Reconstruction de données manquantes dans des séries temporelles de mesures de déplacement par télédétection

Alexandre Hippert Ferrer

▶ **To cite this version:**

Alexandre Hippert Ferrer. Reconstruction de données manquantes dans des séries temporelles de mesures de déplacement par télédétection. Algorithme et structure de données [cs.DS]. Université Savoie Mont Blanc, 2020. Français. NNT : 2020CHAMA027 . tel-03507672

HAL Id: tel-03507672
https://theses.hal.science/tel-03507672

Submitted on 3 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ SAVOIE MONT BLANC

THÈSE

Pour le grade de

DOCTEUR DE L'UNIVERSITÉ SAVOIE MONT BLANC

Spécialité : **STIC Traitement de l'Information**

Arrêté ministériel : 25 mai 2016

Présentée par

Alexandre HIPPERT-FERRER

Thèse dirigée par **Philippe BOLON**
et codirigée par **Yajing YAN**

préparée au sein du **Laboratoire LISTIC**
dans **l'École Doctorale SISEO**

Reconstruction de données manquantes dans des séries temporelles de mesures de déplacement par télédétection.

Thèse soutenue publiquement à Annecy le 16 octobre 2020
devant le jury composé de :

M. Guillaume GINOLHAC
Professeur des Universités, Université Savoie Mont Blanc, Président

M. Laurent FERRO-FAMIL
Professeur des Universités, Université de Rennes 1, Rapporteur

M. Jean-Philippe OVARLEZ
Directeur de recherche, ONERA, Rapporteur

M. Antoine RABATEL
Physicien adjoint CNAP, Université Grenoble Alpes, Examinateur

M. Phillipe BOLON
Professeur des universités, Université Savoie Mont Blanc, Directeur de thèse

Mme Yajing YAN
Maître de conférences, Université Savoie Mont Blanc, Co-Directrice de thèse

À Serena.

Remerciements

Comme le veut l'usage, je me hâte de remercier en premier lieu les personnes qui ont encadré ce travail de thèse et qui ont su, chacun à leur manière, me guider à travers cette belle et sinueuse aventure nommée doctorat. Yajing, il faut le dire haut et fort, tu as été présente du début à la fin, à Annecy et à distance, telle une fidèle mentore. Je me souviens de ce premier papier corrigé au stylo rouge, qui de loin ressemblait davantage à un amas de ratures et autres reformulations syntaxiques ! *Ô désespoir* ! C'était sans compter sur tes immuables encouragements et ta volonté désintéressée de rendre les choses meilleures. Je ne peux que t'exprimer ma plus profonde gratitude. Philippe, avec ton précieux recul, tu as permis à ce travail d'avancer vers des directions sûres. Je retiendrai sans doute ta bonne humeur, ta curiosité scientifique et, agenda dense oblige, ton efficacité !

Je tiens à remercier les membres du jury d'avoir accepté d'évaluer mon travail de recherche et de s'être déplacés malgré les circonstances difficiles. Je remercie Jean-Philippe Ovarlez et Antoine Rabatel d'avoir effectué ce travail important qu'est le rapport du manuscrit. Vos corrections sont une pierre précieuse à l'édifice. Je remercie Laurent Ferro-Famil, qui a traversé la France dans toute sa largeur pour venir m'écouter et apporter des remarques constructives à la discussion. Enfin, je remercie Guillaume Ginolhac qui m'a accompagné en cette fin thèse sur le terrain de l'estimation statistique et m'a aidé à entrouvrir la porte d'une autre communauté scientifique.

Je veux exprimer toute ma reconnaissance aux personnes, au LISTIC et à Polytech, qui ont rendu ce travail possible, tant dans son aspect pratique que technique : Joëlle, Jean-Claude, Emmanuel, Martine, Elsa, Ingrid, Florent ainsi que les personnes chargées du nettoyage. Sans vous, il y aurait des chercheurs mais pas de vie de laboratoire. J'ai une pensée pour Sébastien et Lamia, qui ont fait l'effort (car c'en est un !) de recevoir tout.e.s les doctorant.e.s, y compris moi-même, en offrant leur bureau comme espace de discussion et d'écoute. Votre bienveillance, plus que bienvenue, est nécessaire. Merci. Un autre grand merci aux personnes que j'ai pu croiser lors des conseils du laboratoire et autres commissions locaux : Alexandre, Abdou, Flavien, Sylvie, Frédéric, Nicolas, Éric, Kavé. Ces séances, motivées par votre enthousiasme témoignant de l'envie de progresser dans tous les sens du terme, m'ont permis d'y voir plus clair dans ce millefeuille organisationnel qu'est l'université. Merci à Virginie et Christian de l'école doctorale, qui ont su, par un engouement certain, faciliter mon accueil à l'université. Je n'oublie pas les personnes avec qui j'ai eu l'occasion de collaborer en dehors du laboratoire : Romain et Louise à Grenoble, Nabil et Arnaud à Nanterre. Affaire à suivre.

