Reconstruction de données manquantes par analyse en EOF : application aux séries temporelles de mesures de déplacement InSAR.

Alexandre HIPPERT-FERRER[1], Yajing YAN[1], Philippe BOLON[1]

[1]Laboratoire d'Informatique, Systèmes, Traitement de l'Information et de la Connaissance, Université Savoie Mont Blanc
5 chemin de Bellevue, 74944 Annecy-le-Vieux, France
alexandre.hippert-ferrer@univ-smb.fr

Résumé – Une méthode itérative, nommée EM-EOF (Expectation Maximization-Empirical Orthogonal Functions), est proposée afin de reconstruire des données manquantes au sein de séries temporelles de mesure de déplacement InSAR. La méthode EM-EOF décompose de manière itérative la covariance temporelle d'une série temporelle en différents modes EOF, puis sélectionne le nombre optimal de modes EOF afin de reconstruire la série temporelle. Après une initialisation appropriée des données manquantes, la méthode EM-EOF effectue (i) une minimisation de l'erreur de validation croisée pour estimer le nombre optimal de modes EOF à utiliser dans la reconstruction puis (ii) une mise à jour itérative des valeurs manquantes basée sur un critère de convergence prédéfini. Les résultats sur une série temporelle d'interférogrammes obtenus sur le glacier du Gorner démontre l'efficacité de la méthode pour reconstruire les valeurs manquantes.

Abstract – An iterative method, namely EM-EOF (Expectation Maximization-Empirical Orthogonal Functions) is proposed for the first time to retrieve missing values in InSAR displacement time series. The EM-EOF method iteratively decomposes temporal covariance into different EOFs modes by solving the eigenvalue problem, and then selects an optimal number of EOFs modes to reconstruct the time series. After an appropriate initialization of missing values, the EM-EOF method performs (i) a cross-validation error minimization to find an estimate of the optimal number of EOFs to use in the reconstruction and (ii) an iterative update of missing values which gives the best estimate of missing data points based on a predefined convergence criterion. Results on a time series of unwrapped interferograms over the Gorner Glacier demonstrate the high efficiency of the EM-EOF method at retrieving missing values.

1 Introduction

Les séries temporelles de mesure de déplacement terrestre obtenues par imagerie radar à synthèse d'ouverture (SAR) souffrent fréquemment de données manquantes en temps et en espace souvent liées aux limites techniques des méthodes de calcul de déplacement et/ou au changement de la surface imagée. L'analyse et la reconstruction de séries temporelles a suscité d'importants développements méthodologiques [1–3], en particulier dans la communauté océan-atmosphère. Cependant, aucune étude n'a été menée, à notre connaissance, sur le développement de méthodes de reconstruction de données manquantes au sein de produits issus de l'interférométrie radar (InSAR) comme les séries temporelles d'interférogrammes. Étant donné la spécificité des données InSAR, une telle méthode de reconstruction de données manquantes doit prendre en compte les complexités du signal de déplacement (linéaire, oscillatoire, etc.) et du type de bruit (corrélé spatialement et spatio-temporellement).

Parmis les méthodes existantes [4], les fonctions empiriques orthogonales (EOFs) ont été utilisées pour reconstruire des données manquantes et pour prédire l'évolution spatio-temporelle de signaux [5, 6]. Les principaux avantages de cette famille de méthodes comprennent la facilité d'implémentation et la non nécessité d'information a priori sur la nature du déplacement. L'analyse en EOF repose sur une décomposition de la matrice de covariance d'une série temporelle en modes EOF, permettant ainsi une représentation du signal en terme de modes de variabilité : tendances linéaires, oscillations, et bruit [7, 8]. En initialisant les données manquantes par une valeur pertinente, puis en sélectionnant un nombre approprié de modes EOF, il est ainsi possible d'extraire les caractéristiques principales d'une série temporelle, à partir de laquelle les valeurs manquantes peuvent être reconstruites. En mesure de déplacement InSAR, l'analyse en EOF a récemment été utilisée afin de débruiter et extraire un signal de déplacement dans une série temporelle d'interférogrammes issus d'images Sentinel-1A/B [9]. Les très bons résultats obtenus confirment l'efficacité des méthodes basées sur les EOFs pour l'analyse de séries temporelles de mesures de déplacement InSAR. Il semble donc intéressant d'utiliser ce type d'analyse pour reconstruire les données manquantes au sein de telles séries temporelles.

