Master TRIED

TPB04 : Rapport

Sujet :

Régression : modélisation du diffusiomètre NSCAT

Réalisé par :

Karim ASSAAD

Année universitaire :

2016/2017

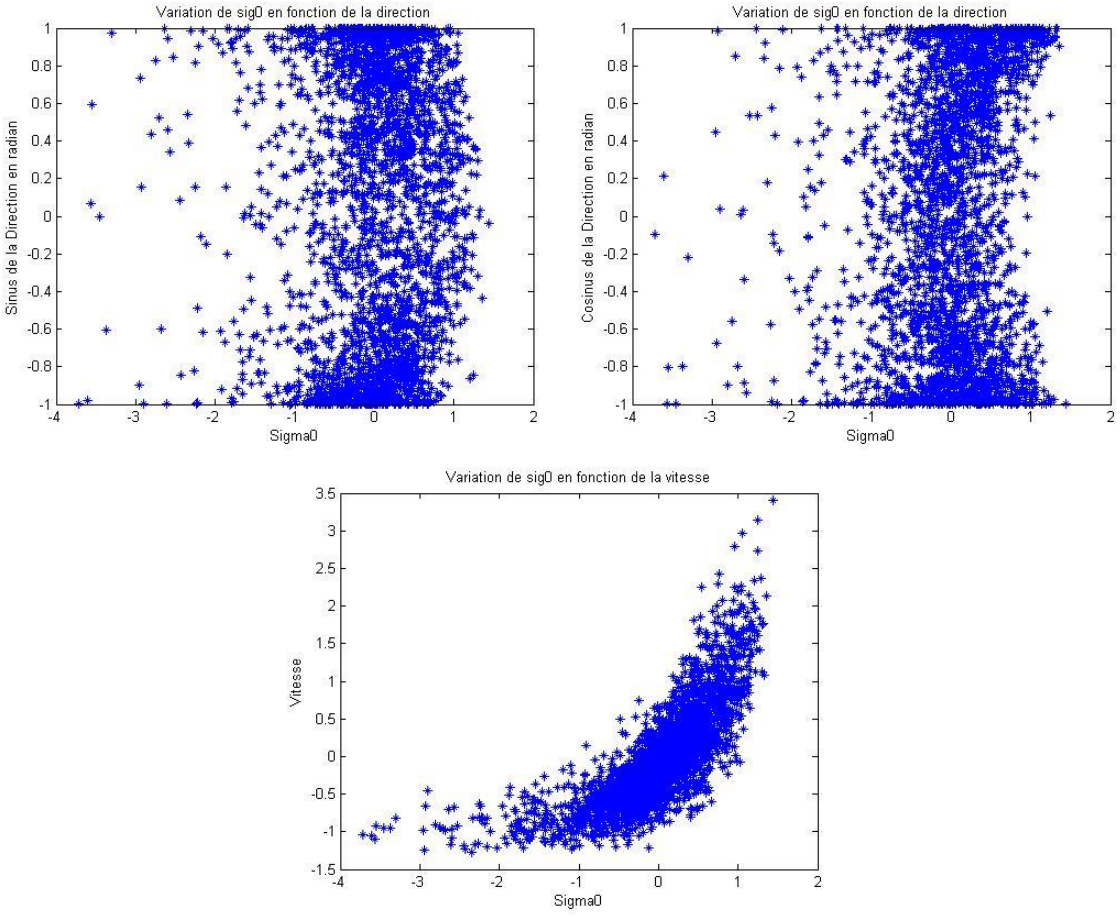
1ère Partie : Modélisation du diffusiomètre NSCAT

1) Préparation des données

Ce Tp consiste à réaliser une régression par PMC pour modéliser la fonction du diffusiomètre NSCAT. Le rôle du PMC sera donc de donner la valeur de sortie Sigma0 en fonction de la direction et de la vitesse du vent.