Last but not least sur le campus d'Annecy-le-Vieux : une pensée émue pour tous et toutes les (post-)doctorant.e.s pour qui j'ai beaucoup de reconnaissance. Après avoir passé une année à partager un bureau avec Héla, qui m'a offert de la joie, du café et des tasses, l'heure de l'*open space* avait enfin sonné. Étienne, Emna, Melissa, Flo', Fanny, Hermanito, Lauranne, Quentin, Charles, Matthias, Davide, Xinyi, oui vous là ! Vous avez rendu l'atmosphère de cette thèse si particulière, entre pauses cafés, jeux de cartes, petit-déjeuners, soirées au bar, vidéos des doctorants... *let's conquer the world* ! Il y a eu entre temps, c'est vrai, un peu de travail. Sans vous, le LISTIC ne serait sans doute pas le même, en tout cas pour moi. Une belle pensée pour Mathias, compagnon de grimpe et ami, sans qui les soirées annéciennes auraient été plus tristes. Je remercie aussi Annaïg, Tigran, Sophie, Olga, les deux Thomas, Aurélien, Thibault, Ludo, qui, à un moment où un autre, ont été présent et ont participé, j'en suis persuadé, à ce travail. Merci à Marine et Nathalie de m'avoir fait découvrir le monde si riche des archives. On se reverra, ça aussi, j'en suis persuadé !

Enfin, je veux embrasser mes parents et mon frère, qui ont été bien plus qu'une épaule durant ces trois années. Alain, Nuria, F-X : votre indéfectible soutien m'a été d'une aide puissante. Lorsqu'on se sent aimé et soutenu, se hasarder (puisque c'est aussi cela, la science) n'est plus qu'une tâche des plus élémentaires. *El millor pel final*, je veux remercier Serena pour avoir été là *à chaque moment* : la distance n'a été qu'un extraordinaire argument pour rendre plus fort ce qui nous lie.

Résumé

Malgré la masse de données (satellitaires et in-situ) disponible en mesure de déplacement, l'incomplétude de données reste toujours un problème fréquemment rencontré. Ce phénomène est principalement dû au changement des propriétés de surface de l'objet observé et/ou aux limites techniques des méthodes de calcul de déplacement terrestre (e.g. interférométrie différentielle, corrélation croisée). Rendant les données discontinues en espace et en temps, l'incomplétude de données constitue un écueil vers la compréhension complète des phénomènes physiques sous-jacents liés au déplacement de surface. Malgré ce constat, l'analyse de données manquantes ne bénéficie pas d'une attention sérieuse et dédiée à la mesure de déplacement. Des méthodes de reconstruction adaptées aux données sont ainsi nécessaires pour gérer la présence de données manquantes spatio-temporelles au sein de séries temporelles de mesure de déplacement. Dans cette thèse, nous proposons trois approches pour l'analyse et la reconstruction de données manquantes en mesure de déplacement par télédétection. Les deux premières approches sont basées sur la décomposition de la covariance temporelle et spatio-temporelle du signal de déplacement en fonctions empiriques orthogonales (EOFs). Ces études ont débouchées sur le développement de deux méthodes, appelées EM-EOF et *extended* EM-EOF, nécessitant d'initialiser les valeurs manquantes avant traitement. La troisième étude, plus prospective, est orientée vers l'estimation robuste de la matrice de covariance du signal de déplacement, sans initialisation préalable des valeurs manquantes. Ces trois approches ont en commun de s'appuyer sur un schéma de résolution itératif de type espérance-maximisation (EM) ainsi que sur la sélection d'un nombre réduit de modes décrivant le maximum de variabilité du signal de déplacement. L'ensemble des cas d'études sur données réelles et synthétiques fournissent des résultats prometteurs, renforçant l'intérêt que porte l'étude des données manquantes en mesure de déplacement par télédétection.

Mots-clefs : Données manquantes, mesure de déplacement, EOF, covariance, algorithme EM, série temporelle.

Abstract

Despite the large volume of available (satellite and in-situ) data in displacement measurement, data incompleteness is still a commonly encountered issue. This phenomena is mainly due to surface property changes of the observed object and/or to technical limitations of the displacement extraction methods (e.g. differential interferometry, offset tracking). By generating time and space discontinuity, data incompleteness can hinder a thorough understanding of underlying physical phenomenon that induce surface displacement. However, missing data analysis in displacement measurement has not been paid significant and dedicated attention. In this context, advanced reconstruction methods, adapted to the data specificities, are necessary for handling spatio-temporal gaps in displacement measurement time series. In this thesis, we propose three approaches for analysing and imputing missing data in remotely sensed displacement measurement time series. The two first approaches are based on the decomposition of the temporal and spatio-temporal covariance of the displacement signal into Empirical Orthogonal Functions (EOFs). These studies have led to the development of two methods, called EM-EOF and extended EM-EOF, both requiring an initialization of the missing values. The third approach intends to explore techniques in robust estimation of the covariance matrix without initialization of the missing values. All approaches rely on an Expectation-Maximization (EM)-type iterative resolution scheme and reckon with the covariance low rank structure, which describe most of the variability of the displacement signal. Both synthetic simulations and real data applications present promising results, bringing to light the interest of the proposed approaches for missing data imputation in remotely sensed displacement measurement time series.

Keywords : Missing data, displacement measurement, EOF, covariance, EM algorithm, time series.

