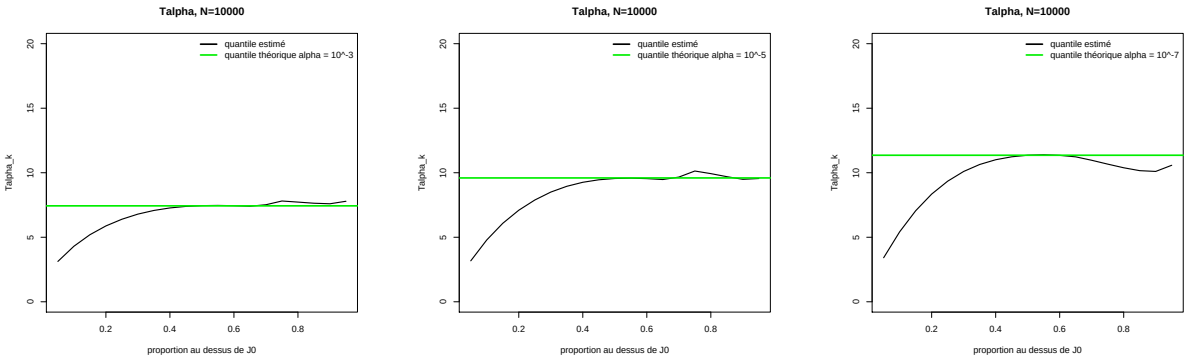


FIGURE 3.4 – Estimation pour une loi de Rayleigh de paramètre $\rho = 2$

Niveaux de probabilité	$\mathcal{N}(0,1)$	$\mathcal{E}(1)$	$R(2)$
$p = 1 - 10^{-3}$	$q_{th} = 6.90\,;\, \widehat{T} = 7.00$	$q_{th} = 3.09\,;\, \widehat{T} = 3.11$	$q_{th} = 7.43\,;\, \widehat{T} = 7.32$
$p = 1 - 10^{-5}$	$q_{th} = 11.31\,;\, \widehat{T} = 11.35$	$q_{th} = 4.26\,;\, \widehat{T} = 4.21$	$q_{th} = 9.60\,;\, \widehat{T} = 9.70$
$p = 1 - 10^{-7}$	$q_{th} = 15.87\,;\, \widehat{T} = 15.82$	$q_{th} = 5.19\,;\, \widehat{T} = 5.13$	$q_{th} = 11.35\,;\, \widehat{T} = 11.52$

TABLE 3.1 – Récapitulatif de l'estimateur $\widehat{T}_\alpha$

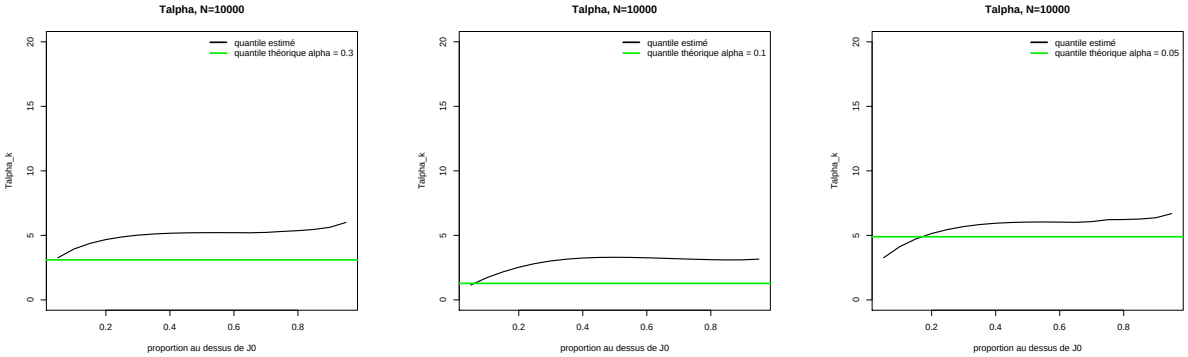


FIGURE 3.5 – Estimation du quantile de niveau $p = 1 - 0.3$, $p = 1 - 0.1$ et $p = 1 - 0.05$

Malgré l'instabilité de l'estimateur observée sur les figures précédentes et la mauvaise estimation de la fonction Λ'', le tableau (3.1) montre des résultats tout à fait satisfaisants pour les trois lois testées. Cependant lorsque l'on regarde les niveaux de probabilité (moins élevés) $p = 1 - 0.3$, $p = 1 - 0.1$ et $p = 1 - 0.05$, on remarque que les estimations sont plus éloignées des valeurs théoriques. En effet, comme le montre la figure (3.5), où est illustré le cas de la loi N(0,1), on surestime la valeur théorique du quantile pour tous les J_0. On a pu observer cela sur les autres lois et pour des niveaux de probabilité encore plus faibles. Le modèle est bon uniquement dans les queues de distributions.

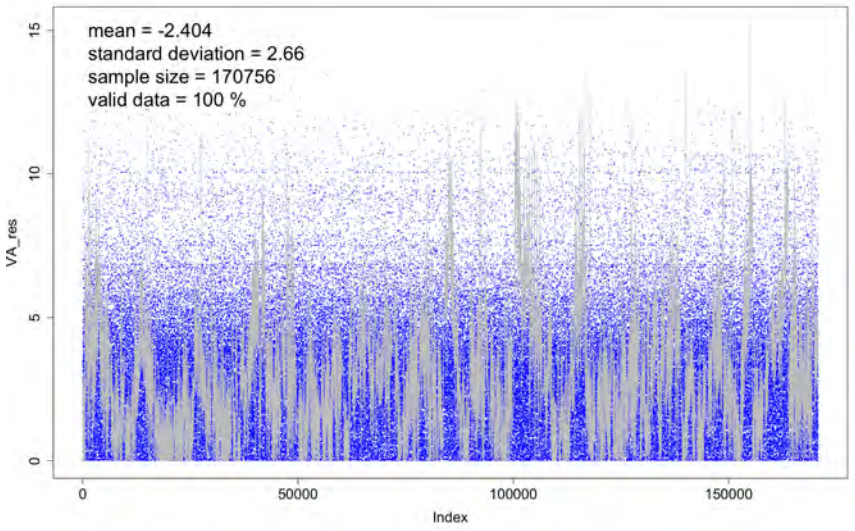
3.5 Conclusion

Nous avons formulé une approximation de la queue d'une distribution à partir de méthodes issues de techniques d'analyse asymptotique. D'un point de vue général, la méthode du point selle présente de bonnes propriétés d'approximation et de complexité calculatoire. Cependant, pour être utilisé dans le modèle d'intégrité de cette étude, l'estimateur construit dans ce chapitre ne paraît pas suffisamment stable et il nécessite une étape de seuillage pour estimer le quantile recherché.
On peut imaginer une procédure de choix de seuil identique à celle qui est présentée dans le dernier chapitre pour la loi des extrêmes. Cet algorithme détecte automatiquement des plages de stabilité au sein des estimateurs qui sont fonctions d'un certain seuil. Un seuil optimal est ensuite calculé selon un critère d'optimisation où sont comparées la loi empirique des observations et la loi estimée par le modèle. Cette comparaison se fait à des niveaux de probabilité pour lesquels le modèle empirique et le modèle d'extrême sont efficaces. Les dernières simulations présentées dans ce chapitre ont montré que l'estimateur construit par approximation point selle n'était bon que dans les queues de distribution. On ne peut donc pas appliquer le même heuristique que la procédure des modèles d'extrêmes.
Bien que peu coûteux pour les logiciels de simulation, l'algorithme de Newton-Raphson utilisé pour la résolution numérique de l'équation du point selle, peut s'avérer être mal adapté au traitement de très grands échantillons auxquels la société HELILEO est confrontée. De plus, il se peut que cette solution de résolution numérique ne fournisse pas de solution explicite au problème et mène donc à un échec de la procédure d'estimation. Pour ces raisons, nous avons privilégié l'estimation de quantile extrême par la méthode POT, présentée dans le chapitre 2.

CHAPITRE 4

Théorème d'indépendance asymptotique

Résumé : Les données de positionnement GPS auxquelles on s'intéresse dans cette étude présentent un fort caractère de dépendance temporelle. Cette dépendance temporelle est principalement due à la récursivité des algorithmes d'estimation de la position qui sont implémentés au sein des récepteurs. Les outils statistiques d'estimation de quantiles extrêmes présentés dans les chapitres précédents sont valables pour des variables aléatoires indépendantes et identiquement distribuées ou présentant une faible dépendance. Nous voulons casser la structure de dépendance temporelle existant au sein des données GPS en appliquant une permutation aléatoire sur les indices temporels de l'échantillon observé. On donne dans ce chapitre une justification théorique de cet heuristique ainsi qu'un théorème d'indépendance asymptotique atteinte après permutation. L'énoncé de ce théorème ainsi que sa preuve y sont donnés.



Sommaire

4.1 Introduction : dépendance au sein des données GPS

La première manifestation de dépendance temporelle peut être observée sur les nuages de points enregistrés par certains récepteurs. En effet on note sur la figure (4.1) suivante la présence de 'bras' en périphérie du nuage. Ce phénomène est typique de la forte corrélation temporelle présente au sein des données en sortie de récepteur. Ces écarts sont dus à la récursivité de l'algorithme d'estimation de la position implémenté dans les récepteurs GPS. A un instant donné, si le récepteur mesure une position éloignée du centre du nuage de points, il faudra attendre quelques secondes (la fréquence d'enregistrement la plus courante est de 1Hz) pour que le récepteur se recentre sur la partie la plus concentrée du nuage.

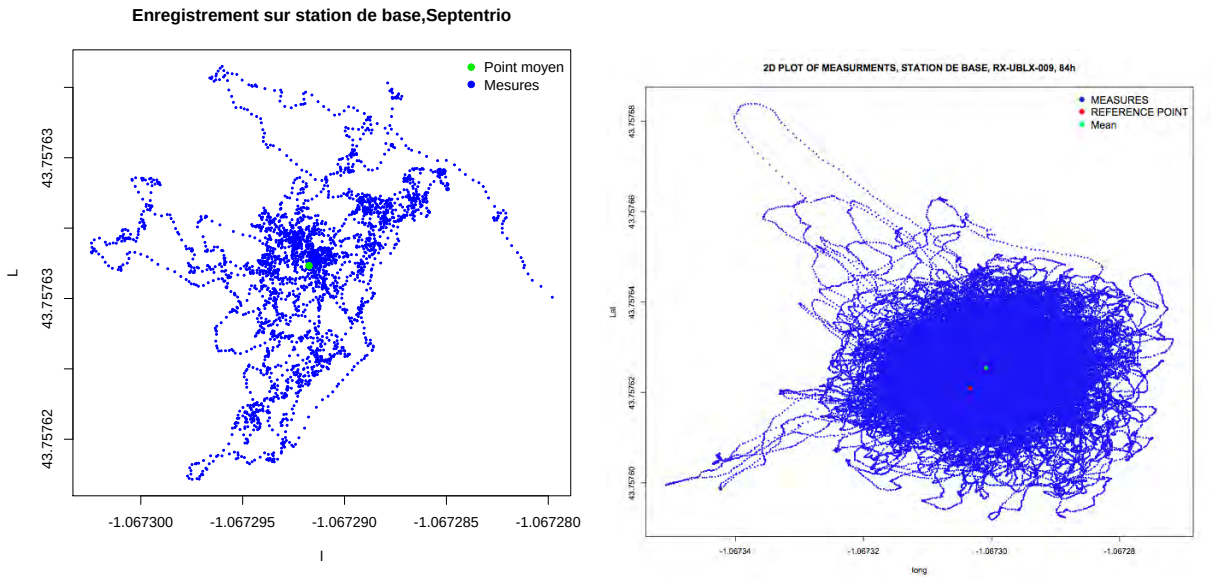


FIGURE 4.1 – Nuages de points caractéristiques

Lorsqu'on observe des enregistrements de taille importante sur une antenne fixe, on peut être confronté à la présence de cycles au sein des données. En effet, la période de révolution de la constellation de satellites GPS est de 12 heures. Sur des enregistrements de plus de 12 heures, on voit alors apparaître une redondance d'information apportée par le

passage successif des mêmes satellites.
Cependant, l'analyse des erreurs de positionnement (horizontale et verticale) ainsi que des Protection Levels, provenant d'enregistrements statiques de tailles importantes, révèle la présence de cycles de 24h au sein des données. En effet, l'estimation du périodogramme des différentes séries nous a permis de trouver des cycles d'une durée de 86785 secondes, soit 24,1 heures (figure 4.2). Ces cycles peuvent aussi être vus sur le corrélogramme des séries comme le montre la figure 4.3 .

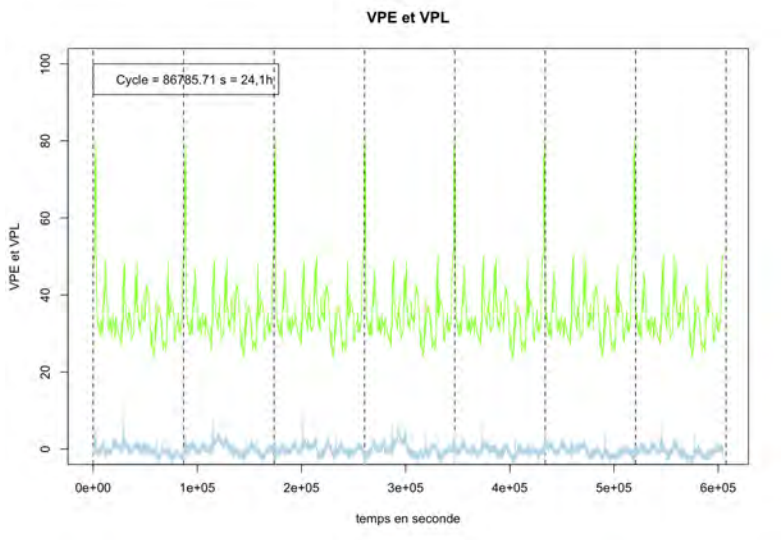


FIGURE 4.2 – Présence de cycles dans VPE et VPL

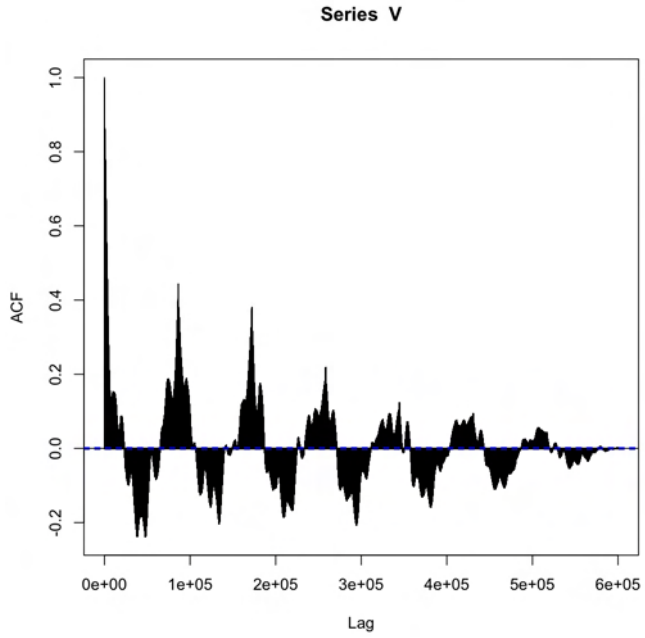


FIGURE 4.3 – Fonction d'autocorrélation de l'erreur verticale

L'objectif de la suite de ce chapitre est de montrer que l'on peut briser cette structure de dépendance présente au sein des données à l'aide d'une permutation aléatoire sur les indices temporels des processus considérés. Une justification théorique de cet heuristique est établie dans les paragraphes qui suivent. Cette preuve nécessite l'utilisation de la fonction caractéristique d'une variable aléatoire, dont on rappelle la définition et certaines propriétés.

4.2 Rappels sur les fonctions caractéristiques et leurs propriétés

Définition : On appelle fonction caractéristique d'une variable aléatoire réelle X, la fonction ψ_X de $\mathbb{R}$ dans $\mathbb{C}$ définie par :

$$\psi_X(u) = \mathbb{E}[e^{iuX}] = \int_{\mathbb{R}} e^{iux} f_X(dx) \tag{4.1}$$

où f_X est la loi de probabilité de la variable aléatoire X. La fonction caractéristique de X ne dépend que de la loi de X. On notera qu'elle correspond à la transformée de Fourier de cette loi.
La transformée de Fourier s'étend au cas multidimensionnel. Si on s'intéresse à des variables aléatoires vectorielles, c'est-à-dire si X est une variable aléatoire à valeur dans $\mathbb{R}^d$, sa fonction caractéristique est la fonction de $\mathbb{R}^d$ dans $\mathbb{C}$ définie par :

$$\psi_X(u) = \mathbb{E}[e^{i<u,X>}] = \int_{\mathbb{R}} e^{i<u,x>} f_X(dx) \tag{4.2}$$

pour tout $u = (u_1, \cdots, u_d) \in \mathbb{R}^d$, où $< u, x >= \sum_{i=1}^{d} u_i x_i$ est le produit scalaire usuel sur $\mathbb{R}^d$.

Propriété : caractérisation de l'indépendance. Une famille $X_1, \cdots, X_d$ de variables aléatoires réelles est indépendante si, et seulement si, pour tout $u = (u_1, \cdots, u_d) \in \mathbb{R}^d$:

$$\psi_{(X_1,\cdots,X_d)}(u_1, \cdots, u_d) = \prod_{j=1}^{d} \psi_{X_j}(u_j). \tag{4.3}$$

4.3 Condition de mélange et énoncé du théorème

Condition de mélange (H) :
Soit $(X_t)_{t\in\mathbb{Z}}$ un processus strictement stationnaire et centré, vérifiant la condition de mélange suivante :

Pour $r \geq 2$ **fixé** et $(t_1^N), (t_2^N), \cdots, (t_r^N) \in \mathbb{Z}$ avec

$$\lim_{N\to\infty} |t_i^N - t_j^N| = +\infty \quad , i \neq j,\ i = 1 \cdots r,\ j = 1 \cdots r$$

Alors,

$$\lim_{N\to\infty} \mathbb{E}\left[\exp\left(i(u_1 X_{t_1^N} + u_2 X_{t_2^N} + \cdots + u_r X_{t_r^N}) \right) \right] = \psi(u_1)\psi(u_2)\cdots\psi(u_r) \tag{4.4}$$

où $\psi(u)$ est la fonction caractéristique marginale du processus X_t définie pour tout $u \in \mathbb{R}$ par :

$$\psi(u) = \mathbb{E}[e^{iuX_t}] \tag{4.5}$$

Cette condition de mélange signifie que plus l'échantillon est grand, si on prend des éléments de l'échantillon initial et qu'on les écarte, alors on tend vers l'indépendance.

Théorème 4.3.1. *On pose $(Y_t^N)_{t=1\cdots N}$ tel que $Y_t^N = X_{\sigma_N(t)}$ où $\sigma_N : t \to \sigma_N(t) \in \{1, \cdots, N\}$ est une permutation aléatoire sur $1, \cdots, N$ de loi uniforme sur Σ_N, ensemble des permutations sur $\{1, \cdots, N\}$.*
Pour $r \geq 2$, on a la convergence en loi suivante quand $N \to \infty$:

$$(Y_{t_1}^N, \cdots, Y_{t_r}^N) \longrightarrow \Big(\mathcal{L}(X_t)\Big)^{\otimes r} \quad \textit{où } t_1, \cdots, t_r \textit{ sont } \textbf{\textit{fixés}} \tag{4.6}$$

La preuve de ce théorème s'effectue essentiellement en manipulant des notions de permutations et de suites. Pour plus de compréhension, nous allons faire la preuve en deux temps. On traitera d'abord un cas simple pour lequel on ne prend en compte que deux points de l'échantillon. On traitera le cas général par la suite.

4.4 Preuve pour $r = 2$

On s'intéresse à la fonction caractéristique du couple $Y_{t_1}^N, Y_{t_2}^N$ donnée par :

$$\begin{aligned}
\psi_{Y_{t_1}^N, Y_{t_2}^N}(u_1, u_2) &= \mathbb{E}\Big[\exp\Big(i(u_1 Y_{t_1}^N + u_2 Y_{t_2}^N)\Big)\Big] \\
&= \mathbb{E}\Big\{\mathbb{E}\Big[\exp\Big(i(u_1 Y_{t_1}^N + u_2 Y_{t_2}^N)\Big) \mid \sigma_N\Big]\Big\} \\
&= \frac{1}{N!}\sum_{\sigma_N \in \Sigma_N} \mathbb{E}\Big[\exp\Big(i(u_1 X_{\sigma(t_1)} + u_2 X_{\sigma(t_2)})\Big)\Big]
\end{aligned}$$

Soit Σ_N^k l'ensemble des permutations telles que deux indices t_1 et t_2 du processus soient espacés de k positions après permutation. k est non nul et prend donc ses valeurs dans l'ensemble $\big\{-(N-1), \cdots, (N-1)\big\}/\big\{0\big\}$. On a

$$\Sigma_N = \bigcup_k \Sigma_N^k \tag{4.7}$$

et Σ_N^k s'écrit

$$\Sigma_N^k = \big\{\sigma_N \in \Sigma_N : \sigma(t_2) = \sigma(t_1) + k\big\} \tag{4.8}$$

On note $l_{N,k} = \#\ \Sigma_N^k$ le cardinal de l'ensemble des permutations Σ_N^k .

On revient à l'expression de la fonction caractéristique et on a donc :

$$\psi_{Y_{t_1}^N, Y_{t_2}^N}(u_1, u_2) = \frac{1}{N!} \sum_{k=-(N-1), k\neq 0}^{N-1} l_{N,k}\ \mathbb{E}\Big[\exp\Big(i(u_1 X_1 + u_2 X_{1+k})\Big)\Big] \tag{4.9}$$

Il faut alors calculer le cardinal $l_{N,k}$ de l'ensemble des permutations Σ_N^k.
On considère les deux cas possibles $k > 0$ et $k < 0$.

1. $k > 0$: on a $\sigma(t_1) = \sigma(t_2) + k$ donc $\sigma(t_2) \in \big\{1, 2, \cdots, (N-k)\big\}$. Il y a $(N-k)(N-2)!$ permutations possibles dans ce cas.
2. $k < 0$: on a $\sigma(t_2) = \sigma(t_1) - k$ donc $\sigma(t_1) \in \big\{1, 2, \cdots, (N+k)\big\}$. Il y a $(N+k)(N-2)!$ permutations possibles dans ce cas.

Finalement, on peut écrire :

Proposition 4.4.1.

$$l_{N,k} = (N - |k|)(N-2)! \tag{4.10}$$

On a montré que $l_{N,k}$ ne dépendait pas de t_1 et t_2.

On revient à la fonction caractéristique et on a maintenant,

$$\begin{aligned} \psi_{Y_{t_1}^N, Y_{t_2}^N}(u_1, u_2) &= \frac{1}{N!} \sum_{k=-(N-1), k\neq 0}^{N-1} l_{N,k}\ \mathbb{E}\Big[\exp\Big(i(u_1 X_1 + u_2 X_{1+k})\Big)\Big] \\ &= \frac{1}{N-1} \sum_{k=-(N-1), k\neq 0}^{N-1} (1 - \frac{|k|}{N})\ \psi_{X_1, X_{1+k}}(u_1, u_2) \end{aligned}$$

Lemme 4.4.1. *Soit une suite $(w_k)_{k\in\mathbb{Z}}$, on suppose que*

$$\lim_{|k|\to\infty} w_k = a\ ,\ a \in \mathbb{R} \tag{4.11}$$

Alors on a,

$$\lim_{N\to\infty} \frac{1}{N-1} \sum_{k=-(N-1), k\neq 0}^{N-1} (1 - \frac{|k|}{N})\ w_k = a \tag{4.12}$$

Ici, w_k apparaît comme la fonction caractéristique $\psi_{X_1, X_{1+k}}(u_1, u_2)$ à u_1 et u_2 fixés. Or si (X_t) vérifie la condition de mélange (H) alors $\psi_{X_1, X_{1+k}}(u_1, u_2) \to \psi(u_1)\psi(u_1)$ quand

$k \to \infty$.

Ainsi on va montrer à l'aide des fonctions caractéristiques que la loi de (X_t) sachant les permutations uniformes sur les indices est la même que la loi de (X_t), si (X_t) vérifie la condition de mélange (H). Autrement dit, on va montrer que $\psi_{Y^N_{t_1}, Y^N_{t_2}}(u_1, u_2)$ et $\psi_{X_1, X_{1+k}}(u_1, u_2)$ ont la même limite. On donne la démonstration du lemme 4.4.1 qui permettra de conclure.

Démonstration. On coupe la somme en deux, et il suffit donc de montrer que :

$$\lim_{N\to\infty} \frac{1}{N-1} \sum_{k=1}^{N-1} (1 - \frac{|k|}{N})\, w_k = \frac{a}{2} \tag{4.13}$$

Le principe de la démonstration sera identique pour les indices négatifs.

On coupe la somme précédente en deux :

$$\begin{aligned} \frac{1}{N-1} \sum_{k=1}^{N-1} (1 - \frac{|k|}{N})\, w_k &= \frac{1}{N-1} \sum_{k=1}^{k_0(\varepsilon)} (1 - \frac{|k|}{N})\, w_k & (4.14) \\ &+ \frac{1}{N-1} \sum_{k=k_0(\varepsilon)+1}^{N-1} (1 - \frac{|k|}{N})\, w_k & (4.15) \end{aligned}$$

où pour $k \geq k_0(\varepsilon) \geq 0$, $w_k - a < \varepsilon$.

On calcule la limite des deux termes obtenus. Pour le premier terme de la somme, w_k ayant une limite finie, on a :

$$\lim_{N\to\infty} \frac{1}{N-1} \sum_{k=1}^{k_0(\varepsilon)} (1 - \frac{|k|}{N})\, w_k = 0 \tag{4.16}$$

Pour le second terme de la somme on utilise l'inégalité suivante :

Soit $\varepsilon > 0$, pour $k \geq k_0(\varepsilon)$,

$$a - \varepsilon \leq w_k \leq a + \varepsilon \tag{4.17}$$

d'où

$$\frac{a-\varepsilon}{N-1} \sum_{k=k_0(\varepsilon)+1}^{N-1} (1 - \frac{|k|}{N}) \leq \frac{1}{N-1} \sum_{k=k_0(\varepsilon)+1}^{N-1} (1 - \frac{|k|}{N})\, w_k \leq \frac{a+\varepsilon}{N-1} \sum_{k=k_0(\varepsilon)+1}^{N-1} (1 - \frac{|k|}{N}) \tag{4.18}$$

On cherche un majorant pour les bornes de l'inégalité.

$$\begin{aligned} \sum_{k=k_0(\varepsilon)+1}^{N-1} (1 - \frac{|k|}{N}) &= \big((N-1) - (k_0(\varepsilon)+1)\big) - \frac{1}{N} \sum_{k=k_0(\varepsilon)+1}^{N-1} k & (4.19) \\ &= (N - 2 - k_0(\varepsilon)) - \frac{1}{N} \sum_{k=k_0(\varepsilon)+1}^{N-1} k & (4.20) \end{aligned}$$

et

$$\sum_{k=k_0(\varepsilon)+1}^{N-1} k = \sum_{k'=1}^{N-1-k_0(\varepsilon)} (k'+k_0(\varepsilon)) \qquad \text{avec } k' = k - k_0(\varepsilon) \tag{4.21}$$

$$= \frac{(N-1-k_0(\varepsilon))(N-k_0(\varepsilon))}{2} + k_0(\varepsilon)(N-1-k_0(\varepsilon)) \tag{4.22}$$

$$= \frac{(N-1-k_0(\varepsilon))(N+k_0(\varepsilon))}{2} \tag{4.23}$$

d'où

$$\frac{1}{N}\sum_{k=k_0(\varepsilon)+1}^{N-1} k = \left(1 - \frac{1}{N} - \frac{k_0(\varepsilon)}{N}\right)\frac{N+k_0(\varepsilon)}{2} \tag{4.24}$$

En ajoutant le facteur $\frac{1}{N-1}$ présent dans l'inégalité on obtient :

$$\frac{1}{N-1}\left(\frac{1}{N}\sum_{k=k_0(\varepsilon)+1}^{N-1} k\right) = \frac{1}{2}\left(1 - \frac{1}{N} - \frac{k_0(\varepsilon)}{N}\right)\left(1 + \frac{k_0(\varepsilon)+1}{N-1}\right) \longrightarrow \frac{1}{2} \quad \text{quand } N \to \infty \tag{4.25}$$

et donc on a pour le terme que l'on cherchait à majorer :

$$\frac{1}{N-1}\sum_{k=k_0(\varepsilon)+1}^{N-1} (1 - \frac{|k|}{N}) = \frac{(N-2-k_0(\varepsilon))}{N-1} - \frac{1}{2}\left(1 - \frac{1}{N} - \frac{k_0(\varepsilon)}{N}\right)\left(1 + \frac{k_0(\varepsilon)+1}{N-1}\right)$$

$$\longrightarrow \frac{1}{2} \quad \text{quand } N \to \infty$$

En conclusion de l'inégalité :

$$\limsup_{N\to\infty} \frac{1}{N-1}\sum_{k=k_0(\varepsilon)+1}^{N-1} (1 - \frac{|k|}{N})\, w_k \leq \frac{a+\varepsilon}{2} \tag{4.26}$$

$$\liminf_{N\to\infty} \frac{1}{N-1}\sum_{k=k_0(\varepsilon)+1}^{N-1} (1 - \frac{|k|}{N})\, w_k \geq \frac{a-\varepsilon}{2} \tag{4.27}$$

Finalement, si on considère tous les termes de la somme initiale que l'on a scindée en deux parties on obtient en reprenant (14),

$$\limsup_{N\to\infty} \frac{1}{N-1}\sum_{k=1}^{N-1} (1 - \frac{|k|}{N})\, w_k \leq \frac{a+\varepsilon}{2} \tag{4.28}$$

$$\liminf_{N\to\infty} \frac{1}{N-1}\sum_{k=1}^{N-1} (1 - \frac{|k|}{N})\, w_k \geq \frac{a-\varepsilon}{2} \tag{4.29}$$

Puisque ε est arbitraire on a la limite suivante :

$$\lim_{N\to\infty} \frac{1}{N-1} \sum_{k=1}^{N-1} (1 - \frac{|k|}{N})\, w_k = \frac{a}{2} \tag{4.30}$$

et donc,

$$\lim_{N\to\infty} \frac{1}{N-1} \sum_{k=-N+1}^{N-1} (1 - \frac{|k|}{N})\, w_k = a \tag{4.31}$$

□

En combinant les résultats précédents, on obtient :

$$\lim_{N\to\infty} \psi_{Y^N_{t_1}, Y^N_{t_2}}(u_1, u_2) = \psi(u_1)\psi(u_2) \tag{4.32}$$

Ce qui achève la preuve pour $r = 2$. La preuve pour le cas le plus simple étant terminée, on s'intéresse maintenant au cas général où on choisit cette fois r points parmi N pour la permutation.

4.5 Preuve pour r points parmi N

Tout comme dans la preuve pour le cas $r = 2$, on s'intéresse à la fonction caractéristique du r-uplets $Y^N_{t_1}, \cdots, Y^N_{t_r}$ donnée par :

$$\begin{aligned}
\psi_{Y^N_{t_1},\cdots,Y^N_{t_r}}(u_1, \cdots, u_r) &= \mathbb{E}\Big[\exp\Big(i \sum_{j=1}^{r} u_j Y^N_{t_j} \Big) \Big] \\
&= \mathbb{E}\Big\{ \mathbb{E}\Big[\exp\Big(i \sum_{j=1}^{r} u_j Y^N_{t_j} \Big) \mid \sigma_N \Big] \Big\} \\
&= \frac{1}{N!} \sum_{\sigma_N \in \Sigma_N} \mathbb{E}\Big[\exp\Big(i \sum_{j=1}^{r} u_j X_{\sigma_N(t_j)} \Big) \Big]
\end{aligned}$$

Soit $\Delta \in \mathbb{R}^r$, et soit Σ^Δ_N l'ensemble des permutations telles que r indices $t_1, \cdots, t_r$ du processus soient espacés d'une distance Δ_j deux à deux après permutation. Δ_j est non nul pour $j = 1, \cdots, r-1$ et $\Delta_r = 0$. Cela signifie que l'on fixe le dernier point des r points considérés parmi N.
Σ^Δ_N est donc définie par :

$$\Sigma^\Delta_N = \big\{ \sigma_N \in \Sigma_N : \sigma_N(t_j) = t_r + \Delta_j, \forall j = 1, \cdots, r-1; \Delta_r = 0 \big\} \tag{4.33}$$

On note $l_{N,\Delta} = \#\ \Sigma^\Delta_N$ le cardinal de l'ensemble des permutations Σ^Δ_N . On revient à l'expression de la fonction caractéristique et on a donc :

$$\psi_{Y^N_{t_1},\cdots,Y^N_{t_r}}(u_1, \cdots, u_r) = \frac{1}{N!} \sum_{\Delta_1 \cdots \Delta_{r-1}} l_{N,\Delta} \mathbb{E}\Big[\exp\Big(i \sum_{j=1}^{r} u_j X_{t_r + \Delta_j} \Big) \Big] \tag{4.34}$$

On calcule le cardinal $l_{N,\Delta}$ de la permutation Σ_N^Δ :

Les Δ_j peuvent prendre leurs valeurs dans $-N+1, \cdots, N+1$ pour $j = 1, \cdots, r-1$. On rappelle que la permutation est définie par :

$$\sigma_N(t_j) = t_r + \Delta_j, \quad \forall j = 1, \cdots, r-1 \tag{4.35}$$

on aura alors :

$$1 - \min_j \Delta_j^- \leq \sigma(t_r) \leq N - \max_j \Delta_j^+ \tag{4.36}$$

Au final, le cardinal de Σ_N^Δ est donc

$$l_{N,\Delta} = (N-r)!(N - \max_j \Delta_j^+ + \min_j \Delta_j^-) \qquad j = 1, \cdots, r-1 \tag{4.37}$$

Revenons à la fonction caractéristique, on pose $\psi_\Delta^r = \mathbb{E}\left[\exp\left(i\sum_{j=1}^r u_j X_{t_r+\Delta_j}\right)\right]$, on peut alors écrire :

$$\psi_{Y_{t_1}^N,\cdots,Y_{t_r}^N}(u_1, \cdots, u_r) = \sum_{\Delta_1 \cdots \Delta_{r-1}} \frac{(N - \max_j \Delta_j^+ + \min_j \Delta_j^-)}{N(N-1)\cdots(N-r+1)} \, \psi_\Delta^r \tag{4.38}$$

Montrons alors que si $N \to \infty$ alors $\psi_{Y_{t_1}^N,\cdots,Y_{t_r}^N}(u_1, \cdots, u_r) \to \prod_{j=1}^r \psi(u_j)$.

