Thomas Corpetti
OSUR, UMR 6554 LETG
Rennes
thomas.corpetti@univ-rennes2.fr

Rapport sur le mémoire de thèse de Madame Liyun HE-GUELTON

Emulation statistique de champs de vents à haute résolution

Le document présenté par Liyun He-Guelton s'intéresse à l'émulation de champs de vents océaniques à haute résolution spatiale à partir de données simulées ECMWF (European Center for Medium Range Weather Forecast) à moyenne résolution et de quelques observations radar de type SAR à très haute résolution. Cela est notamment appliqué à des zones géographiques où cette tâche est très difficile, comme c'est le cas de la zone d'étude considérée : la région de Bergen, sur la côte ouest de la Norvège, dont les côtes et les fjords influencent grandement la circulation et rendent les modèles physiques existants inefficaces.

Face à cette situation, l'auteure fait le choix très raisonné d'une approche par apprentissage statistique car dans ces zones géographiques complexes, les modèles physiques existants ne sont pas satisfaisants. Une méthodologie basée sur les SVR (Support Vector Regression) est proposée et fournit de très bons résultats.

Dans la suite, j'analyse plus en détail le document et je termine par quelques commentaires.

Analyse du document

Le document est clair et agréable à lire. Il commence par une introduction générale qui pose le problème traité et les principaux choix qui seront retenus.

Le *Chapitre 2* se consacre ensuite *à la méthodologie et l'état de l'art.* On y trouve une justification du choix d'une approche statistique pour ce travail (les modèles physiques sont moins efficaces et plus coûteux en temps de calcul). Ensuite, une revue des techniques de downscaling statistique est proposée, en appuyant sur trois méthodes plus communément employées en océanographie : les *méthodes analogues* (détail d'un article qui donne de bons résultats par plus proches voisins), les *méthodes de régression linéaire multiples* (détail d'un article de Goubanova et al, 2010 et de la thèse de Marie Minvielle) et les *méthodes de régression multi-modèles* (détail d'un article de Walmsley et al et de la thèse de Mohamed Bassam Ben Ticha).

Les techniques présentées sont analysées de manière critique où l'auteure démontre les limites actuelles de ces approches : la classification des vents par leur direction n'est pas

envisageable pour de larges zones et les directions ne sont pas les meilleurs discriminants dans le troisième cas ; il serait plus judicieux d'exploiter des relations linéaires par rapport aux distances quadratiques dans le second cas, …

Ce chapitre démontre que Liyun He-Guelton a acquis une bonne connaissance de ces méthodes, de la façon dont elles se situent les unes par rapport aux autres, et de leurs cadres de validité. Même s'il est précisé dès le début que les approches physiques ne sont pas adaptées et ne seront donc pas utilisées dans ce travail, il aurait cependant été intéressant de compléter cet état de l'art par quelques illustrations de techniques de downscaling physique.

L'auteure réécrit ensuite les méthodes de régression présentées en utilisant un noyau, ce qui permet d'unifier le formalisme et d'appréhender ces différentes approches par le choix d'un noyau différent.

Le *chapitre 3 présente les données utilisées*. C'est un chapitre très bien détaillé, où l'auteure a pris soin de préciser dès le début les différences entre « échelle » et « résolution » afin de lever toute ambiguïté. Les différentes sources permettant d'obtenir des données liées à l'état de l'océan sont introduites : modèles numériques et télédétection notamment (en précisant dans ce dernier cas la manière de remonter aux vents à partir du coefficient de rétrodiffusion et en mentionnant les limites de ces approches indirectes telles que la sensibilité aux pluies, aux éléments flottant à la surface de l'océan, …). La zone d'étude est également présentée en illustrant les nombreuses difficultés liées à l'analyse de la dynamique près des côtes, justifiant encore une fois la nécessité de se doter de méthodologies adaptées.

