Proceedings of the Institute of Acoustics

TESTS DE LOIS SUR DES BRUITS PRELEVES EN MER

E. PETIT B. FAURE G. JOURDAIN

CEPHAG U.R.A. C.N.R.S. 346

ENSIEG BP 46 38402 St MARTIN D'HERES - FRANCE

Résumé :
Dans cette étude nous nous intéressons aux propriétés statistiques des bruits et signaux propagés en mer. Nous appliquons sur des échantillons de bruit prélevés en mer des tests non paramétriques ("run above the mean" et Wilcoxon) pour former des échantillons aléatoires et étudier l'homogénéité de la loi. Afin d'identifier la distribution statistique du bruit ambiant, nous proposons un modèle stochastique de type mélange fini de lois gaussiennes et nous utilisons l'algorithme E. M. pour en estimer les paramètres.
Mots-clés : acoustique sous-marine, distribution statistique, tests non paramétriques, loi de Gauss, loi "mélange".

1. INTRODUCTION

Lorsqu'on dispose d'une importante base de données expérimentales, relativement à un caractère mesurable donné (pression acoustique, tension ou courant électrique, ...), la mise en évidence des propriétés statistiques du caractère à étudier exige certaines précautions. De multiples questions se posent quant à la manière de procéder, comme par exemple : comment doit on former un échantillon de qualité ? Quel est le critère de qualité à retenir ? Doit on prendre un maximum de points ? Comment s'assurer de l'homogénéité de la loi de probabilité que l'on cherche à identifier ? Par ailleurs, bien qu'il existe une littérature abondante sur le sujet, cet avantage peut apparaître comme un inconvénient lorsqu'il faut choisir un test statistique parmi tous ceux existant. Pour ce qui est du choix des tests d'indépendance et d'homogénéité, nous avons retenu deux tests, respectivement le test des "longueurs" ("test of the run above the mean"), et celui de Wilcoxon (test des rangs). Ce choix a été dicté par des raisons de simplicité mais aussi par le fait que ces tests sont non paramétriques et n'utilisent par conséquent aucune hypothèse a priori quant à la distribution des données.
Dans un premier temps nous positionnons le problème, puis dans un deuxième temps nous décrivons les diverses méthodes utilisées dans la procédure d'analyse, enfin nous appliquons ces tests à des bruits ambiants captés en mer et nous proposons un modèle stochastique de bruit.
Au cours de cette présentation, nous insistons en particulier sur la notion d'échantillon aléatoire, propriété essentielle sous-jacente à toute étude statistique.

2. POSITION DU PROBLEME

2.1 Rappel : tests d'hypothèse

Les tests statistiques ont pour but de déterminer si une certaine hypothèse Ho concernant les données à étudier est vraie ou fausse. En traitement du signal, un certain nombre de propriétés permettent de simplifier considérablement les calculs nécessaires à des estimateurs de paramètres, ou à des récepteurs optimaux :

* l'indépendance : un échantillon $(x_1, x_2, .., x_n)$ est indépendant si les n variables aléatoires qui le génèrent sont indépendantes
* le caractère aléatoire : $(x_1, x_2, .., x_n)$ est aléatoire s'il est composé des réalisations de n variables aléatoires et de même loi
* l'homogénéité : $(x_1, x_2, .., x_n)$ et $(y_1, y_2, .., y_m)$ sont homogènes s'ils sont aléatoires et suivent une même loi.

TESTS DE LOIS SUR DES BRUITS PRELEVES EN MER.

La méthodologie des tests d'hypothèse demande de construire d'abord une statistique, c'est à dire une fonction des données que l'on comparera à un seuil et dont on connaît la densité de probabilité. On définit le seuil de telle manière que la probabilité de déclarer Ho fausse lorsque elle est vraie soit égale à une certaine valeur α. α est alors appelé niveau de signification du test. En général, la densité de probabilité des données n'est pas connue. C'est pourquoi on utilise des tests non-paramètriques, indépendants de la loi. Notons que ces tests reposent sur le fait que les (n!) arrangements possibles des x_i sont tous équiprobables, cette condition étant garantie par la nature aléatoire des échantillons (condition suffisante). En ce qui concerne les signaux que nous avons à traiter, les hypothèses qui nous intéressent sont l'homogénéité et l'indépendance.