Comme dans le cas $r = 2$, on fait l'hypothèse que les fonctions ψ_Δ^r ont une limite finie ψ^*. On rappelle qu'une fonction caractéristique est toujours bornée par 1 en module. Or si ces fonctions vérifient la condition de mélange (H) (4.4), alors $\psi_\Delta^r \to \prod_{j=1}^r \psi(u_j)$. Ainsi, on va alors montrer que $\psi_{Y_{t_1}^N,\cdots,Y_{t_r}^N}(u_1, \cdots, u_r)$ et ψ_Δ^r ont la même limite.

On note

$$\frac{(N - \max_j \Delta_j^+ + \min_j \Delta_j^-)}{N(N-1)\cdots(N-r+1)} = S_\Delta^N \tag{4.39}$$

On peut écrire

$$\begin{aligned} \psi_{Y_{t_1}^N,\cdots,Y_{t_r}^N}(u_1, \cdots, u_r) &= \sum_{\Delta_1 \cdots \Delta_{r-1}} (S_\Delta^N \psi_\Delta^r + \psi^* - \psi^*) \\ &= \sum_{\Delta_1 \cdots \Delta_{r-1}} S_\Delta^N (\psi_\Delta^r - \psi^*) + \sum_{\Delta_1 \cdots \Delta_{r-1}} S_\Delta^N \psi^* \end{aligned} \tag{4.40}$$

On étudie chacune des deux sommes séparément.

On commence par la somme $\sum_{\Delta_1 \cdots \Delta_{r-1}} S_\Delta^N \psi^*$.

Un des arguments clés est que $\sum_{\Delta_1\cdots\Delta_{r-1}} S_\Delta^N$ tend vers $\frac{1}{N!}\sum_{\sigma\in\Sigma_N}$ et sa limite vaut 1. On a donc par le théorème de convergence dominée

$$\lim_{N\to\infty}\sum_{\Delta_1\cdots\Delta_{r-1}} S_\Delta^N\psi^* = \psi^*. \tag{4.41}$$

On s'intéresse maintenant à la somme $\sum_{\Delta_1\cdots\Delta_{r-1}} S_\Delta^N(\psi_\Delta^r - \psi^*)$.

On coupe cette somme en deux parties pour étudier indépendamment le comportement des petits termes d'une part et des grands termes d'autre part.

$$\sum_{\Delta_1\cdots\Delta_{r-1}} S_\Delta^N(\psi_\Delta^r-\psi^*) = \sum_{|\Delta_j|\le\eta_j} S_\Delta^N(\psi_\Delta^r-\psi^*) + \sum_{|\Delta_j|>\eta_j} S_\Delta^N(\psi_\Delta^r-\psi^*) \quad , j=1,\cdots,r-1. \tag{4.42}$$

On regarde tout d'abord le comportement des petits termes.

On a :

$$\frac{1}{N(N-1)\cdots(N-r+1)} \sim \frac{1}{N^r} \quad \text{quand } N\to\infty$$

et sur $|\Delta_j|\le\eta_j$ pour $j=1,\cdots,r-1$, $\sum_{|\Delta_j|\le\eta_j} S_\Delta^N$ se décompose :

$$\sum_{|\Delta_1|\le\eta_1\cdots|\Delta_{r-1}|\le\eta_{r-1}} \frac{(N-\max_j\Delta_j^+ + \min_j\Delta_j^-)}{N(N-1)\cdots(N-r+1)} \sim \frac{1}{N}\sum_{|\Delta_1|\le\eta_1}\cdots\frac{1}{N}\sum_{|\Delta_{r-1}|\le\eta_{r-1}}\Big(1-\frac{\max_j\Delta_j^+-\min_j\Delta_j^-}{N}\Big)$$

or

$$1-\frac{\max_j\Delta_j^+-\min_j\Delta_j^-}{N}\le 1 \tag{4.43}$$

donc

$$\begin{aligned}\frac{1}{N}\sum_{|\Delta_1|\le\eta_1}\cdots\frac{1}{N}\sum_{|\Delta_{r-1}|\le\eta_{r-1}}\Big(1-\frac{\max_j\Delta_j^+-\min_j\Delta_j^-}{N}\Big) &\le \frac{1}{N}\sum_{|\Delta_1|\le\eta_1}\cdots\frac{1}{N}\sum_{|\Delta_{r-1}|\le\eta_{r-1}} 1\\ &\le 2^{r-1}\frac{1}{N}\eta_1\times\cdots\times\frac{1}{N}\eta_{r-1}\\ &\le \varepsilon,\quad \forall\varepsilon>0.\end{aligned}$$

D'autre part, on a supposé que la fonction ψ_Δ^N avait pour limite ψ^* quand $N\to\infty$. De plus, ψ_Δ^N étant bornée par 1 en module,

$$|\psi_\Delta^r-\psi^*|\le 2 \tag{4.44}$$

En combinant ce résultat avec le résultat précédent, on a que la somme

$$\sum_{|\Delta_j|\le\eta_j} S_\Delta^N(\psi_\Delta^r-\psi^*)$$

a une limite finie pour les Δ_j finis.

On s'intéresse maintenant au comportement des grands termes de l'expression (4.42) décrit par la somme $\sum_{|\Delta_j|>\eta_j} S_\Delta^N(\psi_\Delta^r - \psi^*) \quad , j = 1, \cdots, r-1$.

En réutilisant l'hypothèse selon laquelle la limite de $\sum_{\Delta_1 \cdots \Delta_{r-1}} S_\Delta^N$ vaut 1 lorsque $N \to \infty$, on peut dire que pour les grandes valeurs des Δ_j,

$$\sum_{|\Delta_j|>\eta_j} S_\Delta^N \leq 1 \tag{4.45}$$

En combinant cette inégalité avec l'hypothèse $\psi_\Delta^N \to \psi^*$ quand $N \to \infty$ que l'on peut réécrire $\psi_\Delta^N - \psi^* < \varepsilon',\ \forall \varepsilon' > 0$, on a pour les grands Δ_j

$$\sum_{|\Delta_j|>\eta_j} S_\Delta^N(\psi_\Delta^r - \psi^*) \leq 1 \times \varepsilon' \quad , j = 1, \cdots, r-1. \tag{4.46}$$

Au final, en regroupant les résultats obtenus pour les petits et grands termes on a

$$\begin{aligned} \sum_{\Delta_1 \cdots \Delta_{r-1}} S_\Delta^N(\psi_\Delta^r - \psi^*) &= \sum_{|\Delta_j|\leq\eta_j} S_\Delta^N(\psi_\Delta^r - \psi^*) + \sum_{|\Delta_j|>\eta_j} S_\Delta^N(\psi_\Delta^r - \psi^*) \\ &\leq \varepsilon'', \quad \text{pour } \varepsilon'' > 0. \end{aligned}$$

Afin de conclure cette preuve, on revient à l'expression (4.40) dans laquelle on a décomposé la fonction caractéristique en deux sommes. En appliquant le résultat précédent et celui obtenu à l'expression (4.41), on a part le théorème de convergence dominée que

$$\lim_{N\to\infty} \left\{ \sum_{\Delta_1 \cdots \Delta_{r-1}} S_\Delta^N(\psi_\Delta^r - \psi^*) + \sum_{\Delta_1 \cdots \Delta_{r-1}} S_\Delta^N \psi^* \right\} = \lim_{N\to\infty} \sum_{\Delta_1 \cdots \Delta_{r-1}} S_\Delta^N \psi^* = \psi^* \tag{4.47}$$

d'où

$$\lim_{N\to\infty} \sum_{\Delta_1 \cdots \Delta_{r-1}} S_\Delta^N \psi_\Delta^r = \psi^* \tag{4.48}$$

On peut donc écrire

$$\lim_{N\to\infty} \psi_{Y_{t_1}^N, \cdots, Y_{t_r}^N}(u_1, \cdots, u_r) - \psi^* = \lim_{N\to\infty} \sum_{\Delta_1 \cdots \Delta_{r-1}} S_\Delta^N \psi_\Delta^r - \psi^* = 0 \tag{4.49}$$

Ainsi, on a montré que $\psi_{Y_{t_1}^N, \cdots, Y_{t_r}^N}(u_1, \cdots, u_r)$ et ψ_Δ^r, fonction caractéristique sachant les permutations, ont même limite $\prod_{j=1}^r \psi(u_j)$ par la condition de mélange (H), ce qui achève la preuve.

4.6 Résultats

La figure suivante illustre un échantillon de l'erreur horizontale enregistrée sur la station fixe d'HELILEO (courbe grise) et le même échantillon ayant subit une permutation

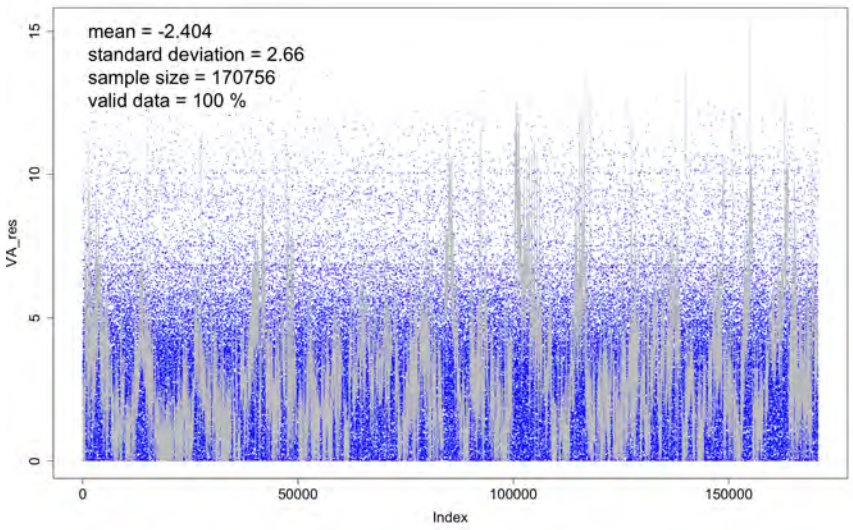


FIGURE 4.4 – Echantillon avant (grey) et après (blue) permutation

aléatoire suivant une loi uniforme (points bleus). On voit que la corrélation temporelle n'apparaît plus au sein des données.
Les courbes de fonctions d'autocorrélation qui suivent permettent de confirmer la disparition de dépendance temporelle après permutation. En effet, sur la figure 4.5 , avant permutation, on remarque la présence de deux pics correspondant à des écarts de 24 et 48 heures.

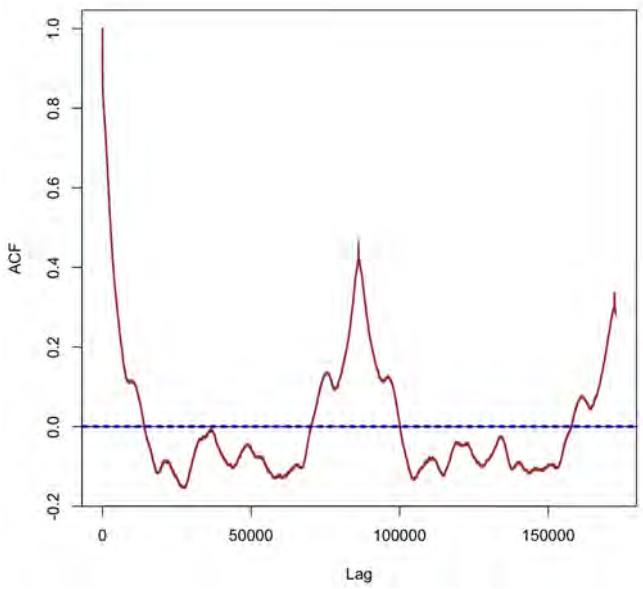


FIGURE 4.5 – Fonction d'autocorrélation avant permutation

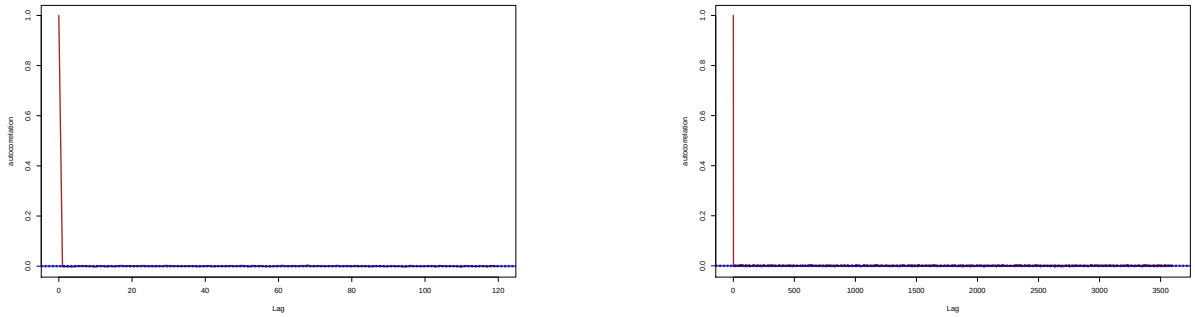


FIGURE 4.6 – Autocorrélation après permutation

Après permutation, la fonction d'autocorrélation est nulle dès le premier écart entre deux instants.

4.7 Conclusion

Nous avons mis en évidence une dépendance temporelle assez forte dans les données GPS étudiées. En effet, les erreurs de positionnement mais aussi les niveaux de protections (Protection Levels) calculés par le récepteur se présentent sous la forme de données cycliques. Une étude des périodogrammes des séries temporelles a permis de déceler la présence de cycles de 24 heures au sein des données. Cette corrélation temporelle remet en cause la validité des estimateurs de quantiles extrêmes présentés dans le chapitre précédent. En effet, les théorèmes limites qui ont permis d'établir les estimateurs POT, Gumbel, etc... sont valables sous l'hypothèse d'observations indépendantes et identiquement distribuées.
Enfin, ce chapitre a permis de fournir une justification théorique de l'heuristique selon lequel il est possible de briser la structure de dépendance existant au sein d'un processus stationnaire en appliquant une permutation aléatoire selon une loi uniforme sur les indices temporels du processus.

CHAPITRE 5

Applications en lien avec les activités de la société HELILEO

Résumé :
Ce dernier chapitre constitue la partie applicative de cette étude. Nous y présentons le protocole d'analyse élaboré au cours de ces trois années. Il est composé des étapes suivantes :

- visualisation et analyse des données : nous étudierons les distributions des données GPS en sortie de récepteur.
- vérification des hypothèses requises par les modèles employés : l'hypothèse d'indépendance étant traitée dans le chapitre 4, nous nous intéresserons à la stationnarité des données.
- sélection du modèle, estimation des paramètres et adéquation du modèle aux données : au travers d'un cas d'étude, un exemple d'application du modèle POT issu de la théorie des extrêmes sera appliqué aux données GPS.
- estimation du quantile recherché et des intervalles de confiance associés : enfin, les estimation de quantiles extrêmes permettront de réaliser une analyse d'intégrité, thème de ce manuscrit.

Pour terminer cette partie pratique, nous présenterons l'outil logiciel qui a été développé durant cette étude. Son architecture ainsi que ses fonctionnalités y seront décrites.

Sommaire

5.1 Introduction

Le GPS a pour objectif de fournir une solution de position à l'utilisateur. Celle-ci doit être la plus fiable possible, en particulier pour les applications aéronautiques pour lesquelles les contraintes d'intégrité sont importantes. On définit l'intégrité comme une mesure de confiance que l'on peut accorder aux informations fournies par le système de positionnement [KH06].
La position est calculée sur la base de mesures fournies par les satellites de la constellation GPS. Or ces mesures peuvent parfois devenir erronées (défaillance des satellites, présence de multi-trajets, d'interférences, etc...), entraînant une dégradation de la position calculée inacceptable pour l'utilisateur. On souhaite fournir une méthodologie permettant la mise en place d'un protocole de test visant à observer les performances d'intégrité du système GPS/EGNOS à un niveau utilisateur, c'est à dire que l'on s'intéresse aux données fournies par les récepteurs, en bout de chaîne de positionnement. EGNOS est le système de satellites géostationnaires européen, dit SBAS (Satellite Based Augmented System), améliorant la précision du positionnement GPS et disposant d'un service d'intégrité. EGNOS transmet aux utilisateurs des informations relatives à la santé des satellites ainsi que des paramètres permettant de borner les erreurs de position. EGNOS sera le principal système d'augmentation du futur système Galileo.
Dans le cadre d'un programme d'évaluation des performances d'EGNOS, la société HELILEO s'est vue confiée par la commission européenne, une vaste campagne de tests héliportés. L'objet de cette campagne était de fournir un rapport de performances suivant différents critères parmi lesquels figuraient la précision de positionnement (horizontale et

verticale) ainsi qu'une mesure d'intégrité. On s'intéresse dans cette étude aux précisions de positionnement ainsi qu'à la mesure d'intégrité pour deux mode de positionnement : statique et dynamique. En plus de ce rapport, HELILEO se devait de fournir une justification de la quantité d'heures de vols engagées dans ce programme.
Il s'agit dans ce chapitre de donner une méthodologie statistique pour répondre aux problématiques suivantes :

- est-ce que les positions délivrées par les récepteurs en vol remplissent les exigences de sécurité imposées par l'OACI (Organisation de l'Aviation Civile Internationale) ?
- combien d'heures de vols sont nécessaires afin de garantir que les conditions de sécurité soient remplies ?

Nous choisirons d'évaluer les conditions les plus strictes imposées par l'OACI. Il s'agit des approches de précision CAT I pour lesquelles le risque d'intégrité est fixé à $1 - 2.10^{-7}$ par approche.
Un des objectifs pratiques de cette étude a été de livrer un outil fonctionnel incluant les études théoriques décrites dans ce manuscrit. Ainsi, une plate forme d'analyse statistique de données a été développée de manière à évaluer les performances du système au niveau utilisateur (en sortie de récepteur). Cette plate forme permet de traiter les données recueillies lors de la campagne d'essais dans leur ensemble ou sous forme de scénario particulier (choix du vols, choix du récepteur etc...). L'utilisateur peut alors choisir de générer des rapports automatiques sur lesquels figurent un ensemble d'indicateurs statistiques représentés graphiquement afin d'évaluer les performances du système. L'architecture ainsi que les fonctionnalités de la plateforme feront l'objet d'une section dans ce chapitre.

5.2 Analyse des données GPS/EGNOS

Il s'agit d'évaluer les performances d'intégrité du système à partir d'un modèle reposant sur des modèles probabilistes issus de la théorie des extrêmes. De manière générale, la mesure d'intégrité sera vue comme un problème d'estimation de quantile extrême. La théorie des extrêmes repose sur des théorèmes asymptotiques qui requièrent que certaines hypothèses sur les observations soient vérifiées afin d'utiliser les outils qui en découlent. Le théorème de Fisher-Tippet [FT28] ainsi que le théorème de Pickands [Pic75] permettant de décrire le comportement des queues de distribution, sont énoncés pour des suites de variables aléatoires indépendantes et identiquement distribuées. Les observations qui nous intéressent seront, d'un point de vue probabiliste, des processus temporels dans lesquels les phénomènes de dépendance/indépendance se manifestent par une corrélation temporelle entre les données plus ou moins forte. On rappelle que l'équivalence entre corrélation et indépendance n'est vraie théoriquement que dans le cas gaussien.
Avant d'analyser les performances du système, il est nécessaire d'établir un diagnostic sur la nature des données fournies par les récepteurs GPS afin d'évaluer la validité du modèle de mesure d'intégrité sur des données réelles. Le chapitre 4, consacré au théorème d'indépendance asymptotique, a montré qu'en présence de données fortement corrélées, les conditions d'indépendances pouvaient être atteintes en appliquant une permutation aléatoire sur les indices temporels du processus étudié. Reste à étudier la stationnarité des données.

5.2.1 Protocole d'acquisition de données

Les positions ont été enregistrées par divers récepteurs GPS disponibles chez HELILEO. Le pool de capteurs est composé d'une gamme de récepteurs représentative des différentes qualités présentes sur le marché. Les protocoles et formats de sortie des constructeurs étant tous différents, nous avons développé une plate forme de mise en forme des données en sortie de récepteur de manière à toutes les convertir aux formats acceptés par la plate forme d'analyse de données.
Les récepteurs sont embarqués dans quatre valises, H-box, réalisées par HELILEO (figure 5.1). Elles sont composées d'un étage récepteur et d'un étage d'alimentation et de stockage des données. Les H-box sont placées dans les soutes des hélicoptères de manière à réaliser des enregistrements en vol.

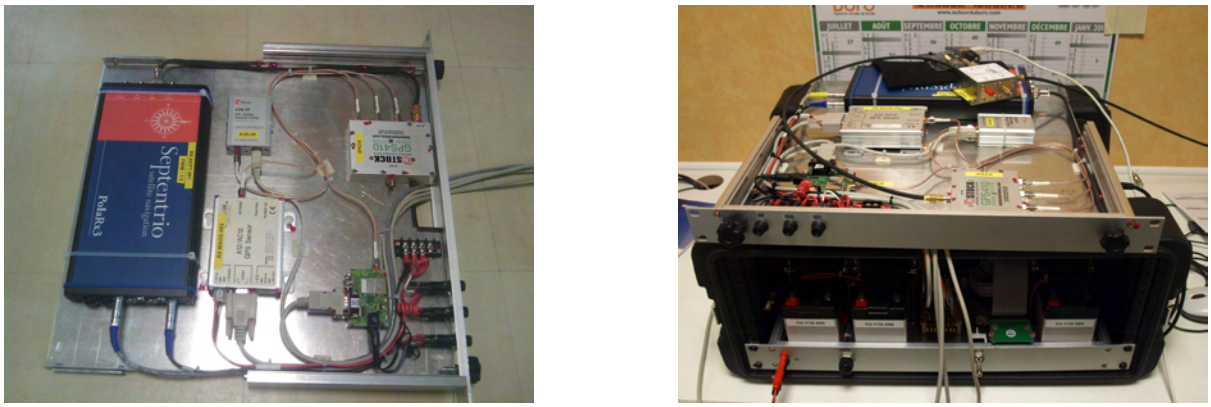


FIGURE 5.1 – H-box

5.2.2 Présentation des données

Les performances du système GPS/EGNOS sont évaluées en terme de précision de positionnement et d'intégrité, au travers des variables suivantes : les erreurs de positionnement (HPE, VPE), les niveaux de protection HPL et VPL (Horizontal ou Vertical Protection Level) et les seuils d'alerte HAL et VAL (Horizontal ou Vertical Alarm Limit). Nous définissons ces variables avant d'établir le modèle probabiliste de mesure d'intégrité. Deux exemples d'échantillon selon deux modes de positionnement, statique puis dynamique, seront ensuite commentés.

5.2.2.1 Définition des variables observées

Les définitions ci-après sont tirées de la norme RTCA/DO229 [DO96] intitulée "Minimum Operational Performance Standards (MOPS) for Global Positionning System/Wide Area Augmentation System (GPS/WAAS) Airborne Equipement". Cette norme aéronautique américaine décrit les processus liés au développement de solutions de positionnement embarquées dans les aéronefs utilisant le système GPS augmenté du système WAAS (équivalent d'EGNOS américain).

Erreurs de positionnement :
On définit les erreurs de positionnement comme la différence entre la position vraie et la

position mesurée. Les erreurs de positionnement sont exprimées sur le plan horizontal et sur l'axe vertical. Le calcul de ces erreurs est fonction du repère dans lequel sont exprimées les coordonnées des positions. On note (X^*, Y^*, Z^*) et (ϕ^*, λ^*, h^*) les coordonnées de la position de référence dans le repère cartésien et dans le repère géodésique respectivement. ϕ désigne la latitude, λ la longitude et h la hauteur au dessus de l'ellipsoïde de référence.

Dans le repère cartésien :
L'erreur de positionnement horizontale, notée HPE (Horizontal Position Error), est définie dans le repère cartésien comme une fonction du temps t par :

$$\mathrm{HPE}(t) = \sqrt{(X^*(t) - X(t))^2 + (Y^*(t) - Y(t))^2} \tag{5.1}$$

L'erreur de positionnement verticale, notée VPE (Vertical Position Error), est définie dans le repère cartésien comme une fonction du temps t par :

$$\mathrm{VPE}(t) = Z^*(t) - Z(t) \tag{5.2}$$

Dans le repère géodésique :
L'erreur de positionnement horizontale HPE est définie par la distance entre deux points dont les coordonnées sont exprimées en radians :

$$\mathrm{HPE}(t) = R\cos^{-1}\Big(\sin(\lambda(t))\sin(\lambda^*(t)) + \cos(\lambda(t))\cos(\lambda^*(t))\cos(\phi^*(t) - \phi(t))\Big) \tag{5.3}$$

Où R représente le rayon de la Terre. Une attention particulière doit être apportée à la valeur de R en fonction de la zone géographique où s'effectuent les mesures à la surface du globe. En effet, on peut envisager de prendre comme valeur pour R, le rayon moyen équatorial, le rayon moyen polaire ou la moyenne des deux selon que l'on se trouve plus proche d'un pôle ou de l'équateur. Les valeurs des approximations de ces différents rayons sont :
Rayon moyen équatorial : $R_{eq} \approx 6378.137$ km.
Rayon moyen polaire : $R_{pol} \approx 6356.752$ km.
Les essais en vols ayant eu lieu en France, dans la région Aquitaine, il est alors raisonnable de considérer la moyenne entre le rayon équatorial et le rayon polaire (figure 5.2). On aura $R_{moy} = 6367.444$ km. Pour exprimer l'erreur de positionnement horizontale en mètres, la valeur de R doit être exprimée dans la même unité, en mètres.
L'erreur de positionnement verticale est définie dans le repère géodésique par :

$$\mathrm{VPE}(t) = h^*(t) - h(t) \tag{5.4}$$

Niveaux de protection PL :
Les niveaux de protection sont définis comme étant les bornes que l'erreur de positionnement ne peut dépasser selon une probabilité donnée.

Niveaux d'alerte AL :
Les niveaux d'alerte sont définis comme étant le maximum acceptable pour l'erreur de positionnement. Il seront fonction de la phase de vol considérée.

FIGURE 5.2 – Aquitaine à l'échelle du globe terrestre

5.2.2.2 Modèle de mesure d'intégrité

On considère deux événements menant à une perte d'intégrité au niveau utilisateur. Le cas où l'erreur de positionnement excède le niveau de protection associé $x\text{PE} > x\text{PL}$, x désigne H ou V selon l'erreur étudiée. Cet événement est qualifié de Misleading Information (MI). Le second événement est le cas où l'erreur de positionnement excède le niveau d'alerte $x\text{PE} > x\text{AL}$; il est appelé Hazardous Misleading Information (HMI).
D'un point de vue probabiliste, le modèle d'intégrité utilisateur sera vu comme un quantile d'ordre $1-p$ dont on rappelle la définition : soit X une variable aléatoire réelle suivant une loi F. Pour p un réel de l'intervalle $[0,1]$, le quantile d'ordre $1-p$ est la valeur $q(p)$ vérifiant :

$$\mathbb{P}(X \leq q(p)) = F(q(p)) = 1 - p$$

ou de façon équivalente,

$$\mathbb{P}(X > q(p)) = 1 - F(q(p)) = p$$

Définition 5.2.1. *On définit le modèle de mesure d'intégrité au niveau utilisateur pour l'événement MI par :*

$$\mathbb{P}(xPE > xPL) = 1 - p$$

ou encore,

$$\mathbb{P}(\frac{xPE}{xPL} < q(p)) = p \tag{5.5}$$

$q(p)$ est vu comme un seuil d'alerte. Si on montre que $q(p)$ est inférieur à 1, on garantit que l'intégrité utilisateur est assurée suivant notre modèle. Le modèle sera identique pour l'événement HMI correspondant à $xPE > xAL$.

La particularité de ce modèle réside dans les niveaux de probabilité auxquels on s'intéresse. En effet, les contraintes de sécurité décrites par les standards aéronautiques [DO96] atteignent des niveaux de probabilité très élevés de l'ordre de $1-2.10^{-7}$. Ces probabilités correspondent à des fréquences d'occurrence d'événements rares, situés dans les queues de distribution. Les quantiles associés à de tels niveaux sont qualifiés de quantiles extrêmes et se situent bien souvent au delà du domaine des observations. Plusieurs méthodes statistiques existent pour l'estimation de quantiles extrêmes. En effet, les méthodes de Monte

Carlo permettent d'atteindre des niveaux de probabilité aussi élevés par simulation. Cependant, des échantillons de tailles très importantes associés à des niveaux de probabilité aussi élevés impliquent un temps de calcul trop important pour considérer cette méthodologie. Dans le cadre d'une mesure d'intégrité, la méthode la plus adaptée est l'utilisation de la théorie des extrêmes spécialement dédiée à l'étude du comportement des queues de distributions [AG09]. Cette théorie fournit une classe de modèles permettant l'extrapolation de l'observé vers le non observé et ainsi la caractérisation des événements rares situés dans les queues de distribution. Comme mentionné dans l'introduction de cette section, les théorèmes asymptotiques qui constituent la clé de voute de cette théorie sont vrais sous des hypothèses de variables aléatoires indépendantes et identiquement distribuées. Les sections suivantes présentent des exemples de données dont nous disposons de manière à évaluer la validité de ces hypothèses.

5.2.2.3 Exemple de données : cas statique

Les enregistrements statiques présentés dans cette étude proviennent de la station de référence d'HELILEO. Cette station enregistre 24h/24h les signaux des différentes constellations GNSS en vue de manière à effectuer un monitoring permanent de ces dernières. De plus, elle constitue un point d'appui pour des positionnements différentiels. Les erreurs de positionnement dans le cas statique seront calculées par rapport à la position de la station. Cette position a été estimée au moyen de techniques avancées (GPS différentiel RTK) permettant d'éliminer une grandes parties des erreurs systématique affectants les mesures de pseudo-distances. Des études spécifiques ont alors permis d'établir un positionnement doté d'une précision de 0.40 cm à 95% (figure 5.3). La figure 5.4 présente un enregistrement de positionnement statique (figure de gauche) ainsi que l'erreur de positionnement horizontale associée (à droite) d'une durée d'environ trois jours.

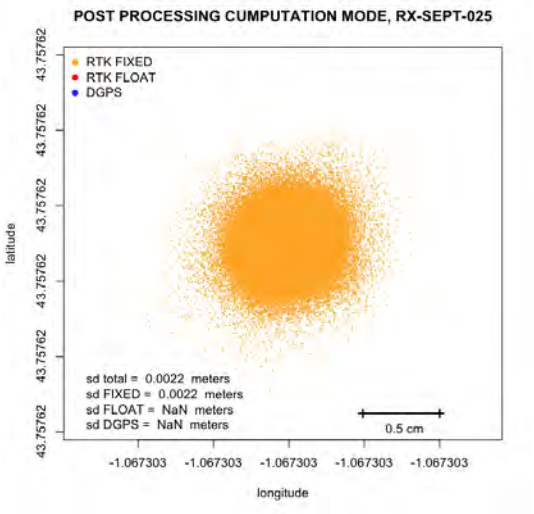


FIGURE 5.3 – Enregistrement statique (post processing)

A plus grande échelle, quatorze jours, on observe sur la figure 5.5 des phénomènes cycliques et certaines valeurs anormalement élevées. De plus le processus n'apparaît pas stationnaire. Ces deux remarques sont contraires aux hypothèses requises pour utiliser les outils issus de la théorie des extrêmes. Cependant, nous verrons dans la section suivante que la stationnarité peut être atteinte pour de grands échantillons. L'hypothèse

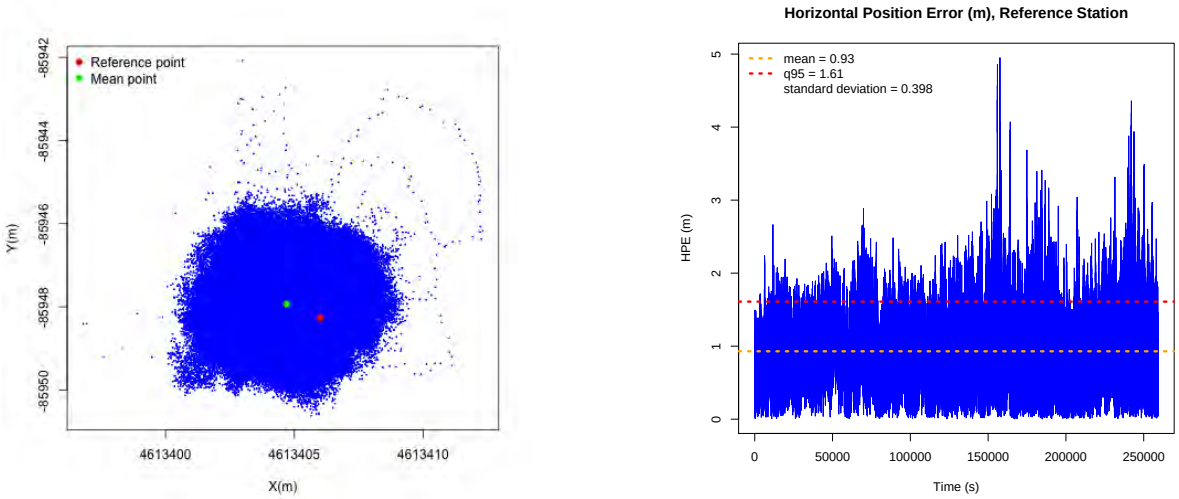


FIGURE 5.4 – Enregistrement statique

d'indépendance des données est traitée dans le chapitre 4 dans lequel nous avons montré que la structure de dépendance d'un processus pouvait être brisée en appliquant une permutation aléatoire sur les indices temporels des observations. Ainsi, pour de grands échantillons, il est possible d'atteindre une indépendance asymptotique.