Notations et acronymes

Généralités

a ou A	Scalaire
$\mathbf{a}$	Vecteur
$\mathbf{A}$	Matrice

Notations

M	Délai spatial
N	Nombre d'observations
P	Nombre de variables
R, r	Nombre optimal de modes, rang d'une matrice
s	Indice espace
t	Indice temps
$\mathbf{M}$	Matrice indicatrice des données manquantes
$\mathbf{X}, \mathbf{Y}$	Champ spatio-temporel, matrice de données
$\mathbf{X}'$	Anomalie du champ spatio-temporel
$\mathbf{Y}_{\text{obs}}$	Partie observée de $\mathbf{Y}$
$\mathbf{Y}_{\text{mis}}$	Partie manquante de $\mathbf{Y}$
$\{\mathbf{y}_i\}$	Ensemble des vecteurs $\mathbf{y}_i$ pour $i = 1, \ldots, N$
$\mathbf{\Sigma}$	Matrice de covariance
$\mathbf{a}$	Composante principale
$\mathbf{u}$	Vecteur propre, EOF
$\boldsymbol{\theta}$	Paramètres d'un modèle paramétrique
$\hat{\boldsymbol{\theta}}$	Estimateur de $\boldsymbol{\theta}$
$\boldsymbol{\theta}^{(m)}$	Estimation de $\boldsymbol{\theta}$ à l'itération m d'un algorithme
$\Omega_{\boldsymbol{\theta}}$	Espace des paramètres
$\mathbb{E}[\cdot]$	Espérance mathématique
$\mathbb{R}$	Ensemble des nombres réels
$\mathcal{C}_k$	Critère de confiance à l'indice k
Λ	Critère du biais de surestimation
$p(\mathbf{Y}\|\boldsymbol{\theta})$	Distribution conditionnelle de $\mathbf{Y}$ sachant $\boldsymbol{\theta}$
$\mathcal{L}(\cdot)$	Fonction de vraisemblance
$\ell(\cdot)$	Logarithme de la fonction de vraisemblance

$\mathcal{Q}(\cdot)$	Fonction surégatoire : espérance de $\ell(\cdot)$
$\mathcal{N}$	Distribution gaussienne
$\mathcal{CG}$	Distribution gaussienne composée
Γ	Distribution Gamma
$\mathcal{O}(\cdot)$	Complexité algorithmique
δ_k	Cross-RMSE
$\delta^2_{\mathcal{SH}_{++}}$	Distance géodésique sur l'espace des matrices hermitiennes définies semi-positives
$\delta^2_{\mathcal{R}++}$	Distance géodésique sur l'espace des réels positifs
μ	Moyenne empirique
σ	Écart-type
τ	Paramètre de texture
$\arg\max_{a\in S} f$	Argument du maximum de la fonction $f : A \to B$ où $S \in A$
$\arg\min_{a\in S} f$	Argument du minimum de la fonction $f : A \to B$ où $S \in A$
$\text{diag}(\cdot)$	Opérateur diagonal
$\mathcal{H}(\cdot)$	Opérateur du point-fixe
$\text{Tr}(\cdot)$	Opérateur trace
$\text{vec}(\cdot)$	Opérateur vectorisation
$\cdot^H$	Opérateur transposée conjugée
$\cdot^T$	Opérateur transposée
$\odot$	Produit de Hadamard

Acronymes

ACP	Analyse en composantes principales
ACPP	Analyse en composantes principales probabiliste
ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer
CV	Cross validation
DINEOF	Data Interpolation with Empirical Orthogonal Functions
EEOF	Extended empirical orthogonal function
EM	Expectation-Maximization
EMV	Estimateur du maximum de vraisemblance
EM-EOF	Expectation-Maximization empirical orthogonal function
EOF	Empirical orthogonal function
ES	Elliptique symétrique
ESA	European Space Agency
ESS	Effective sample size
EVD	Eigenvalue decomposition
FP	Point Fixe
GNSS	Global Navigation Satellite System
IDW	Inverse distance weighting
InSAR	Interférométrie SAR
LOS	Line-of-sight
MNT	Modèle numérique de terrain
M-SSA	Multichannel singular spectrum analysis
NMSE	Normalized mean-square error
NNI	Nearest neighbor interpolation
OVPF	Observatoire volcanologique du Piton de la Fournaise
PC	Composante principale
RMSE	Root-mean-square error
RGI	Randolph Glacier Inventory
SAR	Synthetic Aperture Radar
SCM	Sample covariance matrix
SCN	Spatially correlated noise
SK	Simple kriging
SLC	Single Look Complex
SNR	Signal-to-noise ratio
SPOT	Système Probatoire pour l'Observation de la Terre
SSA	Singular spectrum analysis
STCN	Spatio-temporally correlated noise
SVD	Singular value decomposition
2D-SSA	2 dimension singular spectrum analysis

Table des matières

Résumé v
Abstract vii
Nomenclature ix
Table des matières xiii

Introduction générale **1**

1 Données manquantes en mesure de déplacement par télédétection : aperçu et problématique **5**
1.1 Le problème des données manquantes en télédétection 6
1.1.1 Qu'est-ce qu'une donnée manquante ? 6
1.1.2 Les données manquantes en télédétection 6
1.1.3 Type de données manquantes 7
1.2 Données manquantes en mesure de déplacement 9
1.2.1 L'imagerie SAR 9
1.2.2 L'imagerie optique 15
1.2.3 Le Global Network Satellite System 16
1.3 Méthodes pour l'interpolation de données manquantes 17
1.3.1 Méthodes en télédétection 17
1.3.2 Méthodes en mesure de déplacement 19
1.4 Approches prédictives 22
1.4.1 Les fonctions empiriques orthogonales 22
1.4.2 Les fonctions empiriques orthogonales étendues 26
1.4.3 Sélection du nombre de modes 28
1.4.4 Note sur l'initialisation des données manquantes 30
1.5 Approches paramétriques 30
1.5.1 L'algorithme Espérance-Maximisation 30
1.5.2 Initialiser ou ne pas initialiser 31
1.5.3 Estimation de la matrice de covariance 31
1.6 Synthèse 32