Pour la direction du vent, la normalisation va consister à prendre les sinus et les cosinus de la direction du vent qui est exprimer en radian.
Pour la vitesse on appliquera une normalisation par centrage-réduction.
Pour le coefficient de rétrodiffusion sigma0 on appliquera aussi une normalisation par centrage-réduction.

Les figures qui montre la dépendance entre la variable à expliquer normaliser en fonction des variables expliquer normaliser sont présentés ci-dessous.



On peut voir que les 2 premier graphique ne représente pas une dépendance que le 3eme graphe représente une dépendance proche du linéaire (~ exponentielle). En plus on peut voir que les 2 premier sont borner entre 1 et -1 tandis que le 3eme est borner entre -1.5 et 3.5.

2) Réalisation de la fonction neuronale

Maintenant on va passer à la création du modèle de prédiction en utilisant un PCM avec une architecture de 5 neurones cacher. Pour ce faire on divise notre jeu de données en 2 ensemble. Le premier utiliser pour l'apprentissage et le deuxième pour la validation.

Mesure des erreurs

Pour mesurer la performance du modèle on présente ci-dessous un tableau qui montre plusieurs erreurs pour la prédiction de l'ensemble de validation.

Tableau 1 : Représentation des erreur de l'ensemble de validation

RMS	Biais	RMS relative
2.8210	0.0007	93.1281

Afin d'avoir une meilleure appréciation des résultats, on va faire une prédiction de l'ensemble de toutes les données (données d'apprentissage + données de validation). On présente ci-dessous les erreurs obtenues.

Tableau 2 : Représentation des erreur de l'ensemble d'apprentissage + validation

RMS	Biais	RMS relative
2.8210	0.0007	0.1507

On remarque que la RMS et le Biais n'ont pas changer pour les 2 prédictions, par contre le RMS relative a subi une amélioration radicale.

Cela vient du fait qu'on a utilisé dans le deuxième cas, la moitié des données de prédiction dans l'apprentissage du modèle donc ce n'est pas surprenant que le modèle va bien prédire ces données.

Mesure des erreurs en fonction des intervalles

Pour mieux apprécier les performances du PMC, en fonction de chaque variable explicative, on va présenter dans cette partie deux tableaux d'erreurs qui permettront de voir les erreurs de prédiction en fonction des intervalles de vitesse et de direction.

Tableau 3: Représentation des erreur du PMC en fonction des intervalles de la directions

Direction entre	RMS	BIAIS	RMS Relative	Taille	Moyenne Sig
[0, 60]	3.0883	0.0543	0.1742	511	-14.5047
[60, 120]	2.7332	0.1108	0.1506	650	-16.9277
[120, 180]	2.5131	-0.0792	0.1323	945	-16.0513
[180,240]	3.1451	0.1075	0.1715	482	-16.8073
[240, 300]	3.0631	0.0033	0.1480	680	-17.7333
[300,360]	2.6360	-0.0917	0.1443	830	-14.9014

Les intervalles qui continent les plus de points sont les intervalles ou le modèle fait le moins d'erreur, et contrairement à ça, les intervalles qui continent les moins de points sont les intervalles ou le modèle fait le plus d'erreur. Par exemple l'intervalle [120, 180] qui contient le plus de point (945 points) a une RMS égale a 2.5, un biais égale a -0.07 et une RMS relative égale à 0.13 qui sont le minimum des erreurs de ce tableau.

En plus la moyenne de la variable Sigma0 est -16.1326 et les intervalles qui ont une moyenne proche d'elle ont une erreur plus petite que celle des intervalles avec des moyennes moins proches de la moyenne totale. Cela vient du fait qu'on fait une dé normalisation avant le calcul de l'erreur.

On peut conclure que la taille de l'intervalle et sa moyenne ont un rôle important dans la performance.