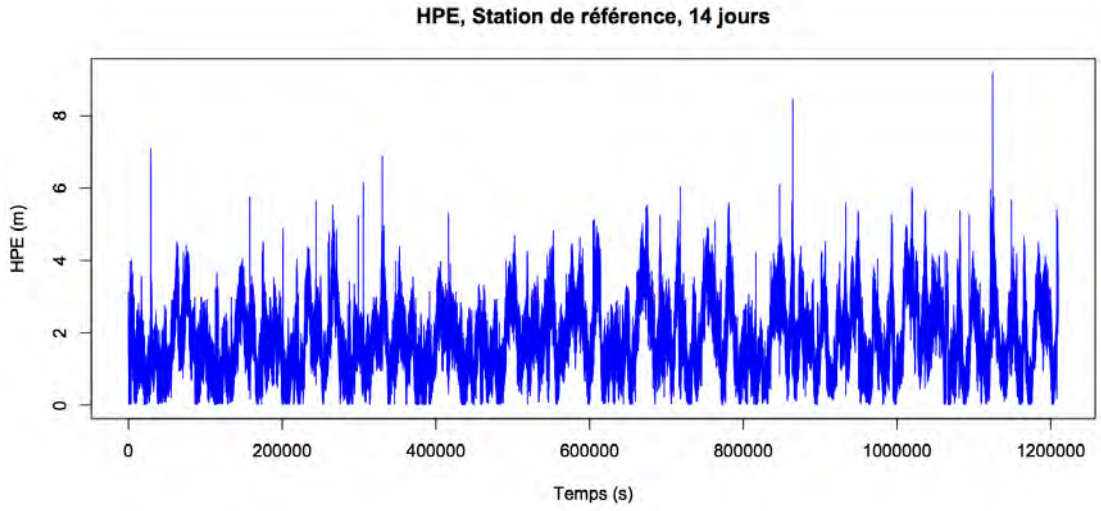


FIGURE 5.5 – HPE, enregistrement statique

La figure 5.6 présente l'évolution des trois variables impliquées dans notre modèle de mesure d'intégrité (pour le cas horizontal) sur une période de quatorze jours. L'événement $x\text{PL} > x\text{AL}$ permet une évaluation de l'intégrité au niveau système SBAS, car les niveaux de protection (PL) sont estimés par les récepteurs en intégrant des paramètres transmis par le système EGNOS. On dira dans ce cas que le service d'intégrité n'est pas opérationnel (FS pour Failed System).

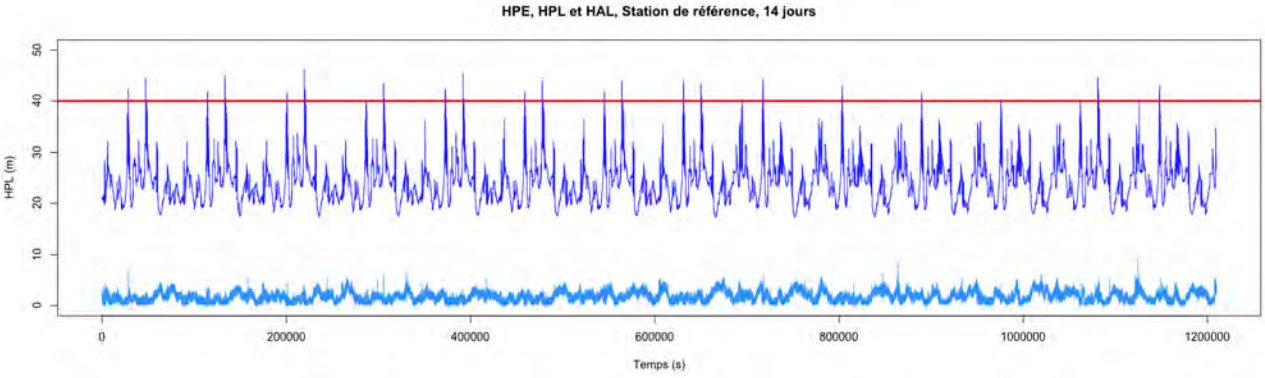


FIGURE 5.6 – HPE, HPL et HAL, enregistrement statique

5.2.2.4 Exemple de données : cas dynamique, en vol

Par analogie avec la figure 5.4 (à gauche), la figure 5.7 montre un ensemble d'enregistrements effectués en vol, lors de la campagne d'essais héliportés. Pour le cas dynamique, les trajectoires de référence (figure 5.7 à droite, points verts) sont obtenues par des techniques de positionnement différentiel RTK en s'appuyant sur la station de référence au sol d'HELILEO.

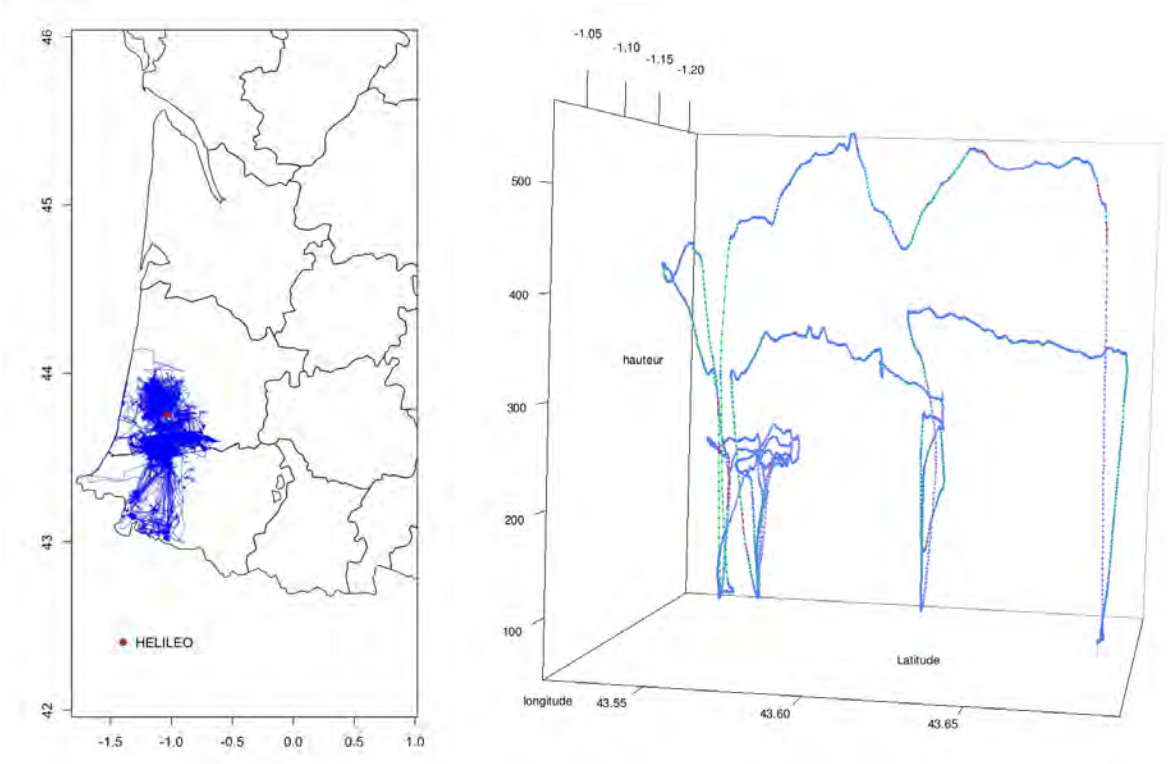


FIGURE 5.7 – Enregistrement dynamique en vol

La figure 5.8 présente les trois variables d'intérêt pour le modèle de mesure d'intégrité.

Ce processus correspond à plusieurs enregistrements effectués en vol sur le même récepteur, concaténés en un seul jeu de données. On note sur cet exemple que l'événement $x\text{PL} > x\text{AL}$ est beaucoup plus fréquent que dans le cas statique.

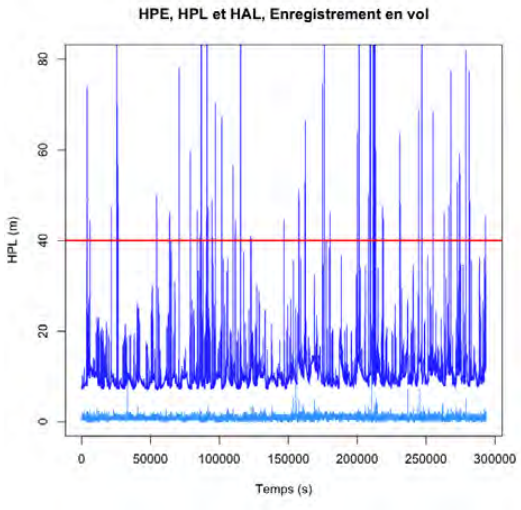


FIGURE 5.8 – HPE, HPL et HAL, enregistrements dynamiques concaténés

5.2.3 Stationnarité asymptotique des erreurs de positionnement

La question de la stationnarité des variables que l'on souhaite étudier est primordiale pour utiliser le modèle de mesure d'intégrité. En effet, la théorie des extrêmes, dont est issu ce modèle, repose sur des théorèmes asymptotiques valables pour des variables aléatoires indépendantes et identiquement distribuées. Nous avons remarqué sur les exemples présentés dans la section précédente, que les enregistrements n'apparaissaient pas à première vue comme des processus stationnaires. On considère la stationnarité d'un processus au sens faible définie par :

Définition 5.2.2. *Un processus $(x_t, t \in \mathbb{Z})$ est dit stationnaire au sens faible, ou stationnaire au second ordre si les trois conditions suivantes sont satisfaites :*

1. $\forall t \in \mathbb{Z}, \quad \mathbb{E}[x_t] = \mu$, *indépendant de* t.
2. $\forall t \in \mathbb{Z}, \quad var(x_t) = \sigma^2$, *indépendant de* t.
3. $\forall (t,h) \in \mathbb{Z}^2, \quad cov(x_t, x_{t+h}) = \mathbb{E}[(x_{t+h} - \mu)(x_t - \mu)] = \gamma(h)$, *indépendant de* t.

Sur de grands échantillons, si on regarde l'évolution temporelle des moments d'ordres 1 et 2 d'une erreur de positionnement (horizontale, cas dynamique par exemple), on voit qu'ils se stabilisent à partir d'un certain temps d'enregistrement (figure 5.9). A plus grande échelle, la courbe de droite montre le même phénomène pour un enregistrement statique de quatorze jours.
De plus, si on regarde cette évolution d'un point de vue des fonctions de densité de probabilité, on observe le même phénomène de convergence où on voit sur la figure 5.10 que les lois des erreurs tendent vers des lois stables en rouge.

On rappelle que l'on estime la densité de probabilité d'une variable aléatoire de la manière

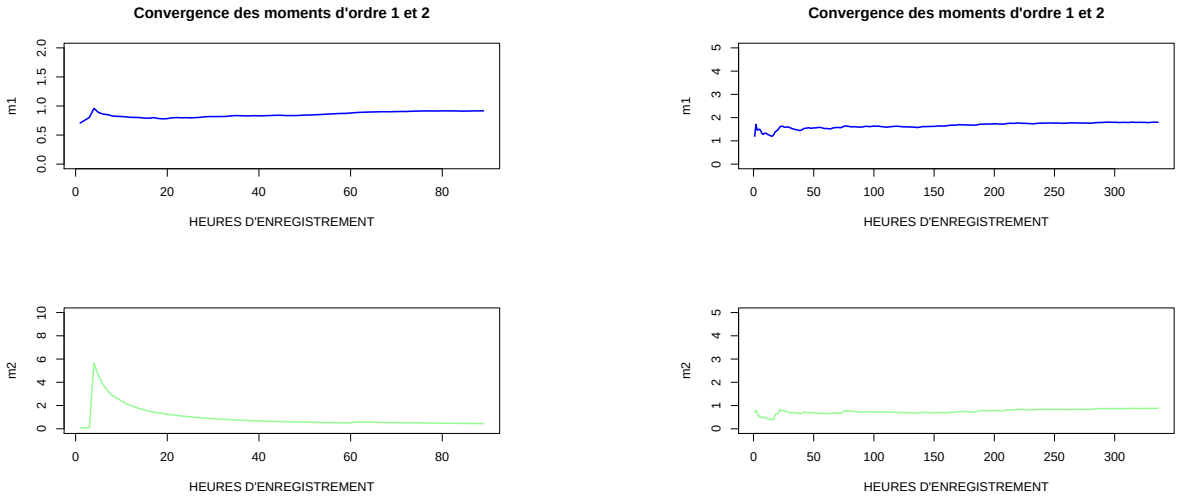


FIGURE 5.9 – Stationnarité asymptotique (a)

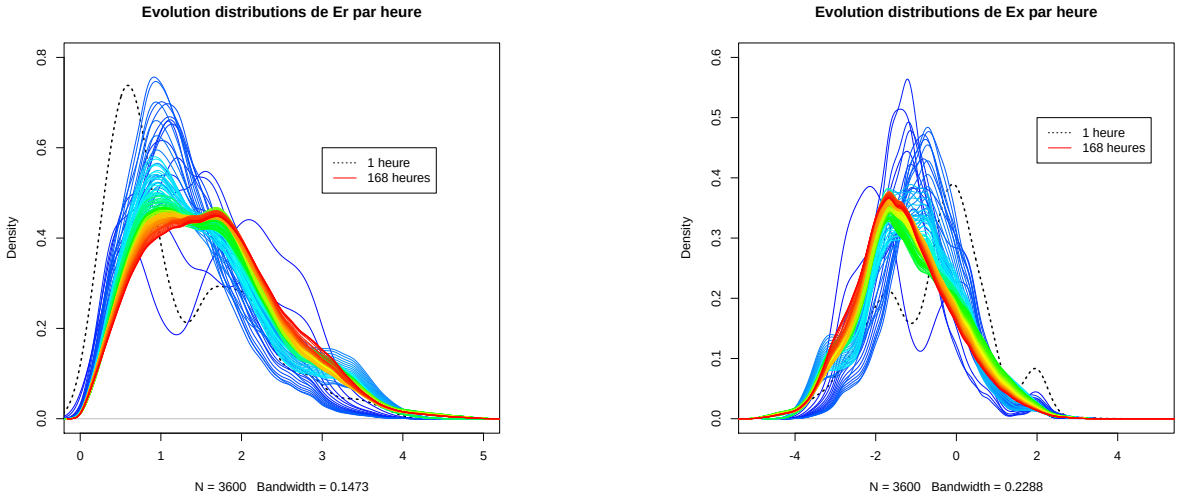


FIGURE 5.10 – Stationnarité asymptotique (b)

suivante [BC+07] :

Soit (X_n) une suite de variables aléatoires indépendantes et de même loi, de densité de probabilité f. On veut estimer f à partir d'un échantillon (X_n). On suppose f bornée, *i.e.* $\forall\, x \in \mathbb{R}, f(x) \leq c$, et f dérivable à dérivée bornée par une constante L.

Méthodes classiques : Estimation non récursive

– Estimateur de *Parzen-Rosenblatt* [Par62] :

$$\widehat{f}_n(x) = \frac{1}{nh_n} \sum_{k=1}^{n} K\Big(\frac{x - X_k}{h_n}\Big)$$

Où (h_n) est une suite de réels positifs qui tend vers 0. K noyau, fonction positive continue bornée vérifiant les conditions standards d'intégrabilité.

Méthodes récursives : Forme générale d'un estimateur récursif à noyau :

$$\widehat{f}_n(x) = (1-\gamma_n)\widehat{f}_{n-1}(x) + \frac{\gamma_n}{h_n}K\Big(\frac{x-X_n}{h_n}\Big)$$

(γ_n) et (h_n) étant deux suites de réels positifs qui tendent vers 0. K noyau. Deux exemples d'estimateurs récursifs :

– avec $\gamma_n = \frac{1}{n}$, on obtient l'estimateur de *Wolverton-Wagner* [WW69] :

$$\widehat{f}_n(x) = \frac{1}{n}\sum_{k=1}^{n}\frac{1}{h_k}K\Big(\frac{x-X_k}{h_k}\Big)$$

Cet estimateur comporte quelques similitudes avec l'estimateur antérieur de *Parzen-Rosenblatt*, notamment par l'utilisation dans les deux cas de noyaux. Cependant, là où l'estimateur de *Parzen-Rosenblatt* est un estimateur à nombre de tirages (ou taille d'échantillon) fixé, l'estimateur de *Wolverton-Wagner* permet, quant à lui, de tenir compte des tirages au fur et à mesure de leur arrivée, sans en avoir fixé le nombre à l'avance, et d'obtenir ainsi une propriété de convergence d'un estimateur enrichi autant qu'on le souhaite.
De nombreux travaux ont exploré de manière assez exhaustive l'estimation de densité non-paramétrique du type *Parzen-Rosenblatt*, affinant les propriétés de convergence de cet estimateur en fonction de la régularité de la densité à estimer. On peut consulter au sujet de l'estimation non-paramétrique de densité les ouvrages de Devroye [DG85] ou Tsybakov [Tsy09] par exemple. On trouvera des théorèmes de convergence des estimateurs à noyau récursifs dans [BC$^+$07]. Les preuves sont construites à partir de martingales, approche originale et peu souvent privilégiée pour ce genre de démonstration.

– avec $\gamma_n = \frac{1}{S_n}h_n$, on obtient l'estimateur de *Deheuvels* [DH80] :

$$\widehat{f}_n(x) = \frac{1}{S_n}\sum_{k=1}^{n}K\Big(\frac{x-X_k}{h_k}\Big)$$

où $S_n = \sum_{k=1}^{n} h_k$.

Les trois noyaux les plus couramment utilisés sont le noyau uniforme, le noyau d'*Epanechnikov* et enfin, le noyau Gaussien dont les expressions sont :

$$\begin{aligned} K_{Unif}(x) &= \frac{1}{2a}\mathbb{1}_{|x|\le a} \\ K_{Epa}(x) &= \frac{3}{4b}\Big(1-\frac{x^2}{b^2}\Big)\mathbb{1}_{|x|\le b} \\ K_{Gauss}(x) &= \frac{1}{\sqrt{2\pi}}\exp\Big(\frac{-x^2}{2}\Big) \end{aligned}$$

5.2.4 Erreurs de positionnement et domaine d'attraction de Gumbel

Dans un cadre d'étude du comportement des valeurs extrêmes, cette section a pour but de montrer que les erreurs de positionnement auxquelles on s'intéresse (plan horizontal et axe vertical) appartiennent au domaine d'attraction de Gumbel.

D'après la littérature, il est souvent dit que les erreurs de positionnement sur chaque axe, doivent suivre une loi normale centrée [KH06], [PS96]. Ceci n'est pas toujours vrai en réalité. Prenons pour exemple l'erreur sur l'axe vertical. On observe sur la figure 5.11 qu'en effet, la distribution de l'erreur considérée est proche d'une loi normale. Cependant, elle n'est pas tout à fait centrée sur 0 (certains cas sont plus flagrants) et on voit sur la figure de droite, que l'adéquation avec le modèle gaussien n'est plus valable pour les queues de distribution. C'est principalement ces zones qui nous intéressent dans l'étude des valeurs extrêmes.

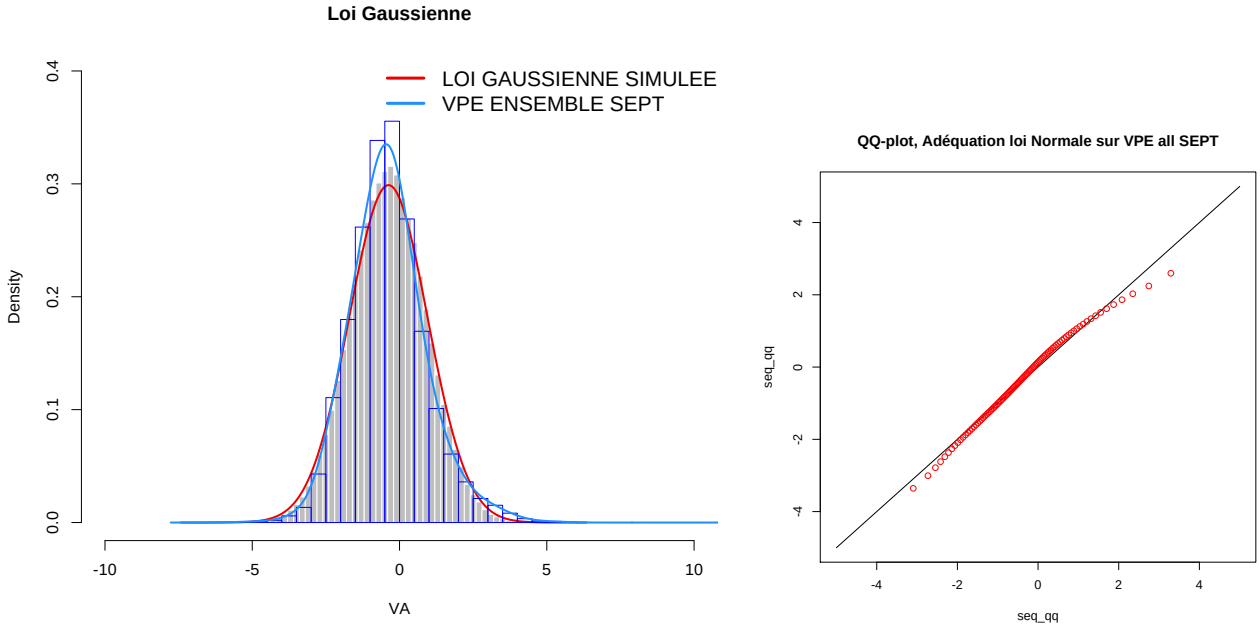


FIGURE 5.11 – Adéquation loi Normale

En ce qui concerne les erreurs horizontales, elles sont distribuées selon une loi proche d'une loi de Rayleigh de paramètre ρ [Far08], où ρ est l'écart type de la variable considérée recentrée sur 0 (figure 5.12).

Définition 5.2.3. *Soit $(X_n), n \in \mathbb{N}$ une suite de variables aléatoires distribuées suivant une loi de Rayleigh de paramètre ρ. La fonction densité de probabilité de cette loi est donnée par :*

$$f_X(x) = \frac{x}{\rho^2} \exp\left(-\frac{1}{2}\frac{x^2}{\rho^2} \right) \mathbb{1}_{x \geq 0} \tag{5.6}$$

Soit l'erreur horizontale HPE définie tel que précédemment :

$$\text{HPE}(t) = \sqrt{(X^*(t) - X(t))^2 + (Y^*(t) - Y(t))^2}$$

On écrit cette erreur en fonction des erreurs sur l'axe X, E_X, et sur l'axe Y ,E_Y :

$$\text{HPE}(t) = \sqrt{(E_X(t))^2 + (E_Y(t))^2} \tag{5.7}$$

On montre que la loi de Rayleigh décrit bien la densité de probabilité du module d'un vecteur isotrope gaussien.

Démonstration. Soit (X, Y) un vecteur formé de deux variables aléatoires gaussiennes, centrées et indépendantes. Pour un peu plus de généralité on commence par supposer que leurs variances ne sont pas égales avant de se réduire au cas isotrope. On a donc $X \sim N(0, \sigma_X)$ et $Y \sim N(0, \sigma_Y)$. X et Y étant supposées indépendantes, leur densité jointe se réduit à :

$$f_{X,Y}(x, y) = f_X(x) \times f_Y(y)$$

On a alors :

$$\begin{aligned} f_{X,Y}(x, y) &= \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\Big(-\frac{1}{2}\frac{x^2}{\sigma_X^2}\Big) \times \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp\Big(-\frac{1}{2}\frac{y^2}{\sigma_Y^2}\Big) \\ &= \frac{1}{2\pi\ \sigma_X\sigma_Y} \exp\Big(-\frac{1}{2}\big(\frac{x^2}{\sigma_X^2} + \frac{y^2}{\sigma_Y^2}\big)\Big) \end{aligned}$$

Dans le cas particulier d'un vecteur isotrope gaussien ($\sigma_X = \sigma_Y$) :

$$f_{X,Y}(x, y) = \frac{1}{2\pi\ \sigma_X^2} \exp\Big(-\frac{1}{2}\frac{x^2 + y^2}{\sigma_X^2}\Big)$$

On considère maintenant la variable aléatoire réelle positive définie par :

$$R = (X^2 + Y^2)^{\frac{1}{2}}$$

On cherche alors la loi de R définie par la fonction de répartition. Pour tout $r \geq 0$,

$$F_R(r) = \mathbb{P}(R \leq r) = \mathbb{P}(X^2 + Y^2 \leq r^2)$$

ou encore,

$$F_R(r) = \int\int_{x^2+y^2\leq r^2} f_{X,Y}(x, y)dxdy$$

On considère alors le changement de variables en coordonnées polaires, défini par la fonction :

$$h : \begin{cases} \mathbb{R}^{*2} & \to \mathbb{R}^{*+} \times [0, 2\pi[\\ (x, y) & \longmapsto (r,\theta\,) \end{cases}$$

et

$$\begin{cases} x &= r\cos\theta \\ y &= r\sin\theta \end{cases}$$

L'application h est bijective de $\mathbb{R}^{*2}$ dans $\mathbb{R}^{*+}\times[0,2\pi[$, d'inverse $h^{-1}(r,\theta\,) = (r\cos\theta, r\sin\theta)$ et de jacobien :

$$J_{h^{-1}} = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r$$

d'où, en appliquant la formule du changement de variables et le théorème de *Fubini*,

$$F_R(r) = \frac{1}{2\pi\ \sigma_X^2}\int_0^r \rho\exp\Big(-\frac{1}{2}\frac{\rho^2}{\sigma_X^2}\Big)d\rho\int_0^{2\pi} d\theta$$

On extrait finalement la densité de probabilité marginale de la variable aléatoire R :

$$f_R(r) = \frac{r}{\sigma_X^2}\exp\Big(-\frac{1}{2}\frac{r^2}{\sigma_X^2}\Big)\mathbb{1}_{r\geq 0} \tag{5.8}$$

Cette densité de probabilité représente la loi de Rayleigh (de paramètre σ_X) d'un vecteur bivarié gaussien isotrope. □

La loi de Rayleigh est une loi issue d'une loi de Weibull elle même issue d'une loi gamma généralisée [Str64]. On trouvera dans cet article différents estimateurs du paramètre ρ ainsi que les intervalles de confiances associés. Cette loi est couramment utilisée pour modéliser des erreurs exprimées sous forme de distance puisque elle apparaît comme la norme d'un vecteur aléatoire gaussien bidimensionnel (dont les composantes sont indépendantes, centrées et de même variance).
Notons que si le vecteur gaussien est légèrement anisotrope (*ie.* $\sigma_X \approx \sigma_Y$), on pourrait exprimer la loi de R sous la forme :

$$f_R(r) \approx \frac{1}{\sigma_X\ \sigma_Y}\ r\exp\Big(-\frac{r^2}{\sigma_X^2+\sigma_Y^2}\Big)$$

On retrouve exactement la loi de Rayleigh dans le cas où $\sigma_X = \sigma_Y$.

On a simulé une loi de Rayleigh de paramètre ρ à partir d'une loi uniforme sur [0.1].

Proposition 5.2.1. *Soit U une variable aléatoire réelle distribuée suivant une loi uniforme sur* [0.1]. *Alors, la variable R telle que :*

$$R = \rho\sqrt{-2\ln(U)} \quad avec \quad U \sim U([0,1])$$

suit une loi de Rayleigh de paramètre ρ. On pourra retenir que $R(\sigma) \sim W(2,\lambda)$ avec $\sigma = \frac{1}{2\lambda}$, où $W(\beta,\lambda)$ est une loi de Weibull de paramètres β et λ.

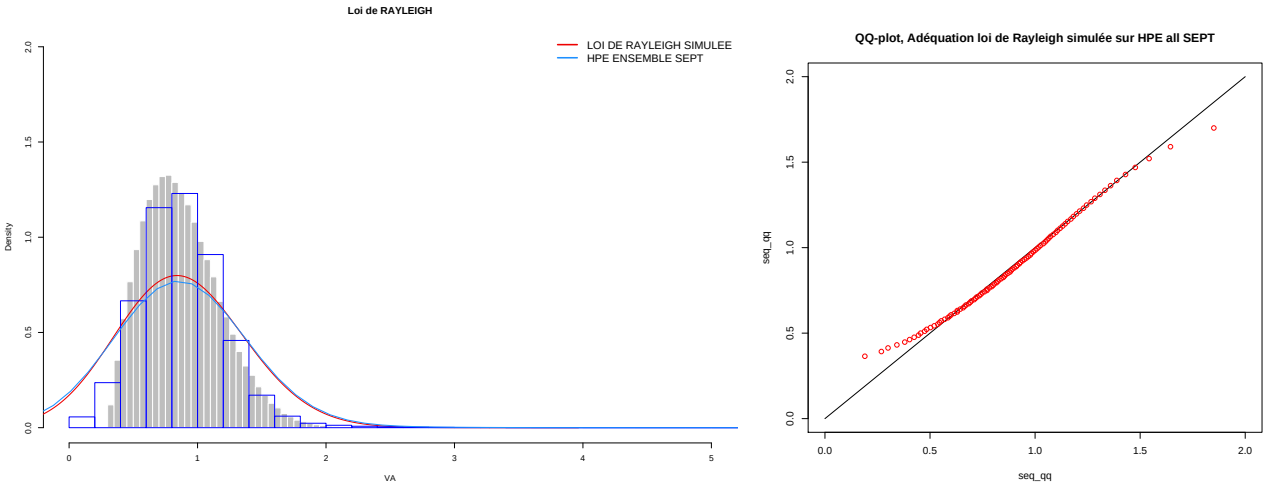


FIGURE 5.12 – Adéquation loi de Rayleigh

La figure 5.13 met en évidence de façon plus prononcée la non adéquation des modèles paramétriques pour les queues de distribution (loi de Rayleigh à gauche et loi normale à droite). Il n'apparaît donc pas raisonnable comme mentionné dans l'introduction générale de ce manuscrit, de considérer une méthode paramétrique pour estimer les quantiles intervenant dans notre modèle de mesure d'intégrité.

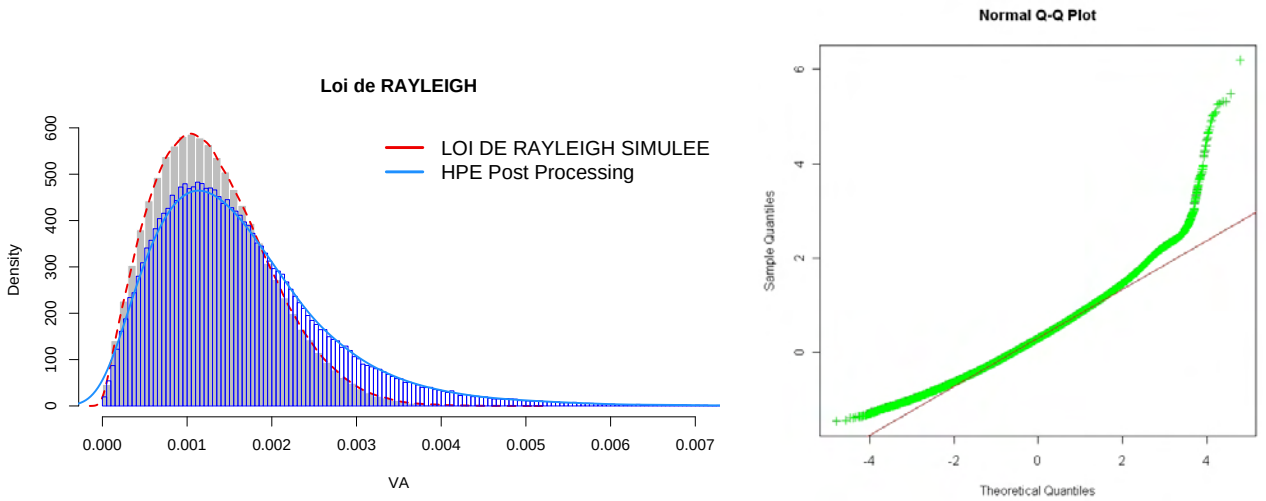


FIGURE 5.13 – Epaisseur des queues de distribution

On appliquera la méthodologie POT présentée dans le chapitre dédié aux valeurs extrêmes pour estimer les quantiles du modèle de mesure d'intégrité. Nous avons établi deux estimateurs de quantiles extrêmes dans le cadre de la méthode POT. Le premier est l'estimateur directement obtenu par inversion de la loi des excès et le second est valable pour le cas où l'indice des valeurs extrêmes γ est nul ; ou de façon équivalente, lorsque la loi des observations appartient au domaine d'attraction de Gumbel. Il est connu que la loi normale appartient au domaine d'attraction de Gumbel [EKM97]. Pour la loi de Rayleigh, il faut le vérifier. Une indication provient

du fait que la loi de Rayleigh est un cas particulier d'une loi Gamma généralisée définie par [Str64] :

$$f(x) = \frac{\alpha}{b^{\frac{\beta+1}{\alpha}}\,\Gamma(\frac{\beta+1}{\alpha})}\; x^{\beta}\,\exp\big(-\frac{x^{\alpha}}{b}\big) \quad \alpha > 0\ ,\ \beta > -1\ ,\ 0 < x < \infty \tag{5.9}$$

dont les cas particuliers sont les suivants :

1. Lorsque $\alpha = 1$, $\beta = \rho - 1$ pour $\rho > 0$: distribution gamma. Particulièrement lorsque $\beta = 0$ on obtient la distribution exponentielle.
2. Lorsque $\alpha = 2$, $\beta = 2$: distribution de Maxwell.
3. Lorsque $\alpha = m$, $\beta = m - 1$: distribution de Weibull dont la distribution de Rayleigh est un cas particulier lorsque $\alpha = 2$.

La loi Gamma fait elle même partie du domaine d'attraction de Gumbel. On rappelle la condition suffisante d'appartenance à ce domaine d'attraction énoncée par Gnedenko (1943) [Gne43] :

Théorème 5.2.1. *Si F est de classe $\mathcal{C}^2$, F appartient au domaine d'attraction de Gumbel si*

$$\lim_{x\to\infty} \frac{(1 - F(x))\partial^2 F(x)}{(\partial F(x))^2} = -1 \tag{5.10}$$

Proposition 5.2.2. *La loi de Rayleigh appartient au domaine d'attraction de Gumbel.*

Démonstration. Soit (X_n) une suite de variables aléatoires distribuées selon une loi de Rayleigh de paramètre σ dont la fonction de répartition est définie par :

$$F(x) = 1 - \exp(-\frac{x^2}{2\rho^2}) \tag{5.11}$$

Afin de vérifier l'appartenance de la loi de Rayleigh au domaine d'attraction de Gumbel on calcule le rapport énoncé dans le théorème 5.2.1. On a :

$$\partial F(x) = \frac{1}{\rho^2} x \exp(-\frac{x^2}{2\rho^2})$$

et,

$$\partial^2 F(x) = \frac{1}{\rho^2}\exp(-\frac{x^2}{2\rho^2}) - \frac{x^2}{\rho^4}\exp(-\frac{x^2}{2\rho^2})$$

donc,

$$\frac{(1 - F(x))\partial^2 F(x)}{(\partial F(x))^2} = \frac{\exp(-\frac{x^2}{2\rho^2})\big[\frac{1}{\rho^2}\exp(-\frac{x^2}{2\rho^2}) - \frac{x^2}{\rho^4}\exp(-\frac{x^2}{2\rho^2})\big]}{\big(\frac{1}{\rho^2} x \exp(-\frac{x^2}{2\rho^2})\big)^2} = \frac{\rho^2}{x^2} - 1$$

On a bien la limite qui vaut -1 quand x tend vers l'infini et la condition est vérifiée. □

On pourra donc utiliser l'estimateur de quantile extrême dans le cas Gumbel pour notre modèle d'analyse d'intégrité.

