

## TPB04 : Régression : modélisation du diffusiomètre NSCAT

### I - Présentation du domaine d'Application

Simulation de la fonction directe d'un diffusiomètre satellitaire à l'aide des Perceptrons Multicouches - PMC et estimation de la variance.

Le diffusiomètre est un radar actif d'observation de la surface océanique (actif car il possède son propre source d'illumination de la surface d'observation, par rapport à d'autres radars dits passifs qui utilisent le rayonnement solaire). Son principe de mesure est basé sur la réflexion diffuse des ondes électromagnétiques qu'il a émises et qui sont diffusées dans toutes les directions par la surface océanique. Le coefficient de rétrodiffusion qu'il mesure est le rapport entre la puissance diffusée dans la direction du radar donc rétrodiffusée et la puissance incidente émise. Ce coefficient de rétrodiffusion (noté  $\sigma_0$  ou  $\sigma_0$ ) dépend de la rugosité de la surface de l'océan, de la géométrie de mesure (angle de visée ou d'incidence  $\theta$ ) et des caractéristiques du radar (la fréquence du signal utilisé, la polarisation des ondes électromagnétiques). La rugosité de surface dépend au premier ordre d'approximation du vent local soufflant sur la surface. Le vent engendre, par l'intermédiaire d'une force de frottement, des vaguelettes de quelques centimètres de  $LO$  (longueur d'onde ou la distance entre 2 crêtes) qui elles interagissent avec les ondes émises par le radar, ondes aussi de quelques centimètres de  $LO$ , selon les caractéristiques radars utilisées.

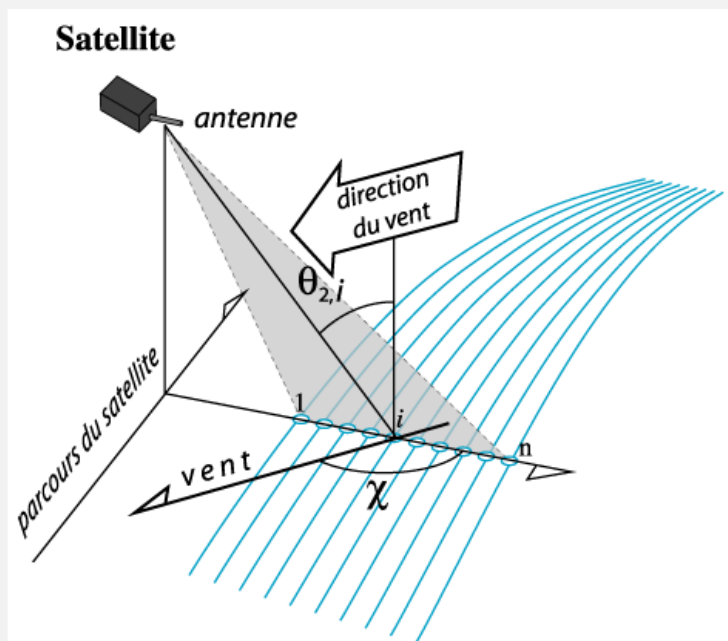
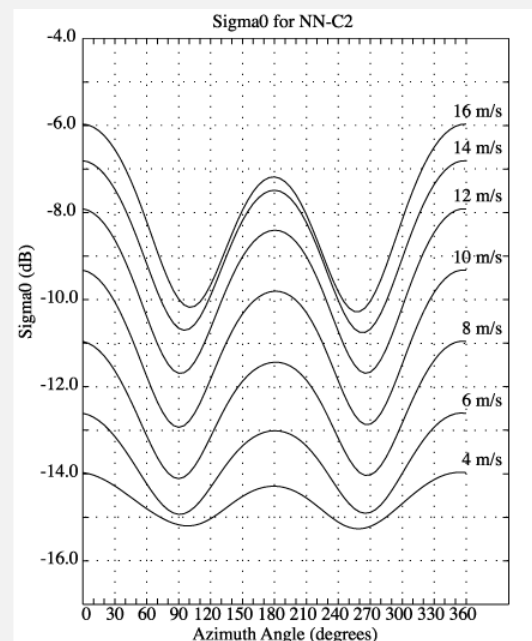


Schéma du capteur NSCAT HH:  $\theta$  représente l'angle d'incidence du signal et  $\chi$  l'angle d'azimut ou direction du vent par rapport au signal.



Courbe théorique pour un angle d'incidence constant.

...