Tableau 4 :

Vitesse entre	RMS	BIAIS	RMS Relative	Taille	Moyenne Sig
[0, 2]	6.4766	-0.5491	0.2489	61	-28.3118
[2, 4]	5.2942	-0.2049	0.2079	410	-23.8866
[4, 6]	3.5607	0.0144	0.1768	777	-19.2987

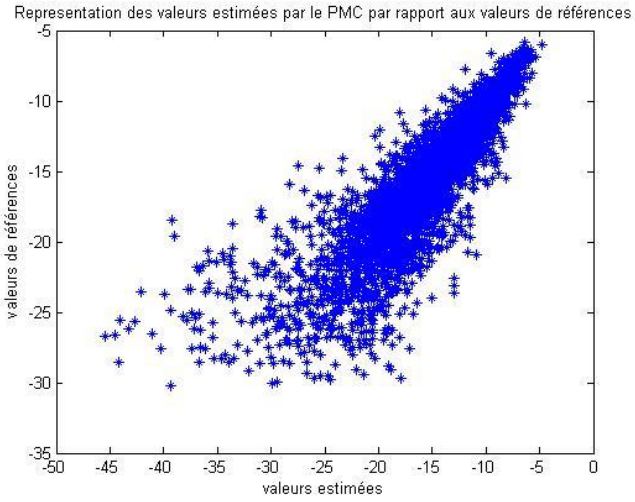
[6,8]	2.1124	0.0042	0.1311	923	-16.4995
[8, 10]	1.7316	0.0368	0.1268	784	-14.4661
[10, 12]	1.5424	0.0992	0.1287	516	-12.7169
[12, 14]	1.4394	-0.0175	0.1330	287	-11.4672
[14, 16]	1.2327	0.1043	0.1313	157	-10.1959
[16, 18]	1.0515	0.1243	0.1229	95	-9.0008
[18, 20]	0.9283	-0.0493	0.1046	55	-8.3124
[20, 22]	1.0394	0.1711	0.1290	20	-8.1295
[22, 24]	0.5958	-0.3265	0.0610	4	-7.9300
[24, 26]	0.6234	0.0537	0.0910	7	-7.6800
[26, 28]	1.3048	1.3048	0.2048	1	-6.3700
[28, 30]	1.1560	1.1560	0.2403	1	-4.8100

On peut faire la même conclusion que celle du tableau précédant et on peut voir que l'intervalle [6,8] qui a le plus grand nombre de point et la moyenne la plus proche que celle de la moyenne de Sigma0, a le moins d'erreurs. Le RMS des intervalles [22,24] [24,26] est meilleur que celui des autres intervalles mais dans ces deux cas le nombre des points sont égale a 4 et a 7 respectivement donc on ne peut pas le prendre comme un cas général.

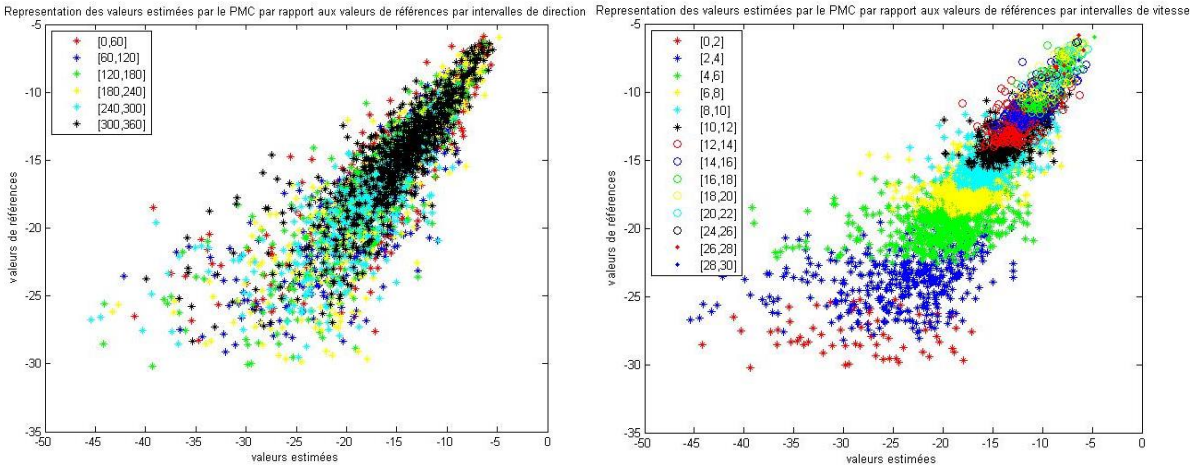
Et d'après ses deux tableaux on peut dire que le découpage de la vitesse en intervalle donne un résultat meilleur que celle du découpage de la direction du vent a des intervalles. Ces résultats nous indiquent que c'est mieux de faire des intervalles sur les variables explicatives qui ont une dépendance avec la variable à expliquer.