5.3 Procédure automatique de choix du seuil u

On souhaite appliquer le modèle de mesure d'intégrité aux données que nous venons de décrire. Le problème de mesure d'intégrité est ramené à un problème d'estimation de quantile associé à des niveaux de probabilité très élevés. Cette raison nous pousse à utiliser les estimateurs issus de la théorie des extrêmes présentés au chapitre 2. On préfèrera la méthodologie POT (Peak Over Threshold) à la méthode bloc maxima. En effet, nous avons montré au chapitre 2 que la méthode POT était mieux adaptée au cadre de cette étude. Les paramètres de la loi des excès, ainsi que le quantile recherché, sont estimés en fonction d'un seuil u. Les exemples du chapitre consacré à l'étude des valeurs extrêmes ont mis en évidence l'impact du seuil u sur la variabilité des estimateurs des paramètres, des quantiles extrêmes ainsi que leurs intervalles de confiance associés. En effet, les valeurs théoriques des quantiles recherchés étaient comprises dans les intervalles de confiances pour des plages de seuils assez restreintes. Ainsi, le choix du seuil u est une étape déterminante pour l'ajustement du modèle de la loi des excès aux données. Jusqu'ici un "contrôle humain" était nécessaire pour dégager des zones de stabilité des estimateurs exprimés en fonction du seuil u, afin de choisir un niveau u satisfaisant la bonne adéquation du modèle aux données.

L'objectif final de cette étude est de développer un outil fonctionnel évitant tout contrôle humain durant les analyses de performances des matériels. De plus, le contexte d'une plateforme d'essais comme HELILEO, implique un très grand nombre d'échantillons de données à analyser. Il peut alors devenir très couteux en terme de temps, d'effectuer "à la main", une analyse de performances (positionnement ou intégrité) sur l'ensemble des données à disposition. Cette partie a pour but de fournir une procédure automatique de choix de seuil construite à partir de l'estimation des paramètres de la loi GPD et des quantiles extrêmes, en fonction du seuil.

5.3.1 Génération d'une séquence de seuils propres aux données

Le premier contrôle humain est le choix de la plage de seuils à parcourir pour les estimation des paramètres de la loi des excès γ et σ ainsi que du quantile recherché $q(p)_{POT}$ ou $q(p)_{\gamma=0}$. On utilise la technique consistant à chercher des valeurs de u pour lesquelles les estimateurs $\hat{\gamma}$, $\hat{\sigma}$ et $\widehat{q(p)}_{POT}$ sont stables. Cependant, afin d'appliquer cette procédure à différents scénarios (erreurs horizontales ou verticales, différents matériels, etc), on génère automatiquement une séquence de seuils u_{auto}. On construit un critère générique en considérant la valeur maximale $X_{n,n}$ de l'échantillon et en retenant la statistique d'ordre telle que 5% des données soient supérieures à cette valeur. Il s'agit du quantile de niveau $p = 1 - 0.05$ que l'on note $X_{0.95.n,n}$ et qui est estimé de façon empirique (annexe A, section A.1). On choisit de conserver dix valeurs au dessus du seuil maximal noté $X_{n-10,n}$ de manière à ne pas bloquer la procédure par manque d'observations au dessus du seuil maximal. La séquence de seuils générés parcours une plage de valeurs allant de $X_{0.95.n,n}$ à $X_{n-10,n}$. Les figures 5.14a et 5.14b présentent des séquences ayant un pas de 0.1. Ce faible pas implique un coût calculatoire important pour la procédure, et ne présente pas de meilleur résultat qu'avec un pas plus large ; on fixe le pas à 0.5.

On illustre les différentes phases de la procédure par un exemple transverse aux sections qui vont suivrent. On simule un échantillon de 360 000 observations distribuées selon une loi de Rayleigh de paramètre $\rho = 2$.

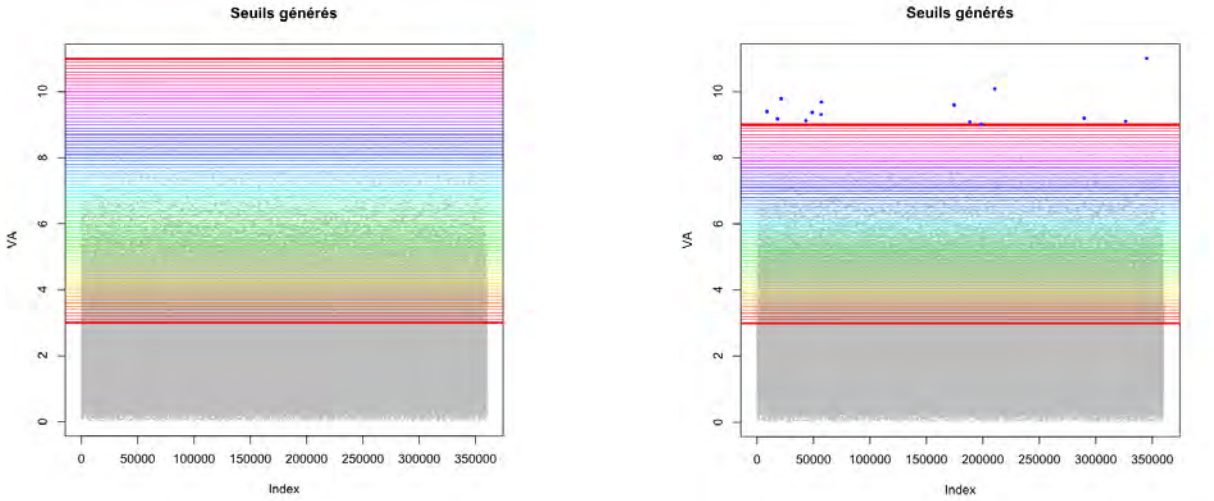


FIGURE 5.14 – Génération automatique d'une séquence de seuils u

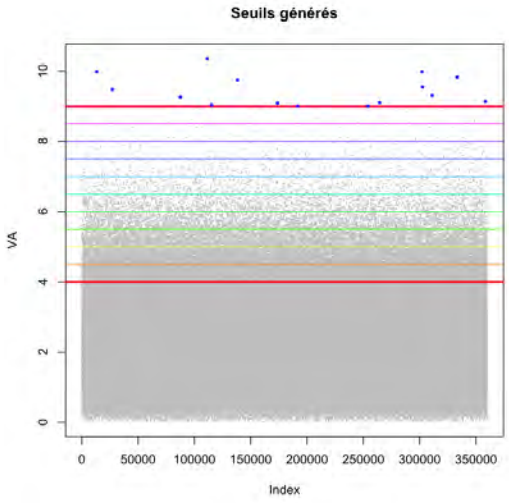


FIGURE 5.15 – Séquence de seuils u, pas $= 0.5$

5.3.2 Sélection d'une plage de stabilité

Pour un échantillon donné, les paramètres de la loi des excès sont estimés en fonction d'une plage de seuils u. Afin de répondre aux problématiques d'analyse de performances (en terme de position ou d'intégrité), on doit fournir une valeur d'un quantile ainsi que son intervalle de confiance associé. Il faut alors trouver le seuil u parmi les seuils testés, qui mènera à la meilleure adéquation du modèle (GPD) aux données. La seconde brique de la procédure de choix de seuil présentée dans cette

section est destinée à la détection de plages de stabilité, continues ou discontinues, au sein des estimateurs des paramètres de la loi des excès ainsi que sur l'estimateur de quantile retenu.
On rappelle les expressions des estimateurs de quantile extrême $\widehat{q}(p)_{\text{POT}}$ et $\widehat{q}(p)_{\gamma=0}$:

$$\widehat{q}(p)_{\text{POT}} = X_{n-n_e,n} + \frac{\widehat{\sigma}}{\widehat{\gamma}}\left[\left(\frac{n_e}{p\,n}\right)^{\widehat{\gamma}} - 1\right] \tag{5.12}$$

$$\widehat{q}(p)_{\gamma=0} = X_{n-n_e,n} + \widehat{\sigma}\log\left(\frac{n_e}{np}\right) \tag{5.13}$$

où les paramètres de la loi des excès γ et σ sont estimés par la méthode du maximum de vraisemblance qui a montré de bonnes propriétés sur les exemples théoriques du chapitre 2.
La sélection d'une plage de stabilité est obtenue grâce à un algorithme développé à cet effet. Cette algorithme utilise un calcul de gradient pour l'estimateur de quantile extrême $\widehat{q}(p)_{POT}$ (ou $\widehat{q}(p)_{\gamma=0}$ dans le cas Gumbel) en fonction de $u(p_{obs})$, pour ensuite stocker les valeurs de $u(p_{obs})$ pour lesquelles le gradient est proche de 0. Toutefois, seront stockées uniquement les valeurs formant une séquence continue d'au moins trois valeurs. Ceci atteste que l'algorithme a décelé une zone de stabilité et non une valeur isolée. Plusieurs séquences discontinues de ce type peuvent être détectées. On obtient ainsi N_u valeurs de $u(p_{obs})$ retenues.
La stabilité des estimateurs $\widehat{\gamma}$, $\widehat{\sigma}$ sur la séquence retenue $u(p_{obs})^i, i = 1, \cdots, N_u$ sera ensuite testée de la même manière par souci de validation. On précise que cette procédure est valable pour les différents estimateurs de γ, σ et $q(p)$. Reste maintenant à déterminer le seuil optimal $u(p_{obs})^{opt}$ pour lequel, la loi GPD sera la mieux ajustée aux données considérées.

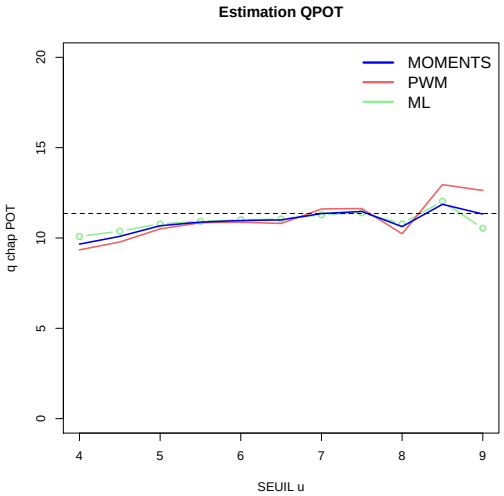


FIGURE 5.16 – Séquence de seuils u, pas $= 0.5$

On remarque sur la figure 5.16 un plateau de stabilité pour les trois estimateurs dans la partie centrale de la courbe. L'amplitude des valeurs des seuils retenus peut paraître large mais pour des échantillons de très grandes tailles, les conditions d'un

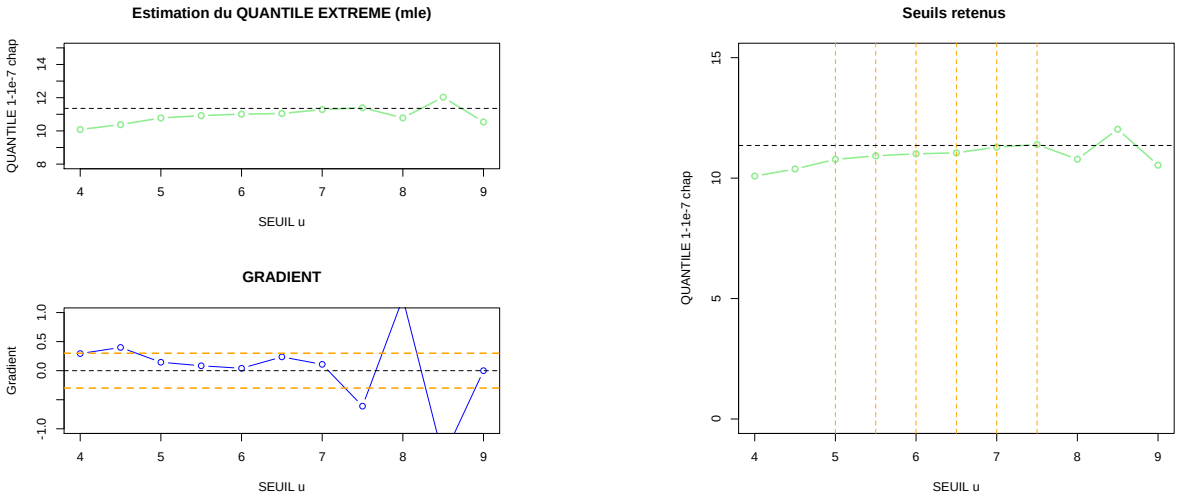


FIGURE 5.17 – Sélection automatique d'une plage de stabilité

u suffisamment grand pour assurer les propriétés asymptotiques du modèle, sont remplies. On retient pour cet exemple l'estimateur du maximum de vraisemblance qui a montré de bonnes propriétés sur les exemples précédents.
La figure 5.17 (à droite) permet d'identifier la plage de seuils retenus, satisfaisant la condition suivante (figure 5.17 en bas à gauche) :

$$\frac{\partial}{\partial p_{obs}} \widehat{q}(p_{obs}) = 0 \pm \delta \tag{5.14}$$

On fixe le paramètre δ à 0.25. Il pourra être modifié en cas d'un nombre trop faible de seuils retenus. Reste à trouver parmi ces seuils, lequel correspond au modèle GPD le mieux ajusté aux observations.

5.3.3 Adéquation du modèle aux observations et sélection de seuil

Les N_u seuils retenus vont permettre de former autant de modèles GPD. Soit p_j, $j \in \{1, \cdots, N_p\}$ une suite de N_p probabilités comprises entre 0.9 et 1. N_p sera fonction du pas choisi pour la séquence.

Remarque : L'adéquation du modèle est évaluée dans les zones qui nous intéressent : les queues de distributions. Pour de gros échantillons, on considère qu'en dessous d'un niveau de probabilité $p = 0.9$ les observations sont suffisamment nombreuses pour que l'ajustement aux observations se fasse à l'aide de l'estimateur du quantile empirique. On va alors combiner deux modèles pour caractériser la totalité de la loi des observations. Il serait tout à fait envisageable d'utiliser un estimateur de quantile paramétrique (ex : a priori gaussien, exponentiel, Rayleigh, ...) pour ajuster la partie centrale de la loi des observations [Gar02].

La figure 5.18 présente deux ensembles de modèles testés pour deux échantillons différents injectés dans la procédure. La courbe en noir est la fonction de répartition

empirique des observations estimée entre 0 et 1.

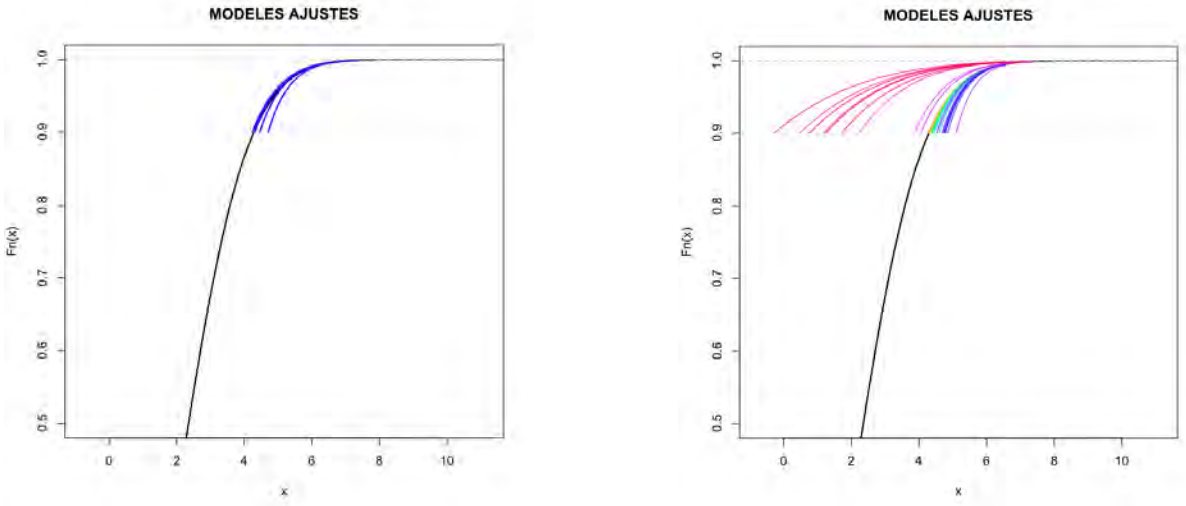


FIGURE 5.18 – Modèles générés en fonction de u

On cherche le modèle qui sera le plus proche de la loi des observations. Ainsi, le seuil optimal u^{opt} cherché est celui qui minimise au sens des moindres carrés la somme des écarts entre le quantile empirique estimé $\widehat{q}(p)_{emp}$ et le quantile estimé $\widehat{q}(p)_{POT}$ par la méthode POT, pour chaque niveau de probabilité p_j compris entre 0.9 et 1, et pour chaque seuil u^i, $i = 1, \cdots, N_u$, issu de la sélection par calcul de gradient. Le quantile empirique $q(p)_{emp}$ est estimé au moyen de la méthode non paramétrique décrite en annexe A. On a le critère d'optimisation suivant :

$$u^{opt} = \underset{u^i}{\operatorname{argmin}} \left\{ \sum_{j=1}^{N_p} \left(\widehat{q}(p_j)_{emp} - \widehat{q}^{u^i}(p_j)_{POT} \right)^2 \right\}, \quad i \in \{1, \cdots, N_u\} \tag{5.15}$$

Le modèle retenu (figure 5.19) correspond à un seuil égal à 7 pour cet exemple. La valeur du quantile associé à un niveau $p = 1 - 10^{-7}$, est estimée à $\widehat{q}^{opt}_{\text{POT}} = 11.31$ avec l'estimateur du maximum de vraisemblance. L'intervalle de confiance, de niveau $\alpha = 0.05$, est obtenu avec la méthode Delta et vaut $IC_{DM}(q^{opt}_{\text{POT}}) = [9.01 \ ; \ 13.59]$. On rappelle la valeur théorique du quantile recherché : $q_{th} = 11.35$. La procédure a aussi été implémentée pour l'estimateur de quantile $q(p)_{\gamma=0}$ dans le cas où la loi des observations appartient au domaine d'attraction de Gumbel. On obtient un résultat similaire pour le même échantillon que précédemment : le seuil retenu est 8, le quantile de niveau $1 - 10^{-7}$ est $\widehat{q}^{opt}_{\gamma=0} = 11.18$ et l'intervalle de confiance correspondant $IC_{th}(q^{opt}_{\gamma=0}) = [10.62 \ ; \ 11.74]$.

Il existe d'autres critères de choix de seuil optimaux. On trouvera dans [Bei04] et dans [DVF06] deux critères construits sur la minimisation de l'erreur quadratique moyenne asymptotique. Le premier implique l'estimation de nouveaux paramètres en fonction du seuil difficiles à sélectionner pour une procédure automatique et le second est dédié aux lois appartenant au domaine d'attraction de Fréchet.

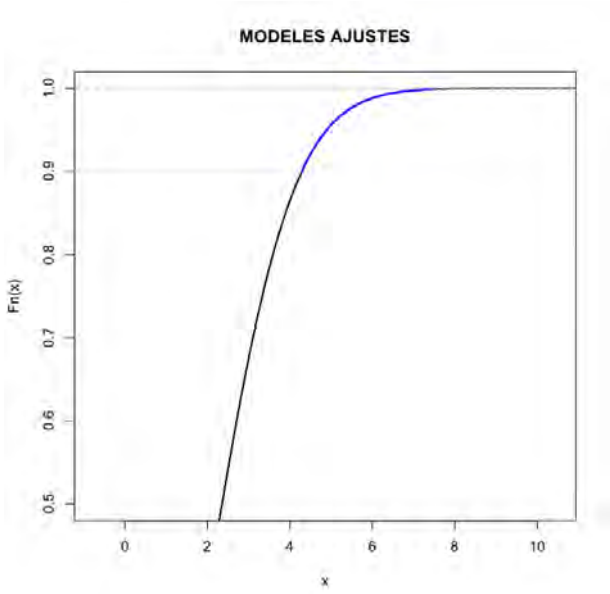


FIGURE 5.19 – Modèle retenu

Un deuxième critère de choix de modèle a été développé dans cette étude pour le cas où la loi des observations appartient au domaine d'attraction de Gumbel. On rappelle que lorsque l'on est dans ce cas particulier, la loi des excès tend vers une loi exponentielle pour de grands échantillons (*cf.* théorème 2.4.1). Ainsi on a construit un test sur l'exponentialité des excès et plus particulièrement sur la linéarité du QQ-plot (figure 5.20) obtenu en traçant les quantiles de la loi exponentielle $-\ln(j/n_e)$ en fonction des excès ordonnés $(X_{n-n_e+j,n_e} - X_{n-n_e,n_e})$, pour $j = 1, \cdots, n_e$.

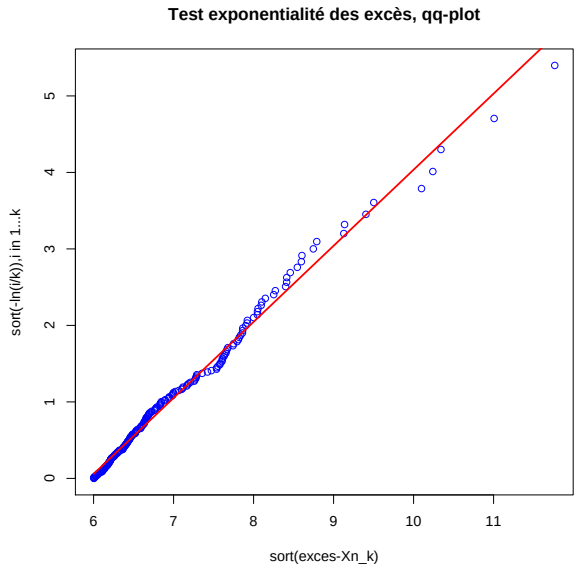


FIGURE 5.20 – Exponentialité des excès

Chaque seuil u^i engendre un nombre d'excès n_e^i pour $i = 1, \cdots, N^u$. Les n_e^i pren-

dront leurs valeurs dans $[10, n_e^{\max}]$. On cherche alors le nombre d'excès n_e^{opt} correspondant au seuil u^{opt} tel que :

$$n_e^{opt} = \underset{n_e^i}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N_u} \Big(-\ln(\frac{j}{n_e^i}) - (X_{n-n_e^i+j,n_e^i} - X_{n-n_e^i,n_e^i}) \Big)^2 \right\} \tag{5.16}$$

pour $i \in \{1, \cdots, N_u\}$ et $j \in \{1, \cdots, n_e\}$.

Sur un tirage différent de l'exemple précédent, toujours distribué suivant une loi de Rayleigh de paramètre 2, le seuil retenu est $u^{opt} = 8$, le quantile de niveau $1-10^{-7}$ est $\widehat{q}_{\gamma=0}^{opt} = 11.75$ et l'intervalle de confiance correspondant $IC_{th}(q_{\gamma=0}^{opt}) = [11.06\ ;\ 12.43]$.

Le paragraphe suivant présente les performances des procédures décrites dans cette section.

5.3.4 Evaluation de la procédure

On souhaite évaluer les performances de la procédure d'ajustement décrite précédemment. Pour cela, on réalise $l = 100$ tirages différents d'un échantillon de données sur lequel on va appliquer la procédure afin d'estimer un quantile extrême. On fixe le niveau de quantile recherché à $p = 1-10^{-7}$ comme dans les exemples de la section précédente. On évalue la procédure pour les deux estimateurs $q(p)_{\mathrm{POT}}$ et $q(p)_{\gamma=0}$. On retiendra comme résultats les l valeurs des $\widehat{\gamma}$ et $\widehat{q}$ estimées, dont les valeurs théoriques sont connues, ainsi que les intervalles de confiances (obtenus par la méthode Delta) associés aux quantiles. Le niveau des intervalles est fixé à $\alpha = 0.05$. On analysera les seuils u retenus ainsi que la probabilité de réussite de la procédure. On considère que l'estimation est un succès lorsque la valeur théorique q_{th} du quantile recherché est contenue dans l'intervalle de confiance, ce qui n'est pas toujours le cas (voir chapitre 2 section 2.3.3). On réalise ce test pour trois lois appartenant au domaine d'attraction de Gumbel : une loi de Rayleigh, une loi normale et une loi exponentielle.

5.3.4.1 Sur une loi de Rayleigh de paramètre 2

Estimateur POT :

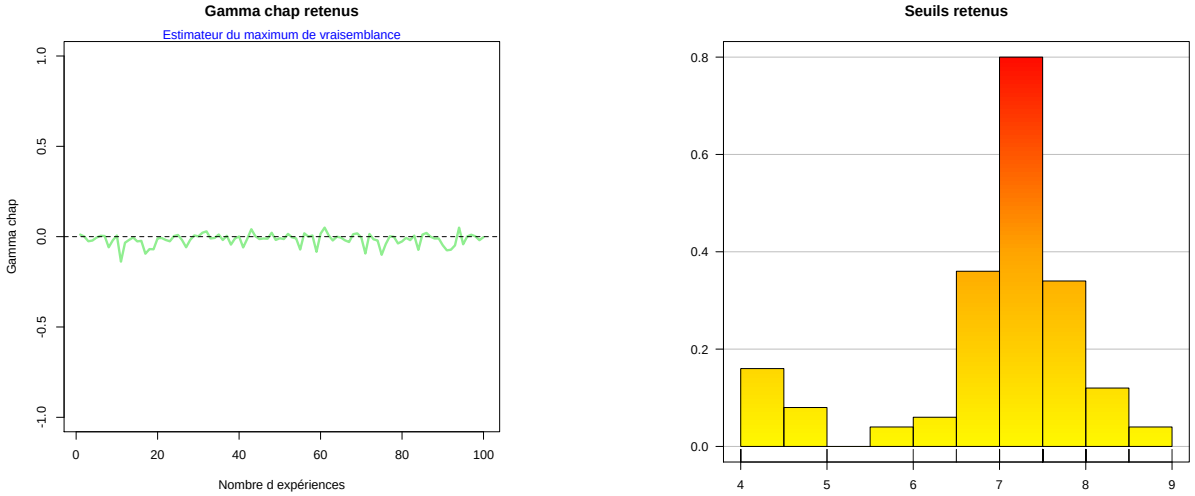


FIGURE 5.21 – Paramètre $\widehat{\gamma}$ estimé sur les $l = 100$ tirages et seuils retenus

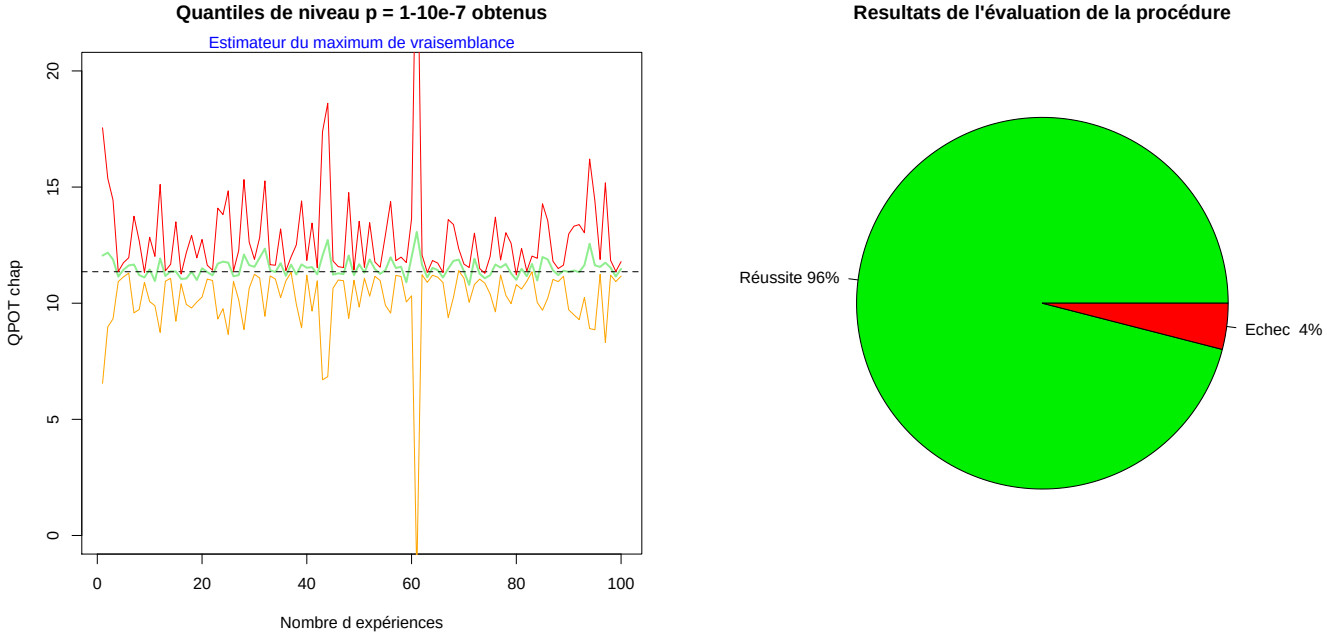


FIGURE 5.22 – Quantile $\widehat{q}(p)_{\text{POT}}$ estimé sur les $l = 100$ tirages et probabilité de succès

Estimateur dans le cas Gumbel :

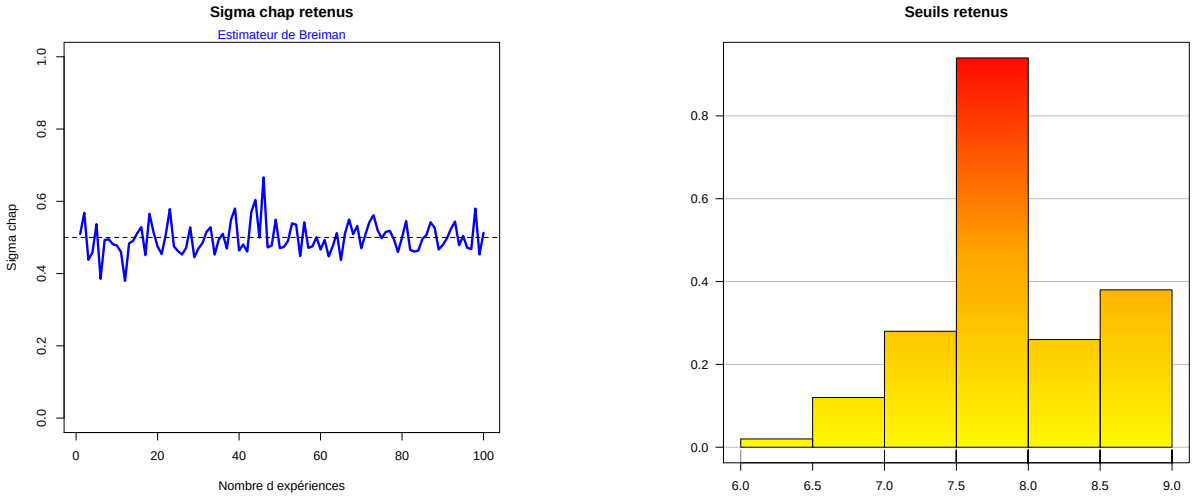


FIGURE 5.23 – Paramètre $\widehat{\sigma}$ estimé sur les $l = 100$ tirages et seuils retenus

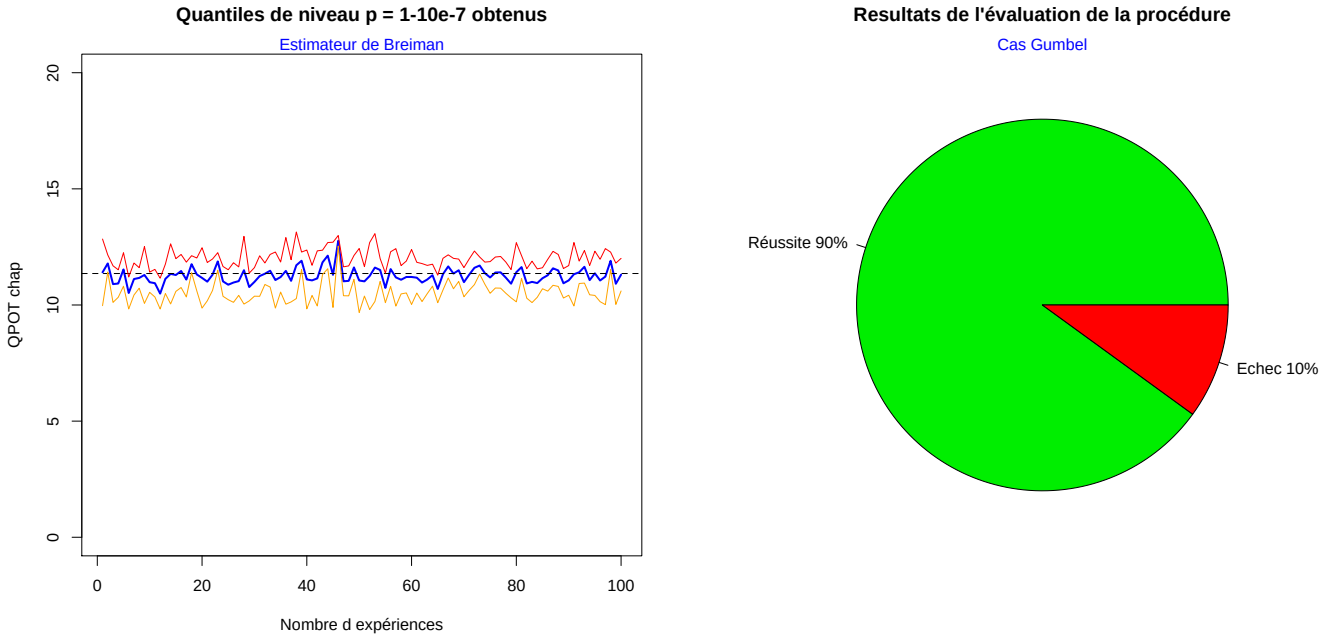


FIGURE 5.24 – Quantile $\widehat{q}(p)_{\gamma=0}$ estimé sur les $l = 100$ tirages et probabilité de succès

5.3.4.2 Sur une loi normale $N(0,1)$

Estimateur POT :