Nous proposons dans cette étude une méthode itérative de reconstruction de données manquantes nommée Expectation Maximization-Empirical Orthogonal Functions (EM-EOF), basée sur l'analyse en EOF et adaptée aux séries temporelles de mesure de déplacement InSAR. Le principe utilisé est équivalent à celui de l'algorithme EM : après initialisation des valeurs manquantes, celles-ci sont mises à jour par leurs valeurs espérées en prenant en compte les valeurs observées (Expectation), puis l'erreur entre les données initiales et reconstruites est minimisée (Maxi-

mization). La méthode EM-EOF se déroule en deux étapes. Afin de considérer le cas où aucune vérité terrain n'est disponible, la technique de validation croisée est proposée pour valider les résultats avec les données existantes. La méthode EM-EOF est ensuite appliquée à une série temporelle d'interférogrammes issus d'images Sentinel-1A/B acquises entre novembre 2016 et mars 2017 sur le glacier du Gorner.

2 Méthodologie

Le principe général de la méthode EM-EOF est illustré par le diagramme en figure 1. Après initialisation des valeurs manquantes, le nombre optimal de modes EOF pour reconstruire la série temporelle est estimé (étape 1). Une mise à jour itérative des valeurs manquantes est ensuite réalisée jusqu'à convergence de l'algorithme selon le principe EM décrit ci-haut (étape 2).

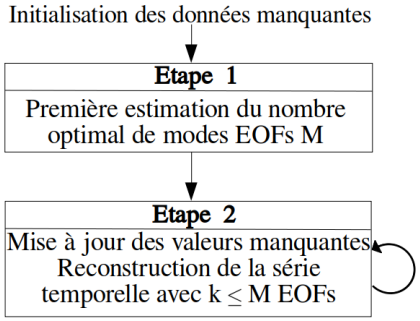

FIGURE 1 – Diagramme la méthode EM-EOF en deux phases. Les données manquantes sont d'abord initialisées. La phase 1 effectue une première estimation du nombre optimal de modes EOF M qui minimise l'erreur de reconstruction. La phase 2 est un affinage itératif des valeurs manquantes pour reconstruire la série temporelle avec un nombre k d'EOFs.

2.1 Notions théoriques

Soit un champ spatio-temporel $X(s,t)$ contenant des données manquantes. Le champ peut être représenté en forme matricielle :

$$X = \left(\mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_n\right) = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{pmatrix} \quad (1)$$

où chaque colonne $\mathbf{x}_t$ est une observation sur p points à un temps t ($t = 1, \dots n$) et chaque ligne est une série temporelle à un point s ($s = 1, \dots p$). La moyenne spatiale du champ à chaque temps lui est soustraite afin de former l'anomalie spatiale X' du champ X :

$$X' = X - \mathbf{1}_n \bar{X} \quad (2)$$

où $\mathbf{1}_n$ est un vecteur ligne de longueur n et $\bar{X}$ contient chaque moyenne spatiale $\bar{\mathbf{x}}_t$ de l'observation $\mathbf{x}_t$. La matrice de covariance temporelle $\hat{C}$ s'exprime alors par :

$$\hat{C} = \frac{1}{p-1} X'^T X' \quad (3)$$

Les vecteurs propres de la matrice $\hat{C}$ sont ensuite obtenus en résolvant l'équation suivante :

$$\hat{C}U = U\Lambda \quad (4)$$

où $\Lambda = diag(\lambda_1, \dots \lambda_n)$ contient les valeurs propres λ de la matrice $\hat{C}$ par ordre décroissant, chacune d'entre elles indiquant la fraction de la variance totale expliquée par le mode EOF qui lui correspond. Chaque colonne $\mathbf{u}_i$ de U est un vecteur propre de $\hat{C}$ correspondant à la valeur propre λ_i. Chaque vecteur propre est orthogonal à l'autre, d'où l'appelation EOF. Chaque mode EOF décrit la variabilité temporelle des motifs spatiaux dans X [8]. L'anomalie peut être reconstruite totalement en sommant les composantes principales (PCs) a_i multipliées par les transposées des vecteurs propres :

$$\hat{X}' = \sum_{i=1}^{n} a_i \mathbf{u}_i^t \quad (5)$$

Les PCs, définies par $a_i = X'\mathbf{u}_i$, représentent les motifs spatiaux ayant une même variabilité temporelle dans la série temporelle, alors que les vecteurs propres montrent comment ces motifs oscillent dans le temps. La troncature de l'équation (5) par un nombre $M \ll n$ permet de représenter le signal par les M premiers modes EOF correspondants aux plus grandes valeurs propres, et ainsi de capturer les principales caractéristiques temporelles du signal, alors que les autres modes EOF représentent souvent d'autres sources de perturbations comme du bruit [9]. Pour obtenir le champ reconstruit $\hat{X}$, il suffit d'ajouter la moyenne spatiale à l'anomalie :