Ensuite, les données disponibles sont introduites. L'auteur démontre que leur exploitation directe n'est pas optimale car il se pose des questions liées à la synchronisation entre les données à haute résolution (ponctuelles) et celles à basse résolution (toutes les 6h). Un protocole original a alors été proposé, en exploitant une bouée qui effectue des mesure à très haute fréquence pour finalement retenir la donnée à basse résolution la plus proche en éliminant les données aberrantes (qui se situent notamment là où il y a présence de plateformes pétrolières).

Le *chapitre 4* propose une *analyse statistique conjointe* : on essaie de comprendre la manière dont les données sont organisées, dans le but de définir la méthodologie qui sera proposée au chapitre suivant. Une analyse des erreurs entre les champs à haute et basse résolution (interpolés) est présentée, illustrant une fois de plus les difficultés qui existent près des fjords et des zones côtières sur différents critères (moyenne, écarts types des différences, rose des vents, diagrammes de dispersion, niveaux d'intensité, …).

Ce chapitre est important et démontre que l'auteure a pris soin de ne pas exploiter les techniques développées par la suite de manière « aveugle » mais Liyun He-Guelton a cherché à comprendre les interactions existantes entre ces données, ce qui permettra au chapitre suivant de définir judicieusement les variables explicatives.

Dans le *chapitre 5*, la *technique d'émulation* est développée. C'est le cœur méthodologique du travail proposé. L'auteure rentre dans le détail des méthodes de régression présentées dans le chapitre 2. La sélection des variables explicatives est en premier lieu discutée puis les techniques de régression elles-mêmes sont présentées.

Trois types d'informations sont retenues pour les variables explicatives : globales, locales, non-locales.

- Les informations globales sont obtenues sur la base d'une Analyse en Composante Principale, où, si j'ai bien compris, les modes sont choisis surtout pour s'ajuster à la variabilité au large puisque les informations aux zones côtières ne sont pas représentatives. On aurait aimé plus de détails sur cette partie, et notamment plus de formalisme : même si l'ACP est une technique bien connue, le fait de contraindre le choix des modes (si c'est bien ce qui est fait) au large aurait mérité plus de détails.
- Les variables explicatives locales sont quant à elles choisies par le nombre de points à basse résolution dans un voisinage local du point étudié.
- La sélection des variables non locales est originale et très bien expliquée. L'idée d'utiliser les points à faible entropie conditionnelle semble en effet bien adaptée aux situations géographiques complexes (zones côtières, fjords, …). Cette partie est bien détaillée, tous les concepts nécessaires à la définition de l'entropie sont clairement présentés et illustrés pour les 3 types de points rencontrés (au large, fjords, côtier). Cette étape nécessite de discrétiser les intensités et directions du vent. L'auteur fait un choix de 3 intensités de vents et 4 directions. Bien que cette division soit tout a fait raisonnable et logique pour l'application choisie, une discussion sur ce choix et sur l'éventuelle influence du nombre de classes de vents aurait été intéressante ici.

Cette étape de sélection de variables joue un rôle fondamental dans le processus de régression. Les choix effectués ici semblent très pertinents et démontrent que l'auteure a cerné les principaux processus mis en jeu pour expliquer les vents à haute résolution.

Ensuite, les 3 méthodes de régression présentées dans le chapitre 2 sont détaillées de manière plus formelle, en décrivant plus profondément et très pédagogiquement (avec illustrations sur des données synthétiques) la technique SVR qui sera utilisée par la suite. La manière dont l'auteure fixe les hyerparamètres, en s'appuyant sur un article de Cherkassky et Ma, est également présentée et intéressante. Je n'ai cependant pas bien compris quelle technique est finalement utilisée pour ces hyperparamètres (car de la validation croisée est appliquée au chapitre 6. Est-ce juste une approche pour contraindre l'espace des valeurs possibles ?).