Le diffusiomètre effectue des mesures à incidence oblique ( $\theta$  entre  $20^\circ$  et  $50^\circ$ ), plus le vent est fort, plus la diffusion est importante et donc plus la mesure radar est forte. En résumé : à vent fort, diffusion plus importante dans toutes les directions et signal mesuré fort, tandis qu'à vent faible, faible diffusion dans la direction du radar et donc signal mesuré plus faible. Dans le cadre de ce projet, le diffusiomètre est NSCAT de la NASA. Il travaille en bande *Ku* ( $14\text{ GHz}$ ) et effectue des mesures en polarisations verticale (VV) et horizontale (HH) de  $\sigma_0$ .

Ce TP a été inspiré par la thèse sur la « contribution à l'étude des diffusiomètres NSCAT et ERS-2 par modélisation neuronale Influence de la hauteur des vagues sur le signal diffusiométrique », soutenue par TRAN Vinh Bao Ngan. Elle est accessible par le lien suivant :

[https://www.locean-ipsl.upmc.fr/~mmsa/pages/documents/Tran\\_These\\_1999/Tran\\_These\\_1999.pdf](https://www.locean-ipsl.upmc.fr/~mmsa/pages/documents/Tran_These_1999/Tran_These_1999.pdf)

## II - Les Objectifs

Le 1<sup>er</sup> but de ce projet est de modéliser la fonction de transfert permettant de déterminer la polarisation horizontale (HH) de  $\sigma_0$  et ceci en fonction du vent (direction et module).

Le but suivant sera de mettre en œuvre une méthode d'estimation de la variance du bruit en fonction des données en entrées. Cette méthode pourra être mise en œuvre sur un exemple de démonstration puis aussi et principalement sur la modélisation objet du 1<sup>er</sup> but. L'estimation de la variance sera utilisée pour attacher des intervalles de confiances aux résultats des régressions.

Incidemment, outre les objectifs applicatifs proprement dit, on aborde dans ce travail des problèmes récurrents au traitement des données réelles comme par exemple les calculs d'erreurs ou la normalisation/dénormalisation qu'il faut savoir bien maîtriser. Si la normalisation permet aux systèmes d'apprentissages statistiques d'être plus efficient, seules les valeurs dénormalisées exprimées dans une unité physique peuvent intéresser l'expert du domaine concerné.

=====

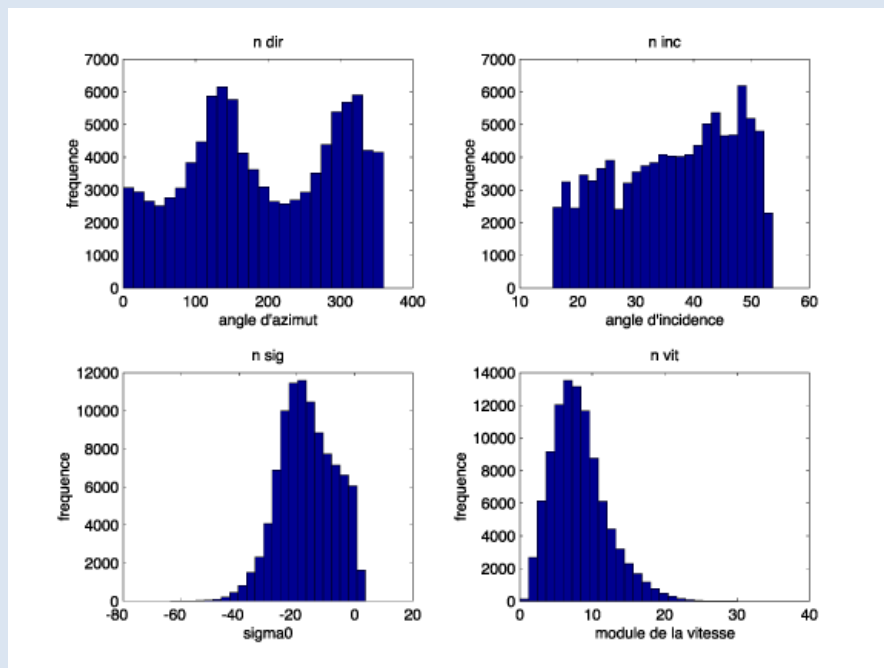
*Le rapport de TP devra être synthétique. Il doit montrer la démarche suivie, et ne faire apparaître que les résultats nécessaires. Il s'agit de quantifier les résultats tout en rédigeant un rapport qui les analyse et les commente. Les paramètres utilisés devront être indiqués, Les graphiques des expériences doivent être insérés dans le rapport. Les résultats présentés devront être analysés et commentés.*