$$\hat{X} = \hat{X}' + \mathbf{1}_n \bar{X} \quad (6)$$

2.2 Initialisation des valeurs manquantes

La valeur d'initialisation joue un rôle clef dans le processus de reconstruction, puisque c'est elle qui permet d'interpoler en utilisant l'analyse en EOF. Cette valeur doit être choisie de manière à être en accord avec la distribution des données observées, au risque de perturber l'énergie du système initial. Il existe dans la littérature plusieurs types d'initialisation : par 0 [5, 10] et par des valeurs voisines aux données manquantes [11]. Malgré ces résultats, il n'existe pas d'étude comparative de l'impact du paramètre d'initialisation sur la performance de la reconstruction. Nos simulations numériques ont montré qu'une initialisation des valeurs manquantes par un bruit centré de distribution gaussienne et un bruit corrélé spatialement ne change pas la performance de reconstruction par rapport à une initialisation par 0. De plus, il a été montré que plus la distribution des valeurs d'initialisation diffère de celle des données observées, plus le temps de convergence augmente. Nous choisissons donc d'initialiser les valeurs manquantes par 0, puisque cela ne nécessite pas de disposer d'information sur la nature du déplacement.

2.3 Étape 1 : première estimation du nombre optimal de modes EOF

Après initialisation des données manquantes, les équations (3) et (4) sont calculées, puis la série temporelle est reconstruite (équation (5)) en ajoutant successivement les modes EOF un à un. A chaque étape, l'erreur moyenne quadratique de validation croisée (cross-RMSE) [12] est calculée :

$$E(k) = \frac{1}{N}\sqrt{\sum_{i=1}^{N} |\hat{x}_i^k - x_i|^2} \tag{7}$$

où k désigne le nombre de modes EOF utilisés dans la reconstruction, $\{x_i\}_{1\leq i\leq N}$ est une collection de N éléments de X, et $\{\hat{x}_i^k\}$ est la même collection après reconstruction de X avec k EOFs. Les points x_i sont choisis aléatoirement parmis les données existantes puis retirés avec leurs valeurs gardées en mémoire. Après reconstruction, les points x_i sont comparés à $\hat{x}_i^k$ selon l'équation (7) afin d'établir une mesure de l'erreur de reconstruction. Le nombre de points N doit être choisi de manière à représenter statistiquement les données tout en omettant l'effet de dégradation des données reconstruites (les points étant rétirés des données, on augmente la quantité de données manquantes). L'utilité de la cross-RMSE se manisfeste particulièrement lorsqu'aucune vérité terrain n'est disponible pour valider les résultats, ce qui est fréquent en mesure de déplacement terrestre par télédétection. Le nombre optimal de modes EOF M est ensuite obtenu en recherchant le minimum de $E(k)$:

$$\underset{M\in[1,n]}{\arg\min}\, E(k) \tag{8}$$

où $k = 1, \dots n$ et n est le nombre maximum de modes EOF (ici la dimension temporelle). Des tests sur signaux synthétiques ont montré qu'un bruit fortement corrélé en temps et en espace présent dans les données peut engendrer une surestimation du nombre optimal de modes EOF, qui s'explique par le fait que ce type de bruit possède un comportement similaire au signal de déplacement, rendant ainsi la séparation du bruit et du signal de déplacement plus délicate.

2.4 Étape 2 : mise à jour des valeurs manquantes

La deuxième phase consiste en une mise à jour des données manquantes : pour chaque nombre de modes EOF k, les équations (3), (4) et (5) sont calculées de manière itérative. A l'itération i, la reconstruction est effectuée en utilisant les nouvelles valeurs des données manquantes obtenues à l'itération $i-1$. La cross-RMSE est calculée à chaque itération. La convergence de cette erreur est obtenue seulement si la différence $\Delta E = E_i(k) - E_{i-1}(k)$, où $E_i(k)$ désigne la cross-RMSE à l'itération i avec k modes EOF, est strictement inférieure à une valeur prédéfinie α. Une fois ce critère de convergence vérifié, un nouveau mode EOF est ajouté dans la reconstruction et la mise à jour itérative est renouvelée. La procédure est arrêtée si et seulement si la cross-RMSE augmente, et continue sinon jusqu'à ce que le nombre optimal de modes EOF M soit atteint.

3 Application à une série temporelle d'interférogrammes

16 interférogrammes ont été obtenus à partir de paires d'images Sentinel-1A/B acquises entre novembre 2016 et mars 2017 sur le glacier du Gorner (Suisse). Les données manquantes sont spatio-temporellement corrélés et la série temporelle contient quatre interférogrammes manquants.