2.2 Discussion

Nous abordons cette étude par la voix statistique, cependant il est utile de mener conjointement une approche temporelle en examinant la stationnarité des signaux traités. On examine dans un premier temps la stationnarité au sens large qui met en jeu les moments statistiques d'ordre 1 et 2 (moyenne, variance, ou de façon moins directe autocorrélation et densité spectrale de puissance). Une façon progressive de mesurer la stationnarité est alors de tracer, par exemple, la moyenne et la variance en fonction du temps en utilisant une "fenêtre glissante". Si l'on pousse l'étude jusqu'au moment d'ordre 4, on obtient alors des informations sur le caractère gaussien ou non du signal étudié (skew = 0 et kurtosis = 3 si la loi est gaussienne). On procède ensuite au tracé de l'histogramme pour caractériser de façon plus complète la distribution du bruit. Généralement, les histogrammes sont moyennés sur un certain nombre de réalisations. Dans ce cas, il faut veiller à ce que les propriétés statistiques du processus soient invariantes au cours du temps, d'une part, et d'autre part, à ce que les échantillons traités contiennent bien toute l'information statistique.
Revenons à l'idée première d'échantillons aléatoires, et naturellement à l'étude de l'indépendance statistique. Rappelons, que si deux variables aléatoires X_1 et X_2 sont indépendantes, leur densité de probabilité conjointe est égale au produit des densités des probabilités marginales. Nous disposons d'une suite d'observations ($x_1, x_2, .. ,x_n$), qui sont considérées comme réalisations des variables aléatoires $X_1, X_2, .. ,X_n$. S'il existe un entier k tel que les variables aléatoires X_i et X_{i+k} soient indépendantes alors on peut former, à partir de la suite aléatoire d'observations, un échantillon aléatoire en "sautant" k données entre deux valeurs successives. Il reste maintenant à évaluer le "saut temporel" k. Dans le cas d'un bruit gaussien la non corrélation entraîne l'indépendance statistique et il suffit d'examiner la fonction d'autocorrélation pour en déduire les valeurs possibles de k. On peut également obtenir un ordre de grandeur de ce paramètre à partir de la connaissance spectrale du signal aléatoire.
Cependant, comme en pratique le bruit n'est jamais tout à fait gaussien ni tout à fait stationnaire, il est recommandé de procéder à un test statistique sur l'échantillon en question. C'est précisément le rôle du test du "run above the mean" dont l'hypothèse testée est " les v. a. sont indépendantes et suivent une même loi" . Notons, que ce test d'hypothèse pourrait également servir à tester l'homogénéité de la loi. Toutefois dans ce but, nous utiliserons le test de Wilcoxon appliqué à deux échantillons aléatoires espacés dans le temps.
Cette étude débouche sur la présentation d'un modèle statistique de type mélange fini de lois gaussiennes. Pour identifier les paramètres de ce modèle, nous utiliserons l'estimateur du maximum de vraisemblance. La procédure itérative qui en découle est en fait l'application de l'algorithme E. M. (Expectation Maximisation) au cas particulier du mélange fini de lois gaussiennes.

TESTS DE LOIS SUR DES BRUITS PRELEVES EN MER.

3. METHODOLOGIE

Les étapes décrites ci-dessous retracent les grandes lignes de l'étude du bruit ambiant. Les résultats obtenus à partir des données expérimentales reçues en a. s. m. seront présentés dans les parties 4 et 5 :

Etape 1 : formation des échantillons

Soit la suite aléatoire issue du signal reçu (échantillonné à la fréquence $1/T_e$). On forme alors des échantillons de taille N par des sauts de k valeurs dans la suite aléatoire. En d'autres termes, cette opération revient à sous-échantillonner le signal reçu avec une période : $T_e' = (k+1)T_e$.
De cette manière, à partir d'une même suite, il est tout à fait possible de créer des échantillons comparables du point de vue statistique et sur lesquels seront appliqué un ou plusieurs tests statistiques d'hypothèses.