○ Diagrammes de dispersions

Le diagramme de dispersion de sigma0 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.

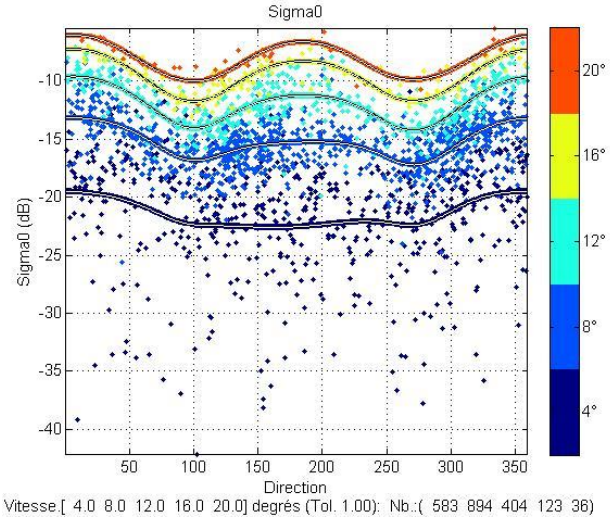


Ce graphe nous montre que la dépendance entre les valeurs estimer et les valeurs référence est linéaire. En plus elle suit la première bissectrice donc on peut dire que dans la plupart des cas les valeurs estimer sont bien prédite.



Comme on a conclu precedament (dans la partie Mesure des erreurs en fonction des intervalles), on peut voir que c'est mieu de faire des prediction sur plusieur intervalles decouper selon la vitesse puisque dans ce cas les classe sont bien ordonnes contrerement au cas des intervalles decouper selon la direction du vent.

Donc pour mieux visualiser la régression qui a 2 variables explicative (Vitesse et Direction) en 2 dimension, on va présenter une figure présentant en valeur dénormalisée et pour chacune des vitesses [4 8 12 16 20], les données de Sigma0 en fonction de la Direction et les courbe de régression du PMC correspondant a chaque valeur prise pour la Vitesse.



Comme prévu, les courbes de régression sont bien ordonnés. Cela reconfirme notre conclusion dans la partie précédente. En plus on peut voir que les courbe de correspondant aux Vitesse égale a 8, 12, 16 et 20 ont une allure fortement sinusoïdale. Tandis que l'effet sinusoïdale est moins présent dans le cas de la vitesse égale a 4.

2ème Partie : Encadrement par approximation de la variance

1) Le cas où la direction du vent et la vitesse sont considérer comme variables explicative

