Prédiction d'images par krigeage et ACP sur base d'ondelettes, avec application à un modèle d'inondation côtière

Tran Vi-Vi Élodie Perrin, Olivier Roustant, Jeremy Rohmer, Olivier Alata

Prédiction d'images par krigeage et ACP sur base d'ondelettes, avec application à un modèle d'inondation côtière

Tran Vi-vi Élodie PERRIN[1], Olivier ROUSTANT[1], Jérémy ROHMER[2], Olivier ALATA[3]

[1]Mines Saint-Étienne, Univ. Clermont Auvergne, CNRS, UMR 6158 LIMOS, Institut Henri Fayol, F–42023 Saint-Étienne, France

[2]BRGM, 3 av. Claude Guillemin, BP 36009, 45060 Orléans Cedex 2, France

[3]Lab. Hubert Curien, UMR CNRS 5516, UJM-Saint-Étienne, IOGS, Univ. de Lyon, 42023, Saint-Étienne, France

elodie.perrin@emse.fr, roustant@emse.fr
J.Rohmer@brgm.fr, olivier.alata@univ-st-etienne.fr

Résumé – Cette recherche est motivée par l'étude d'inondations côtières au moyen d'un grand simulateur numérique. Chaque appel au simulateur renvoie une image, correspondant à une carte de hauteur d'eau inondée, en fonction de paramètres d'entrée liés à la mer. L'objectif est de prévoir l'image correspondant à un nouveau jeu de paramètres. Dans ce contexte, une approche classique consiste en 1) Réduire la dimension de l'image, vue comme un vecteur de pixels, par analyse en composantes principales (ACP); 2) Construire un modèle de régression par processus gaussiens - ou krigeage - sur les premières composantes principales; 3) Prédire avec ce modèle. Cependant, l'étape 1) est difficilement applicable au-delà d'un trop grand nombre de pixels. En outre, la régularité spatiale de l'image n'est pas prise en compte. Pour pallier ces inconvénients, nous proposons de modifier l'étape 1) en réalisant une ACP fonctionnelle sur base d'ondelettes de l'image, vue cette fois comme une fonction. Nous montrons comment fixer un nombre unique de coefficients d'ondelettes, commun à toutes les images de l'ensemble d'apprentissage, permettant de garantir une qualité d'approximation suffisante. La méthodologie est appliquée au modèle d'inondation sur des images de 4096 pixels. Elle permet de prédire des images 5 fois plus rapidement qu'avec l'ACP classique, pour une qualité d'approximation au moins équivalente.

Abstract – This research is motivated by the study of coastal flooding with a time-consuming numerical simulator. Each simulator run provides an image, corresponding to a map of flooded water level, depending on input parameters linked to the sea. The aim is to predict an image for a new set of input values. In this framework, a standard approach consists of 1) To reduce dimension of the image, viewed as a vector of pixels, by principal component analysis (PCA); 2) To build a Gaussian process regression model - or Kriging model - on the first principal components; 3) To predict with this model. However, step 1) is hardly applicable for a large number of pixels. Furthermore, it does not account for the spatial regularity of the image. To address these issues, we propose to modify step 1) by doing a functional PCA based on wavelets on the image, now viewed as a function. We show how to select a unique number of wavelets, common to all images of the learning set, in order to guarantee a sufficient approximation quality. The methodology is applied to the coastal flooding model on images with 4096 pixels. Predictions are 5 times faster as for standard PCA, for at least equivalent predictive performances.