Etape 2 : analyse classique des moments jusqu'à l'ordre 4

On calcule la densité spectrale de puissance, l'autocorrelation et les différents moments statistiques d'ordre 1, 2, 3, et 4 au cours du temps, soit encore : la moyenne, la variance, le "skew" et le "kurtosis" . Cette analyse préalable nous donne des informations sur le caractère gaussien du signal, sur la nature plus ou moins stationnaire de la loi statistique et sur la relation de dépendance entre les valeurs observées.

Etape 3 : test d'indépendance statistique

Le test d'hypothèse utilisé est le test du "Run above the mean" défini de la façon suivante :
à partir d'un échantillon $(x_1,x_2,..,x_n)$, on forme la suite $u_1,u_2,..,u_n$ telle que :

$$u_i = 1 \text{ si } x_i > \mu_x$$

$$u_i = 0 \text{ si } x_i < \mu_x$$

Une épreuve est ainsi constituée de n_0 '0' et n_1 '1' .
Exemple : {0,0,1,1,1,0,1,0,0,1,1,1,1,1,0 }

On note R le nombre de longueurs de l'épreuve, c'est à dire le nombre de suites maximales de '1' ou '0' (dans notre exemple R vaut 7).
Les étapes du test de signification se décomposent de la sorte :

1) l'énoncé de l'hypothèse H_0, que l'on accepte provisoirement,
H_0 : les v. a. sont indépendantes et suivent une même loi,

2) le recours à une v. a. dont on connaît la loi de probabilité quand H_0 est exacte,
nous admettrons [2] que R suit une loi normale $N(\mu,\sigma^2)$ avec :

$$\sigma^2 = \frac{2n_0n_1(2n_0n_1 - n_0 - n_1)}{(n_0 + n_1)^2(n_0 + n_1 - 1)} \quad \text{et} \quad \mu = 1 + \frac{2n_0n_1}{n_0 + n_1}$$

3) la détermination d'une valeur critique calculée d'après cette loi de probabilité,
et d'après le seuil de signification α :

$$\text{Proba } (r_1 \le R \le r_2) = 1 - \alpha$$

$$\Rightarrow \quad \text{Proba } (-\zeta \le Z \le \zeta) = 1 - \frac{\alpha}{2} \quad \text{si} \quad Z = \frac{R-\mu}{\sigma}$$

La valeur critique ζ se déduit des tables de la loi normale
Par exemple pour $\alpha = 0.1$ on a $\zeta = 1.65$ d'où $r_1 = \mu - \sigma\zeta$ et $r_2 = \mu + \sigma\zeta$

4) le calcul, à partir de l'échantillon observé, de la valeur prise par la statistique R,

5) la décision qui rejettera H_0 ou l'acceptera,
si R appartient à l'intervalle $[r_1, r_2]$ alors on accepte l'hypothèse H_0.

Remarque :
si R n'appartient pas à l'intervalle $[r_1, r_2]$ on rejette H_0, car si H_0 était exacte on aurait une proportion de $(1-\alpha)$ chance de trouver R dans cet intervalle.

