



Sylvie Thiria
Professeur à l'Université VSQ
E-Mail : thiria@locean-ipsl.upmc.fr

Paris le 30 avril 2013

Rapport sur la thèse de Liyun He Guelton

Emulation statistique de champs de vent à haute résolution

La thèse de Liyun He Guelton porte sur l'élaboration de carte de vent à haute résolution spatio-temporelle (HR) en utilisant la connaissance des sorties de modèle numérique et l'information satellitaire. Il s'agit donc de trouver un bon compromis entre les avantages et les inconvénients de deux sortes d'information. D'une part, les modèles qui produisent des vents dont la couverture spatiale et temporelle est très régulière, mais dont la résolution spatiale a une basse résolution (BR), d'autre part les données satellitaires capables de produire des vents avec une résolution spatiale de l'ordre de la micro-échelle. L'intérêt de la thèse est donc évident puisque l'obtention de cartes de vents Haute Résolution (HR) à couverture spatio-temporelle élevée permettrait d'étudier des dynamiques petites échelles qui sont conditionnées par la situation méso-échelle. Ces dynamiques sont pour la plupart déterminées par la situation géographique du lieu étudié et la disposition méso-échelle des vents autour de ce lieu.

La thèse de Liyun He Guelton repose sur l'hypothèse que pour déterminer un vent HR en un lieu déterminé, on peut introduire la connaissance physique en considérant une « situation BR » autour de ce point. Cette situation peut être obtenue à partir des sorties d'un modèle numérique, celle-ci pouvant être comparée à des vents HR dont la localisation est proche au sens de l'espace ou du temps. Il s'agit donc d'un problème de downscaling statistique.

L'introduction de la BR agit alors comme une contrainte physique et l'utilisation de méthodes statistiques peut permettre de retrouver la valeur du vent HR au point considéré. Etant donné que l'on peut utiliser les modèles numériques en mode prévision, il devient possible d'envisager d'obtenir des cartes HR en prévision si la méthode d'inférence statistique le permet. Il s'agit donc d'un travail pluri-disciplinaire dans lequel il faut dominer à la fois le domaine de la télédétection et celui de l'apprentissage statistique. La clef de voute de l'apprentissage statistique réside dans la création d'une base de données réaliste dans laquelle l'information doit être la plus proche possible de la réalité.

Le manuscrit de thèse est très clair et permet de bien appréhender la démarche de Liyun He Guelton : il s'agit après avoir analysé le problème à résoudre de choisir le meilleur apprentissage statistique capable par régression de réaliser l'appariement : situation BR - vent local HR. En particulier l'introduction présente en peu de pages la problématique et les difficultés à prendre en compte: imperfection des modèles BR, limitation des données HR

dues aux techniques de mesure et à la résolution temporelle. Tous ces effets amènent très clairement à rechercher des fonctions de transfert non linéaires liant BR et HR, la méthode proposée est donc recherchée parmi les méthodes de régression non linéaire, l'apprentissage étant réalisé séparément pour chaque point de grille BR. Etant donné le caractère méthodologique et prospectif de la thèse cela est tout à fait acceptable. Il faut cependant bien considérer que de cette manière la cohérence physique du champ retrouvé peut ne pas être conservé et que cela devrait faire partie des vérifications à effectuer. En particulier, il serait pertinent d'utiliser des critères statistiques capables d'apprécier cette cohérence.

Le deuxième chapitre présente l'état du downscaling statistique dans le domaine des sciences de l'océan et de l'atmosphère et les méthodes de régression linéaire et non linéaire qui vont être comparées. Trois méthodes multi dimensionnelle sont sélectionnées : la méthode des analogue très employée en météorologie, la méthode de régression linéaire multiple et la méthode SVM. Les méthodes sont bien présentées et le manuscrit sera utile en tant que document pédagogique. Le choix de SVM est pertinent puisqu'il s'agit d'une méthode qui a montré son efficacité pour traiter d'un grand nombre de problèmes non-linéaire. Le choix des autres méthodes proposées est cohérent et permet de bien comprendre ce qui différencie les différentes méthodes d'apprentissages actuelles. On ne peut cependant pas dire que SVM est une bonne méthode à cause de ses propriétés d'optimisation et de généralisation. En effet, presque toutes les méthodes statistiques optimisent un critère. L'efficacité d'une méthode va dépendre en premier lieu de la complexité du problème lui même, du nombre de variables explicatives, de la dimension de la base de données. La généralisation découle de l'adaptation de la méthode au problème traité.