1 Introduction

Cette recherche est motivée par l'étude d'inondations côtières au moyen d'un grand simulateur numérique. Chaque appel au simulateur renvoie une image, correspondant à une carte de hauteur d'eau inondée, en fonction de paramètres d'entrée liés à la mer. L'objectif est de prévoir l'image correspondant à un nouveau jeu de paramètres. Dans ce contexte, une approche classique [4] consiste en 1) Réduire la dimension de l'image, vue comme un vecteur de pixels, par analyse en composantes principales (ACP) ; 2) Construire un modèle de régression par processus gaussiens - ou krigeage - sur les premières composantes principales ; 3) Prédire avec ce modèle.
En pratique cependant, l'étape 1) est difficilement applicable au-delà d'un trop grand nombre de pixels. En outre, la régularité spatiale de l'image n'est pas prise en compte. Pour pallier ces inconvénients, nous proposons de modifier l'étape 1) en réalisant une ACP fonctionnelle sur base d'ondelettes de l'image, vue cette fois comme une fonction. L'ACP fonctionnelle revient ici à effectuer une ACP classique, non pas sur les pixels, mais sur les coefficients d'ondelettes.
L'utilisation d'ondelettes pour étudier des grands simulateurs n'est pas nouvelle, voir par ex. [5]. Cependant l'utilisation combinée de l'ACP et des ondelettes, obtenue au moyen de l'ACP fonctionnelle, semble être originale. En outre, nous utilisons ici un critère d'énergie pour fixer un nombre unique de coefficients d'ondelettes, commun à toutes les images de l'ensemble d'apprentissage, permettant de garantir une qualité d'approximation suffisante.
L'article est organisé comme suit. La section 2 présente une brève introduction du krigeage et de l'ACP fonctionnelle. La section 3 précise la méthodologie. Dans la section 4, la performance de la méthode proposée est comparée avec celle utilisant l'ACP standard.

2 Krigeage et ACP fonctionnelle

2.1 Krigeage

Le krigeage est une technique d'interpolation spatiale. Issu de la géostatistique pour prédire des concentrations de minerais dans le sous-sol, elle est maintenant généralisée pour des espaces d'entrée non limités à $\mathbb{R}^2$, et connue sous le nom de régression par processus gaussiens [1].
Pour fixer les idées, on considère une fonction

$$f : \mathcal{X} \subseteq \mathbb{R}^d \to \mathbb{R}$$

et un ensemble d'apprentissage formé d'un plan d'expériences $x^{(1)}, \ldots, x^{(n)}$ et d'observations correspondantes $y_i = f(x^{(i)})$ $(i = 1, \ldots, n)$. Dans l'interprétation probabiliste du krigeage, la fonction f est vue comme la réalisation d'un processus gaussien $Y(x)$ de moyenne $m(x)$ et de fonction de covariance - ou noyau - $C(x, x')$, contenant la dépendance spatiale entre x et x'. Un exemple de noyau est donné par :

$$C(x, x') = \sigma^2 \exp\left(-\frac{1}{2\theta^2}||x - x'||^2\right).$$

La prédiction en un nouveau point $x^\star$ est alors obtenue sous la forme d'une loi de probabilité, gaussienne, correspondant à la loi conditionnelle de $Y(x^\star)$ sachant $Y(x^{(i)}) = y_i$ $(i = 1, \ldots, n)$. En particulier, on peut calculer de façon analytique la moyenne (prédiction pontuelle) et sa variance (incertitude) :

$$\begin{array}{rcl} \hat{y}(x^\star) & = & m(x^\star) + C(x^\star)^\top \mathbf{C}^{-1} y \\ \sigma_y^2(x^\star) & = & C(x^\star, x^\star) - C(x^\star)^\top \mathbf{C}^{-1} C(x^\star) \end{array} \quad (1)$$

où $\mathbf{C} = (C(x^{(i)}, x^{(j)}))_{1 \leq i,j \leq n}$ est la matrice de covariance aux points du plan d'expériences, et $C(x^\star) = (C(x^\star, x^{(i)}))$ est le vecteur des covariances entre le nouveau point et le plan d'expériences. On peut voir que la prédiction (ponctuelle) $\hat{y}(x^\star)$ s'obtient comme la moyenne des valeurs observées y pondérée par des poids faisant intervenir la corrélation spatiale entre le nouveau point et les points d'apprentissage.