### III - Les Données

Les données mises à disposition sont constituées d'une base de données issue du radar NSCAT du satellite SeaScat. Leur taille est d'un peu plus de 4000 exemples. 4 fichiers sont inclus :

- Diffu\_Dir.dat : Angle d'azimut (en degré) ou direction du vent ( $\chi$ )
- Diffu\_Inc.dat : Angle d'incidence (en degré) du signal ( $\theta$ )
- Diffu\_Vit.dat : Module de la vitesse du vent : V (en m/s).
- Diffu\_Sig.dat : Sigmao (ou  $\sigma_0$ ), composante HH uniquement (en dB)

Cette base n'est pas équilibrée (voir histogrammes de fréquences ci-dessous), il y a bien plus d'exemples pour les basses vitesses que pour les hautes. Certains angles d'azimut sont mieux représentés que d'autres.



Attention : la figure montre l'histogramme de valeurs pour la base de données complète (100 000 mesures). Ici nous allons travailler avec un sous-ensemble, à angle d'incidence fixe, autour de 35 degrés (4 098 mesures). De ce fait, l'information relative à l'angle d'incidence peut être écartée des données d'entrée.

### IV - Éléments pour la réalisation du TP

La plupart des fonctions nécessaires à la réalisation de ce TP sont disponibles. Dans le cas contraire il faudra les réaliser. On pourra ponctuellement donner des indications sur les fonctions à utiliser sans pour autant entrer dans le détail de la programmation qu'il vous incombera de prendre à votre charge. (Avant d'utiliser une fonction, il est important de se référer à sa documentation pour avoir des précisions sur son utilisation.)

## V – 1<sup>ère</sup> Partie : Modélisation du diffusiomètre NSCAT

Nous vous demandons de réaliser une régression par PMC pour modéliser la fonction du diffusiomètre NSCAT. Le rôle du PMC sera donc de donner la valeur de sortie  $\sigma_0$  en fonction de la direction et de la vitesse du vent. **Les fichiers de données à utiliser sont donc Diffu\_Dir.dat et Diffu\_Vit.dat, pour les données d'entrées et Diffu\_Sig.dat pour celles de sortie.** Cependant, Le PMC ne devra pas travailler directement avec les données brutes mais avec des données normalisées dont on va préciser les méthodes. Après normalisation, les matrices de données devront être séparées en ensemble d'apprentissage et ensemble de validation à concurrence de 2/3 pour l'un et 1/3 pour l'autre.

### Travail à faire

#### 1) Préparation des données :

► Après lecture des fichiers de données, il faudra procéder à leur normalisation de la façon suivante :

- Pour la direction du vent, la normalisation va consister à prendre les sinus et les cosinus de la direction du vent qu'il faudra donc bien penser à exprimer en radian préalablement (*La relation de la variable de sortie,  $\sigma_0$ , par rapport à la direction du vent a une périodicité de  $360^\circ$ . Un codage linéaire pour cette variable où l'intervalle de 0 à 360 serait codé entre -1 et 1, ne permettrait pas de préserver la relation de périodicité. Un codage par une fonction sinusoïdale est conseillé*).

- Pour la vitesse on appliquera une normalisation par centrage-réduction (avec un coefficient de 2/3). (*Le module de la vitesse du vent varie dans l'intervalle de 0 à 30 m/s approximativement. La relation avec la variable de sortie est proche du linéaire. Un codage linéaire est donc approprié. Un codage classique est la normalisation par **min/max**. Les données sont codées entre -1 et 1. Nous proposons cependant une normalisation par **centre-réduction** des données. Cette méthode permet de ramener la moyenne des données ainsi traitées à 0 et son écart type à 1. Empiriquement, dans ces conditions, l'apprentissage se fait dans des « bonnes conditions » pour éviter de tomber en saturation dans l'état des cellules internes, et en évitant l'écrasement des valeurs produit lors d'une normalisation par min/max d'une série de données à distribution gaussienne*).

- Pour le coefficient de rétrodiffusion  $\sigma_0$  (qui varie entre -45 et 5 dB approximativement) on appliquera là aussi une normalisation par centrage-réduction avec un coefficient de 2/3.

► Constituer les matrices d'entrée et de sortie du PMC pour l'ensemble d'apprentissage et pour celui de validation. (Selon son niveau de programmation, on pourra écrire les fonctions ou utiliser les fonctions des TP précédents.)