2 La méthode EM-EOF **35**
2.1 Introduction 36
2.2 La méthode EM-EOF 36
2.2.1 Organisation des données 37
2.2.2 Décomposition de la covariance 37
2.2.3 Reconstruction des données 38
2.2.4 Estimation du nombre optimal de modes 39
2.2.5 Initialisation des données manquantes 40
2.2.6 L'algorithme EM-EOF 40
2.3 Simulations numériques 41

2.3.1 Type de champ de déplacement . . . 42
2.3.2 Type de perturbation . . . 43
2.3.3 Type de données manquantes . . . 43
2.3.4 Paramètres de simulations . . . 44
2.3.5 Résultats et discussions . . . 45
2.4 Application sur données réelles . . . 54
2.4.1 Glacier du Gorner . . . 55
2.4.2 Glacier de Miage . . . 59
2.4.3 Étude d'un cas limite : le glacier d'Argentière . . . 59
2.4.4 Comparaison avec les méthodes NNI et krigeage . . . 62
2.5 Conclusion . . . 62

3 La méthode EM-EOF étendue 65
3.1 Introduction . . . 66
3.2 La méthode EM-EOF étendue . . . 67
3.2.1 Organisation et augmentation des données . . . 67
3.2.2 Estimation et décomposition de la covariance spatio-temporelle . . . 68
3.2.3 Reconstruction de la covariance spatio-temporelle . . . 69
3.2.4 Sélection du nombre optimal de modes . . . 70
3.2.5 Détermination du décalage spatial . . . 73
3.2.6 Synthèse de la méthode EM-EOF étendue . . . 74
3.3 Simulations numériques . . . 75
3.3.1 Type de champ de déplacement . . . 76
3.3.2 Type de perturbation et type de données manquantes . . . 76
3.3.3 Paramètres de simulations . . . 77
3.3.4 Résultats et discussion . . . 78
3.3.5 Bilan de l'étude synthétique . . . 91
3.4 Application sur données optiques : le cas du glacier Fox . . . 93
3.5 Conclusion et perspectives . . . 98

4 Vers une estimation robuste de la matrice de covariance de données incomplètes 101
4.1 Introduction . . . 102
4.2 Modèles statistiques et type de données manquantes . . . 102
4.2.1 Modélisation statistique . . . 103
4.2.2 Type de données manquantes . . . 106
4.3 Principe de l'algorithme EM . . . 107
4.3.1 Estimation du maximum de vraisemblance . . . 107
4.3.2 L'algorithme EM . . . 108
4.4 Estimation de la matrice de covariance en modèle Gaussien . . . 108
4.4.1 Estimation avec données manquantes de forme générale . . . 109
4.4.2 Estimation en rang faible . . . 111
4.4.3 Simulations numériques . . . 112
4.5 Estimation robuste de la matrice de covariance . . . 113
4.5.1 Estimation avec données manquantes en bloc . . . 114
4.5.2 Estimation en rang faible . . . 117
4.5.3 Simulations numériques . . . 117
4.6 Comparatif avec la méthode EM-EOF dans le cas Gaussien . . . 122
4.6.1 Estimation de la matrice de covariance sur données synthétiques . . . 122
4.6.2 Reconstruction de données manquantes sur données réelles . . . 123
4.6.3 Discussion . . . 130
4.7 Synthèse . . . 131

Conclusion générale **133**

Bibliographie **I**

Table des figures **XVII**

Liste des tableaux **XXV**

Liste des publications **XXVII**

A Génération d'un bruit corrélé **XXIX**
A.1 Génération d'un bruit spatio-temporel XXIX

B Opérateur Sweep et données GNSS **XXX**
B.1 L'opérateur sweep XXX
B.2 Données GNSS XXXI

Introduction générale

> Toute vision du monde a une singulière tendance à se considérer comme la vérité dernière sur l'univers.
>
> – Carl Gustav Jung, *L'Âme et la Vie*

> Il est clair que l'idée d'une méthode fixe, ou d'une théorie fixe de la rationalité, repose sur une conception trop naïve de l'homme et de son environnement social. Pour ceux qui considèrent la richesse des éléments fournis par l'histoire et qui ne s'efforcent pas de l'appauvrir pour satisfaire leurs bas instincts – leur soif intellectuelle, sous forme de clarté, précisions, "objectivité", "vérité" –, pour ceux-là, il devient clair qu'il y a un seul principe à défendre en toutes circonstances et à tous les stades du développement humain. C'est le principe : *tout est bon*.
>
> – Paul Feyerabend, *Contre la méthode*

La planète Terre est aujourd'hui surveillée. À distance. Des milliers de capteurs, embarqués à bord de plateformes satellitaires propulsées à des vitesses vertigineuses à plus de 400 kilomètres au-dessus de nos têtes, scrutent, imagent, sondent chaque mer, montagne, vallée et ville atteignable. Ces plateformes ont permis, depuis le lancement du satellite soviétique Sputnik en 1957, d'observer le déclin de l'étendue de la glace de mer en Arctique à partir des années 70 [Comiso2002], l'augmentation du niveau des océans [Cazenave2004] ou encore la fonte et le retrait des glaciers [Vaughan2013], autant de variables étant considérées comme des indicateurs du réchauffement climatique selon le groupe intergouvernemental d'experts sur l'évolution du climat (GIEC).