Le krigeage est largement utilisé pour l'étude de grands simulateurs numériques, pour deux raisons principales. D'une part, sa nature probabiliste permet d'obtenir à la fois une prévision et une incertitude de prévision, ce qui est utile pour les problèmes de calcul de risque (en particulier pour les inondations côtières). D'autre part, la méthode est paramétrée par une fonction (un noyau de covariance), ce qui la rend flexible, et permet d'intégrer des informations métier. Par conséquent, ce sera la méthode de prédiction privilégiée dans cet article, même si la méthodologie est transférable à d'autres modèles statistiques tels que les modèles linéaires, forêts aléatoires, etc.

2.2 ACP fonctionnelle

L'analyse en composantes principales (ACP) est une technique de réduction de dimension dans $\mathbb{R}^N$. Elle résume l'information contenue dans un ensemble de points en cherchant des axes orthogonaux, ou composantes principales, sur lesquels la variance des projections est la plus grande. Ces axes sont obtenus en diagonalisant la matrice de covariance empirique.
Cette technique peut bien sûr s'appliquer à des images, vue comme un vecteur de N pixels. Cependant, la diagonalisation de la matrice de covariance est coûteuse numériquement lorsque N est de l'ordre de quelques milliers, voire impossible au-delà. De plus une image contient souvent une régularité qui n'est pas prise en compte en travaillant pixel par pixel.
L'ACP fonctionnelle généralise l'ACP à des fonctions (ici des images). En effet, le problème résolu par l'ACP utilise des ingrédients de nature géométrique qui sont transposables aux fonctions de carré intégrable. On peut donc également résumer l'information apportée par un ensemble de fonctions à quelques composantes principales, obtenues en diagonalisant un opérateur de covariance, analogue de la matrice de covariance. En pratique, la méthode s'applique à des sous-espaces fonctionnels de dimension finie, obtenus avec une base de fonctions, comme des splines, des polynômes orthogonaux, des ondelettes, etc. Lorsque la base est orthonormée, on montre que réaliser une ACP fonctionnelle revient à effectuer une ACP classique effectuée sur les coefficients des fonctions dans la base. Autrement dit, pour une image exprimée dans une base d'ondelettes, effectuer une ACP fonctionnelle revient à faire une ACP classique, non pas sur les pixels, mais sur les coefficients d'ondelettes. Pour plus de détails, on renvoie à [2].

3 Méthodologie

Sans perte de généralité, on supposera que les images sont définies sur le domaine spatial $[0, 1]^2$. Le simulateur est vu comme une fonction :

$$\begin{array}{rcccc} f & : & \mathcal{X} \subseteq \mathbb{R}^d & \longrightarrow & \mathbb{L}^2([0,1]^2) \\ & & x & \mapsto & y_x(z) \end{array} \quad (2)$$

où x est le vecteur des entrées, et $y_x(z)$ est la valeur de l'image en z renvoyée par le simulateur en x. On suppose que l'on connaît n simulations de f : $\left\{(x^{(i)}, y_{x^{(i)}}(z)), i = 1, \ldots, n\right\}$. On cherche à prédire l'image $f(x^\star)$ en un nouveau point $x^\star$.

Dans la représentation 2, chaque image est vue comme une fonction de $L^2([0, 1])$. En pratique, il est nécessaire de se ramener à la dimension finie. Au lieu de discrétiser l'image en espace, en considérant les pixels, on considère un sous-espace de dimension finie de fonctions au moyen d'une base de fonctions. Pour les images, il est naturel de travailler avec une base d'ondelettes notée dans la suite $\phi_j(z), j = 1, \ldots, K$. Ainsi, on a pour tout $x \in \mathcal{X}$:

$$y_x(z) = \sum_{j=1}^{K} \beta_j(x)\phi_j(z) \quad (3)$$

où les $(\beta_j(x))_{j=1,\ldots,K}$ sont les coefficients d'ondelettes.