► Produire les nuages de points des variables normalisées. On présentera 3 figures pour lesquelles on mettra en abscisse la variable d'entrée et en ordonnée la sortie  $\sigma_0$ .

...

## 2) Réalisation de la fonction neuronale

Pour l'initialisation puis l'apprentissage du PMC, vous devez choisir certains paramètres. Pour les fonctions d'activation, nous vous conseillons l'usage de la fonction tangente hyperbolique pour les neurones cachés et linéaire pour les neurones de sortie. Le nombre de neurones sur la couche cachée ne devrait pas être trop élevé (de l'ordre de 5). Un nombre d'itérations entre quelques centaines ou quelques milliers devrait convenir. Les autres paramètres peuvent être laissés à leur valeur par défaut.

On précise ici, en accord avec ce qui a été présenté, que la matrice d'entrée du PMC doit être constituée de 3 vecteurs colonne correspondant :

- au sinus de la direction du vent exprimée en radian
- au cosinus de la direction du vent exprimée en radian
- à la vitesse du vent centrée-réduite

La matrice de sortie est pour sa part, constituée du vecteur  $\sigma_0$  en valeur normalisée par centrage-réduction.

**Dans la réalité ces fonctions sont mises au point sur un ensemble de données très important (plusieurs centaines de millier).** Afin d'avoir une meilleure appréciation des résultats, nous les présenterons en utilisant l'ensemble de toutes les données (réunion de l'ensemble d'apprentissage et de validation) et en utilisant les poids au minimum de l'erreur en validation.

Après apprentissage, vous devrez produire (et commenter) :

► Deux tableaux d'erreurs qui permettront d'apprécier les performances du PMC, l'un en fonction de la vitesse et l'autre de la direction. On travaillera pour estimer les résultats sur des intervalles de vitesse et de direction. On prendra des intervalles de **2m/s** pour la vitesse et de **60°** pour la direction. Les erreurs à produire sont la RMS, la RMS relative et le BIAIS. Le nombre de points et la moyenne par intervalle peuvent être des indications intéressantes. Que peut-on conclure de ces informations ?

► Le diagramme de dispersion de  $\sigma_0$  en valeur dénormalisée. C'est le nuage de points des valeurs estimées par le PMC par rapport aux valeurs de références (sorties désirées) après dénormalisation. On devra de plus produire 2 deux autres diagrammes où les points devront être associés à une couleur en fonction de la vitesse pour l'un et en fonction de la direction pour l'autre.

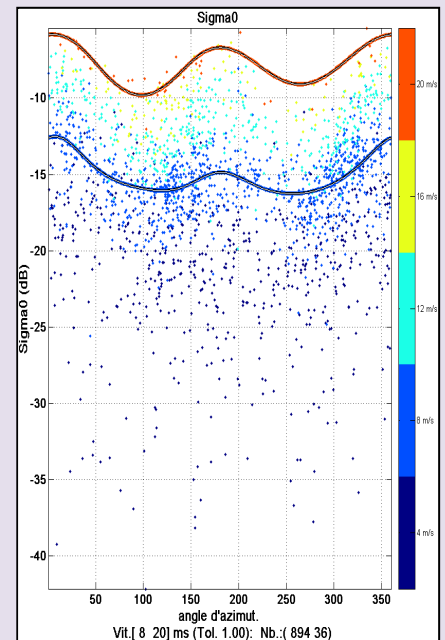
...

- Une figure présentant en valeur dénormalisée et pour chacune des vitesses [4 8 12 16 20] :
- les données proches, avec une tolérance de **1m/s**, autour de la vitesse considérée
  - « la » courbe de régression du PMC. On en représentera les abaques en fonctions des vitesses.

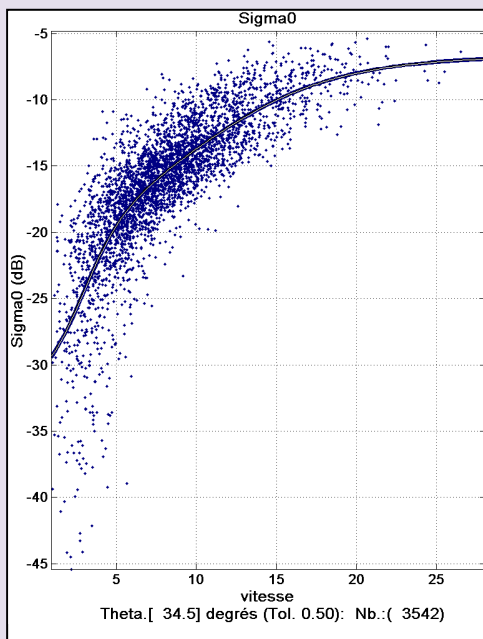
Sur la figure, on aura :

- en abscisse : l'angle d'azimut en degré en abscisse. Pour cela on échantillonnera l'axe de  $1^\circ$  à  $360^\circ$  par pas de  $1^\circ$ .
- en ordonnée : sigma0
- les courbes d'abaque de la régression (PMC) associées à chaque vitesse.