Les satellites sont aussi capables de calculer et de surveiller avec précision les déplacements terrestres, comme les glaciers, les séismes, les phénomènes de subsidence en milieu urbain ou les glissements de terrain. Parmi les satellites placés en orbite, on distingue les satellites imageurs passifs (imageurs optiques) des satellites imageurs actifs (imageurs radar). Le premier programme satellite à imagerie optique, Landsat, a permis de collecter des images à 80 mètres de résolution depuis le début des années 70. Dès la fin des années 70, le développement des radars à ouverture de synthèse (SAR, *Synthetic Aperture Radar*) avec le satellite américain Seasat, puis avec les familles de satellites ERS, ENVISAT, RADARSAT, etc., dans les années 90, a permis de s'affranchir des contraintes rencontrées dans le domaine optique, comme la présence de nuages ou l'absence de lumière.

Le suivi des déplacements de surface est fondamental pour mieux comprendre les phénomènes de déformation de la croûte terrestre liés aux forces de compression et de tension à la limite des

plaques continentales. Dans le cas des glaciers, le suivi continu de leur vitesse d'écoulement permet de mieux représenter leur dynamique interne et constitue également une variable climatique essentielle (VCE) selon le *Global Climate Observing System* (GCOS). La variété des méthodes dérivées de l'imagerie optique et de l'imagerie SAR, comme la corrélation d'images ou l'interférométrie radar (InSAR) rendent à présent possible le suivi temporel de la vitesse des déformations terrestres (volcaniques, sismiques, écoulements glaciaires) avec une précision atteignant le millimètre par an.

Plus récemment, la mise à disposition au grand public d'images radar et optique, notamment dans le cadre du projet Copernicus de l'Agence Spatiale Européenne (ESA), permet l'utilisation massive par le plus grand nombre des images acquises par la famille de satellites d'observation européens Sentinel. Ainsi, fin 2018, le système d'accès aux données Sentinel publiait en ligne la quantité astronomique de 16 TB de données par jour. Il est désormais possible, grâce à cette masse de données transmise et aux combinaisons d'observations issues de différentes plateformes, de voir évoluer les déformations en quasi-temps réel.

Malgré l'existence de ces techniques, les observations de déplacement par télédétection souffrent fréquemment d'un problème d'incomplétude, que l'on désigne par l'expression données manquantes [Shen2015]. Ce phénomène se produit lorsqu'il n'est pas possible de mesurer la variable désirée, c'est-à-dire le déplacement. Parmi les causes très diverses qui engendrent l'incomplétude de données, on peut citer par exemple : les événements naturels, comme les intempéries, survenant avant ou pendant la mesure, les limites des techniques de mesure du déplacement ou les défaillances d'un ou plusieurs capteurs. Plus précisément, il peut s'agir d'une forte chute de neige venant perturber la mesure du déplacement d'un glacier par interférométrie radar, ou d'un arrêt ponctuel de fonctionnement d'un instrument pour maintenance, créant ainsi une discontinuité temporelle de l'observation du déplacement.

Pourquoi vouloir analyser et reconstruire les données manquantes ? Premièrement, les données manquantes empêchent une compréhension à la fois précise et globale du déplacement de surface par l'obscuration de certaines zones, parfois très vastes. De plus, le déplacement de surface étant lié à d'autres paramètres physiques sous-jacents, l'observation réduite du premier peut conduire à une connaissance partielle des seconds, puisqu'une modélisation géophysique rigoureuse des déformations requiert des observations spatialement et temporellement résolues. Deuxièmement, les données manquantes peuvent être sources d'erreurs récurrentes dans l'interprétation des données de déplacement, comme les interférogrammes ou les cartes de corrélation d'image.

Ce travail de thèse entend aborder cette problématique en se consacrant entièrement à l'analyse des données manquantes en mesure de déplacement par télédétection. Un intérêt tout particulier est porté au développement de méthodes et d'algorithmes d'interpolation spatio-temporelle pour reconstruire des séries temporelles de cartes de déplacement incomplètes. Le but est d'apporter une alternative robuste aux méthodes d'interpolation existantes, dédiée à la mesure de déplacement, ce qui, à notre connaissance, n'a pas été porté à l'étude jusqu'à maintenant.

Les méthodes proposées sont mises à l'épreuve à travers des applications concrètes de mesures de déplacement de glaciers alpins situés dans différentes régions du monde, calculés à partir d'images Sentinel par corrélation d'image SAR, interférométrie différentielle et corrélation d'images optique où le phénomène de données manquantes est récurrent. Une application à vocation exploratoire est également proposée sur des mesures de déplacement incomplètes en zone volcanique calculées par Global Network Satellite System (GNSS), dont le système émetteur-récepteur permet de mesurer des déformations terrestres.

Ce mémoire s'organise en cinq chapitres et deux annexes dont le contenu est le suivant.

Le **chapitre 1** pose d'abord le problème de l'incomplétude de données en mesure de déplacement, puis dresse un inventaire des différentes méthodes de reconstruction de données manquantes en télédétection et en particulier en mesure de déplacement, ce qui permet de cerner le positionnement de cette thèse.