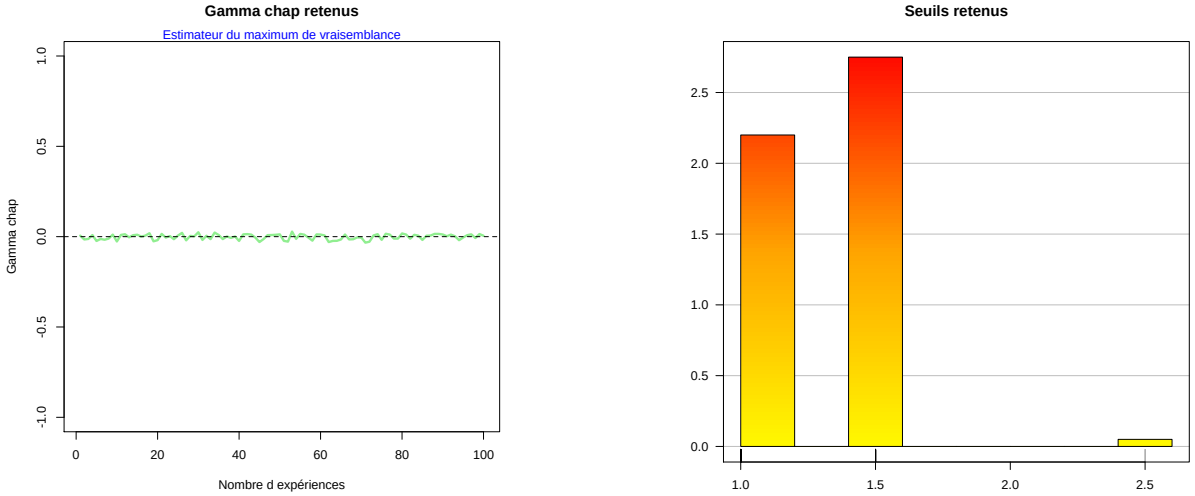


FIGURE 5.25 – Paramètre $\widehat{\gamma}$ estimé sur les $l = 100$ tirages

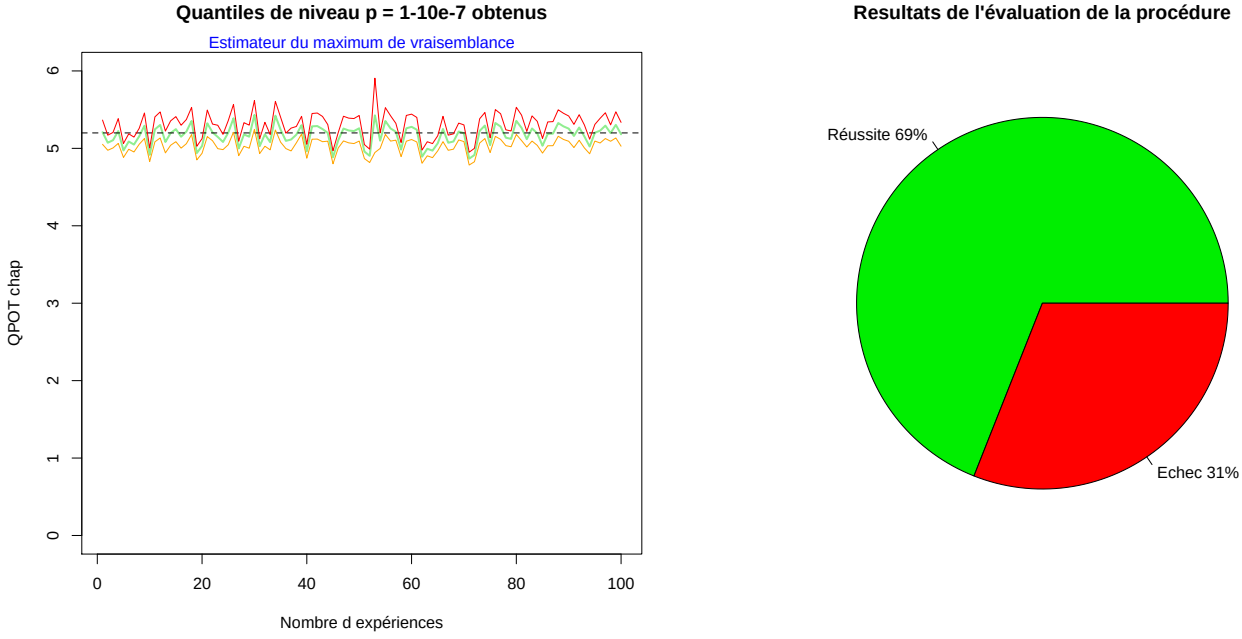


FIGURE 5.26 – Quantile $\widehat{q}(p)_{\text{POT}}$ estimé sur les $l = 100$ tirages et probabilité de succès

Estimateur Gumbel :

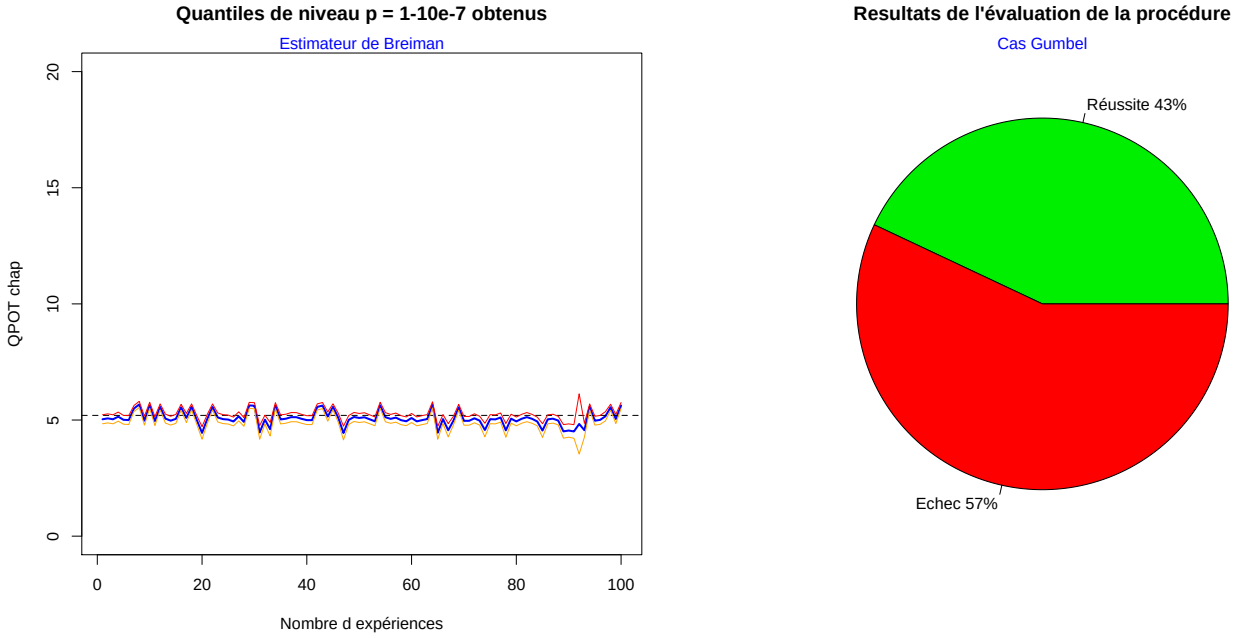


FIGURE 5.27 – Quantile $\widehat{q}(p)_{\gamma=0}$ estimé sur les $l = 100$ tirages et probabilité de succès

5.3.4.3 Sur une loi exponentielle

Estimateur POT :

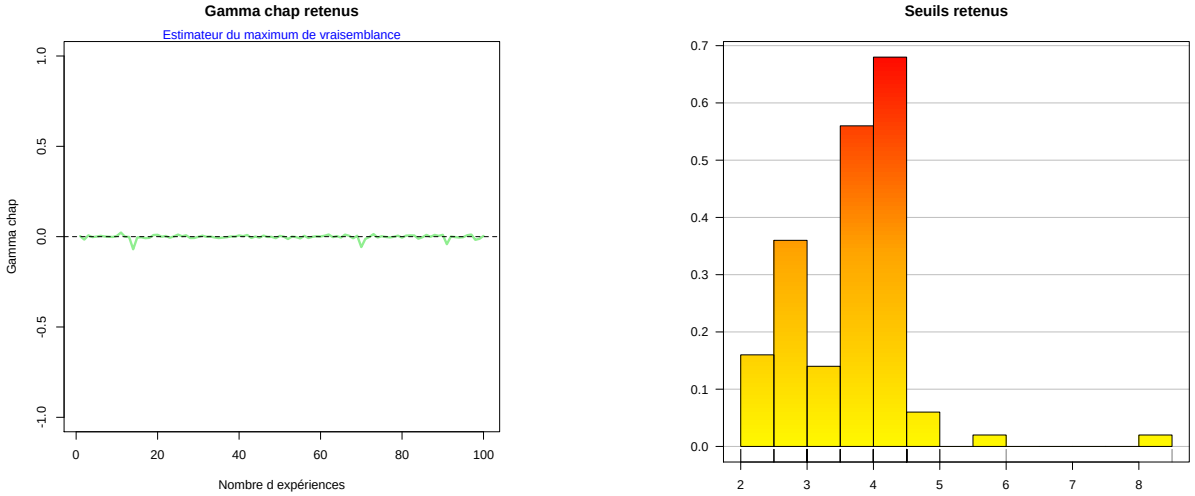


FIGURE 5.28 – Paramètre $\widehat{\gamma}$ estimé sur les $l = 100$ tirages et seuils retenus

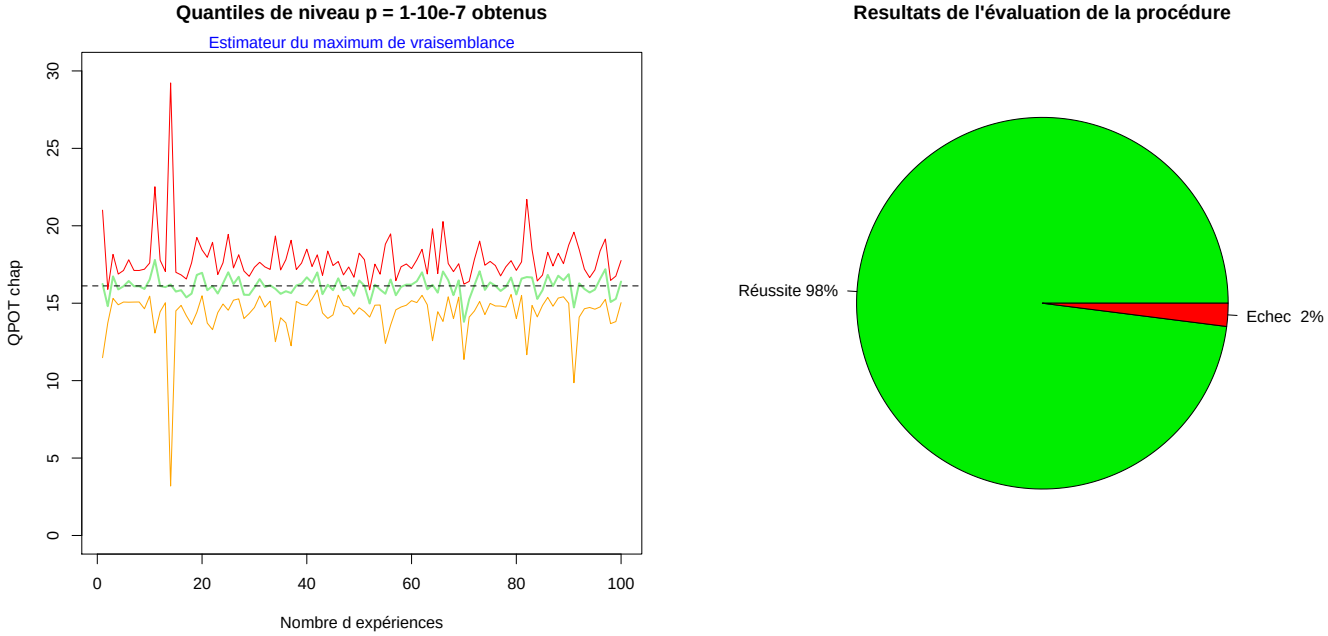


FIGURE 5.29 – Quantile $\widehat{q}(p)_{\text{POT}}$ estimé sur les $l = 100$ tirages et probabilité de succès

Estimateur Gumbel :

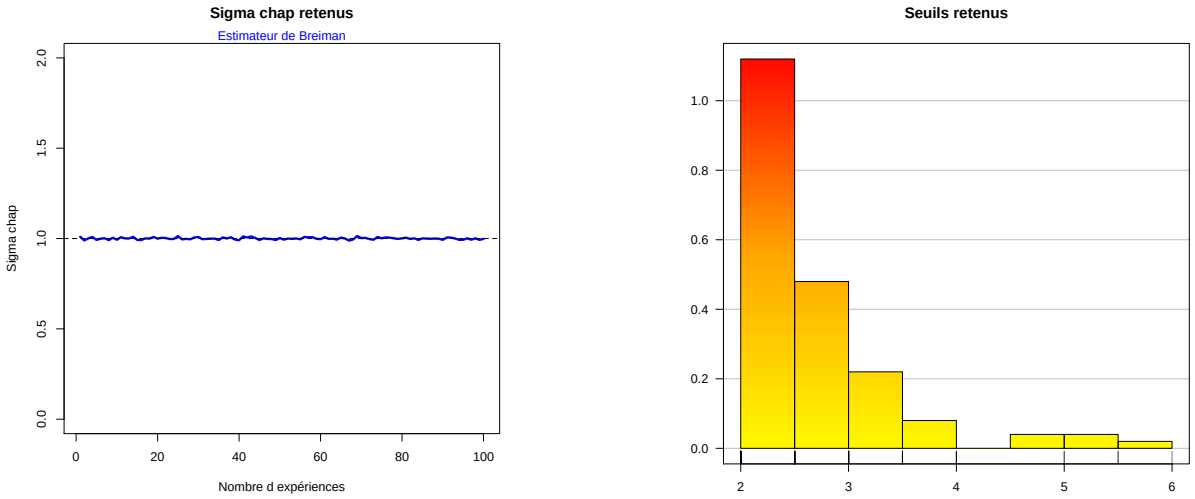


FIGURE 5.30 – Paramètre $\widehat{\sigma}$ estimé sur les $l = 100$ tirages et seuils retenus

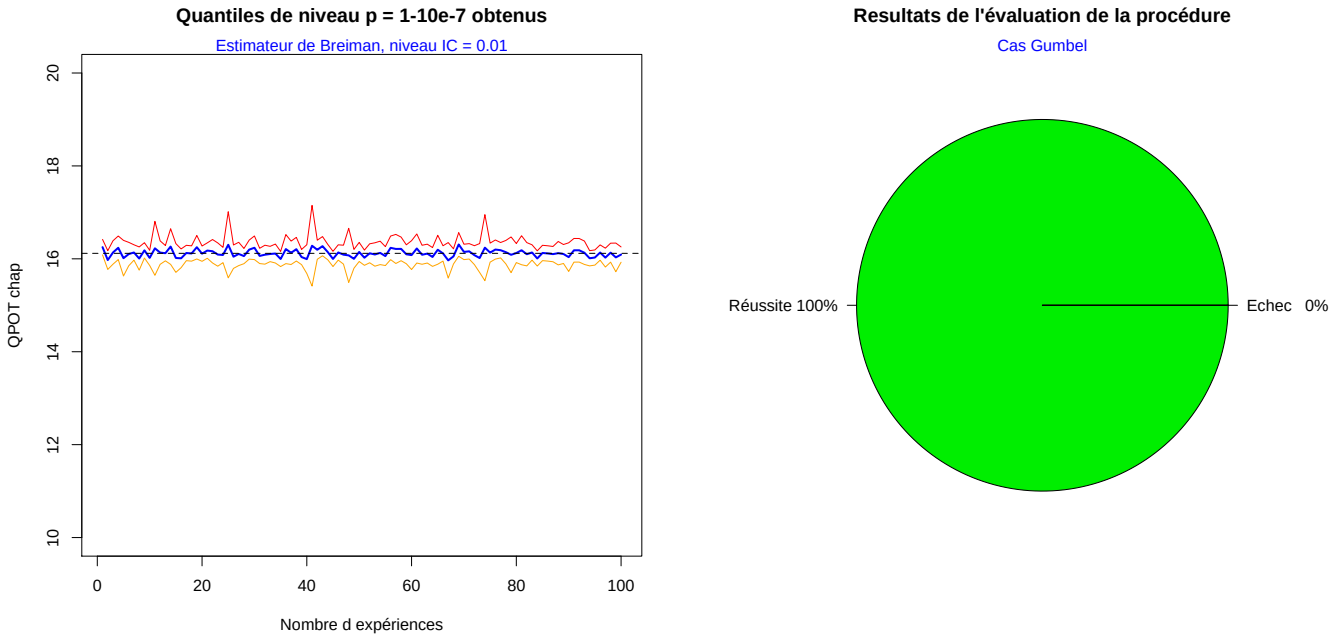


FIGURE 5.31 – Quantile $\widehat{q}(p)_{\gamma=0}$ estimé sur les $l = 100$ tirages et probabilité de succès

La procédure démontre des performances très correctes pour les lois de Rayleigh et exponentielle, que l'on utilise l'estimateur $\widehat{q}(p)_{\text{POT}}$ ou $\widehat{q}(p)_{\gamma=0}$. En revanche, pour une loi symétrique, comme la loi normale, les estimations sont plus sensibles et les performances ne sont pas acceptables pour utiliser cette procédure de façon autonome. Cependant, si on observe les courbes des figures 5.26 et 5.27 attentivement, on note que les estimations évoluent non loin des valeurs théoriques. La non réussite du test provient alors du fait que les intervalles de confiance sont trop étroits.

Pour pallier cette carence de réussite, on choisit d'étudier les valeurs absolues des variables aléatoires distribuées selon des lois symétriques. Dans ce cas, le pourcentage de réussite passe à 92% au lieu de 69%. Afin d'augmenter les performances de réussite de la procédure, on peut élargir les intervalles de confiance en fixant le niveau à $\alpha = 0.01$. Ainsi, la valeur théorique du quantile recherché aura plus de chance d'être contenu dans l'intervalle et le pourcentage de réussite s'en verra augmenté.
Cette procédure pourra être utilisée pour l'estimation de quantile extrême présent dans le modèle de mesure d'intégrité. De plus, elle a fait l'objet d'un développement logiciel au sein d'une plate forme d'analyse de données qui sera présentée en fin de chapitre.

5.4 Analyse d'intégrité offline

On propose de faire une évaluation de performance d'intégrité suivant le modèle de mesure d'intégrité présenté précédemment. On s'intéresse aux trois variables xPE/xPL, xPE/xAL et xPL/xAL caractérisant les trois événements menant à une perte d'intégrité. Cette analyse est faite dans une approche offline, une fois les enregistrements terminés. Le protocole mis en place pour cette analyse est issu de la partie théorique sur l'étude des valeurs extrêmes, et plus particulièrement de la méthode POT. A titre d'exemple, les paramètres de la loi des excès sont estimés à partir de la méthode du maximum de vraisemblance et le niveau des intervalles de confiance est fixé à $\alpha = 0.05$. Ils seront calculés avec la méthode Delta. On présente tout d'abord un cas statique où les erreurs de positionnement ont été calculées par rapport à la position de référence de la station d'HELILEO. Un second cas d'étude mettra en évidence un jeu de données provenant d'enregistrements en vol. Pour le cas dynamique, les trajectoires de référence sont obtenues par des techniques de positionnement différentiel RTK.

5.4.1 Cas statique

On considère un enregistrement de 14 jours sur la station d'HELILEO. On s'intéresse dans un premier temps à l'événement Misleading Information (MI) pour lequel l'erreur de positionnement PE doit excéder le niveau de protection PL pour une probabilité très faible. On fixe le risque d'intégrité à $p = 1 - 2.10^{-7}$. Ce niveau correspond au niveau maximal auquel on peut être confronté dans le cas d'une approche de précision [DO96]. Les seuils d'alerte AL seront fixés à 40 mètres en horizontal et 10 mètres en vertical.
On rappelle le modèle de mesure d'intégrité établi précédemment :

$$\mathbb{P}(\frac{x\text{PE}}{x\text{PL}} < q(p)) = p \quad ; \quad \mathbb{P}(\frac{x\text{PE}}{x\text{AL}} < q(p)) = p$$

où x est "H" ou "V" selon l'erreur considérée.
On veut estimer le quantile $q(p)$ pour évaluer si l'intégrité est assurée ou non pour ce jeu de données. On utilise la procédure de choix de seuil (*cf.* section précédente) afin de déterminer le modèle d'extrêmes le mieux ajusté aux données. Cette procédure intègre la fonction de permutation aléatoire permettant d'obtenir un jeu de

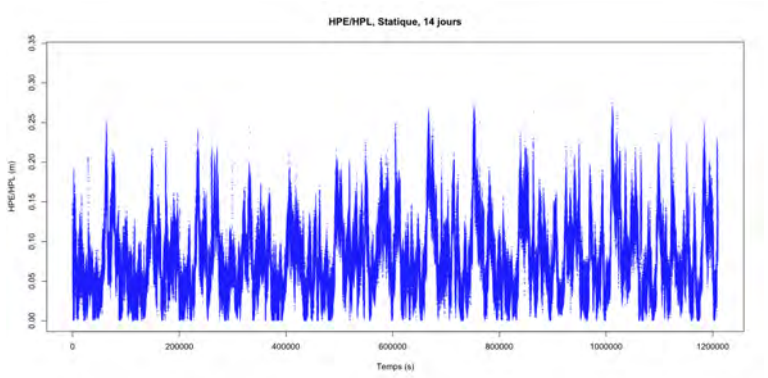


FIGURE 5.32 – Processus Misleading Information

données indépendantes. En effet, on voit sur la figure 5.33 de gauche que les observations restent fortement corrélées après 2 heures ($lag = 7200$) d'enregistrement. Après permutation (courbe de droite), la structure de dépendance a complètement disparu.

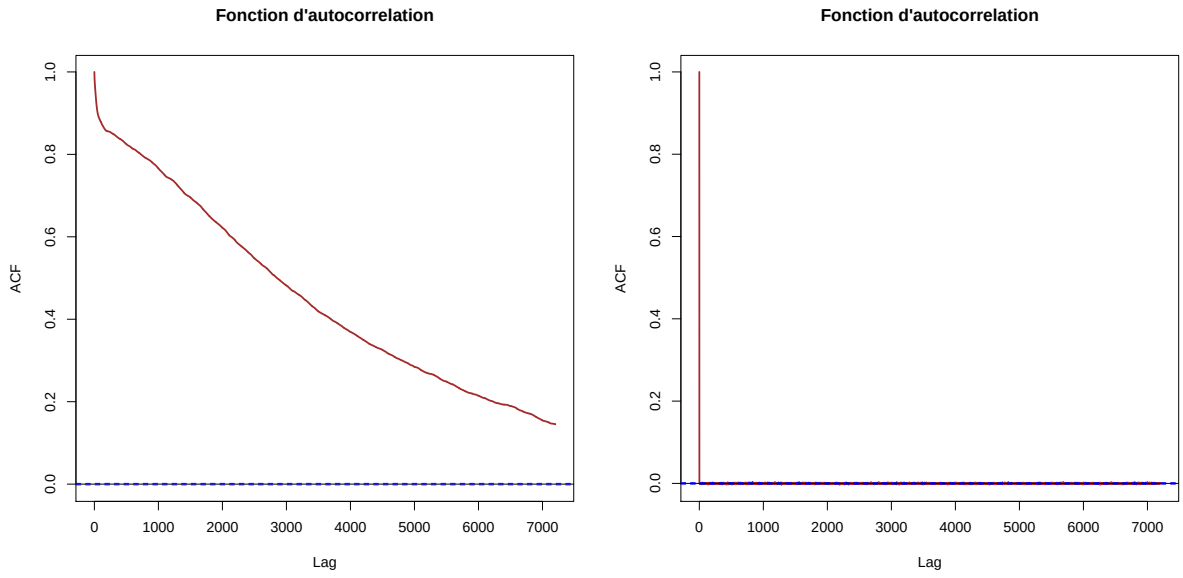


FIGURE 5.33 – Fonctions d'autocorrélation avant et après permutation

Le nouvel échantillon décorrélé est illustré sur la figure 5.34.
La procédure de choix de seuil renvoie le résultat suivant :

Seuil retenu	Quantile estimé	Intervalle de confiance
$u^{opt} = 0.17$	$\widehat{q}(p)_{POT} = 0,316$	$IC_{DM}^{opt} = [0.311; 0.321]$

TABLE 5.1 – Résultat MI, statique, horizontal

On vérifie le fonctionnement de la procédure sur des données réelles en regardant l'estimation du quantile pour une plage de seuils et le modèle choisi automatiquement (figure 5.35). On observe que l'estimation du quantile selon la méthode du

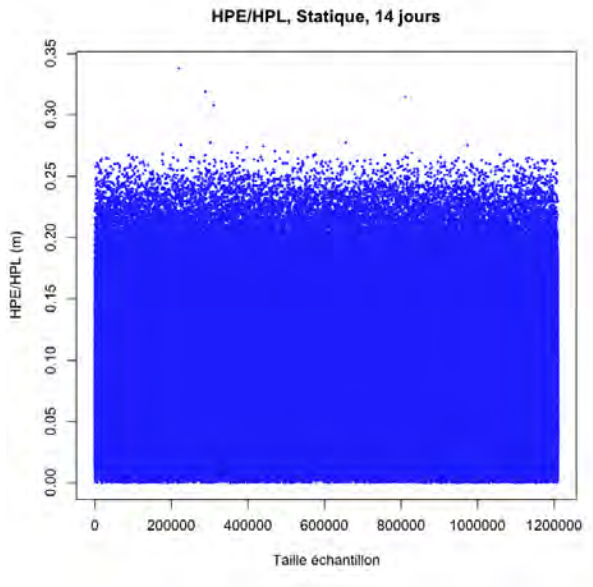


FIGURE 5.34 – Processus Misleading Information après permutation

maximum de vraisemblance est stable en fonction des différents seuils testés. De plus, cette figure met en évidence la bonne approximation de la loi de notre échantillon (en noir) par la loi des excès (en bleu).

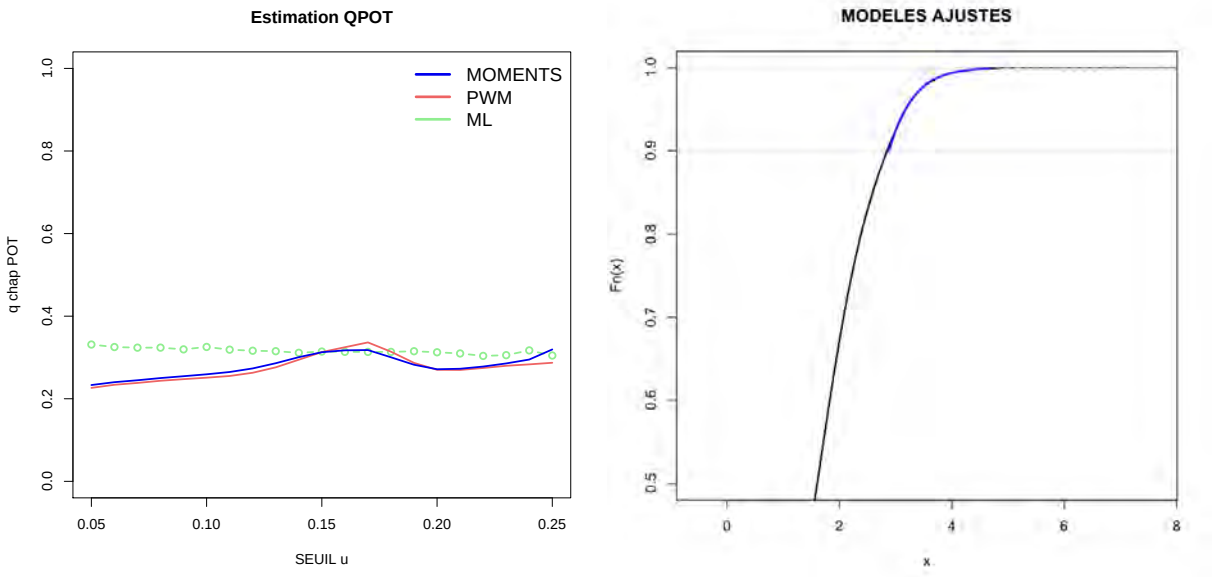


FIGURE 5.35 – Vérification

Pour le cas vertical, on obtient le résultat suivant :

Seuil retenu	Quantile estimé	Intervalle de confiance
$u^{opt} = 0.08$	$\widehat{q}(p)_{POT} = 0,273$	$IC_{DM}^{opt} = [0.259; 0.288]$

TABLE 5.2 – Résultat MI, statique, vertical

On s'intéresse maintenant à l'autre événement pouvant mener à une perte d'intégrité. Il s'agit de l'erreur de positionnement PE qui dépasse le seuil d'alerte AL avec le risque d'intégrité $p = 1 - 2.10^{-7}$.

Seuil retenu	Quantile estimé	Intervalle de confiance
$u^{opt} = 0.70$	$\widehat{q}(p)_{POT} = 0,191$	$IC_{DM}^{opt} = [0.189; 0.194]$

TABLE 5.3 – Résultat HMI, statique, horizontal

Seuil retenu	Quantile estimé	Intervalle de confiance
$u^{opt} = 0.40$	$\widehat{q}(p)_{POT} = 1.078$	$IC_{DM}^{opt} = [0.973; 1.184]$

TABLE 5.4 – Résultat HMI, statique, vertical

Sur le plan horizontal, les quantiles estimés, c'est à dire le seuil *sa* du modèle, sont inférieurs à 1 pour les deux événements MI et HMI, l'intégrité suivant notre modèle est donc assurée pour ce jeu de données. Sur l'axe vertical, le quantile estimé pour l'événement HMI est supérieur à 1. Cet événement mène à une perte d'intégrité. L'intégrité sur l'axe vertical n'est donc pas validée.
On regarde maintenant l'événement pour lequel le niveau de protection PL dépasse le seuil d'alarme AL. On dira dans ce cas que le système n'est plus disponible. Le niveau de probabilité lié à cette disponibilité est de $p = 1 - 10^{-5}$.

Seuil retenu	Quantile estimé	Intervalle de confiance
$u^{opt} = 0.70$	$\widehat{q}(p)_{POT} = 1.403$	$IC_{DM}^{opt} = [1.385; 1.421]$

TABLE 5.5 – Résultat disponibilité, statique, horizontal

Seuil retenu	Quantile estimé	Intervalle de confiance
$u^{opt} = 5$	$\widehat{q}(p)_{POT} = 8.480$	$IC_{DM}^{opt} = [8.461; 8.500]$

TABLE 5.6 – Résultat disponibilité, statique, vertical

Les quantiles estimés sont supérieurs à 1, les performances en terme de disponibilité du système et de service d'intégrité ne sont pas atteintes sur cet enregistrement.

5.4.2 Cas dynamique

Nous étudions un jeu de données recueillies suite à des essais en vol (figure 5.36). Cet échantillon est obtenu par la concaténation de fichiers enregistrés lors de 80 vols pour une durée totale de 314687 secondes. Afin de vérifier les résultats issus de la procédure automatique qui vont suivre, les figures suivantes (5.37, 5.38, 5.39) montrent les quantiles estimés en fonction d'une plage de seuils pour les trois événements MI, HMI et la perte de disponibilité, pour le cas horizontal puis vertical.

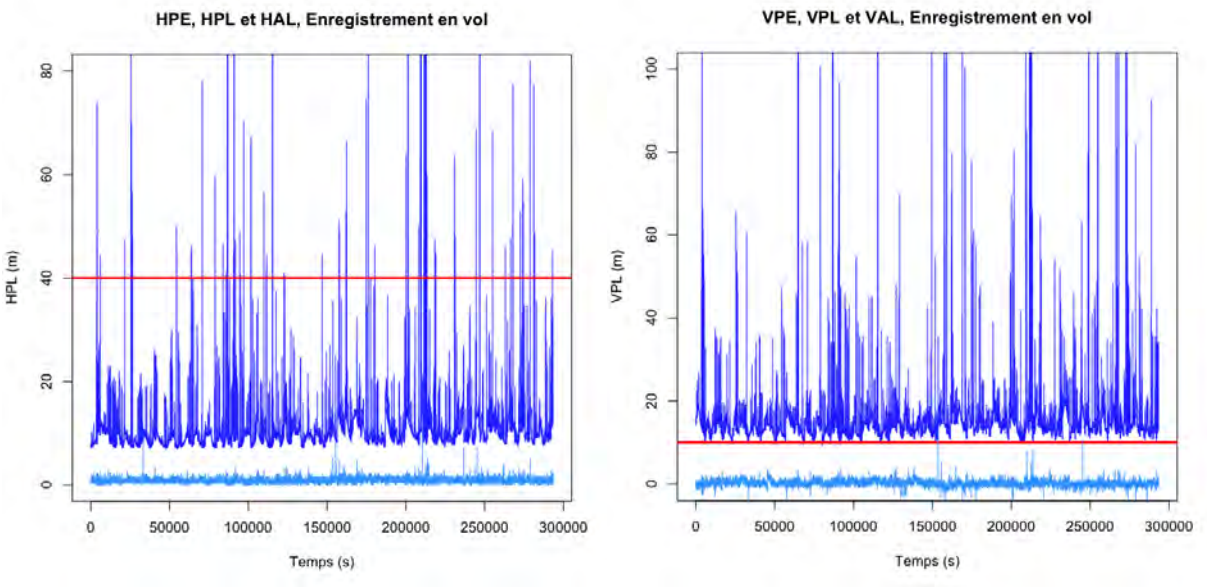


FIGURE 5.36 – Echantillons pour les cas horizontal et vertical

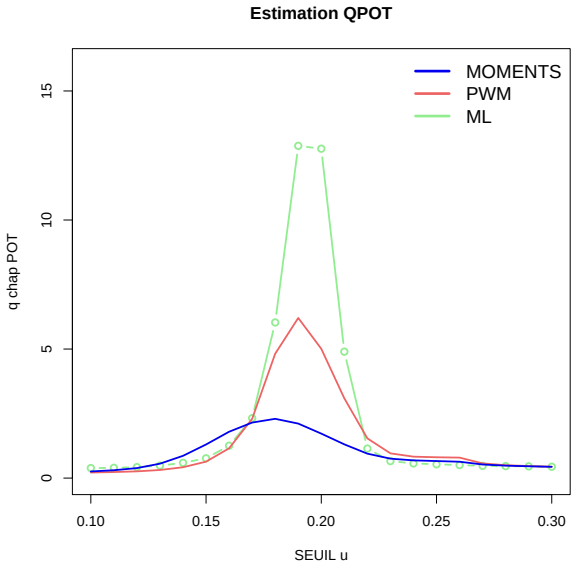


FIGURE 5.37 – Estimation de quantile pour l'événement MI, horizontal

Les résultats obtenus par la procédure automatique sont donnés dans les tables suivantes. Les plages de stabilité des estimateurs ont correctement été isolées par la procédure.