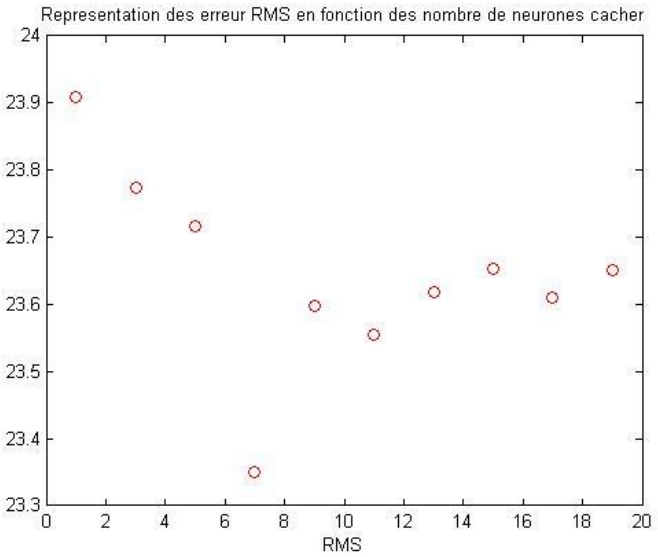
○ PMC1

Afin d'attacher des intervalles de confiance aux résultats trouvés, on va estimer la variance du bruit en fonction des données.

Pour ce faire un premier modèle PMC1 va modéliser Sigmas0 en fonction de la vitesse et la direction. 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)). Donc on va prédire de nouveau les valeurs de Sigma0 en prenons en compte des erreurs quadratique de chaque individu.

Architecture

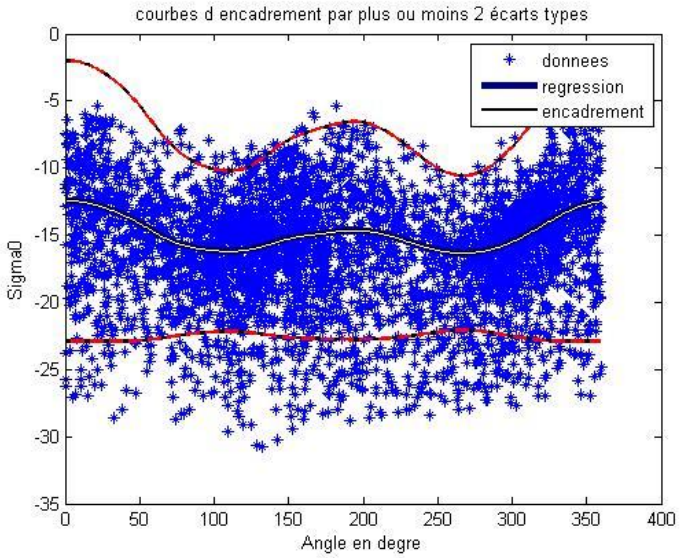
Afin de trouver la meilleure architecture j'ai fait plusieurs modèles avec plusieurs neurones cacher qui correspond à 1, 3 ,5 ,7,9, 11, 13, 15, 17, 19 neurones cacher et puis j'ai calculer leur erreurs RMS qui sont représenter ci-dessous.



On peut voir que la meilleure architecture est celle qui a 7 neurones cacher. Donc par la suite on va utiliser le modèle de PMC avec 7 neurones cacher.

Représentation

La figure ci-dessous est une représentation simultanée des données, de la régression de PMC1 et de l'encadrement à plus ou moins 2 écarts types.



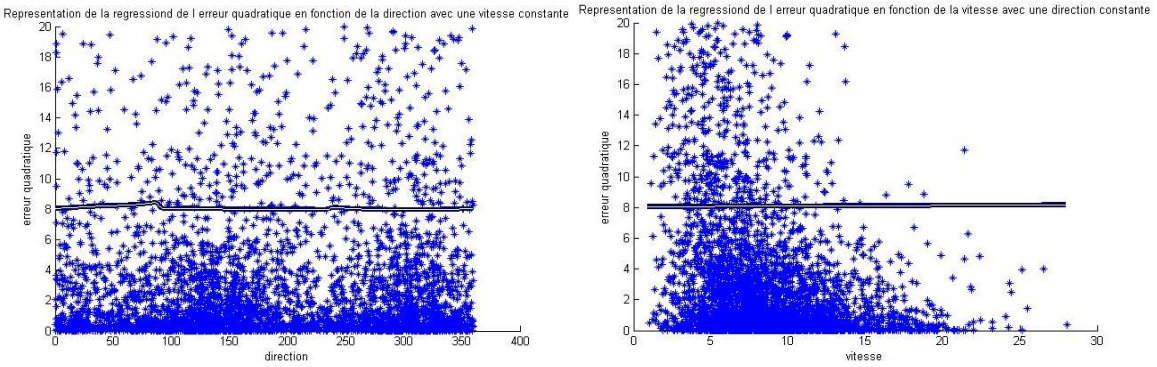
90% des données sont compris entre les 2 intervalles donc on peut dire que les intervalles choisit sknt bien adapté au probleme.