Le **chapitre 2** est consacré au développement d'une méthode de reconstruction de données manquantes reposant sur l'analyse en fonctions empiriques orthogonales (EOFs) et adoptant le formalisme de l'algorithme Espérance-Maximisation (EM). Cette méthode utilise notamment l'information temporelle des données pour reconstruire les valeurs manquantes. Trois applications sur champs de déplacement incomplets obtenus à partir d'images Sentinel-1 sur les glaciers du Gorner, de Miage et d'Argentière sont étudiées afin de saisir les avantages et inconvénient de la méthode proposée.

Le **chapitre 3** propose une extension de la méthode proposée au chapitre 2, en incluant l'information spatiale en plus de l'information temporelle pour la reconstruction de grandes quantités de données manquantes. Une application sur un champ de vitesses de surface obtenues à partir d'images Sentinel-2 sur le glacier Fox est proposée.

Le **chapitre 4** est dédié à l'exploration d'une approche différente pour l'analyse de données manquantes. Cette approche se base sur la connaissance d'un modèle statistique décrivant le comportement des données pour procéder à une inférence des paramètres statistiques. Une application, dont les résultats sont préliminaires, est proposée sur des mesures de déplacement reçues par un réseau de stations GNSS sur le Piton de la Fournaise.

Le **dernier chapitre** effectue une synthèse générale des contributions méthodologiques apportées dans cette thèse, et dresse un certain nombre d'ouvertures faisant suite au travail effectué pendant ces trois années.

Enfin, ce manuscrit est complété par des annexes. L'**annexe 1** décrit une procédure de calcul d'un bruit corrélé simulant les perturbations atmosphériques communes au images SAR. L'**annexe 2** présente le principe de l'opérateur Sweep, dont nous préciserons l'utilité au chapitre 4, et présente les séries temporelle de mesures de déplacement GNSS étudiées au chapitre 4.

L'objectif principal de ce travail de thèse est de reconstruire des séries temporelles de champs de déplacement incomplets. Chaque jeu de données d'applications possède des caractéristiques particulières en terme de complexité du champ de déplacement, de quantité et type de données manquantes ainsi que de niveau de bruit présent dans les données [1]. L'analyse se concentre sur les données suivantes :

1. Une série de 13 interférogrammes sur le glacier du Gorner dans la partie ouest du massif du Mont Rose calculés à partir d'images Sentinel-1 (chapitre 2) ;
2. Une série de 16 interférogrammes sur le glacier de Miage dans le massif du Mont Blanc calculés à partir d'images Sentinel-1 (chapitre 2) ;
3. Une série de 65 champs de déplacement sur le glacier d'Argentière dans le massif du Mont Blanc calculés par corrélation d'amplitude à partir d'images Sentinel-1 (chapitre 2) ;
4. Une série de 13 champs de déplacement calculés par corrélation d'images optiques Sentinel-2 sur le glacier Fox en Nouvelle-Zélande issue du travail de [Millan2019] (chapitre 3) ;

1. Les descriptions précises des données sont fournies dans les chapitres concernés.

5. Des séries temporelles contenant 1086 mesures de déplacement sur 23 stations GNSS au Piton de la Fournaise situé sur l'île de la Réunion (chapitre 4).

Les principales caractéristiques des données d'application sont regroupés dans le tableau 2.

Type de données	**Plateforme**	**Application**	**Taille de la série**	**Traitement**
D-InSAR	Sentinel-1	Glacier du Gorner	13	[Prébet2019]
D-InSAR	Sentinel-1	Glacier de Miage	16	Y. Yan
Corrélation	Sentinel-1	Glacier d'Argentière	16	Cette étude
Corrélation	Sentinel-2	Glacier Fox	12	[Millan2019]
GNSS	GNSS	Déformation du sol	1086	[Smittarello2019b] [Smittarello2019a]

Tableau 2 – Principales caractéristiques des jeux de données étudiés dans cette thèse.

1

Données manquantes en mesure de déplacement par télédétection : aperçu et problématique

Ce chapitre pose les bases contextuelles et les motivations de l'étude des données manquantes en mesure de déplacement. Pour cela, les problèmes que posent l'incomplétude de données sont abordés, puis un examen de la littérature des méthodes existantes est dressé selon deux grandes approches : prédictives et paramétriques. Les positionnements méthodologiques et/ou scientifiques vis-à-vis des problèmes posés sont également identifiés.

Sommaire

1.1 Le problème des données manquantes en télédétection 6
1.1.1 Qu'est-ce qu'une donnée manquante ? 6
1.1.2 Les données manquantes en télédétection 6
1.1.3 Type de données manquantes 7
1.2 Données manquantes en mesure de déplacement 9
1.2.1 L'imagerie SAR 9
1.2.2 L'imagerie optique 15
1.2.3 Le Global Network Satellite System 16
1.3 Méthodes pour l'interpolation de données manquantes 17
1.3.1 Méthodes en télédétection 17
1.3.2 Méthodes en mesure de déplacement 19
1.4 Approches prédictives 22
1.4.1 Les fonctions empiriques orthogonales 22
1.4.2 Les fonctions empiriques orthogonales étendues 26
1.4.3 Sélection du nombre de modes 28
1.4.4 Note sur l'initialisation des données manquantes 30
1.5 Approches paramétriques 30
1.5.1 L'algorithme Espérance-Maximisation 30
1.5.2 Initialiser ou ne pas initialiser 31
1.5.3 Estimation de la matrice de covariance 31
1.6 Synthèse 32