Seuil retenu	Quantile estimé	Intervalle de confiance
$u^{opt} = 0.14$	$\widehat{q}(p)_{POT} = 0.592$	$IC_{DM}^{opt} = [0.532; 0.651]$

TABLE 5.7 – Résultat MI, dynamique, horizontal

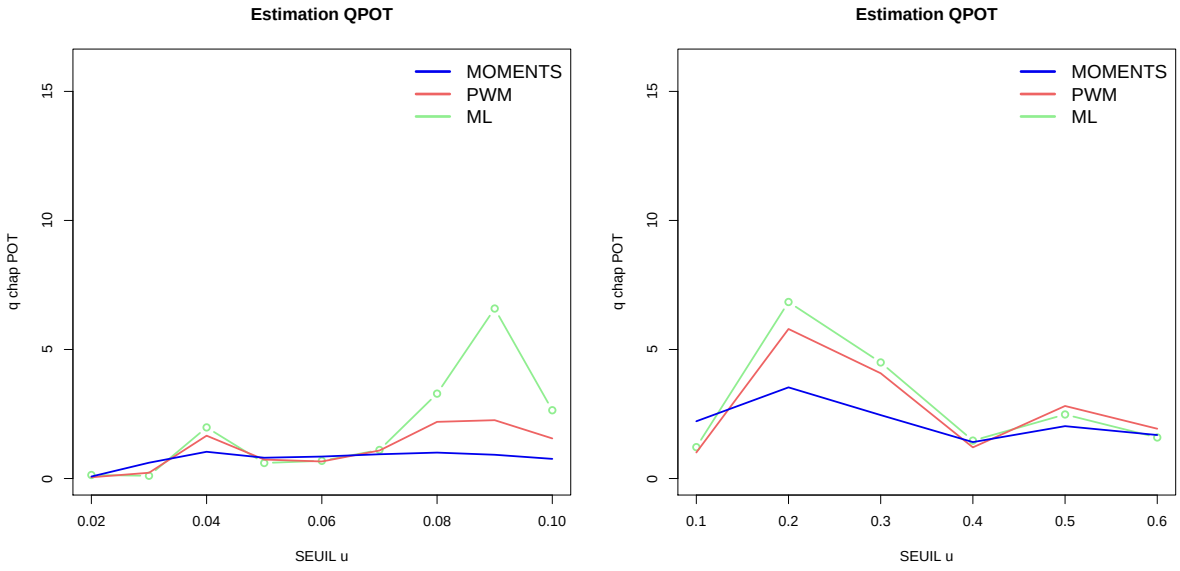


FIGURE 5.38 – Estimation de quantile pour l'événement HMI, horizontal et vertical

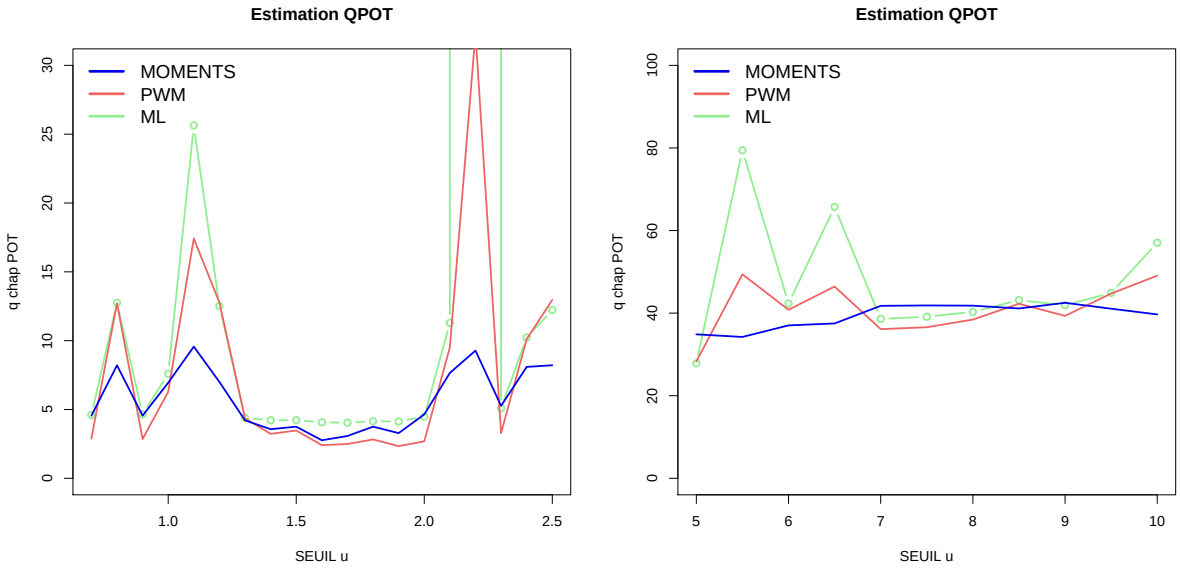


FIGURE 5.39 – Estimation de quantile pour l'événement disponibilité, horizontal et vertical

Seuil retenu	Quantile estimé	Intervalle de confiance
$u^{opt} = 0.35$	$\widehat{q}(p)_{POT} = 0.671$	$IC_{DM}^{opt} = [0.499; 0.842]$

TABLE 5.8 – Résultat MI, dynamique, vertical

Seuil retenu	Quantile estimé	Intervalle de confiance
$u^{opt} = 0.06$	$\widehat{q}(p)_{POT} = 0.685$	$IC_{DM}^{opt} = [-2.312; 3.682]$

TABLE 5.9 – Résultat HMI, dynamique, horizontal

Seuil retenu	Quantile estimé	Intervalle de confiance
$u^{opt} = 0.60$	$\widehat{q}(p)_{POT} = 1.682$	$IC_{DM}^{opt} = [-0.462; 3.828]$

TABLE 5.10 – Résultat HMI, dynamique, vertical

Seuil retenu	Quantile estimé	Intervalle de confiance
$u^{opt} = 1.8$	$\widehat{q}(p)_{POT} = 4.15$	$IC_{DM}^{opt} = [1.819; 6.482]$

TABLE 5.11 – Disponibilité, dynamique, horizontal

Seuil retenu	Quantile estimé	Intervalle de confiance
$u^{opt} = 8.00$	$\widehat{q}(p)_{POT} = 40.302$	$IC_{DM}^{opt} = [-216.9; 297.530]$

TABLE 5.12 – Disponibilité, dynamique, vertical

Les résultats précédents montrent que pour ce jeu de données, l'intégrité suivant notre modèle est assurée sur le plan horizontal, en revanche, elle ne l'est pas sur l'axe vertical. Les performances de disponibilité ne sont pas atteintes non plus.

Nous avons fourni au travers de ce cas d'étude, un exemple d'application du modèle de mesure d'intégrité ainsi que de la procédure automatique de sélection de seuil dans le cadre de la méthodologie Peak Over Threshold (POT). On a pu constater que l'algorithme de sélection isolait correctement des plages de stabilité au sein des estimateurs et que l'adéquation à la loi des excès pour nos jeux de données était souvent très bonne pour les modèles sélectionnés, ceci assurant que nous disposions de suffisamment de données pour tirer de bonnes conclusions sur les quantiles à estimer. De plus, nous avons évoqué une forte corrélation temporelle présente au sein des données. Cette structure de dépendance pourrait altérer l'estimation de quantile. C'est pourquoi, nous avons choisi d'injecter dans la procédure automatique, l'heuristique destiné à briser la structure de dépendance d'un processus au moyen de permutations aléatoires sur les indices temporels de ce dernier (chapitre 4).
Sur les deux exemples traités, nous avons mis en évidence que les performances d'intégrité relatives au système GPS/EGNOS spécifiées par l'ICAO dans les normes aéronautiques [DO96], n'étaient pas atteintes pour les cas les plus restrictifs comme les approches de précision. Les spécifications de performances d'intégrité intègrent une notion temporelle que nous n'avons pas abordé. En effet, le modèle doit être validé par tranche de 150 secondes. Cette durée a été établie relativement aux temps de vol des phases critiques (décollage, approche et atterrissage). Le modèle de mesure d'intégrité est construit à partir d'outils issus de la théorie des extrêmes. Or ces outils ne sont valides et performants que pour des échantillons de grandes tailles. Nous avons donc choisi de faire une analyse globale sur la durée totale des échantillons à disposition.
La plate forme d'analyse de données (*cf* section 5.6) s'est avérée être un outil efficace

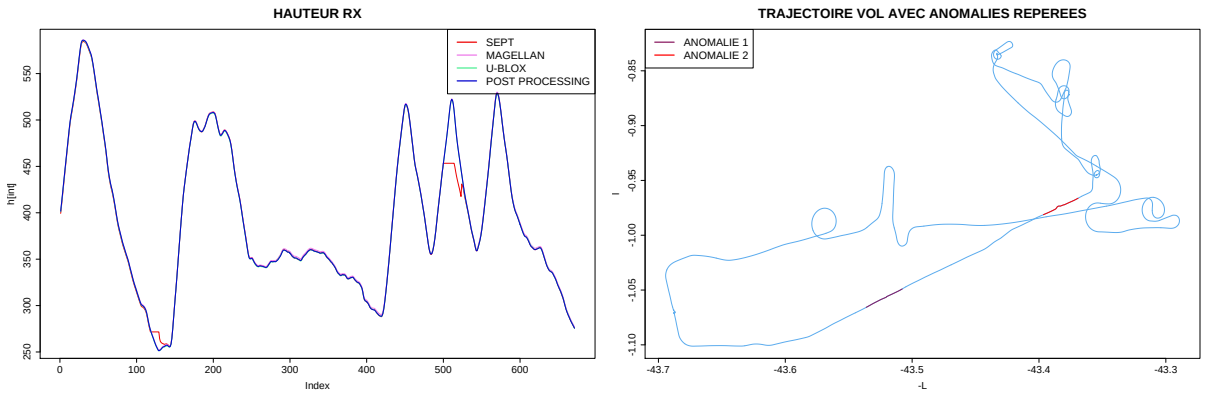


FIGURE 5.40 – Exemple de recherche d'anomalie, décrochage récepteur

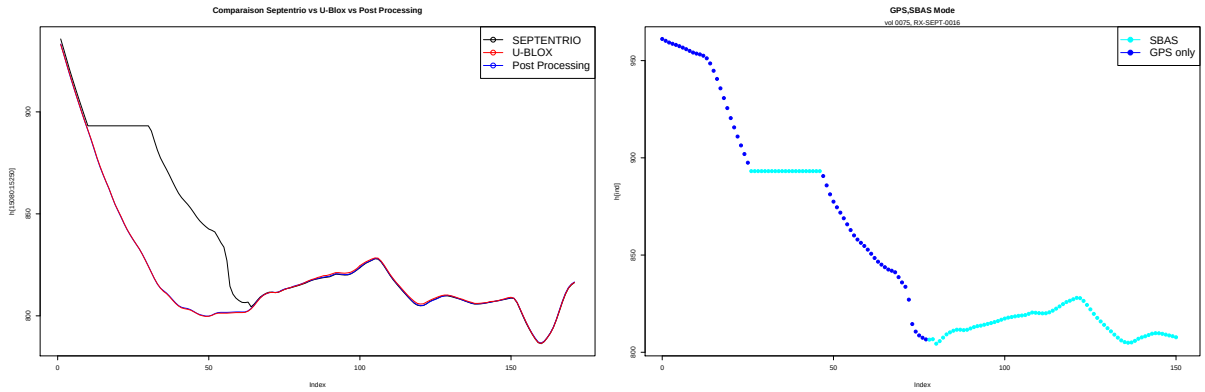


FIGURE 5.41 – Exemple de recherche d'anomalie 2, décrochage récepteur

pour mettre en place des démarches d'investigations de manière à comprendre les raisons de ces pertes d'intégrité. Nous avons pu voir qu'elles étaient le plus souvent dues à des mesures anormales résultant non pas d'une défaillance du système GPS mais d'un dysfonctionnement au niveau des récepteurs embarqués. Nous avons alors cherché à isoler et à caractériser ces phénomènes puis constitué une bibliothèque d'anomalies recensées. Après avoir contacté les fabricants, cette base de données a été d'une grande aide pour comprendre les comportements défaillants des capteurs et donner naissance à des collaborations pour développer des mises à jour logicielles éradiquant ces phénomènes anormaux.

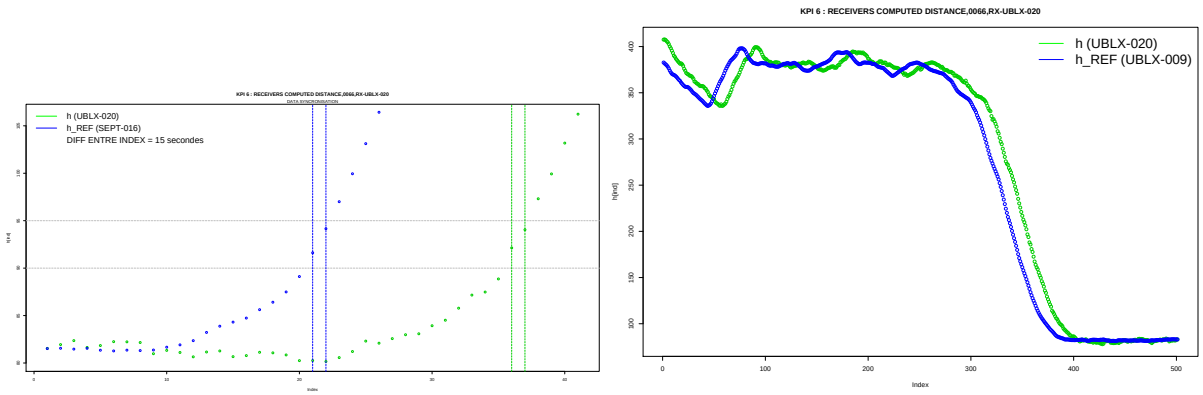


FIGURE 5.42 – Exemple de recherche d'anomalie 3, désynchronisation des horloges

5.5 Temps d'enregistrement nécessaire pour garantir une performance

Un des critères majeurs de performance des récepteurs GNSS est la précision de positionnement qu'ils fournissent. Cette précision est donnée suivant plusieurs critères statistiques : quantiles à 95%, CEP, 2drms, etc [DO96], [ION97]. Un point fondamental est le temps d'enregistrement nécessaire pour justifier de la pertinence et de la significativité de ces indicateurs. Il est dit dans les normes [ION97] que les protocoles de test de précision doivent être réalisés sur une période d'au moins 24h de manière à voir défiler la constellation de satellites au minimum deux fois. La durée de ces tests est évaluée d'un point de vue système et permet dans le cas d'un élément défaillant (panne satellite par exemple) de prendre en compte les mesures erronées dans les échantillons recueillis. Les normes indiquent que la précision de positionnement doit être fournie suivant un quantile de niveau $p = 1 - 0.05$, ou de manière équivalente, à 95%. Dans la plupart des cas, les distributions des erreurs n'ont pas eu le temps de converger en 24 heures (*cf.* section 5.2.3) ; il est alors périlleux d'annoncer une performance sans avoir capturé la totalité de la distribution de la variable observée, au risque d'omettre certaines erreurs élevées.
Il est question dans cette section de fournir une méthodologie permettant de déterminer les temps d'enregistrement minimal nécessaire pour calculer un critère de précision de positionnement. Premièrement, on fournit une méthode générale pour les critères ayant des niveaux de probabilité peu élevé : 50% pour l'indicateur CEP (Circular Error Probable), et 95%. Dans un second temps, on propose de regarder les temps de convergence pour des critères nécessitant la stationnarité des données ; on s'intéressera aux quantiles extrêmes du modèle de mesure d'intégrité de la section précédente. Ainsi, on pourra répondre à la question : combien de temps d'enregistrement faut-il afin de garantir une précision correspondant à des niveaux de probabilité très élevés ?
La question précédente comporte plusieurs termes importants. Le fait que l'on souhaite étudier des quantiles à des niveaux de probabilité élevés nous oriente directement vers l'utilisation d'outils issus de la théorie des extrêmes. Le terme "garantir" a son importance ; il renvoie le statisticien à la notion de confiance et ainsi d'intervalle de confiance. Le critère de convergence que l'on regarde est alors la convergence des

intervalles de confiance en fonction du temps. Ces intervalles de confiances seront associés à l'estimateur de quantile extrême correspondant au niveau de probabilité que l'on se fixe. On souhaite évaluer le temps nécessaire pour appliquer le modèle de mesure d'intégrité vu précédemment. On fixe alors le niveau de quantile à $p = 1 - 10^{-7}$ et le niveau des intervalles de confiance à $\alpha = 0.01$.

5.5.0.1 Enregistrement statique

Pour un enregistrement statique, les moments d'ordres 1 et 2 se stabilisent en moins de 100 heures d'enregistrement (figure 5.43).

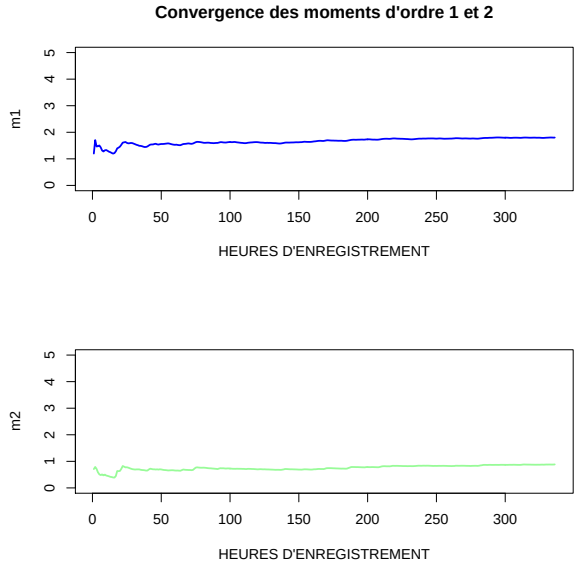


FIGURE 5.43 – Convergence des moments d'ordre 1 et 2, statique

Cependant la largeur des intervalles de confiance du quantile extrême estimé en fonction du temps, passe en dessous d'un seuil $\varepsilon = 0.5$ (que l'on s'est fixé) et sans le franchir à nouveau, à partir de 140 heures (figure 5.44). Ces résultats ont été obtenus en utilisant la procédure de choix de seuil établie précédemment.

5.5.0.2 Enregistrement dynamique

Pour le cas dynamique, les moments d'ordre 1 et 2 mettent plus de temps à se stabiliser. Sur l'exemple suivant, il faut attendre environ 300 heures pour que la variance, moment d'ordre 2, converge (figure 5.45 en bas).
On observe sur la figure 5.46 que la convergence des intervalles de confiance n'est pas atteinte après 90x5 heures d'enregistrement. Ceci s'explique par le fait que le jeux de données observées provient de plusieurs échantillons obtenus après de multiples vols.
On considère plusieurs modèles de récepteurs identiques ayant enregistré plus d'une centaine d'heures de vols chacun. Une analyse comportementale de chaque récepteur a été menée et a montré que l'on pouvait considérer que tous les jeux de données proviennent d'un seul et même récepteur. Attention, ceci est valable pour des modèles haute précision de récepteur dont le fabriquant ne sera pas cité, mais pas pour

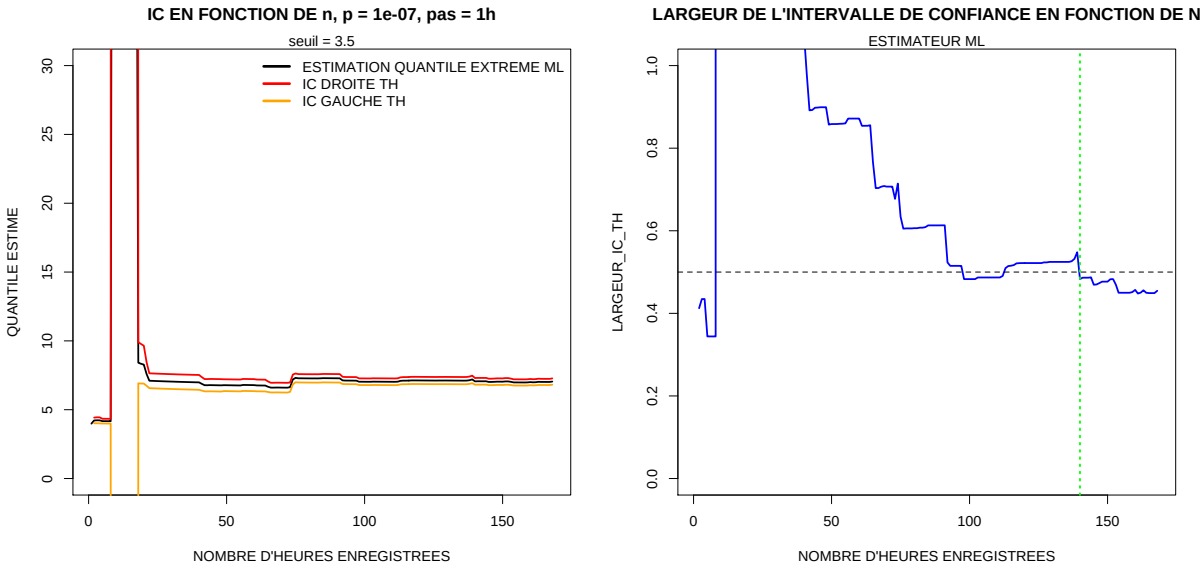


FIGURE 5.44 – Convergence des intervalles de confiance, statique

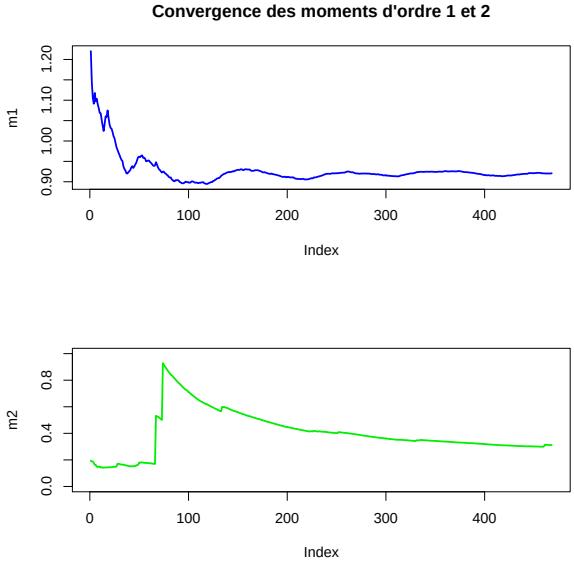


FIGURE 5.45 – Convergence des moments d'ordre 1 et 2, dynamique

tout récepteur GNSS. Ainsi on constitue un échantillon de près de 3000 heures de vols. Afin d'évaluer plusieurs cas et de garantir l'efficacité de la procédure de choix de seuil, on effectue l'estimation du quantile extrême et des intervalles de confiance pour une séquence de seuils retenus par l'algorithme de sélection de seuil (figure 5.47). Les pics vers le haut et vers le bas sont probablement dus à une défaillance de la procédure qui a retenu une valeur trop élevée pour le seuil u. On rappelle qu'il doit faire l'objet d'un compromis entre biais et variance de l'estimateur. La conséquence est une mauvaise adéquation du modèle et une explosion de la variance de l'estimateur. Les pics en bleu à droite de la courbe traduisent l'apparition d'un biais important ; la valeur du seuil sélectionné par la procédure a dû être trop faible

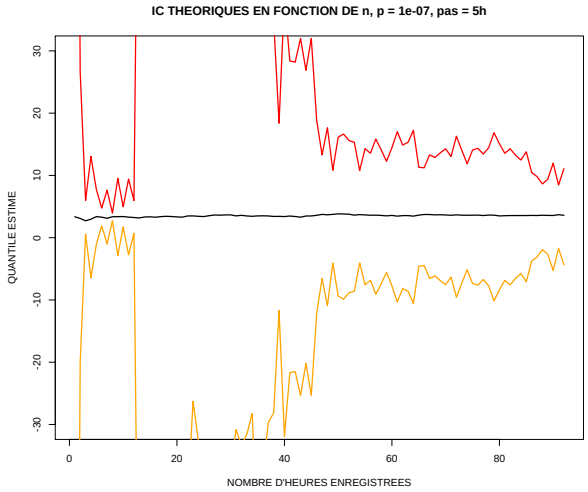


FIGURE 5.46 – Convergence non atteinte des intervalles de confiance, dynamique

(contradiction avec les propriétés asymptotiques du modèle POT impliquant un seuil le plus grand possible).
Pour que chaque largeur d'intervalle de confiance passe en dessous du seuil $\varepsilon = 1$, il faut plus de 3x1000 heures de vol enregistrées (figure 5.48).

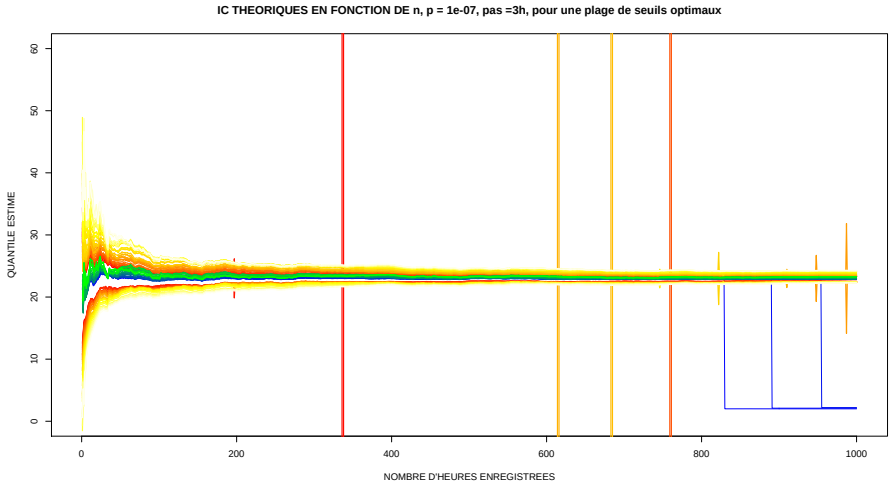


FIGURE 5.47 – Convergence des intervalles de confiance, dynamique

Ces résultats de convergence ont aussi été évalués avec l'estimateur du cas Gumbel et seront fournis en annexe.

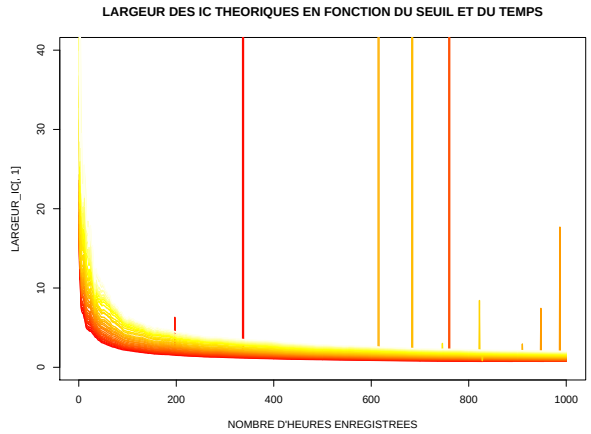


FIGURE 5.48 – Largeur des intervalles de confiance, dynamique

5.6 Plateforme d'analyse de données

La plate forme d'analyse de données a été développée afin de répondre à la demande d'HELILEO qui devait présenter à la commission européenne des rapports d'essais obtenus suite à une vaste campagne d'essais héliportés. Face à la quantité de données à traiter à l'issue de cette campagne, HELILEO a souhaité se doter d'un outil d'analyse statistique intégrant diverses fonctionnalités qui sont décrites dans cette section.
Afin de fournir un outil complet regroupant les différents outils présentés dans ce manuscrit, la fonction de mesure d'intégrité à partir de la théorie des extrêmes y figure également. Cet outil logiciel a été entièrement implémentée avec le logiciel R.

5.6.1 Fonctionnalités

La plate forme se compose de deux blocs indépendants : le premier fournit une génération automatique d'indicateurs, appelés KPI (Key Performance Indicators), définis par la commission européenne. Le second est destiné à l'analyse d'intégrité. La première fenêtre qui apparaît à l'ouverture de la plate forme permet de sélectionner le bloc que l'utilisateur désire utiliser (figure 5.49).

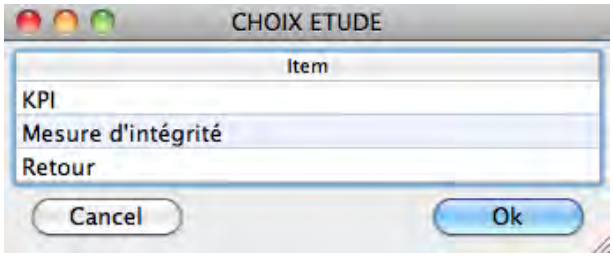


FIGURE 5.49 – Choix de l'étude à réaliser

5.6.1.1 Bloc 1 : génération des KPI

Le premier bloc de la plateforme est destiné à générer automatiquement les indicateurs KPI pour un scénario sélectionné par l'utilisateur. La configuration du scénario se fait à l'aide d'une suite de choix décrivant les différents tests réalisés lors de la campagne d'essais.
La première étape est le choix de l'étude d'un vol isolé ou d'un ensemble de vols. Les vols ont été numérotés, la sélection du vol s'effectue donc à partir d'une liste de numéros correspondants aux vols effectués. Le menu des sélections de vols autorise une sélection discontinue si l'utilisateur le souhaite.
L'étape suivante consiste à choisir le récepteur à étudier parmi l'ensemble des récepteurs embarqués. La gamme de récepteurs testés durant la campagne a été répartie dans cinq H-box ; l'utilisateur devra donc choisir la H-box utilisée. Les compositions des H-Box étant fixes, les listes des récepteurs présents dans les quatre H-box ont été préalablement enregistrées dans le logiciel. Ainsi, lorsque l'utilisateur sélectionne une H-box à étudier, une fenêtre comportant la liste des récepteurs correspondants à cette H-box, s'ouvre en suivant.
Une fois la configuration du scénario achevée, l'utilisateur sélectionne le KPI à générer. Une dernière fenêtre s'ouvre et permet un retour à toute étape antérieure.
La figure 5.51 présente la suite de choix successifs à réaliser pour obtenir la sortie désirée.

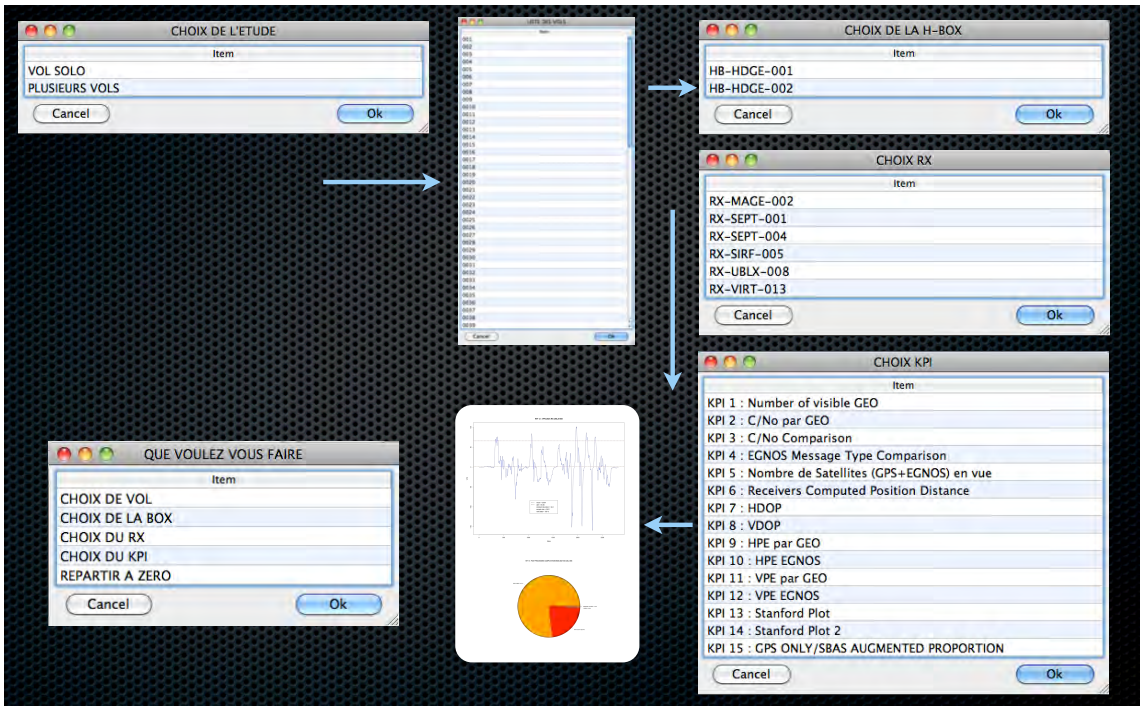


FIGURE 5.50 – Configuration de scénario

Description des KPI
Les KPI (Key Performance Indicators) sont un ensemble d'indicateurs décrivant plusieurs variables témoignant des performances du système GPS/EGNOS pour des essais héliportés. Ces indicateurs ont été définis par la commission européenne et sont au nombre de 14 :

- **KPI 1** : Number of visible GEO.
  Le KPI 1 décrit le nombre de satellites géostationnaires EGNOS visibles pendant l'essai en vol. Ce nombre est censé être égal à 3 signifiant que le récepteur reçoit les signaux de l'ensemble de la constellation EGNOS. Ce KPI est représenté par un camembert sur lequel figurent les proportions de temps où le récepteur a eu 1, 2 ou 3 satellites en vue.
- **KPI 2** : C/N0 per GEO.
  Le KPI 2 décrit le niveau de signal sur bruit du signal reçu par le récepteur, pour les trois satellites EGNOS.
- **KPI 3** : C/N0 comparison.
  Le KPI 3 décrit le niveau de signal sur bruit du signal reçu par le récepteur, pour chacun des trois satellites EGNOS.
- **KPI 4** : EGNOS message type comparison.
  Le KPI 4 décrit la proportion des messages reçus contenant les informations communiquées par les satellites EGNOS au récepteur.
- **KPI 5** : Satellites in view.
  Le KPI 5 donne le nombre de satellites vus par le récepteur.
- **KPI 6** : Receivers computed position distance.
  Le KPI 6 décrit les différences de coordonnées (sur les trois axes dans le repère géodésique) entre le récepteur considéré et un récepteur de référence présent dans la H-box à sélectionner.
- **KPI 7 et 8** : HDOP, VDOP.
  Les KPI 7 et 8 décrivent les DOP horizontaux et verticaux calculés par le récepteur pendant le(s) vol(s).
- **KPI 9 et 11** : HPE, VPE per GEO.
  Les KPI 9 et 11 décrivent les erreurs de positionnement horizontales et verticales calculées à partir des positions estimées en utilisant les informations d'un des satellites EGNOS au lieu des trois.
- **KPI 10 et 12** : HPE, VPE.
  Les KPI 10 et 12 décrivent les erreurs de positionnement horizontales et verticales.
- **KPI 13 et 14** : Stanford plot
  Les KPI 13 et 14 décrivent les performances d'intégrité suivant la représentation Stanford Plot sur l'axe vertical et sur le plan horizontal.