Le nombre optimal de modes EOF pour reconstruire la série temporelle est estimé à 3. La figure 2 illustre des exemples de la reconstruction : le cas n° 1 (1ère ligne) contient 14.6% de données manquantes ; le cas n° 2 (2ème ligne) est un interférogramme manquant et le cas n° 3 (3ème ligne) contient 27.4% de données manquantes. Dans le cas n° 1, le signal de déplacement montre des motifs de déplacement conformes dans les zones de données manquantes, avec un déplacement lissé dans les zones observées en comparaison avec les données initiales, ce qui indique que la reconstruction ne dégrade pas les valeurs observées. En plus d'être centrés en 0 sur la majeure partie du glacier, les résidus ne montrent aucun signal de déplacement apparent. Des résidus importants sont observés sur les bords du glacier, où il existe des erreurs localisées de déroulement de phase dûes à la transition brute entre le rocher statique sur la rive et la glace en mouvement sur le glacier. Afin de reconstruire l'interférogramme manquant (cas n° 2), la moyenne temporelle est ajoutée à l'anomalie reconstruite (équation (6)) car la moyenne spatiale ne peut être calculée. Le déplacement reconstruit est également en accord avec les autres déplacements observés au sein de la série temporelle. La reconstruction du cas n° 3 montre également un motif de déplacement cohérent avec le champ initial. L'interférogramme initial est affecté par de nombreuses pertes de cohérence, dégradant ainsi sa qualité globale. Aucun signal de déplacement n'est toutefois observé au sein du résidu, et les zones de données manquantes sont reconstruites avec succès, ce qui démontre l'efficacité de la méthode EM-EOF face à de plus grandes quantités de données manquantes.

4 Conclusion

EM-EOF est une méthode itérative de reconstruction de données manquantes au sein de séries temporelles de champs de déplacement issus de l'imagerie SAR. Cette méthode ne nécessite pas d'information a priori, est adaptée aux données et peu coûteuse en temps de calcul. L'application sur une série temporelle d'interférogrammes sur le glacier du Gorner a permis de valider l'efficacité de la méthode en reconstruisant un signal de déplacement sans dégrader les zones observées. La méthode a également pu être appliquée avec succès afin de reconstruire un interférogramme manquant en utilisant la moyenne temporelle. Cela montre qu'il est ainsi possible d'augmenter la taille effective d'une série temporelle afin d'améliorer les connaissances des phénomènes physiques observés, en particulier lorsque les données manquantes sont une problématique fréquente. Dans un futur travail, une covariance spatio-temporelle (au lieu de temporelle) pourra