Etape 4 : test d'homogénéité

On considère ici la statistique de Wilcoxon qui teste aussi l'hypothèse : les v. a. sont indépendantes et suivent une même loi .
L' élément nouveau est la nature supposée aléatoire des échantillons testés. Cette condition étant acquise, on ne met en évidence que les variations d'homogénéité de la loi.
Nous appliquons donc ce test en prenant deux par deux des échantillons aléatoires appartenant à deux suites aléatoires finies, espacées dans le temps. La procédure est la suivante :
soient deux échantillons aléatoires $(x_1,x_2,..,x_m)$ et $(y_1,y_2,..,y_n)$. On range les valeurs par ordre croissant dans un même échantillon $\{z_i\}$ de taille (n+m). La statistique de Wilcoxon est alors la somme T des rangs occupés par les z_i. On montre [2] que la v. a. T tend asymptotiquement vers la loi de probabilité gaussienne $N\{E(T),VAR(T)\}$ avec :

$$E(T) = \frac{m(m+n+1)}{2} \quad \text{et} \quad VAR(T) = \frac{mn(m+n+1)}{12}$$

Fort de cette loi de probabilité valable lorsque H_0 est exacte, on peut à nouveau faire le test d'hypothèse et adopter un critère de décision.

Etape 5 : identification de la loi de probabilité

Il est certain que l'identification de la distribution statistique présente d'autant plus d'intérêt que le phénomène statistique possède de bonnes qualités de reproductibilité. Cette étape est donc conditionnée au bon déroulement des tests précédents (étapes 3 et 4).
Pour approcher au mieux la loi réelle, nous avons choisi un modèle de type mélange fini de lois gaussiennes. Il s'avère, en effet, que ce modèle est particulièrement bien approprié au bruit ambiant sous-marin dont la loi présente des propriétés de symétrie, d'unimodalité, etc ... assez proches de la courbe en "cloche" du modèle gaussien.
Pour définir la loi "**mélange**", nous partons du principe que **le bruit est créé par un ensemble de K sources distinctes, émettant chacune un signal aléatoire gaussien de moyenne μ_j et de variance $a_j\sigma^2$ (j = 1...K), mais dont l'activité est aussi un phénomène aléatoire régi par la probabilité d'activité p_j.**
La densité de probabilité du bruit s'exprime donc, pour chaque observation y_i par :

$$P_Y(y_i) = \sum_{j=1}^{K} p_j G_j\left(y_i, a_j\sigma^2\right) \quad \text{avec} \quad G_j\left(y_i, a_j\sigma^2\right) = \frac{1}{\sqrt{2\pi\left(a_j\sigma^2\right)}} \exp\left(-\frac{y_i^2}{2\,a_j\sigma^2}\right)$$

Remarques :

* a_j est un coefficient de pondération associé à la variance σ^2 de l'échantillon.
* Les proportions p_j vérifient la relation :

$$\sum_{j=1}^{K} p_j = 1$$

- Pour garantir l'unimodalité de la distribution, nous prendrons pour moyenne : $\mu_j = \mu = 0$.

Pour estimer les K coefficients a_j et les K proportions p_j de la loi, nous utiliserons le critère du maximum de vraisemblance, qui rend maximale la probabilité de réalisation des évènements y_i. Le calcul conduit à des expressions implicites dont découle la procédure itérative décrite dans [3] (Algorithme E. M.).

Remarques :