- PMC2

Il s'agit maintenant d'optimiser un second réseau PMC2 pour modéliser la fonction Var(d/x)»E[Eapp2/x] à partir d'une autre base d'apprentissage où la sortie désirée est l'erreur quadratique. 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.

Donc après avoir créé un vecteur d'erreur quadratique en utlisant PMC1, on va créer un deuxième modèle PMC2 qui va modéliser l'erreur quadratique en fonction de la vitesse et de la direction. Dans ce modèle on prend comme fonction d'activation 2 la fonction exponentiel (tandis que dans tous les cas précèdent on a pris la fonction linéaire)

Les 2 figures ci-dessous montrent les données des erreurs quadratiques en fonction de la Direction et de la Vitesse respectivement, et la régression faite par PMC2. La première prend la variable Vitesse comme une constante est trace la régression de l'erreur quadratique en fonction de la Direction et la deuxième prend la variable Direction comme une constante est trace la régression de l'erreur quadratique en fonction de la Vitesse.

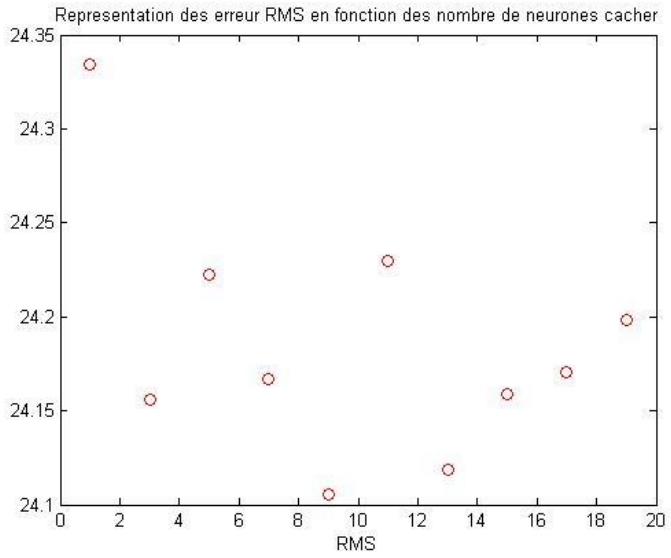


La régression dans les deux cas est presque linéaire, cela vient du fait que les données sont éparpillées et ca explique la grandeur de l'erreur obtenu.

2) Le cas où l'incidence et la vitesse sont considérer comme variables explicative

Dans cette partie on va faire le même travaille fait précédemment dans la partie (Le cas où la direction du vent et la vitesse sont considérer comme variables explicative) mais en prenant comme variables explicatives la vitesse et l'incidence. Les résultats obtenus sont montrer ci-dessous.