1.1 Le problème des données manquantes en télédétection

D'abord rencontré en analyse statistique de données [Rubin1976,Little1987,Preisendorfer1988], le problème des données manquantes en télédétection n'est pas nouveau. Aujourd'hui, celui-ci continue de susciter un vif intérêt de la part de la communauté scientifique. En effet, une recherche rapide des mots clefs "*missing data*" sur deux bases de données en lien direct avec la recherche en télédétection, les bases IEEE Xplore et GeoRef, permet de prendre la mesure de l'engouement scientifique : à ce jour, ce sont respectivement 13718 et 12016 résultats concernant des articles de recherche [1].

1.1.1 Qu'est-ce qu'une donnée manquante ?

La page Wikipédia "Données manquantes" énonce la définition suivante :

> [...] les *données manquantes* ou les *valeurs manquantes* se produisent lorsqu'aucune valeur de données n'est représentée pour une variable pour une observation donnée. Les données manquantes sont courantes et peuvent avoir un effet significatif sur l'*inférence*, les performances de *prédiction* ou toute autre utilisation faite avec les données

Nous reviendrons plus tard sur la signification des termes inférence et prédiction. La plupart du temps en télédétection, les données sont insérées dans une matrice, appelée matrice de données : on peut choisir de représenter les variables en ligne et les observations en colonne, ou l'inverse. Par variable, nous entendons la propriété ou la caractéristique d'un objet observé, alors que l'observation désigne la mesure de cette propriété. En télédétection, des milliers de variables peuvent être mesurées : température de surface de la mer, concentration d'ozone, déplacement de surface, etc. Lorsqu'il y a absence de données, certaines valeurs situées à certaines positions de la matrice de données sont manquantes, alors que les autres sont observées. Dans les situations de données manquantes, on emploiera plus généralement l'expression *incomplétude de données*.

1.1.2 Les données manquantes en télédétection

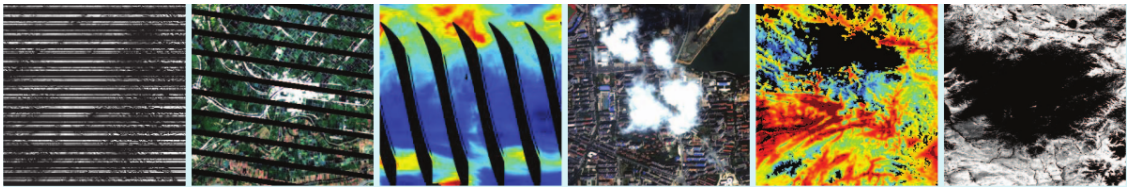
La donnée manquante étant inhérente à la production de données, la télédétection n'échappe pas à ce phénomène. Il est même difficile de trouver une application qui ne soit pas concernée par l'incomplétude de données. À cause d'un capteur défecteux ou de conditions atmophériques difficiles, les données acquises sont souvent dégradées, voire inutilisables. Parmi les nombreux cas existants en télédétection, [Shen2015] cite par exemple : le non fonctionnement de certains détecteurs du spectroradiomètre MODIS à bord du satellite Aqua, la défaillance du capteur SLC sur la plateforme Landsat [2] ou encore une anomalie d'acquisition de l'*ozone monitoring instrument* (OMI) à bord du satellite Aura.

Alors que 35% en moyenne de la surface terrestre est couverte par les nuages à un instant donné [Lin2013], ces derniers constituent une cause majeure d'incomplétude des données issues de capteurs passifs [Melgani2006, Lin2014, Wu2018, Zhang2018]. La continuité des données étant un gage de confiance dans la plupart des applications faisant usage des données mesurées à distance (classification, détection de changement, mesure de paramètres physiques, surveillance des milieux naturels, etc.), il semble logique que le développement et la mise en place de méthodes d'estimation des valeurs manquantes prennent alors un intérêt tout particulier. Il n'est pas anodin qu'un effort important pour développer des méthodes d'interpolation ait été produit pour des

1. Ceci sans compter les quelques 98632 résultats sur la base BioMed Central, une des plus grandes bases de données de la recherche bio-médicale.
2. Ce problème, appelé "SLC-off", a suscité le développement de dizaines de méthodes par des équipes du monde entier pour y remédier.

applications géoscientifiques telles que les sciences de l'atmosphère, l'océanographie, sciences de la végétation ou l'hydrologie [Beckers2003, Kondrashov2006, Alvera-Azcarate2007, Hocke2009, Verger2013, Gerber2018], puisque ces dernières utilisent et dépendent souvent des données de télédétection.

La figure 1.1 montre quelques cas de données manquantes sur des images satellitaires dues aux nuages ou à des capteurs défectueux. Notons que les valeurs manquantes adviennent directement sur la valeur du pixel de l'image ou sur la valeur du paramètre physique observé, ce qui est une conséquence indirecte d'une cause antérieure (présence de nuage, capteur défectueux, intempéries) au calcul du produit quantitatif (carte de déplacement, température de surface de la mer, NDVI [3], réflectance, etc.).