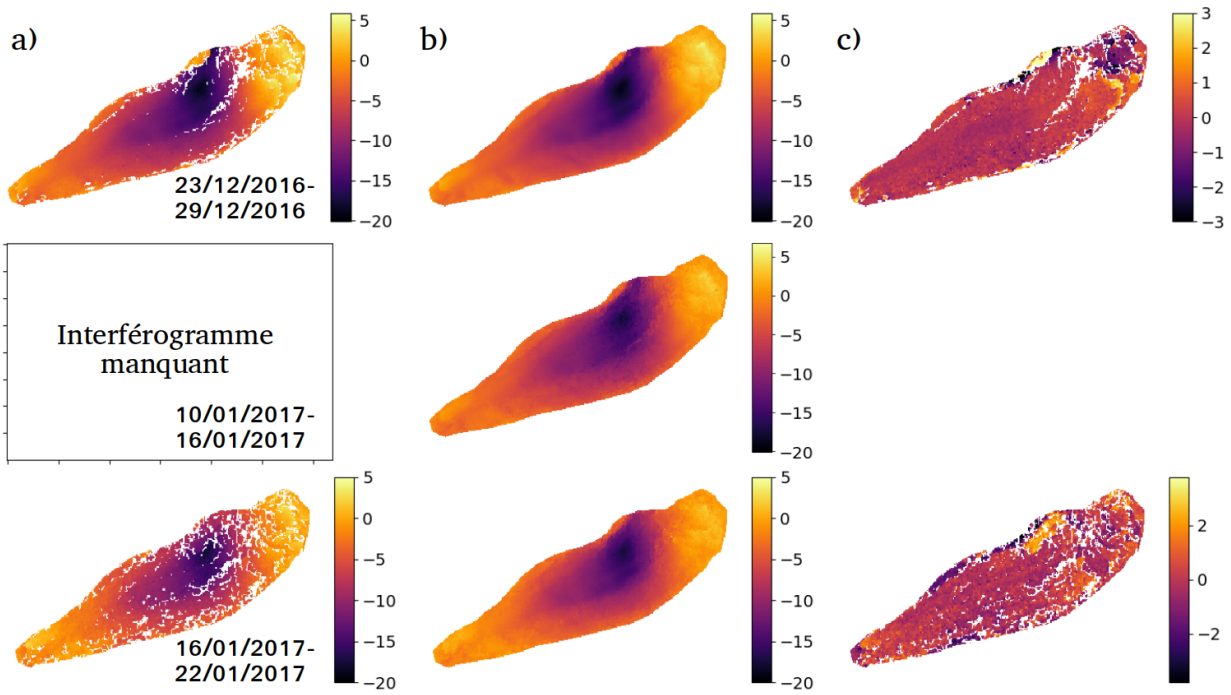

FIGURE 2 – Interférogramme a) initial, b) reconstruit et c) residus (reconstruit-initial) en géometrie radar sur la glacier du Gorner à trois intervalles de temps (23/12/2016-29/12/2016, 10/01/2017-16/01/2017, 16/01/2017-22/01/2017). Les valeurs de déplacement sont exprimées en centimètres dans la ligne de visée (LOS) du radar.

être estimée afin d'affiner la séparation du bruit corrélé et du signal de déplacement, et ainsi mieux identifier les types de déplacements observés. On pourra également s'intéresser à reconstruire directement des interférogrammes complexes afin de réduire les potentielles erreurs de déroulement de phase.

Remerciements Ce travail a été soutenu par le Programme National de Télédétection Spatiale (PNTS, http ://www.insu.cnrs.fr/pnts), projet n° PNTS-2019-11, et par le projet SIRGA financé par l'Université Savoie Mont Blanc.

Références

[1] D. Kondrashov and M. Ghil, "Spatio-temporal filling of missing points in geophysical data sets," *Nonlinear Processes Geophys.*, vol. 13, pp. 151–159, 2006.

[2] A. Alvera-Azcarate, A. Barth, J.-M. Beckers, and R. H. Weisberg, "Multivariate reconstruction of missing data in sea surface temperature, chlorophyll, and wind satellite fields," *J. Geophys. Res.*, vol. 112, no. C03008, 2007.

[3] F. Gerber, R. de Jong, M. E. Schaepman, G. Schaepman-Strub, and R. Furrer, "Predicting missing values in spatio-temporal remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2841–2853, 2018.

[4] H. Shen, X. Li, Q. Chen, C. Zeng, G. Yang, H. Li, and L. Zhang, "Missing information reconstruction of remote sensing data : A technical review," *IEEE Geosci. Remote Sens. Mag.*, vol. 3, pp. 61–85, 2015.

[5] J. M. Beckers and M. Rixen, "EOF calculations and data filling from incomplete oceanographics datasets," *J. Atmos. Oceanic Technol.*, vol. 20(12), pp. 1836–1856, 2003.

[6] C. Xu, "Reconstruction of gappy GPS coordinate time series using empirical orthogonal functions," *J. Geophys. Res. Solid Earth*, vol. 121, pp. 9020–9033, 2016.

[7] M. Ghil, M. Allen, M. D. Dettinger, K. Ide, D. Kondrashov, M. Mann, A. Robertson, A. Saunders, Y. Tian, F. Varadi, and P. Yiou, "Advanced spectral methods for climatic time series," *Review of Geophysics*, vol. 40, pp. 1–41, 2002.

[8] A. Hannachi, I. Jollife, and D. Stephenson, "Empirical orthogonal functions and related techniques in atmospheric science : A review," *Int. J. Climatol.*, vol. 27, pp. 1119–1152, 2007.

[9] R. Prébet, Y. Yan, M. Jauvin, and E. Trouvé, "A data-adaptative eof based method for displacement signal retrieval from InSAR displacement measurement time series for decorrelating targets," *IEEE Trans. Geosci. Remote Sens.*, 2019. Accepted.

[10] T. Schneider, "Analysis of incomplete climate data : Estimation of mean values and covariance matrices and imputation of missing values," *J. Climate*, vol. 14, pp. 853–871, 2001.

[11] B. Walczak and D. L. Massart, "Dealing with missing data : Part I," *Chemom. Intell. Lab. Syst.*, vol. 58, pp. 15–27, 2001.

[12] J.-M. Brankart and P. Brasseur, "Optimal analysis of in situ data in the western Mediterranean using statistics and cross-validation," *J. Atmos. Oceanic Technol.*, vol. 13, pp. 477–491, 1995.