En pratique on choisit les coefficients d'ondelettes au moyen

de la décomposition de l'énergie :

$$||y_x||_2^2 = \iint_{[0,1]^2} y_x(z)^2 dz = \sum_{j=1}^{K} \beta_j(x)^2, \qquad (4)$$

en retenant ceux pour lesquels le ratio $\frac{\beta_j(x)^2}{\sum_{j=1}^K \beta_j(x)^2}$ est le plus important. Cependant, les coefficients $j(x)$ retenus par ce critère dépendent de x, ce qui est un problème pour prédire en un nouveau point $x^\star$. L'idée est alors de considérer la proportion d'énergie moyenne

$$\lambda_j = \mathbb{E}_x \left[\frac{\beta_j(x)^2}{\sum_{j=1}^K \beta_j(x)^2} \right] \qquad (5)$$

et de sélectionner les indices j, maintenant indépendants de x, pour lequel λ_j est le plus important. En pratique λ_j est approché par la moyenne empirique sur les images d'apprentissage, qui est une bonne approximation lorsque les expériences $x^{(1)}, \ldots, x^{(n)}$ sont réparties uniformément dans $\mathcal{X}$.

Nous pouvons maintenant décrire la méthodologie proposée, constituée des étapes suivantes :

1. Faire la décomposition en ondelettes des images de l'ensemble d'apprentissage $y(x^{(i)})$, $i = 1, \ldots, n$.
2. Sélectionner les $k \ll K$ coefficients d'ondelettes les plus importants selon le critère (5), notés $\beta^k(x^{(i)})$.
3. Effectuer une ACP dans $\mathbb{R}^k$ des $\beta^k(x^{(i)})$ $(i = 1, \ldots, n)$. On notera $t_1(x^{(i)}), \ldots, t_d(x^{(i)})$, les d premières composantes principales.
4. Pour chaque composante principale $\ell = 1, \ldots, d$, calculer la prédiction par krigeage $t_\ell(x^\star)$ en $x^\star$, à partir des observations $t_\ell(x^{(i)})$ $(i = 1, \ldots, n)$.
5. Estimer les coefficients d'ondelettes de $y_{x^*}(z)$:
   — Pour les coefficients retenus à l'étape 2, par le vecteur de $\mathbb{R}^k$ correspondant à $t(x^\star)$.
   — Pour les autres : par la moyenne empirique de ces coefficients sur les images d'apprentissage.
6. Reconstituer l'image $y_{x^*}(z)$ à partir des coefficients d'ondelettes estimés à l'étape précédente (Eq. 3).

4 Application aux inondations côtières

4.1 Modèle d'inondation

La méthodologie décrite dans la section 3 est appliquée à un cas d'inondation marine par processus de débordements. Le site d'étude est celui des Bouchôleurs (côte Atlantique française, proche de La Rochelle), qui a été touché lors de la tempête Xynthia en 2010 (voir Fig. 1).
L'inondation marine est simulée à l'aide du code numérique en différences finies MARS ([3]), auquel des adaptations ont été apportées par le BRGM pour tenir compte des spécificités liées aux processus locaux d'inondations (processus hydrauliques autour des connections comme les buses, déversoirs, etc. et les phénomènes de bréchification).
Nous nous focalisons sur l'impact des paramètres de forçages au large sur l'évolution spatiale de la hauteur d'eau maximale à terre après inondation. Les variables d'entrées du simulateur correspondent aux paramètres de l'évolution temporelle (simplifiée) de la marée et de la surcote (voir Fig. 1), à savoir $x = (T, S, t_0, t_+, t_-)$ avec T la pleine mer (variant entre 0.95 et 3.70 m), S le pic de surcote (variant entre 0.65 et 2m), t_0, la différence entre l'instant de T et celui de S (variant entre -6 et +6 heures), t_+ et t_- les durées de montée et de descente du signal (supposé triangulaire) de la surcote (variant entre 0.5 et 12 heures). La sortie du simulateur correspond à une carte de 4096 pixels (chaque pixel ayant une résolution de 25m).
Comme le code est coûteux en temps de calcul ($\sim$0.5-1 heure par simulation), un nombre limité de simulations a été réalisé en tirant aléatoirement 500 configurations parmi les bornes de variations de x selon une séquence de Sobol.