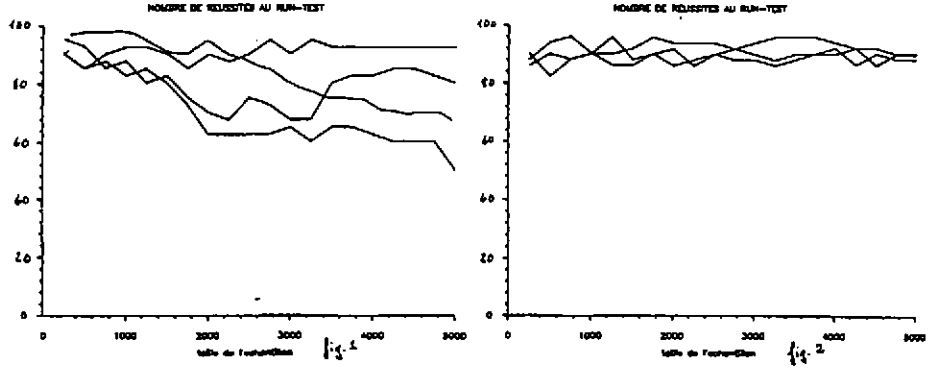
- La formation préalable des échantillons aléatoires prend ici tout son intérêt. D'une part la propriété d'indépendance statistique entre les v. a. permet de simplifier considérablement l'algorithme, d'autre part la dimension relativement modeste des échantillons considérés entraine une réduction significative du temps calcul.
- Initialement, on peut donner au paramètre K (nombre de sources) une valeur élevée de l'ordre de 4 ou 5 et si après estimation certaines composantes ont soit des poids faibles, soit des coefficients pondérateurs a_j proches, alors il est naturel d'abaisser la valeur de K de sorte qu'aucun des deux cas susdits ne se rencontre.
- Enfin il ne faudrait pas croire que ce modèle est en concurrence avec le modèle gaussien, au sens où en tout état de cause, il est soit meilleur soit identique. Il se peut notamment que l'on arrive à la situation : $a_j = 1$ et $p_j = p = 1/K$ pour $j = 1 \ldots K$; qui n'est autre que le cas gaussien $N(0,\sigma^2)$.

4. RESULTATS DES TESTS D'HYPOTHESES

Nous présentons ici une partie des résultats relatifs à des signaux prélevés dans l'océan Atlantique.

4.1 Test d'indépendance statistique

Nous avons représenté sur les figures 1 et 2 les résultats du test du "run above the mean" obtenus sur des échantillons issus d'un même signal mais dont la formation diffère par le paramètre k (voir étape 1) :
fig. 1 k = 39 ; fig. 2 k = 49. Le paramètre k étant fixé, nous formons k+1 échantillons de taille N sur lesquels nous réalisons le test d'indépendance. Le nombre de réussites au test est décrit par l'axe vertical, la variable N par l'axe horizontal. Pour une même figure, les différentes courbes sont obtenues en traitant les mêmes signaux mais pris à des instants différents, et en effectuant la même procédure.



Interprétation :
Plus la taille de l'échantillon augmente, plus le test devient puissant, et donc apte à rejeter l'hypothèse d'échantillons aléatoires. Ceci se traduit par un effondrement des courbes lorsque cette condition n'est pas respectée.

TESTS DE LOIS SUR DES BRUITS PRELEVES EN MER.

Lorsque k prend des valeurs importantes, on observe en règle générale, un taux de réussite au test plus élevé. Ce résultat n'a rien de surprenant si l'on examine la fonction d'autocorrélation. Rappelons qu'en théorie, pour un risque d'erreur toléré $\alpha = 0.1$, si l'hypothèse H_0 est vraie, alors le test décide H_0 à 90%, et cela quelque soit la taille N fixée. Or il se trouve que pour les 3 courbes de la figure 2, qui correspondent à $k = 49$, le pourcentage de réussites au test est presque constant et très proche de 90%. Nous sommes donc en droit de conclure au caractère aléatoire des échantillons traités.

4.2 Test d'homogénéité

On considère des échantillons constants de dimension égale à 500, formés en tenant compte de $k = 49$. On compare alors les 50 échantillons de la première suite aléatoire avec les 50 autres de la deuxième suite, puis de la troisième, etc ... de façon à couvrir le signal entier (correspondant aux 3 minutes d'enregistrement). La figure 3 illustre le résultat du test appliqué au signal précédent traité avec $k = 49$. Pour cette valeur nous avons établi précédemment que les échantillons étaient décorrélés (voir fig. 2). Au regard de la figure 3 nous pouvons maintenant conclure que la loi de probabilité est homogène sur la durée d'enregistrement. Nous vérifions en effet que le pourcentage de réussite au test se maintient autour d'une valeur moyenne de 90% conformément au niveau de signification du test fixé à 0.1.