Par ailleurs, l'auteure fournit de nombreux détails sur la technique SVR mais très peu sur le choix noyau associé, alors que ce noyau joue au final un rôle très important. Il aurait été intéressant de discuter avec autant de rigueur ce choix du noyau. Le reproduisant Gaussien utilisé ici est très certainement le meilleur et de nombreuses théories existantes sur ce noyau le confirment. Ce choix apparaît donc très adapté. Cependant ici, une illustration sur les bénéfices de ce noyau, les limites des autres et leur rôle dans un contexte de régression, aurait éclairé le lecteur. Dans un cadre de classification par exemple, un noyau Gaussien conduit à un espace de dimension infinie où la séparabilité entre classes est optimale ; dans un cadre de régression, quel rôle joue-t-il ?

Enfin, le *chapitre 6* applique la technique précédente à *l'émulation de champs de vitesses*. Le protocole expérimental est en premier lieu présenté. Des comparaisons très complètes, en se focalisant sur des points caractéristiques, en étudiant l'influence de la nature des variables explicatives, en testant si une classification préalable améliore les résultats, en comparant l'apport des différentes variables explicatives, … sont proposées. Les difficultés pratiques liées au choix des paramètres, aux temps de calculs, …, sont discutées.

Une rigueur certaine a été appliquée ici dans le processus de validation afin d'étudier très précisément le rôle de chaque variable, l'influence de chaque étape, … En pratique, l'émulation de champs de vents est très performante et en particulier près des zones difficiles (zones côtières, fjords). On termine convaincu de la performance de la technique proposée !

Ce document se termine par une conclusion et des perspectives montrant que l'auteure a acquis une très bonne maitrise de son sujet ce qui lui permet d'envisager des suites très intéressantes à ce travail (prise en compte d'autres sources d'informations, application à d'autres domaines, ...).

Comme tout travail de thèse, un certain nombre de commentaires/interrogations sont soulevés à la lecture du document. Parmi ceux-ci, j'aurais notamment des interrogations sur la manière dont les composantes (u,v) peuvent être manipulées conjointement. Le choix de considérer ces deux composantes indépendamment se comprend parfaitement du point de vue pratique mais une discussion plus précise sur ce point aurait été intéressante. Ces champs de vents étant majoritairement à divergence nulle, était-ce possible de ne travailler que sur la vorticité par exemple ? En termes de validation, ne peut-on pas extraire certaines propriétés physiques sur les champs de vents émulés (zones caractéristiques, fronts, dynamique, …) pour avoir une autre indication de la qualité de ces champs ? Toujours du point de vue physique, est-il possible combiner à une approche de régression statistique certaines connaissances de base sur les interactions entre résolutions (distributions, …) pour améliorer les résultats ?

UMR 6554 CNRS
LETG
Terre, Écosystèmes et Sociétés
Observatoire des Sciences de l'Univers de Rennes

Synthèse

Le sujet abordé a nécessité de s'attaquer à des notions d'océanographie, de télédétection et de statistiques notamment. On sent que Liyun He-Guelton s'est bien investie dans ces trois axes scientifiques ce qui est un point très important lui ayant permis de mener à bien ce travail. Par ailleurs, l'auteure ne s'est pas contentée de proposer une méthodologie validée sur quelques exemples mais son application pratique à de larges bases de données océanographiques a été effective. Cela est déjà plus rare et on imagine que l'investissement, en termes de manipulation de données, d'adaptation des méthodes et de calcul intensif, a du être conséquent.

Les différents aspects dans le choix des données ou de la méthode de régression sont discutés rigoureusement sans omettre des cas plus obscurs comme le problème difficile de synchronisation des données, de valeurs aberrantes ou d'hyperparamètres. Cela rend le manuscrit très complet et il est possible de reproduire la méthode à partir des informations contenues dans le document.

Pour toutes ces raisons (sur la qualité du document, sur le fait d'avoir mené jusqu'au bout un problème alliant plusieurs disciplines, pour la rigueur dégagée dans la démarche), j'émets un avis très favorable à la soutenance de la thèse de Liyun He-Guelton.

Fait à Rennes le 15 mai 2014