En fonction de la nature des variables à étudier, les KPI sont illustrés de différentes façons : on trouve des représentations sous forme de courbe temporelle, d'histogramme ou de camembert. Afin de compléter la description des échantillons que l'on veut étudier, les indicateurs statistiques suivants figurent sur les graphiques : moyenne, écart-type, quantile à 95%, taille de l'échantillon et pourcentage de données valides. Un exemple de chaque KPI est donné en annexe B.

5.6.1.2 Bloc 2 : mesure d'intégrité

La seconde fonctionnalité de la plate forme est l'étude de la mesure d'intégrité selon le modèle issu de la théorie des extrêmes. Ceci constitue l'application pratique de l'étude présentée précédemment.
Le jeu de données à étudier peut être issu de la configuration d'un scénario comme

dans le bloc 1 de génération des KPI ou être sélectionné indépendamment via une boîte de dialogue spécifiant un chemin d'accès. L'estimateur de quantile extrême utilisé est sélectionné parmi les modèles suivants :

- configuration automatique : méthode POT. Les paramètres de la loi des excès sont alors estimés avec l'estimateur du maximum de vraisemblance, et l'adéquation du modèle est faite avec la procédure automatique de choix de seuil. Les intervalles de confiances fournis sont obtenus avec la méthode Delta.
- loi GEV et méthode bloc maxima : estimation de l'indice des valeurs extrêmes avec l'estimateur de Hill et estimation du quantile avec l'estimateur présenté dans la section 2.2.4.
- loi GPD : méthode POT. L'utilisateur a le choix pour estimer les paramètres de la loi des excès entre la méthode des moments ou l'estimateur du maximum de vraisemblance. Deux estimateurs de quantiles sont alors proposés : $\widehat{q}(p)_{\mathrm{POT}}$ et $\widehat{q}(p)_{\gamma=0}$.

Dans les exemples du chapitre consacré à la théorie des extrêmes, les intervalles de confiance obtenus à l'aide de la méthode Delta avaient une largeur supérieure à celle des intervalles de confiance asymptotiques. Ce comportement, avait pour effet de contenir de manière plus sûre la valeur théorique du quantile à estimer. Pour cette raison, nous avons choisi la méthode Delta pour le calcul des intervalles de confiance dans la plate forme.
L'utilisateur doit ensuite donner le niveau de probabilité du quantile à estimer ainsi que celui de l'intervalle de confiance associé via la boîte de dialogue suivante :

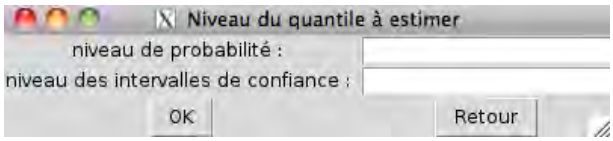


FIGURE 5.51 – Niveaux de probabilité

Les valeurs à rentrer dans ces champs sont p telle que $q(p) = 1 - p$ pour le niveau de probabilité du quantile à estimer. α tel que le niveau de l'intervalle soit $1 - \alpha$.
La fenêtre de résultats (figue 5.52) est un résumé de l'analyse d'intégrité donnant les différents choix de méthode d'estimation faits par l'utilisateur et les valeurs associées. On trouvera les expressions théoriques menant à ces résultats dans le chapitre consacré aux valeurs extrêmes.

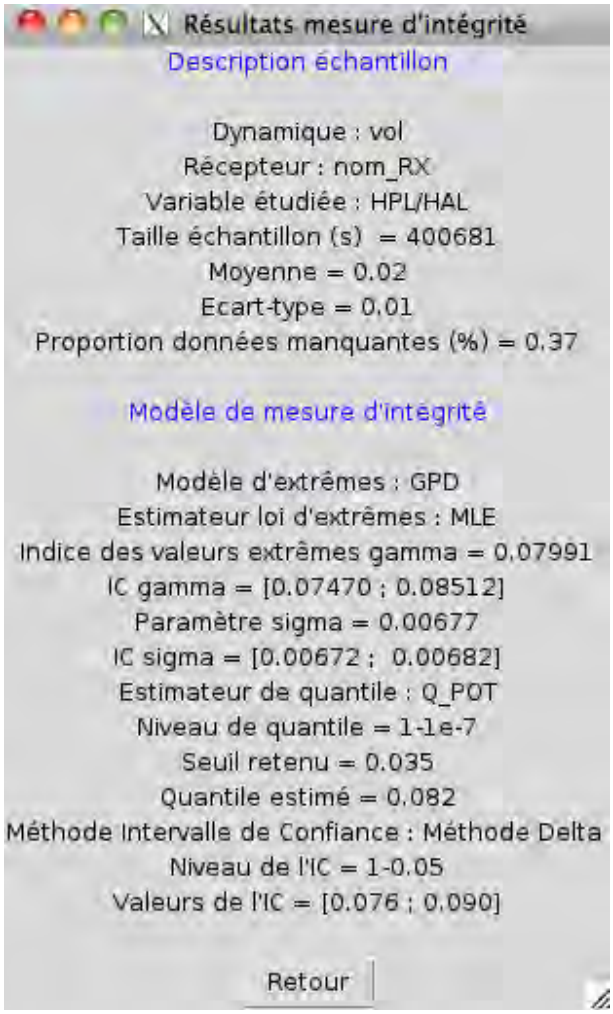


FIGURE 5.52 – Résumé de l'analyse

5.7 Conclusion

Ce cinquième et dernier chapitre constitue la partie pratique des travaux théoriques réalisés au cours de cette étude. Nous avons mis en place un protocole d'analyse afin d'appliquer le modèle de mesure d'intégrité aux données GPS. Ce protocole comporte les étapes suivantes :

- visualisation et analyse des données : nous avons montré que les erreurs de positionnement étaient distribuées selon des lois proches des modèles paramétriques cités dans la littérature. Toutefois, l'adéquation n'est pas satisfaite dans les queues de distribution. Il est donc risqué de faire des hypothèses paramétriques sur la loi des données lorsque l'on manipule des niveaux de probabilité élevés.
- vérification des hypothèses requises par les modèles employés : l'hypothèse d'indépendance étant traitée dans le chapitre 4, nous nous sommes intéressés à la stationnarité des données et nous avons montré que pour de grands échantillons, la stationnarité asymptotique pouvait être atteinte selon trois critères (convergence des moments d'ordre 1 et 2, et convergence des distributions).
- sélection du modèle, estimation des paramètres et adéquation du modèle aux données : au travers de ce cas d'étude, nous avons fourni un exemple d'application

du modèle POT issu de la théorie des extrêmes appliquée aux données GPS. Le problème de choix de seuil a été résolu par l'intermédiaire d'une procédure automatique de sélection reposant sur plusieurs critères dont un critère d'optimisation d'adéquation du modèle.
– estimation du quantile recherché et des intervalles de confiance associés : enfin, cette procédure menant à l'estimation de quantile extrême a montré des performances tout à fait correctes sur les cas théoriques. Les cas d'étude sur des données réelles ont eux aussi mis en évidence de bonnes performances de la procédure aux vues de la stabilité des résultats.

Une fois ce protocole établi et validé, il a alors été possible de répondre aux deux questions posées au début de chapitre :

1. Est-ce que les exigences de fiabilité les plus strictes imposées par les normes aéronautiques sont satisfaites pour un positionnement d'hélicoptère utilisant les signaux GPS/EGNOS ? Le tableau 5.13 ci-dessous regroupe les résultats correspondant obtenus sur deux cas d'étude (statique et en vol) pour des niveaux de fiabilité requis pour des approches de précision CAT 1.

Essai	MI		HMI		Disponibilité	
	Horiz	Vert	Horiz	Vert	Horiz	Vert
Statique	ok	ok	ok	proche de 1	pas ok	pas ok
En vol	ok	ok	ok	pas ok	pas ok	pas ok

TABLE 5.13 – Résultats de l'analyse d'intégrité sur deux cas (statique et en vol)

On rappelle que l'on veut évaluer si les deux événements MI et HMI (*ie.* PE<AL et PE<PL) sont assurés pour un risque de niveau $1 - 2.10^{-7}$. On observe aussi l'événement PL>AL correspondant à une indisponibilité du service d'intégrité et menant aussi à une perte d'intégrité. Le risque dans ce cas est fixé à 10^{-5}.
Pour ce cas d'étude notre modèle montre que les conditions de fiabilité sont remplies pour presque tous les cas concernant les événements MI et HMI. On considère le cas HMI statique sur l'axe vertical comme validé car la valeur de $q(p)$ est très proche de 1 et l'intervalle de confiance associé est étroit. En revanche, les performances correspondant à la disponibilité du service d'intégrité ne sont validées dans aucun des cas traités. On voit très nettement sur la figure 5.36 que les bornes inférieures des niveaux de protections (en bleu foncé) ne sont pas cohérentes avec les niveaux d'alerte (ligne horizontale rouge). Les niveaux de protection sont issus d'un algorithme de contrôle autonome d'intégrité (RAIM) [Mar08], [Fau11] implanté dans les récepteurs. Il se peut que les récepteurs embarqués durant la campagne n'aient pas été configurés pour satisfaire des niveaux de sécurité aussi élevés.

2. Nous avons fourni une méthodologie permettant d'évaluer le temps d'enregistrement nécessaire pour garantir une performance donnée. Cette méthodologie s'appuie sur différents critères en fonction du niveau de performance que l'on

souhaite caractériser. Pour les niveaux des indicateurs standards (*ie.* 50% et 95%) les critères choisis reposent sur la convergence des moments d'ordre 1 et 2 des variables observées (décorrélées) attestant que la stationnarité est atteinte. Pour des niveaux de probabilités extrêmes, on s'intéresse à la convergence des intervalles de confiance associés au quantile estimé. Cette convergence assure que l'on a recueilli suffisamment d'observations pour caractériser correctement la loi des données dans les queues de distribution.

Enfin, la dernière partie de ce chapitre était consacrée à la présentation d'un outil fonctionnel d'analyse de données GPS ayant permis à la société HELILEO de matérialiser ses campagnes de tests et ainsi analyser les performances accomplies par les capteurs embarqués dans les hélicoptères.

Annexes

CHAPITRE A

Estimation de quantile par des méthodes classiques

Cette annexe présente les méthodes les plus couramment utilisées pour estimer un quantile ainsi qu'une comparaison avec les deux estimateurs de quantiles issus de la théorie des extrêmes utilisés dans le protocole d'analyse d'intégrité. Cette comparaison sera faite sur un échantillon distribué suivant une loi de Rayleigh de paramètre $\rho = 2$, comme dans les exemples de ce manuscrit. On estimera alors un quantile de niveau $p = 1 - 10^{-7}$ et le niveau des intervalles de confiance sera fixé à 0.05.

Parmi les techniques d'estimation dites "classiques", on recense trois familles distinctes :

- Les techniques de rééchantillonnage de type Monte-Carlo. Elles sont très délicates à mettre en oeuvre pour des niveaux de probabilité aussi élevés tel que $p = 1 - 10^{-7}$. En effet le coût calculatoire serait beaucoup trop élevé et les tailles d'échantillons dont on dispose trop faibles pour espérer une précision décente.

- La méthode non paramétrique du quantile empirique. La précision de cette technique dépend fortement de la taille de l'échantillon et peut avoir une grande variance pour des valeurs de p proches de 0 ou de 1.

- La méthode paramétrique. On suppose a priori que l'échantillon observé suit une loi paramétrique (par exemple une gaussienne ou une exponentielle). Cette méthode est trop sensible au choix de la famille paramétrique.

Les deux sections qui suivent présentent les techniques d'estimation non paramétrique et paramétrique suivant deux a priori, gaussien et exponentiel. La troisième et dernière section de cette annexe est consacrée à un tableau comparatif des estimateurs évoqués dans cette partie.

A.1 Estimation non paramétrique

La méthode du quantile empirique est la plus commune et la plus simple des méthodes d'estimation de quantile à laquelle on peut penser. Cette méthode consiste à chercher la valeur pour laquelle, la proportion d'observations en-dessous de cette valeur soit égale à $1 - p$. On considère un échantillon de taille n de variables aléatoires i.i.d. On note $X_{1,n} \leq \cdots \leq X_{n,n}$ le réordonnement croissant de l'échantillon. On parlera de statistiques d'ordre. Pour tout $p \in [0, 1]$, l'estimateur du quantile empirique est défini par $\hat{q}(p)_{emp} = X_{p,n}$.
Le problème d'un tel estimateur est qu'il est fortement dépendant du nombre de

données dont on dispose. En effet, si $n < \frac{1}{p}$, il n'y a pas assez d'observations pour apprécier des évènements de probabilité d'occurrence p. De plus, la variance d'un tel estimateur peut être importante. En effet, elle s'exprime par :

$$Var(\widehat{q}(p)_{emp}) = \frac{p(1-p)}{n[F'(q(p))]^2} \tag{A.1}$$

L'intervalle de confiance d niveau $1-\alpha$ associé à cet estimateur est :

$$IC_{1-\alpha}(q(p)_{emp}) = \left[\hat{q}(p)_{emp} \pm \frac{\sqrt{p(1-p)}}{\sqrt{n}\, F'(q(p))} z_{1-\alpha/2}\right] \tag{A.2}$$

où $z_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ d'une loi normale centrée réduite. On note F_n la fonction de répartition empirique de la variable X et on estime F' par :

$$\widehat{F'_n}(q(p)) \simeq \frac{F_n(q(p)) - F_n(h)}{q(p) - h} \qquad \text{avec } h \text{ très proche de } q(p) \tag{A.3}$$

A.2 Méthodes paramétriques

On suppose que la loi de la variable étudiée appartient à une famille paramétrique de loi. Nous allons donner deux exemples suivant deux a priori pour nous rendre compte que l'emploi de modèles paramétriques peut s'avérer très périlleux quand la loi initiale de l'échantillon observé est inconnue.

A.2.1 A priori gaussien

"La loi normale prend une part très importante dans les méthodes de décisions statistiques. Ce modèle statistique a en effet l'intérêt incontestable de pouvoir décrire avec simplicité un grand nombre de phénomènes aléatoires naturels. Son caractère n'en est pas universel pour autant et on constate que bon nombre d'applications statistiques ne peuvent se satisfaire de l'hypothèse de normalité."

On suppose tout d'abord que X suit une loi gaussienne de paramètres μ et σ. On exprime à l'aide d'un changement de variable le quantile d'ordre $1-p$:

$$\hat{q}(p)_{gaussien} = \hat{\mu} + \hat{\sigma} z_{1-p} \tag{A.4}$$

où z_{1-p} désigne le quantile d'ordre $1-p$ d'une loi normale centrée réduite. Dans ce modèle on note que l'estimation du quantile se réduit à l'estimation des paramètres μ et σ que l'on estime par la moyenne et l'écart-type empiriques classiques :

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{et} \qquad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2 \tag{A.5}$$

Pour un quantile gaussien, l'intervalle de confiance de niveau $1-\alpha$ est donné par :

$$IC_{1-\alpha}(q(p)_{gaussien}) = \left[\hat{\mu} + \hat{\sigma} z_{1-p} \pm \frac{\hat{\sigma}}{\sqrt{n}} z_{1-\alpha/2}\right] \tag{A.6}$$

A.2.2 A priori exponentiel

Lorsqu'on observe des échantillons à queues lourdes c'est à dire avec beaucoup de données loin de la moyenne, dans les queues de distribution, on peut supposer que la variable observée suit une loi exponentielle de paramètre λ (loi à queue lourde). Il s'agit alors de calibrer le paramètre λ estimé par :

$$\hat{\lambda}^{-1} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{A.7}$$

puis d'estimer le quantile d'ordre $1-p$ voulu avec l'expression suivante obtenue par inversion de la fonction de répartition de la loi exponentielle :

$$\hat{q}(p)_{exp} = -\frac{log(p)}{\hat{\lambda}} \tag{A.8}$$

L'intervalle de confiance de niveau $1-\alpha$ de cet estimateur est donnée par :

$$IC_{1-\alpha}(q(p)_{exp}) = \Big[\hat{q}(p)_{exp}\big(1 \pm \frac{z_{1-\alpha/2}}{\sqrt{n}}\big)\Big] \tag{A.9}$$

A.3 Résultats

On rappelle la valeur théorique du quantile à estimer pour un niveau $p = 1-10^{-7}$: $q_{th} = 11.35$. La taille de l'échantillon simulé est $N = 360000$.

estimation	intervalle de confiance (5%)	|erreur relative|
$\widehat{q}^{\text{ML}}_{\text{POT}} = 11.32$	$IC_{th} = [9.48; 12.76]$	0.3%
$\widehat{q}_{\gamma=0} = 11.36$	$IC_{th} = [10.87; 11.86]$	0.08%
$\widehat{q}_{\text{emp}} = 9.65$	$IC_{emp} = [9.63; 9.67]$	15%
$\widehat{q}_{\text{gauss}} = 9.336$	$IC_{gauss} = [9.332; 9.341]$	18%
$\widehat{q}_{\text{exp}} = 40.37$	$IC_{exp} = [40.24; 40.50]$	250%

TABLE A.1 – Comparaison avec des estimateurs classiques

Cet exemple montre bien que pour les niveaux de quantiles auxquels nous sommes confrontés dans l'étude de l'intégrité, les méthodes dites classiques ne sont pas appropriées.

C H A P I T R E B

Description et exemple des Key Performance Indicators (KPI)

Cette annexe est une description illustrée des Key Performance Indicators (KPI) mentionnés au chapitre 5. Ces indicateurs ont été définis par la commission européenne pour évaluer les performances du système GPS/EGNOS au moyen de récepteurs embarqués dans des hélicoptères. On en compte 14.

– **KPI 1** : Number of visible GEO.
  Le KPI 1 décrit le nombre de satellites géostationnaires EGNOS visibles pendant l'essai en vol. Ce nombre est censé être égal à 3 signifiant que le récepteur reçoit les signaux de l'ensemble de la constellation EGNOS. Ce KPI est représenté par un camembert sur lequel figurent les proportions de temps où le récepteur a eu 1, 2 ou 3 satellites en vue.

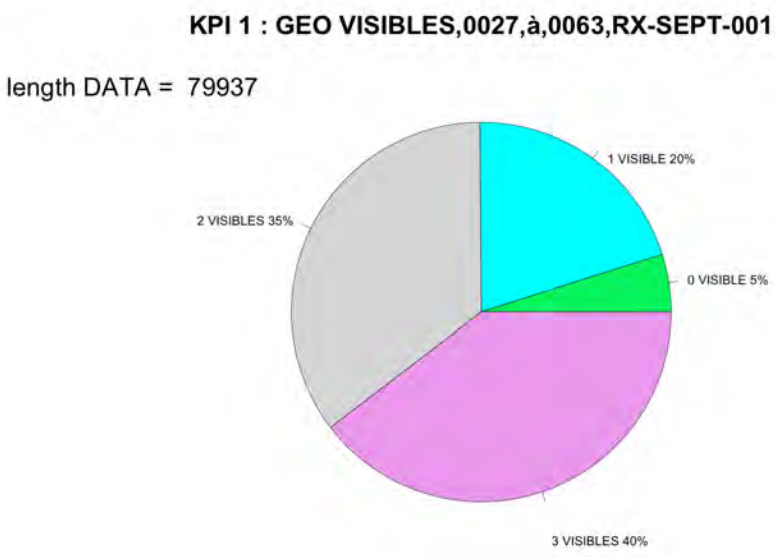


FIGURE B.1 – KPI 1

– **KPI 2** : C/N0 per GEO.
Le KPI 2 décrit le niveau de signal sur bruit du signal reçu par le récepteur, pour les trois satellites EGNOS.

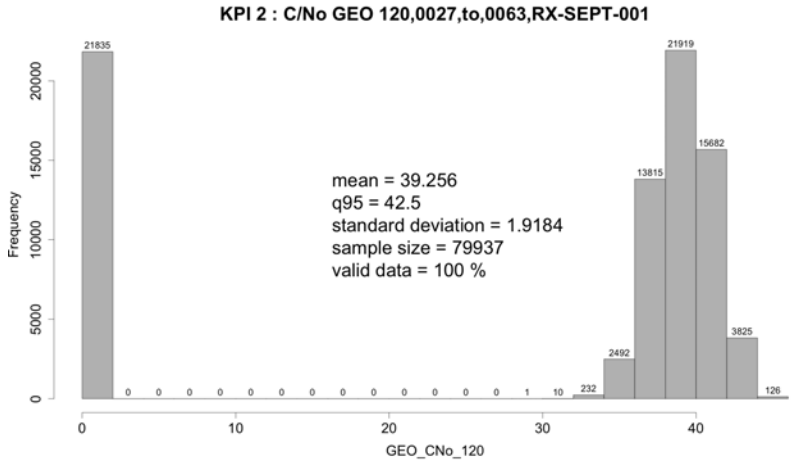


FIGURE B.2 – KPI 2-1 : satellite 1 EGNOS

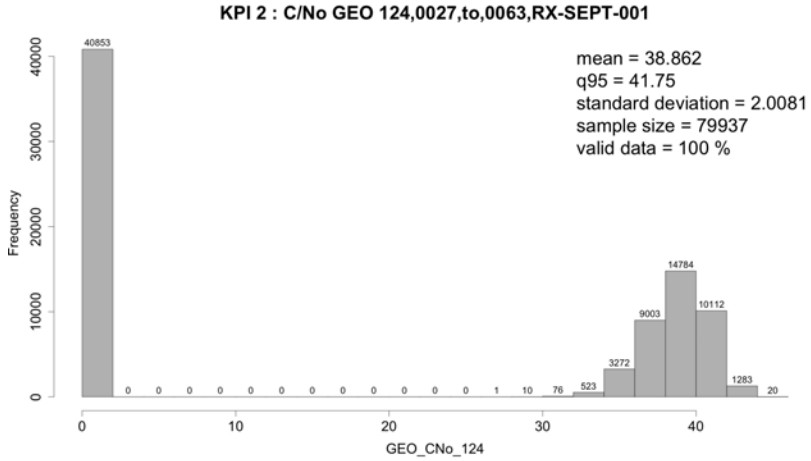


FIGURE B.3 – KPI 2-2 : satellite 2 EGNOS

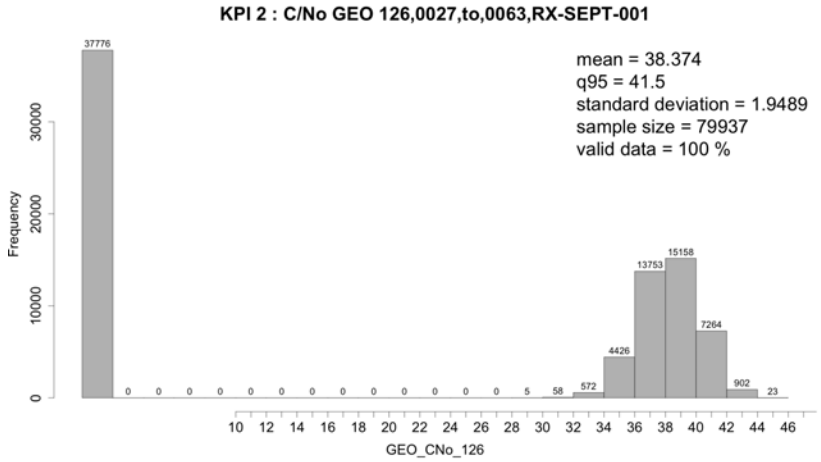


FIGURE B.4 – KPI 2-3 : satellite 3 EGNOS

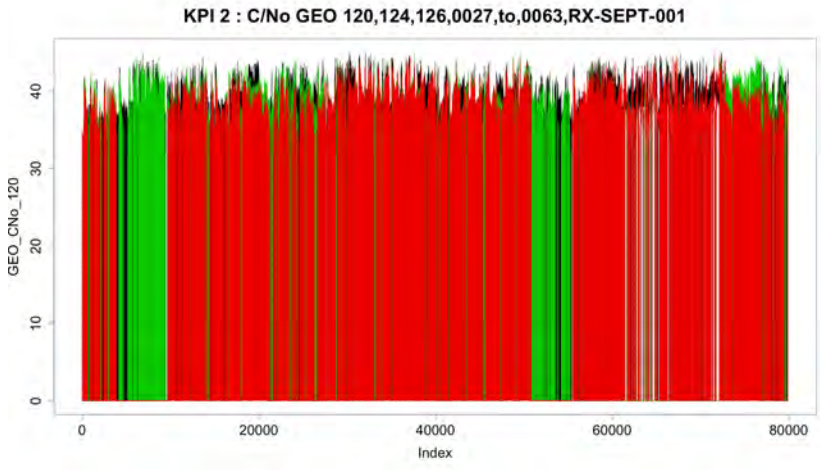


FIGURE B.5 – KPI 2-4

– **KPI 3** : C/N0 comparison.
Le KPI 3 décrit le niveau de signal sur bruit du signal reçu par le récepteur, pour chacun des trois satellites EGNOS.

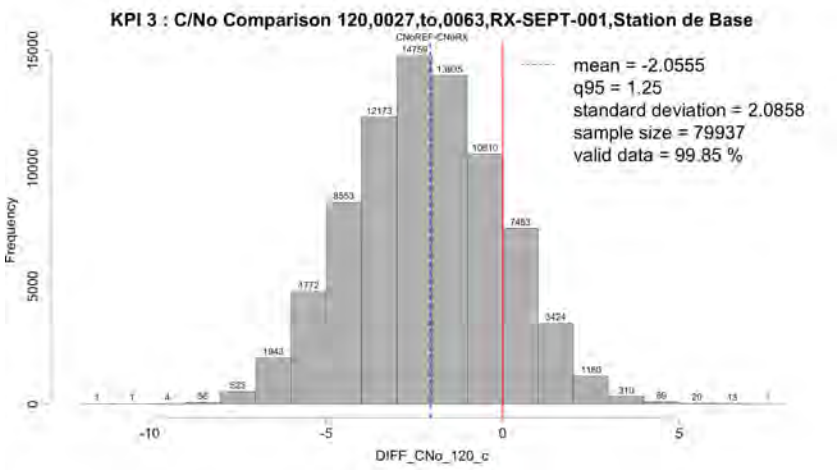


FIGURE B.6 – KPI 3-1 : satellite 1 EGNOS

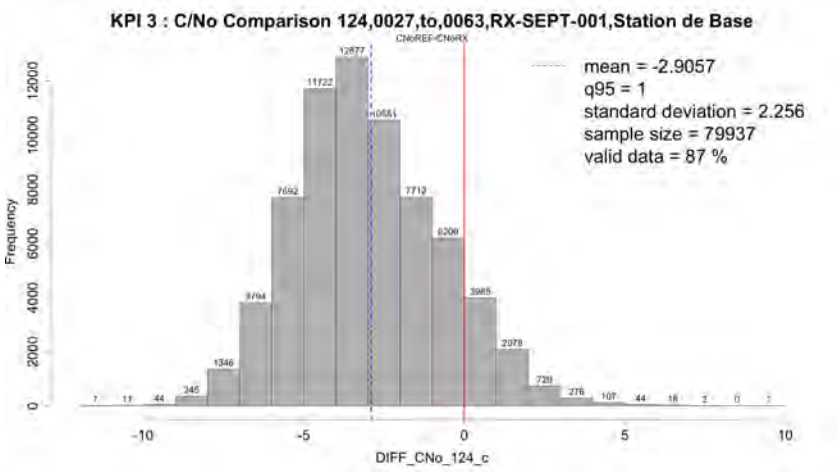


FIGURE B.7 – KPI 3-2 : satellite 2 EGNOS

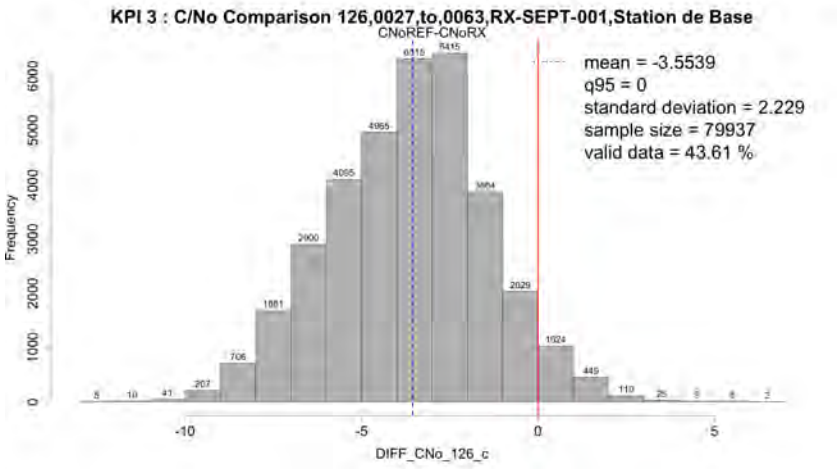


FIGURE B.8 – KPI 3-3 : satellite 3 EGNOS

– **KPI 4** : EGNOS message type comparison.

Le KPI 4 décrit la proportion des messages reçus contenant les informations communiquées par les satellites EGNOS au récepteur.

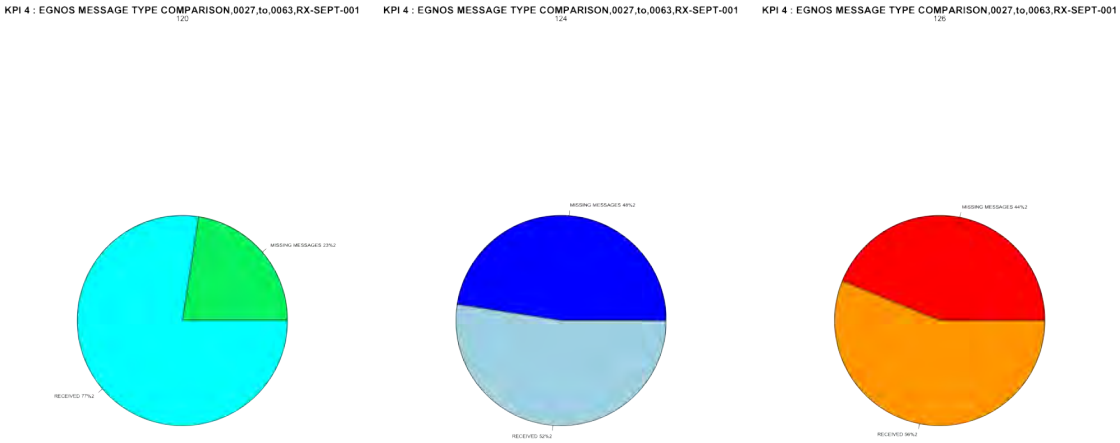


FIGURE B.9 – KPI 4

– **KPI 5** : Satellites in view.
Le KPI 5 donne le nombre de satellites vus par le récepteur.

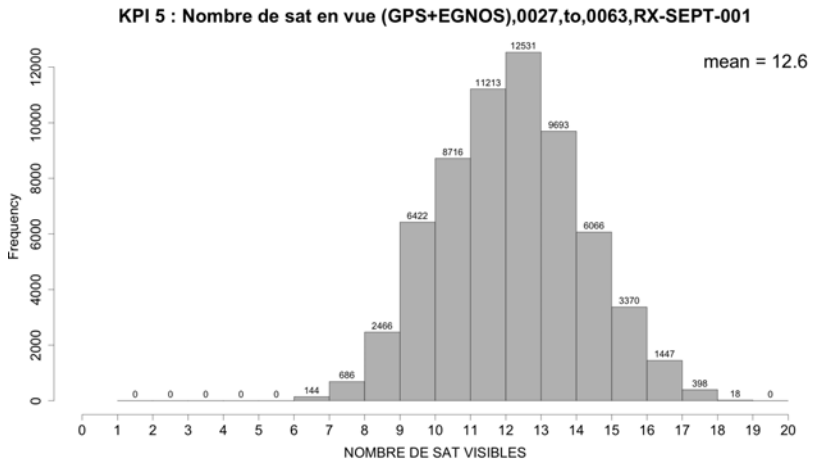


FIGURE B.10 – KPI 5

– **KPI 6** : Receivers computed position distance.
Le KPI 6 décrit les différences de coordonnées (sur les trois axes dans le repère géodésique) entre le récepteur considéré et un récepteur de référence présent dans la H-box à sélectionner.

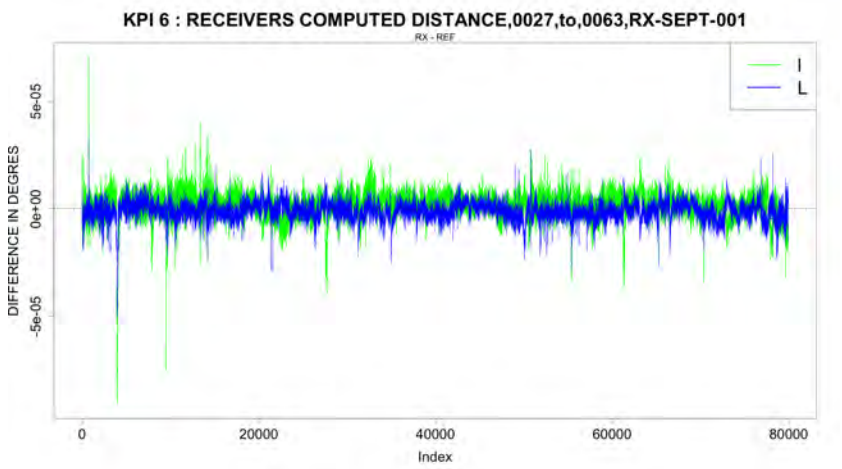


FIGURE B.11 – KPI 6-1 : Latitude, longitude

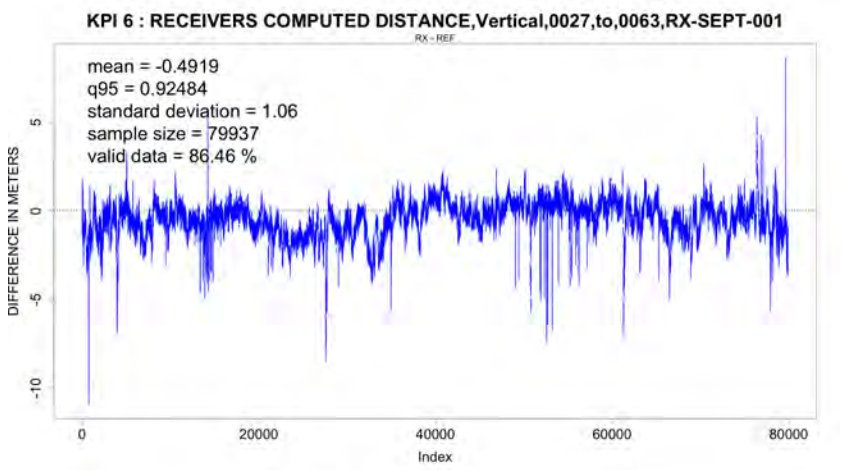


FIGURE B.12 – KPI 6-2 : Hauteur

– **KPI 7 et 8** : HDOP, VDOP.
Les KPI 7 et 8 décrivent les DOP horizontaux et verticaux calculés par le récepteur pendant le(s) vol(s).