La figure ci-contre est un exemple, limité aux vitesses [8 20] de ce à quoi vous devriez à peu près aboutir.



### Un autre type d'affichage



On pourra aussi afficher, après apprentissage, une figure comme celle ci-contre.

Sur la figure, on a une seule courbe qui est tracée pour **Théta**=[34.5] avec une tolérance de 0.5, alors que dans ce qui a été demandé précédemment correspondait à un tracé par vitesses ([4 8 12 16 20]) avec une tolérance de 1.0.

On remarque que ce sont les vitesses qui sont en abscisse, et non pas les directions (angles d'azimut).

Ici, on ne se préoccupe pas de la direction. On se sert de l'angle d'incidence, Théta.



## VI - 2<sup>ème</sup> Partie : Encadrement par approximation de la variance

Afin d'attacher des intervalles de confiance aux résultats trouvés, on propose d'estimer la variance du bruit en fonction des données. La méthode proposée a déjà été mise en œuvre dans un TP précédant (TPB01) à la normalisation près :

Un premier réseau (PMC1) a été optimisé pour modéliser une fonction  $y=f(x)$ , à partir d'une base d'apprentissage  $\{(x_i, d_i) \ i=1, N\}$ .

Après apprentissage, la sortie de PMC1 donne une estimation de l'espérance  $E(d/x)$ . On note  $Eapp$  l'erreur commise par ce 1er réseau :  $Eapp = (y - E(d/x))$ .

Il s'agit maintenant d'optimiser un second réseau (PMC2) pour modéliser la fonction  $Var(d/x) \approx E[Eapp^2/x]$  à partir d'une autre base d'apprentissage où la sortie désirée est l'erreur quadratique :  $\{(x_i, Eapp^2_i) \ i=1, N\}$ . Après apprentissage, la sortie de ce second réseau donnera une estimation de l'espérance  $E[(d/x - E[d/x])^2]$ , c'est à dire, une estimation de la variance de  $d/x$  pour chaque valeur de  $d$ .

Dans le cas de ce TP, les PMC travaillant en valeur normalisée, il faudra penser à dénormaliser la sortie de PMC1 avant de calculer l'erreur quadratique  $Eapp^2$ , puis re-normaliser cette dernière pour construire l'ensemble d'apprentissage de PMC2. Au final les présentations de résultat devront être faites en valeur dénormalisée.

Le travail à réaliser pour cette partie devra porter sur 2 cas :

- le premier cas où l'on a :
  - $y$  = la polarisation horizontale (HH) de  $\sigma_0$ ,
  - $x$  = angle d'incidence ( $\theta$ ) et module du vent ( $V$ ).
- l'autre cas de la 1<sup>ère</sup> partie de ce TP avec :
  - $y$  = la polarisation horizontale (HH) de  $\sigma_0$ ,
  - $x$  = direction du vent ( $\chi$ ), et module du vent ( $V$ ).

Etant donné que la variance est un estimateur du second ordre, il est important d'avoir un maximum de données pour estimer les paramètres : pour estimer PMC2 on prendra l'ensemble des données pour l'apprentissage (on travaillera donc sans ensemble de validation ni de test). On rappelle par ailleurs que pour ce type de modèle, il faut faire attention à ne pas générer de variances négatives.

...

## **Travail à faire**

Les étapes nécessaires sont les suivantes :

1. Générer les bases d'apprentissage nécessaires à ce second réseau
2. Choisir l'architecture du réseau, puis ajuster les poids du réseau.
3. Présenter des figures :
  - qui montrent les erreurs quadratiques et la régression qui en est faite par PMC2. (A faire soit-même en se basant sur ce qui a été utilisé dans la partie précédente.)
  - qui donne une représentation simultanée des données, de la régression (de PMC1) et de l'encadrement à plus ou moins 2 écarts types. On indiquera le pourcentage de données compris dans l'encadrement.

Pour éviter d'avoir des figures trop confuses, il est demandé de faire une figure pour  $\theta$  et pour chaque vitesse définie précédemment avec leur tolérance respective.