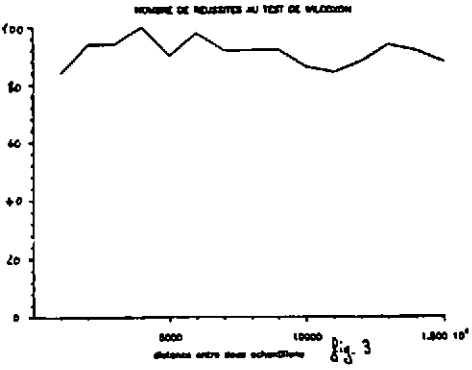


Fig. 3

5. IDENTIFICATION DE LA LOI DE PROBABILITE

L'identification est effectuée conformément à l'étape 5 (cf partie 3) en considérant un mélange de deux lois gaussiennes centrées. Nous nous intéressons ici, plus spécifiquement à l'identification de la distribution statistique de deux signaux aléatoires : l'un prélevé en Méditerranée (signal 1) et l'autre prélevé dans l'océan atlantique (signal 2).

Remarques :
Les histogrammes ont une surface normalisée égale à 1 de telle sorte qu'ils soient homogènes à une d.d.p. . On peut alors les comparer aux d.d.p. estimées à partir du même échantillon mais suivant un modèle mathématique caractérisé notamment par des paramètres variables.
Sans remettre en cause l'estimation de ces paramètres, l'histogramme doit nous aider à mesurer l'adéquation du modèle statistique avec la loi de probabilité du caractère mesurable que nous étudions (Il est clair, que si celle-ci présente deux maxima, l'hypothèse de loi unimodale n'est pas respectée et donc aucun des deux modèles définis précédemment ne peut convenir).
Notons que le caractère mesurable appartient à un phénomène physique aléatoire et continu. Nous pouvons en conséquence admettre que sa d.d.p est une fonction continue et qui plus est de forme régulière . Il en découle qu'un histogramme bien construit doit également présenter une forme régulière.
Pour finir, précisons que l'algorithme E. M. utilisé pour identifier les paramètres des modèles statistiques pourrait aussi s'appliquer à des mélanges de lois non gaussiennes telles que des lois exponentielles, gamma, ...

Interprétation des graphes :
Les figures 4 et 5 mettent en évidence une distribution statistique différente pour nos deux signaux. Pour s'en convaincre, il suffit d'examiner les valeurs des paramètres de la loi "mélange" , en effet :

TESTS DE LOIS SUR DES BRUITS PRELEVES EN MER.

Signal 1 (origine Méditerranée) : $\sigma^2 = 6300$ $a_1 = 0.8$ $a_2 = 1.2$
Signal 2 (origine Océan atlantique) : $\sigma^2 = 28000$ $a_1 = 0.6$ $a_2 = 1.4$

Notons que les deux composantes de la loi "mélange" interviennent avec à peu près la même proportion, $p_1 = p_2 = 0.5$ dans les deux cas.
Au niveau de la forme, l'écart entre les coefficients a_i se traduit sur les résultats relatifs au signal 2 par une distribution statistique plus pointue au centre et plus forte sur les parties latérales. En effet, la valeur du couple (a_1, a_2) s'éloigne du couple (1, 1) qui correspond au modèle gaussien. De cette façon, la loi mélange (courbe 2) se superpose rigoureusement à l'histogramme. Pour confirmer les résultats des tests d'indépendance et d'homogénéité, nous avons refait des tests sur d'autres échantillons issus du même signal. Nous vérifions bien que les résultats sont reproductibles dans la mesure où les échantillons sont soigneusement sélectionnés (étapes 1, 3, et 4). Si cette sélection n'est pas rigoureusement faite, le modèle estimé (loi "mélange") ne correspond plus à la distribution réelle du bruit.