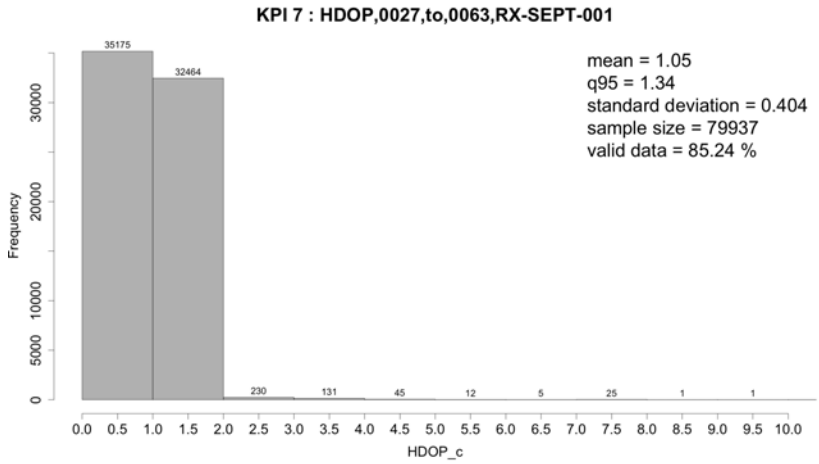


FIGURE B.13 – KPI 7-1

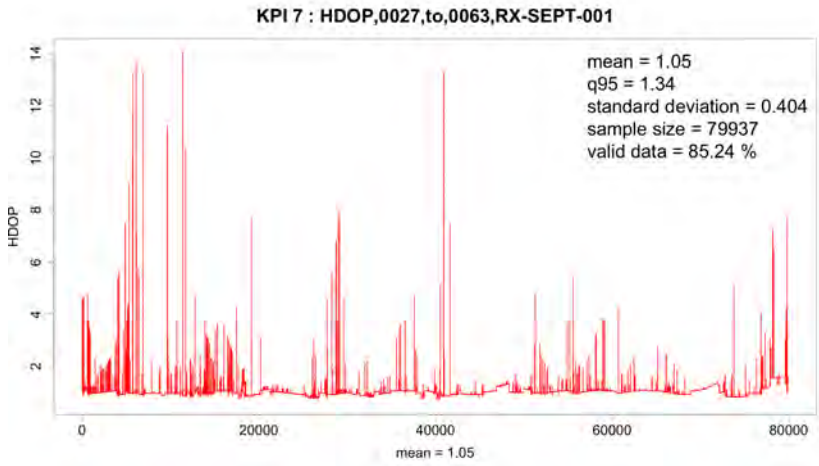


FIGURE B.14 – KPI 7-2

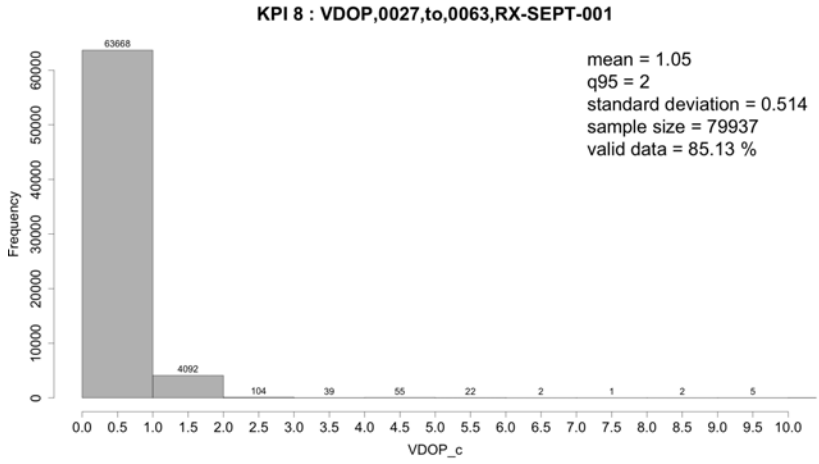


FIGURE B.15 – KPI 8-1

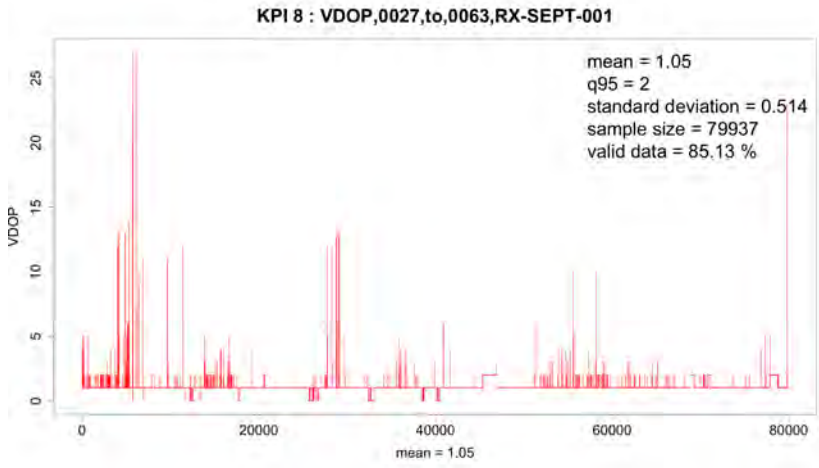


FIGURE B.16 – KPI 8-2

– **KPI 9 et 11** : HPE, VPE per GEO.
Les KPI 9 et 11 décrivent les erreurs de positionnement horizontales et verticales calculées à partir des positions estimées en utilisant les informations d'un des satellites EGNOS au lieu des trois. Ces KPI n'ont finalement pas été utilisés pour le rapport rendu à la commission.

– **KPI 10 et 12** : HPE, VPE.
Les KPI 10 et 12 décrivent les erreurs de positionnement horizontales et verticales.

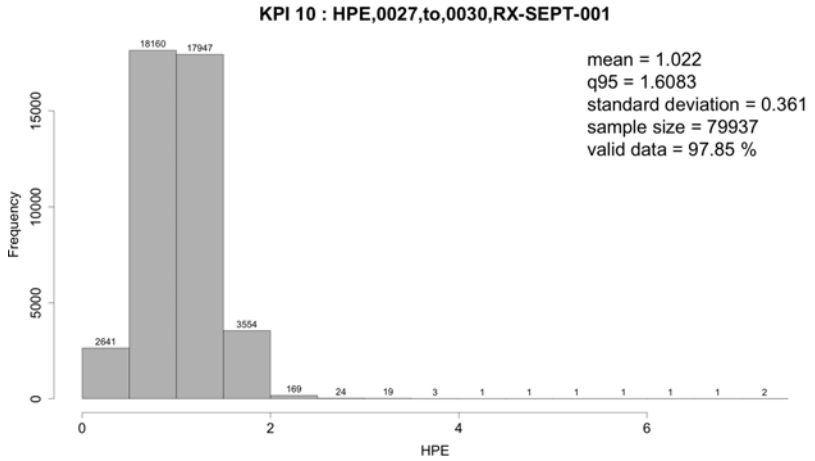


FIGURE B.17 – KPI 10-1

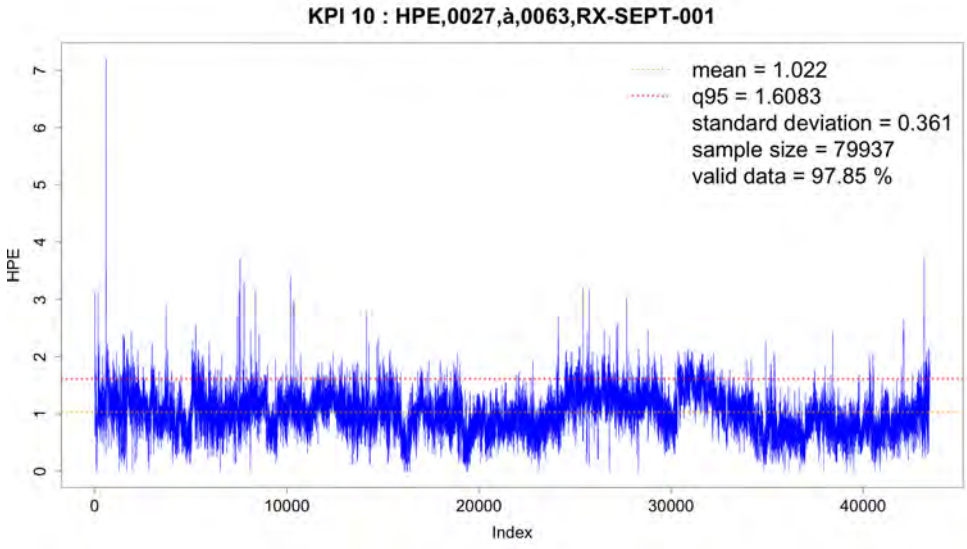


FIGURE B.18 – KPI 10-2

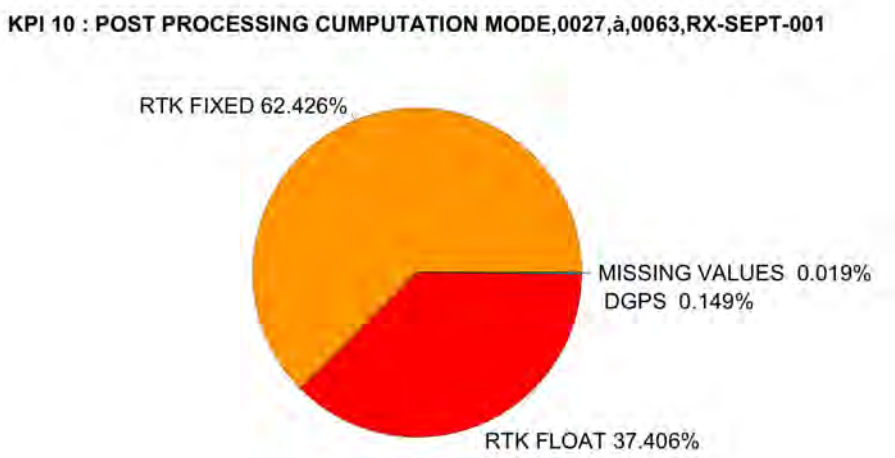


FIGURE B.19 – KPI 10-3

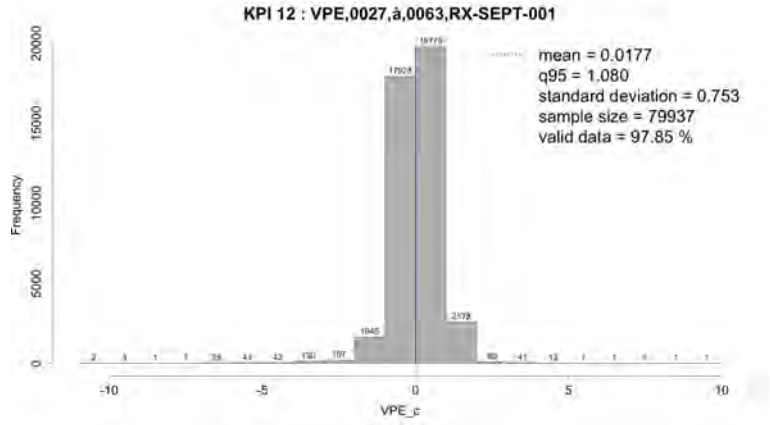


FIGURE B.20 – KPI 12-1

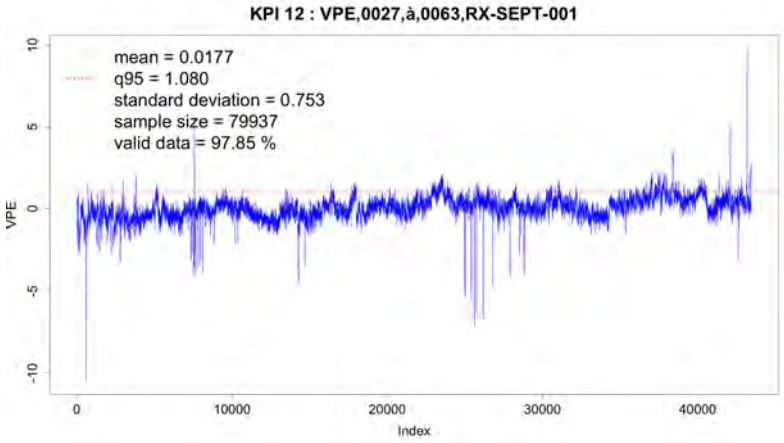


FIGURE B.21 – KPI 12-2

– **KPI 13 et 14** : Stanford plot
Les KPI 13 et 14 décrivent les performances d'intégrité suivant la représentation Stanford Plot [TST+07] sur l'axe vertical et sur le plan horizontal. Les erreurs HPE et VPE sont sur les axes des abscisses tandis que les niveaux de protection HPL et VPL sont sur les ordonnées. Ainsi on peut voir graphiquement les niveaux d'occurrence des événements MI et HMI. De plus, les niveaux d'alerte, AL, figurent aussi sur le graphique, permettant de dissocier les opérations de type CAT1 ou LPV.

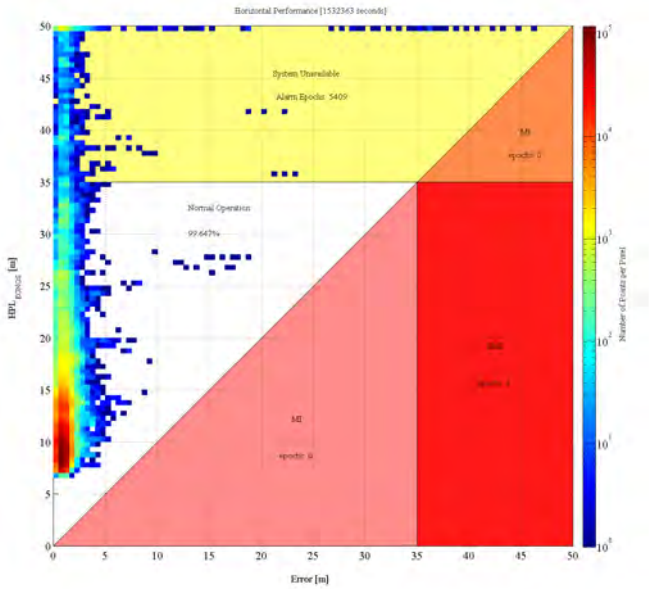


FIGURE B.22 – KPI 13

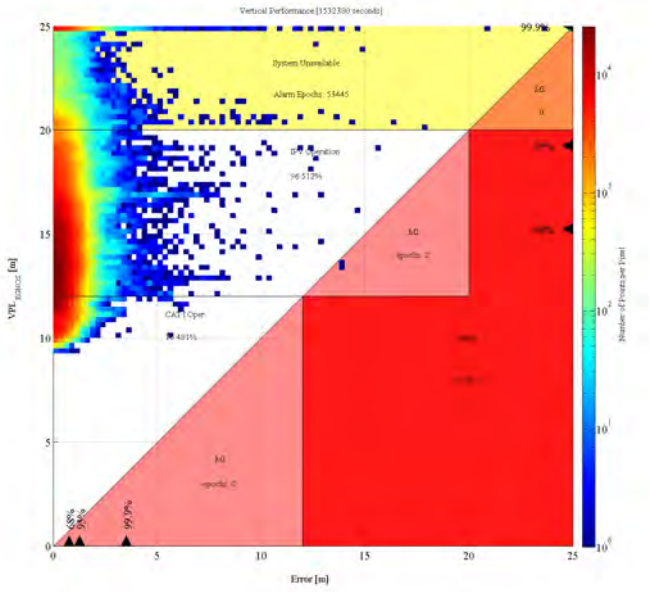


FIGURE B.23 – KPI 14

Bibliographie

[AG] J.M. Azaïs and S. Gadat. Evt-siam : A tool based on extreme-value theory for the assessment of accuracy and integrity assessment of sbas.

[AG09] J.M. Azaïs and S. Gadat. Gnss integrity achievement by using extreme value theory. 2009.

[BC+07] B. Bercu, D. Chafaï, et al. Modélisation stochastique et simulation-cours et applications. 2007.

[BDH74] A.A. Balkema and L. De Haan. Residual life time at great age. *The Annals of Probability*, 2(5) :792–804, 1974.

[Bei04] J. Beirlant. *Statistics of extremes : theory and applications*. John Wiley & Sons Inc, 2004.

[BN78] Barndorff-Nielsen. *Information and exponential families : in statistical theory*. Wiley New York, 1978.

[Bor07] K. Borre. *A Software-Defined GPS and Galileo Receiver : A Single-Frequency Approach*. Applied and Numerical Harmonic Analysis. Birkhäuser, 2007.

[BSK90] L. Breiman, C.J. Stone, and C. Kooperberg. Robust confidence bounds for extreme upper quantiles. *Journal of Statistical Computation and Simulation*, 37(3-4) :127–149, 1990.

[But07] R.W. Butler. *Saddlepoint approximations with applications*, volume 22. Cambridge Univ Pr, 2007.

[BVT96] J. Beirlant, P. Vynckier, and J.L. Teugels. Tail Index Estimation, Pareto Quantile Plots, and Regression Diagnostics. *Journal of the American Statistical Association*, 91(436), 1996.

[CBV+04] V. Calmettes, M. Bousquet, W. Vigneau, F. Legrand, and J. Lemorton. Analyse des brouillages non-intentionnels sur les systèmes de navigation par satellite et des techniques permettant d'en réduire les effets. 2004.

[CDM85a] S. Csorgo, P. Deheuvels, and D. Mason. Kernel estimates of the tail index of a distribution. *The Annals of Statistics*, 13(3) :1050–1077, 1985.

[CDM85b] S. Csörgő, P. Deheuvels, and D. Mason. Kernel estimates of the tail index of a distribution. *The Annals of Statistics*, 13(3) :1050–1077, 1985.

[CH88] M. Csörgő and L. Horváth. Central limit theorems for L p-norms of density estimators. *Probability theory and related fields*, 80(2) :269–291, 1988.

[CH04] E. Castillo and A.S. HADI. Extreme value & related models with applications in engineering & science. *Recherche*, 67 :02, 2004.

[Col73] TW Cole. Periodicities in solar activity. *Solar physics*, 30(1) :103–110, 1973.

[Col01] S. Coles. *An introduction to statistical modeling of extreme values*. Springer Verlag, 2001.

[CV98] S. Csörgo and L. Viharos. Estimating the tail index. *Asymptotic Methods in Probability and Statistics*, pages 833–881, 1998.

[Dan54] H.E. Daniels. Saddlepoint approximations in statistics. *The Annals of Mathematical Statistics*, 25(4) :631–650, 1954.

[Dan87] H.E. Daniels. Tail probability approximations. *International Statistical Review*, pages 37–48, 1987.

[DDH89] A.L.M. Dekkers and L. De Haan. On the estimation of the extreme-value index and large quantile estimation. *The Annals of Statistics*, pages 1795–1832, 1989.

[DDHR00] H. Drees, L. De Haan, and S. Resnick. How to make a hill plot. *The Annals of Statistics*, 28(1) :254–274, 2000.

[DEDH89] A.L.M. Dekkers, J.H.J. Einmahl, and L. De Haan. A moment estimator for the index of an extreme-value distribution. *The Annals of Statistics*, 17(4) :1833–1855, 1989.

[DFDH04] H. Drees, A. Ferreira, and L. De Haan. On maximum likelihood estimation of the extreme value index. *Annals of Applied Probability*, pages 1179–1201, 2004.

[DG85] L. Devroye and L. Györfi. *Nonparametric Density Estimation : The L B1 S View*. Wiley, 1985.

[DGR05] J. Diebolt, A. Guillou, and P. Ribereau. Asymptotic normality of the extreme quantile estimator based on the POT method. *Comptes Rendus Mathematique*, 341(5) :307–312, 2005.

[DH80] P. Deheuvels and P. Hominal. Estimation automatique de la densité. *Revue de Statistique Appliquée*, 28(1) :25–55, 1980.

[DHF06] L. De Haan and A. Ferreira. *Extreme value theory : an introduction*. Springer Verlag, 2006.

[DHM88] P. Deheuvels, E. Haeusler, and D.M. Mason. Almost sure convergence of the Hill estimator. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 104, pages 371–381. Cambridge Univ Press, 1988.

[dHR93] L. de Haan and H. Rootzen. On the estimation of high quantiles. *Journal of Statistical Planning and Inference*, 35(1) :1–13, 1993.

[dHR98] L.E. dE Haan and S. Resnick. On asymptotic normality of the Hill estimator. *Stochastic Models*, 14(4) :849–866, 1998.

[DO96] R. DO. 229 do-229a. *Minimum Operational Performance Standards for Global Positioning System/Wide Area Augmentation System Airborne Equipment*, 1996.

[DR84] R. Davis and S. Resnick. Tail estimates motivated by extreme value theory. *The Annals of Statistics*, pages 1467–1487, 1984.

[Dre98] H. Drees. Optimal rates of convergence for estimates of the extreme value index. *The Annals of Statistics*, 26(1) :434–448, 1998.

[DS90] A.C. Davison and R.L. Smith. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(3) :393–442, 1990.

[Duq97] H. Duquenne. Le modèle de quasi-géoïde français qgf96 et la surface de référence d'altitude raf96. *IGN/LAREG et ESGT, rapport interne*, page 46, 1997.

[DVF06] D.J. Dupuis and M.P. Victoria-Feser. A robust prediction error criterion for pareto modelling of upper tails. *Canadian Journal of Statistics*, 34(4) :639–658, 2006.

[Efr79] B. Efron. Bootstrap methods : another look at the jackknife. *The annals of Statistics*, 7(1) :1–26, 1979.

[EKM97] P. Embrechts, C. Kl
"uppelberg, and T. Mikosch. *Modelling extremal events for insurance and finance*. Springer Verlag, 1997.

[ET93] B. Efron and R. Tibshirani. *An introduction to the bootstrap*, volume 57. Chapman & Hall/CRC, 1993.

[Far08] J. Farrell. *Aided navigation : GPS with high rate sensors*. McGraw-Hill New York, NY, USA :, 2008.

[Fau11] F. Faurie. Algorithmes de contrôle d'intégrité pour la navigation hybride gnss et systèmes de navigation inertielle en présence de multiples mesures satellitaires défaillantes. 2011.

[Fré27] M. Fréchet. Sur la loi de probabilité de l'écart maximum. *Ann. de la Soc. polonaise de Math*, 6 :93–116, 1927.

[FT28] R.A. Fisher and L.H.C. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, pages 180–190. Cambridge Univ Press, 1928.

[Gar02] M. Garrido. *Modélisation des évènements rares et estimation des quantiles extrêmes, méthodes de sélection de modèles pour les queues de distribution.* PhD thesis, 2002.

[Gar03] L. Gardes. *Estimation d'une fonction quantile extrême.* PhD thesis, 2003.

[GDCT04] A. Giremus, A. Doucet, V. Calmettes, and J.Y. Tourneret. A rao-blackwellized particle filter for ins/gps integration. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 3, pages iii–964. IEEE, 2004.

[GLMW79] J.A. Greenwood, J.M. Landwehr, N.C. Matalas, and J.R. Wallis. Probability weighted moments : definition and relation to parameters of several distributions expressable in inverse form. *Water Resources Research*, 15(5) :1049–1054, 1979.

[Gne43] BV Gnedenko. Sur la distribution limite du terme maximum d'une serie aleatoire. *The Annals of Mathematics*, 44(3) :423–453, 1943.

[GTC07] A. Giremus, J.Y. Tourneret, and V. Calmettes. A particle filtering approach for joint detection/estimation of multipath effects on gps measurements. *Signal Processing, IEEE Transactions on*, 55(4) :1275–1285, 2007.

[GTD10] A. Giremus, J.Y. Tourneret, and A. Doucet. A fixed-lag particle filter for the joint detection/compensation of interference effects in gps navigation. *Signal Processing, IEEE Transactions on*, 58(12) :6066–6079, 2010.

[Gum58] E.J. Gumbel. *Statistics of extremes.* Columbia University Press, 1958.

[Hil75] B.M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5) :1163–1174, 1975.

[Hop71] H.S. Hopfield. Tropospheric effect on electromagnetically measured range : Prediction from surface weather data. *Radio Science*, 6(3) :357–367, 1971.

[Hos85] JRM Hosking. Algorithm AS 215 : Maximum-likelihood estimation of the parameters of the generalized extreme-value distribution. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 34(3) :301–310, 1985.

[Hsi93] T. Hsing. Extremal index estimation for a weakly dependent stationary sequence. *The Annals of Statistics*, 21(4) :2043–2071, 1993.

[HW87] J.R.M. Hosking and J.R. Wallis. Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, 29(3) :339–349, 1987.

[HWW85] JRM Hosking, J.R. Wallis, and EF Wood. Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, 27(3) :251–261, 1985.

[ICA06] J. ICAO. International standards and recommended practices, annex 10 to the convention on international civil aviation, radio navigation aids, 2006.

[ION97] ION Institute Of Naivigation. Recommended test procedures for gps receivers. 1997.

[KA61] M.G. Kendall and S. Alan. The advanced theory of statistics. Vols. II and III. 1961.

[KH06] E.D. Kaplan and C.J. Hegarty. *Understanding GPS : Principles And Applications*. Artech House Mobile Communications Series. Artech House, 2006.

[Klo87] J.A. Klobuchar. Ionospheric time-delay algorithm for single-frequency gps users. *Aerospace and Electronic Systems, IEEE Transactions on*, (3) :325–331, 1987.

[Lea74] M. R. Leadbetter. On extreme values in stationary sequences. *Probability Theory and Related Fields*, 28 :289–303, 1974. 10.1007/BF00532947.

[Lea83] MR Leadbetter. Extremes and local dependence in stationary sequences. *Probability Theory and Related Fields*, 65(2) :291–306, 1983.

[Lea91] MR Leadbetter. On a basis for [] Peaks over Threshold'modeling. *Statistics & probability letters*, 12(4) :357–362, 1991.

[LR80] R. Lugannani and S. Rice. Saddle point approximation for the distribution of the sum of independent random variables. *Advances in Applied Probability*, pages 475–490, 1980.

[LRF] J.C. Lévy, B. Rols, and T.A.S. France. GNSS Integrity Achievement by using Extreme Value Theory.

[Mar08] A. Martineau. Etude de la performance du contrôle autonome d'intégrité pour les approches à guidage vertical. 2008.

[MMC02] J.P. Malet, O. Maquaire, and E. Calais. Le gps en géomorphologie dynamique. application à la surveillance de mouvements de terrain (super-sauze, alpes du sud, france)/gps in geomorphological studies. application to the survey of landslides (super-sauze, south france). *Géomorphologie : relief, processus, environnement*, 8(2) :165–179, 2002.

[MP97] Y. Maesono and S.I. Penev. Higher order relations between cornish-fisher expansions and inversions of saddlepoint approximations. *COMPUTING SCIENCE AND STATISTICS*, pages 242–247, 1997.

[Mul06] A. Muller. *Comportement asymptotique de la distribution des pluies extrêmes en France*. PhD thesis, 2006.

[MXWL12] R. Ma, J. Xu, W. Wang, and J. Lei. The effect of 27 day solar rotation on ionospheric f2 region peak densities (nmf2). *Journal of Geophysical Research*, 117(A3) :A03303, 2012.

[NCSS12] A. Nina, V. Cadez, VA Sreckovic, and D. Sulic. The influence of solar spectral lines on electron concentration in terrestrial ionosphere. *Arxiv preprint arXiv :1201.2282*, 2012.

[Par62] E. Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3) :1065–1076, 1962.

[Pha07] Huyên Pham. *Optimisation et contrôle stochastique appliqués à la finance*, volume 61. Springer, 2007.

[Pic75] J. Pickands. Statistical inference using extreme order statistics. *the Annals of Statistics*, 3(1) :119–131, 1975.

[Pié06] J.M. Piéplu. *GPS et Galileo : Systèmes de navigation par satellites*. Eyrolles, 2006.

[PS96] B.W. Parkinson and J.J. Spilker. *Global positioning system : theory and applications*. Number vol. 1 ;vol. 163 in Progress in astronautics and aeronautics. American Institute of Aeronautics and Astronautics, 1996.

[Res97] S.I. Resnick. Heavy tail modeling and teletraffic data. *The Annals of Statistics*, 25(5) :1805–1849, 1997.

[Res07] S.I. Resnick. *Extreme values, regular variation and point processes*. Springer Verlag, 2007.

[SAR99] I.G. SARP. Draft standards and recommended practices for the global navigation system. *Satellite-Based Augmentation Systems (SBAS). Appendix B*, 1999.

[Sma10] C.G. Small. *Expansions and asymptotics for statistics*, volume 115. CRC Press, 2010.

[Smi84] R.L. Smith. Threshold methods for sample extremes. *Statistical extremes and applications*, 621 :638, 1984.

[Smi85] R.L. Smith. Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, 72(1) :67, 1985.

[Smi86] R.L. Smith. Extreme value theory based on the r largest annual events. *Journal of Hydrology*, 86(1-2) :27–43, 1986.

[Smi87] R.L. Smith. Estimating tails of probability distributions. *The annals of Statistics*, 15(3) :1174–1207, 1987.

[Smi89] R.L. Smith. Extreme value analysis of environmental time series : an application to trend detection in ground-level ozone. *Statistical Science*, 4(4) :367–377, 1989.

[Str64] T. Strodka. Estimateurs et intervalles de confiance d'un certain paramètre dans la distribution gamma généralisée de maxwell et de weibull. *Revue de Statistique Appliquée*, 12(2) :79–83, 1964.

[SW94] R.L. Smith and I. Weissman. Estimating the extremal index. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(3) :515–528, 1994.

[TC10] MZ Tian and N.H. Chan. Saddle point approximation and volatility estimation of value-at-risk. *Statistica Sinica*, 20(3) :1239–1256, 2010.

[Tou08] G. Toulemonde. *Estimation et tests en théorie des valeurs extrêmes*. PhD thesis, 2008.

[TST+07] M. Tossaint, J. Samson, F. Toran, J. Ventura-Traveset, M. Hernandez-Pajares, JM Juan, J. Sanz, and P. Ramos-Bosch. The stanford-esa integrity diagram : A new tool for the user domain sbas integrity assessment. *Navigation-Los Angeles and Washington*, 2007.

[Tsy09] A.B. Tsybakov. *Introduction to nonparametric estimation*. Springer Verlag, 2009.

[VdV00] A.W. Van der Vaart. *Asymptotic statistics*. Cambridge Univ Pr, 2000.

[WBB93] A.T.A. Wood, J.G. Booth, and R.W. Butler. Saddlepoint approximations to the cdf of some statistics with nonnormal limit distributions. *Journal of the American Statistical Association*, pages 680–686, 1993.

[Wei78] I. Weissman. Estimation of parameters and larger quantiles based on the k largest observations. *Journal of the American Statistical Association*, 73(364) :812–815, 1978.

[WW69] CT Wolverton and T.J. Wagner. Recursive estimates of probability densities. *Systems Science and Cybernetics, IEEE Transactions on*, 5(3) :246–247, 1969.

Résumé

Parmi les applications GNSS (Global Navigation Satellite System) existantes ou en développement, certaines dont l'aviation, nécessitent de hautes performances en termes de précision de positionnement et de fiabilité.
Ces performances critiques sont évaluées à l'aide d'outils probabilistes et le problème d'appréciation de la précision ou de la fiabilité du système (intégrité) peut être vu comme une estimation de quantile. Ce problème inverse nécessite la connaissance de la fonction de répartition des observations, ce qui n'est pas le cas lorsque l'on travaille sur des données réelles. Il faut alors utiliser des techniques statistiques pour l'estimer.
Les exigences spécifiques à certaines applications, comme par exemple l'atterrissage d'un avion, nécessitent des niveaux de quantiles très élevés atteignant des probabilités de l'ordre de 10^{-7}. Ces probabilités correspondent à des fréquences d'occurrence d'événements rares, situés dans les queues de distribution. Les quantiles associés à de tels niveaux de probabilité sont qualifiés de quantiles extrêmes et se situent le plus souvent au-delà du domaine des observations.
Nous proposons dans cette thèse deux méthodes d'estimation de quantiles extrêmes peu employées dans le domaine du GNSS. La première est une application des modèles issus de la théorie des extrêmes et plus particulièrement du modèle à dépassement de seuil POT (Peak Over Threshold). Cette théorie fournit une classe de modèles permettant l'extrapolation de l'observé vers le non observé et ainsi la caractérisation des événements rares qui peuvent ne jamais avoir été observés. La deuxième méthode fournit une approximation de la décroissance de la queue d'une distribution au moyen de techniques analytiques adaptées à un cadre statistique : il s'agit de la méthode du point selle.
Ces deux techniques de caractérisation des fonctions de répartition sont valables sous certaines hypothèses de stationnarité et d'indépendance des observations ; or les données GPS ne vérifient pas toujours ces conditions. Dans ce travail, nous proposons des méthodes statistiques pour stationnariser les données afin d'utiliser les modèles d'estimation de quantiles extrêmes dans un cadre adéquat.
A partir des outils décrits dans cette thèse, nous fournissons un protocole d'analyse statistique d'intégrité. Les problématiques de calibration de ces outils sont traitées par des processus automatisés dans une plateforme d'analyse de données, support logiciel développé pour cette étude.

Mots-clefs : GNSS, GPS, intégrité, théorie des valeurs extrêmes, quantile extrême, dépendance temporelle, approximation point selle.