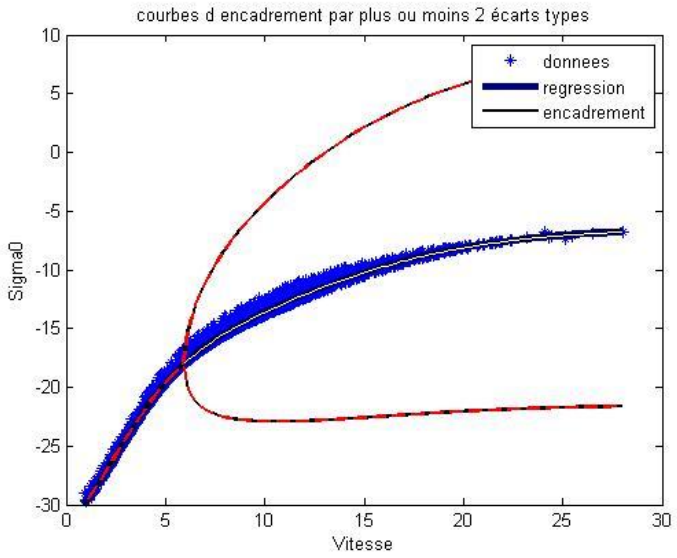
o PMC1



On peut voir que la meilleure architecture est celle qui a 9 neurones cacher. Donc par la suite on va utiliser le modèle de PMC avec 9 neurones cacher.

o *Représentation*

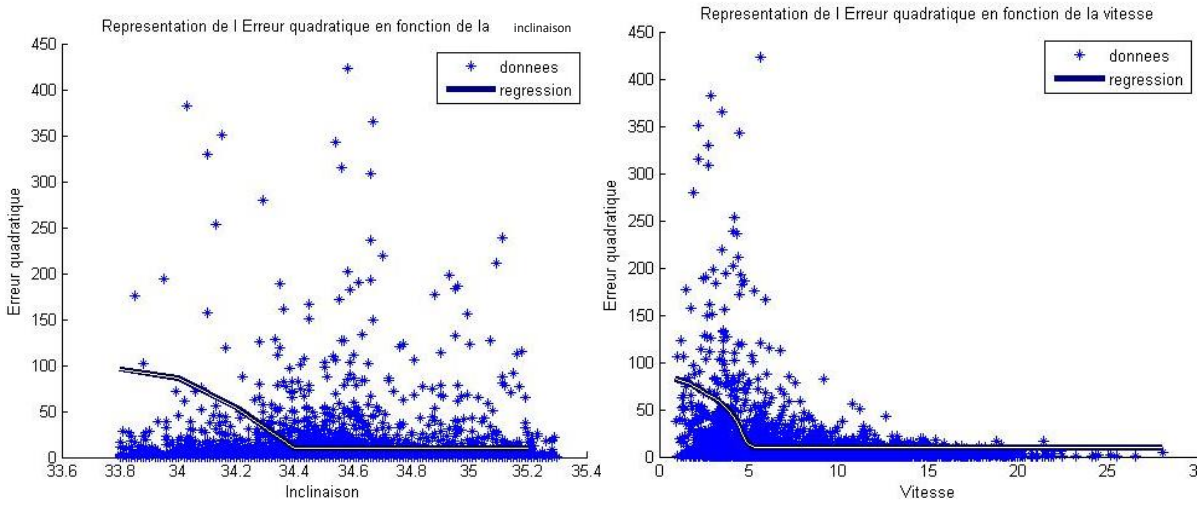
La figure ci-dessous est une représentation simultanée des données, de la régression (de PMC1) et de l'encadrement à plus ou moins 2 écarts types.



70% des données sont compris entre les 2 intervalles. Donc on peut dire que qu'on pouvait faire mieux dans le choix des intervalles surtout que notre modèle est assez performent.

- PMC2

Les 2 figures ci-dessous montrent les erreurs quadratiques et la régression qui en est faite par PMC2, la première prend la variable Vitesse comme une constante et trace la régression de l'erreur quadratique en fonction de l'inclinaison et la deuxième prend la variable Inclinaison comme une constante et trace la régression de l'erreur quadratique en fonction de la Vitesse.



Cela nous montre que le PMC2 de cette partie est plus complexe que celui de la partie précédente. Ce dernier donne des meilleurs résultats que le modèle de la partie précédente car nous sommes dans un problème ou la dépendance des variables ne sont pas linaire (Sigma0 a une dépendance avec les 2 variables explicative).