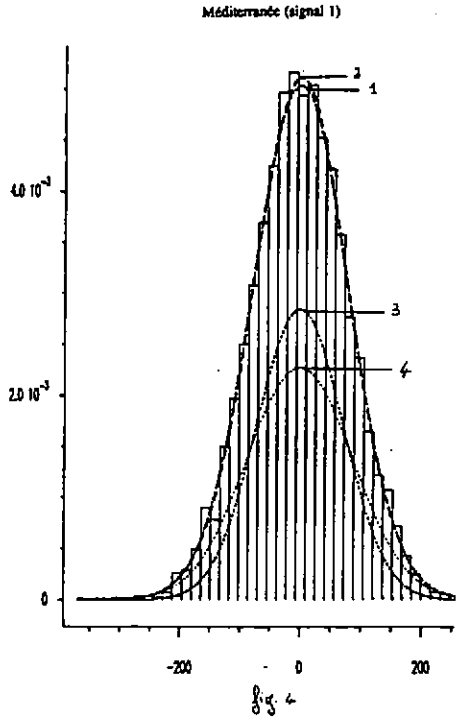


Fig. 4

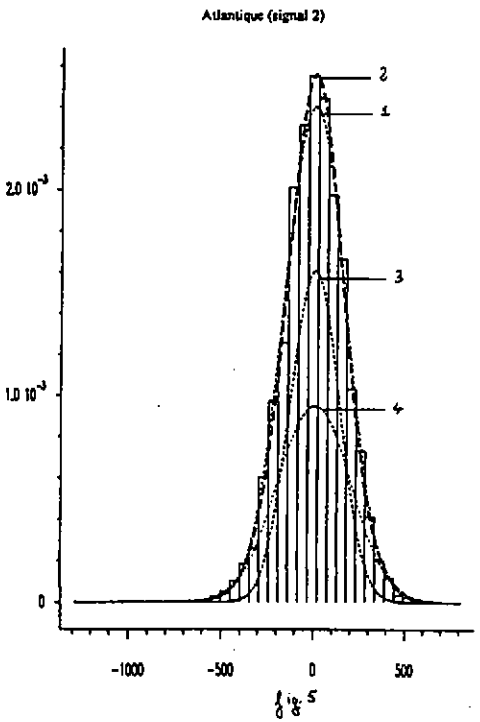


Fig. 5

Fig. 4 et 5 : Tracé de l'histogramme (courbe en escalier), de la densité de probabilité du bruit estimée à partir du modèle gaussien (courbe n°1) et à partir du modèle de type loi "mélange" (courbe n°2). Les courbes n°3 et n°4 correspondent aux d.d.p. associées aux deux composantes de la loi "mélange" (la somme des deux donnant la courbe n°2). La dimension des échantillons traités (signal 1 et 2) est égale à 5000.

6. CONCLUSION

Au cours de cette étude, nous avons proposé de façon concrète une méthodologie pour l'analyse statistique de bruit. Celle-ci a été appliquée dans le domaine de l'acoustique sous-marine sur un bruit ambiant prélevé in situ par un capteur. En s'appuyant sur les résultats expérimentaux nous avons montré la validité des tests ainsi que leur complémentarité et nous avons mis en évidence le fait que la nature aléatoire du bruit ambiant change suivant le lieu géographique. En particulier, on a pu constater que la distribution statistique du bruit capté en Atlantique ne correspondait pas tout à fait à une loi gaussienne mais à un mélange de deux lois gaussiennes. Cette constatation justifie que l'on s'intéresse à ces modèles de bruit en vue de les introduire, lors du traitement de signaux reçus, dans des systèmes de détection - estimation [4].

Remerciements : Ce travail a été fait en partie grâce à une convention DCN - CEPHAG.

7. BIBLIOGRAPHIE

[1] L. LEBART, A. MORINEAU, J. P. FENELON, ' **Traitement des données statistiques** ' DUNOD - PARIS -

[2] P. JAFFARD, **'Initiation aux méthodes de la statistique et du calcul des probabilités'** MASSON - PARIS -

[3] B. S. EVERITT AND D. J. HAND, **'Finite Mixture Distributions'** CHAPMAN AND HALL

[4] E. PETIT, rapport CEPHAG n° 40/91, **'Application de tests statistiques de lois en acoustique sous-marine'.**