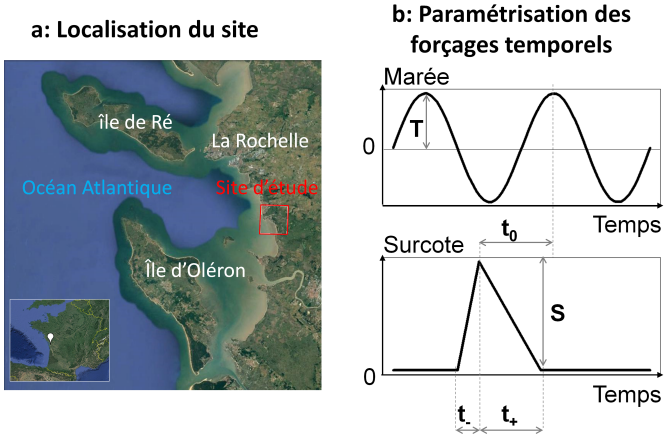

FIGURE 1 – a) Localisation du site, b) Paramétrisation des forçages temporels

4.2 Résultats

Dans cette section, on va comparer les performances de la méthode de prédiction d'images par ACP classique (voir l'introduction), ou par ACP fonctionnelle (voir section 3) sur 500 simulations. Les images de sortie sont de dimension 64×64. Pour la comparaison, on utilise la validation croisée 10-fold. L'échantillon d'observations est divisé en 10 sous-échantillons (soient constitués de 50 simulations). Les sorties de l'un d'entre eux sont supposées inconnues, et sont estimées par ACP ou ACP fonctionnelle, en se basant sur les observations des 9 autres sous-échantillons. La procédure est répétée pour chaque sous-échantillon. La pertinence des modèles est quantifiée avec le critère Q^2, calculé en chaque localisation de la sortie spatiale. En reprenant les notations de la section 3, on définit les Q^2 locaux par (6).

$$Q^2(z) = 1 - \frac{\mathbb{E}_X[(y_X(z) - \hat{y}_X(z))^2]}{\mathbb{V}\mathrm{ar}_X[y_X(z)]}, \ \forall z \in D_z \qquad (6)$$

$\hat{y}_X(z)$ est l'estimation de $y_X(z)$ obtenue par PG-ACP ou PG-ACPF. Le critère Q^2 compare la performance du modèle avec la moyenne, $Q^2(z) \in]-\infty, 1]$:

- Si $Q^2(z) = 0$, alors le modèle est aussi pertinent que la moyenne.
- Si $Q^2(z) < 0$, alors le modèle est moins pertinent que la moyenne.
- Plus $Q^2(z)$ est proche de 1, plus le modèle est considéré pertinent.

On calcule les cartes de Q^2 pour chaque sous-échantillon. Puis, on fait la moyenne des cartes obtenues.

Pour l'ACP fonctionnelle sur base d'ondelettes, on a choisi les ondelettes de Daubechies avec 6 niveaux de décomposition. On note n_{PC} le nombre de composantes principales et p la proportion d'énergie. On estime n_{PC} et p par validation croisée 10-fold : pour différentes valeurs de n_{PC} et p, on calcule la carte Q^2 moyenne obtenue par validation croisée 10-fold, puis on regarde la valeur médiane de cette carte (exemple Fig. 2). On choisit n_{PC} et p correspondant au Q^2 médian à partir desquels le modèle se stabilise et pour une valeur Q^2 qui tend vers 1. Dans la Figure 2, on a estimé $n_{PC} = 1$ pour les deux méthodes, ce qui explique 94% de la variance, et $p = 0.99$, ce qui correspond à prendre les 790 coefficients d'ondelettes les plus importants.
On observe que comparé à l'ACP classique, la prédiction d'image par ACP fonctionnelle a une capacité prédictive au moins équivalente, et est 5 fois plus rapide.