Le troisième chapitre présente les données utilisées, il s'agit d'un chapitre primordial puisqu'il détermine la précision que l'on peut attendre de la co localisation BR et HR. Le modèle choisi est celui de ECMWF à $0,5^\circ$ avec une résolution temporelle de 6 heures, la zone d'étude sur la côte ouest de la Norvège autour de Bergen à une topographie complexe qui permet de tester le downscaling pour des configurations difficiles. Les vents SAR sont ceux produits CMOD4. Le choix de CMOD4 n'est pas justifié, il aurait été intéressant de donner ses limites : quelle sont les précisions de la vitesse et de la direction du vent calculés par CMOD4. La grande variabilité observée des caractéristiques du vent à différents points justifie pour Liyun HE Guelton le fait d'envisager une fonction de régression par point de grille. Afin d'améliorer l'échantillonnage, on aurait pu envisager de regrouper les points présentant des caractéristiques communes. Le choix effectué parmi différentes interpolations pour effectuer les correspondances BR-HR tout en gardant le maximum de variabilité est bien argumenté et convainquant pour la co localisation en temps. La co localisation en espace est elle moins bien documentée, elle joue pourtant aussi sur le réalisme de l'appariement.

Le quatrième chapitre porte sur une analyse conjointe entre BR et HR pour la zone d'étude. Elle permet de distinguer trois zones de difficultés croissantes en fonction des écarts observés entre les vents ECMWF et les vents SAR : les zones off-shore, les zones côtières et les fjords. Les différences sont interprétées géographiquement avec beaucoup de soin et justifient le fait que des relations distinctes BR-HR vont être recherchées selon la position géographique. Pour cette étude, il manque peut être une discussion sur la précision de CMOD4 à reproduire les vents SAR dans les zones de fjords.

Le cinquième chapitre expose clairement les différentes méthodes d'apprentissage proposées et montrent toujours d'une manière synthétique l'intérêt de chacune. En particulier les similarités entre les différentes méthodes sont bien développées. La détermination des variables d'entrée est faite avec soin de manière à pouvoir déterminer le contexte spatial BR à prendre en compte. Ce contexte est déterminé de différentes manières : globalement, voisinage local ou non local. La méthode de détermination d'un voisinage non local par calcul d'entropie conditionnelle est très intéressante en elle même, même si dans ce problème précis de downscaling elle apporte peu d'améliorations par rapport à l'utilisation d'un par voisinage local simple. Cette méthode pourrait servir dans d'autres problèmes pour lesquels une information spatiale doit être prise en compte.

Le dernier chapitre présente une évaluation expérimentale qui compare 4 méthodes différentes et 3 types de variable explicatives (information globale, locale et non locale), l'expérience étant menée sur les douze points des trois zones. Les expériences sont bien menées et SVR apparaît effectivement comme la méthode la plus apte à trouver la cohérence de la base d'apprentissage. La méthode multi modèle proposée qui consiste à regarder l'effet de la direction du vent n'est pas très probante, cela peut être dû au fait que l'effet du vent est plus méso-échelle et qu'il faudrait regarder si considérer des « types de mer » peut améliorer la restitution. L'évaluation des performances est un exercice très difficile. Les différences dans les comparaisons pour les journées du 4/05/2008 et du 4/03/2009 peuvent provenir, comme cela est très justement mentionné, de l'échantillonnage de la base de données qui devrait être mieux explicité dans le chapitre qui décrit sa constitution.

En conclusion, il s'agit d'un très bon travail prospectif, bien mené et bien documenté, qui pourra servir de référence à des validations ultérieures. Le travail pourra être poursuivi dans différentes directions pour permettre d'améliorer la mise en œuvre opérationnelle. Liyun He Guelton a une excellente connaissance des méthodes qu'elle utilise et des problèmes d'application qu'elle traite. Sa recherche a demandé un très gros travail de mise en œuvre, chaque méthode proposée nécessitant pour son application un investissement théorique et un traitement des données très important. Je juge l'ensemble de la recherche tout à fait digne d'être soutenu par Liyun He Guelton en vue d'obtenir le titre de Docteur de l'Université européenne de Bretagne.

S. THIRIA