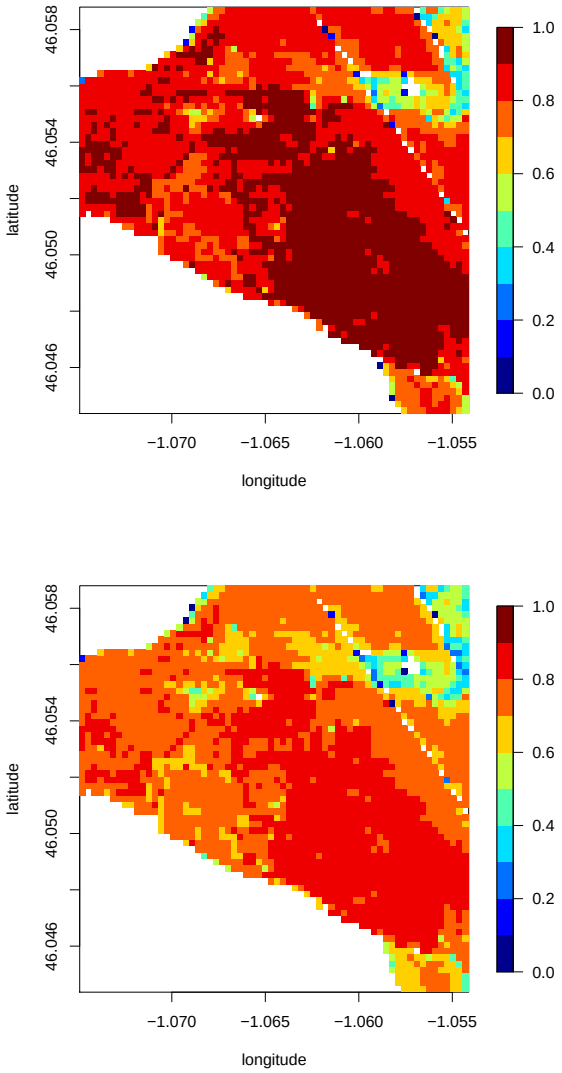

FIGURE 2 – Moyenne des Q^2 spatiaux, PG-ACPF, carte du haut, et PG-ACP, carte du bas. Avec $n_{PC} = 1$, pour PG-ACP et PG-ACPF, et $p = 0.99$ (soit ≈ 790 coefficients d'ondelettes), pour PG-ACPF. La zone blanche en bas à gauche des carte correspond à la mer. Celle en haut à gauche correspond à la terre, mais n'est jamais inondée.

Références

[1] C.E. Rasmussen. C.K.I. Williams. *Gaussian Process for Machine Learning*. The MIT Press, 2005.

[2] J. Ramsay. B.W. Silverman *Functional data analysis*. Springer Science & Business Media, 2005.

[3] P. Lazure. F .Dumas. *An external-internal mode coupling for a 3d hydrodynamical model for applications at regional scale (mars)* Advances in Water Resources, 31 :233-250. 2008.

[4] T. Chen. K. Hadinoto. W. Yan. Y. Ma *Efficient meta-modelling of complex process simulations with time–space-dependent outputs* Computers & chemical engineering, 35(3) :502–509. 2011.

[5] A. Marrel. B. Iooss. M. Jullien. B. Laurent. E. Volkova. *Global sensitivity analysis for models with spatially dependent outputs* Environmetrics, 22(3) :383-397. 2010.