(a) (b) (c) (d) (e) (f)

Figure 1.1 – Exemples de données manquantes en télédétection. (a) réflectance avec capteur défecteux sur la bande 6 d'Aqua MODIS ; (b) le problème SLC-off sur la plateforme Landsat ETM+ ; (c) concentration en ozone par Aura OMI ; (d) image IKONOS-2 en présence de nuages ; (e) mesure de *Land Surface Temperature* par MODIS en présence de nuages ; (f) NDVI en présence de nuage (MODIS). Image issue de [Shen2015] ©2017 IEEE.

1.1.3 Type de données manquantes

Formes des données manquantes

Comme l'illustre la figure 1.2, les données manquantes peuvent prendre plusieurs formes dans la série temporelle d'images. Cette forme dépend du mécanisme responsable de leur cause. Les données manquantes peuvent être distribuées aléatoirement, corrélées spatialement, temporellement, spatio-temporellement. Par "aléatoire", on entend que les positions des données manquantes n'ont pas de dépendance spatiale ni temporelle. "Corrélée" signifie que les positions des données manquantes possèdent une dépendance spatiale, temporelle ou spatio-temporelle.

Souvent, plusieurs de ces distributions sont présentes au sein d'une série temporelle. Prenons quelques exemples simples : la distribution aléatoire peut-être due à une intempérie imprévisible, provoquant ainsi une dégradation de certains pixels de l'image. Dans le cas de la mesure de déplacement, une telle configuration est possible si des changements de terrains aléatoires ont eu lieu entre deux ou plusieurs acquisitions, ce qui rend le calcul du champ de déplacement difficile. Les distributions corrélées peuvent être dues à des phénomènes similaires non aléatoires : dans le cas des glaciers alpins, il peut s'agir de chutes de neige spatialement ou temporellement localisées sur une zone ou un intervalle précis. Selon la technique utilisée, le calcul du déplacement peut être très sensible à ce type de changement, ce qui engendre des erreurs dans le résultat final, comme des valeurs aberrantes ou atypiques qui sont ensuite retirées artificiellement, créant ainsi une incomplétude des données.

Il est également possible qu'une ou plusieurs images soient manquantes, par exemple si le satellite effectue un changement momentané de l'angle de visée. Dans ce cas, la forme des données manquantes est spatio-temporellement corrélée. Enfin, lorsqu'on combine plusieurs séries

3. *Normalized difference vegetation index.*

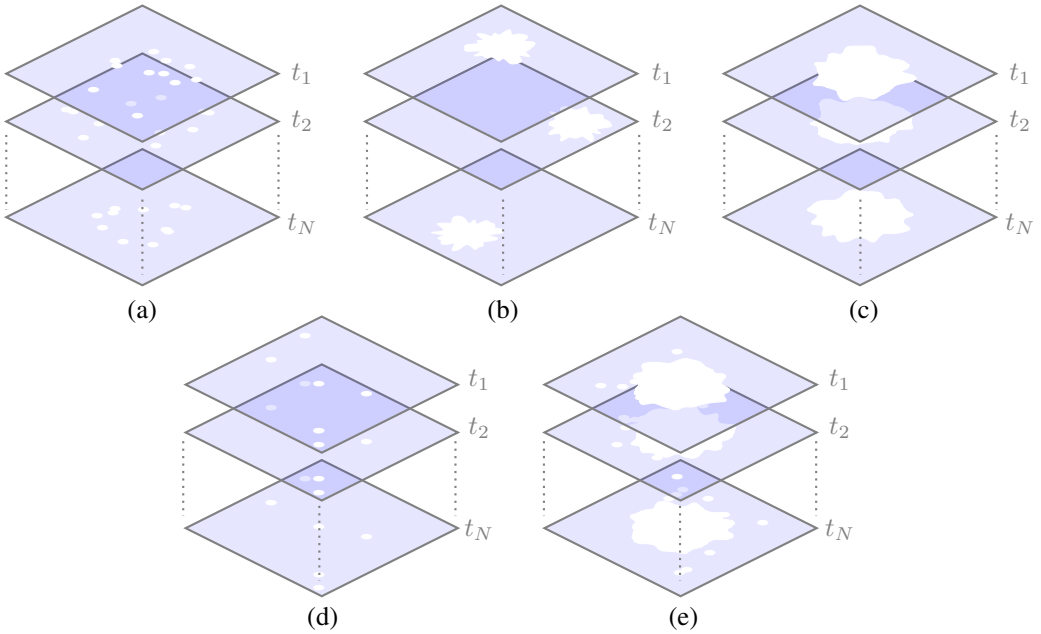


Figure 1.2 – Exemples de formes de données manquantes dans une série temporelle d'images aux dates $t_1, t_2, \ldots, t_N$, où N est le nombre d'images : (a) aléatoire (sans dépendance spatiale ni temporelle) ; (b) corrélée spatialement ; (c) corrélée spatio-temporellement ; (d) corrélée temporellement ; (e) aléatoire et corrélée spatio-temporellement.

temporelles issues d'instruments de satellites différents [Nakamura2007], on se confronte également au temps de mission de chaque plateforme, paramètre susceptible de créer une discontinuité temporelle des acquisitions.

Mécanismes responsables de l'incomplétude de données

Avant d'effectuer toute reconstruction de données manquantes, il est nécessaire de savoir pourquoi ces données manquent. Si l'on raisonne en termes de probabilités, on distingue trois raisons pour lesquelles des données peuvent manquer (voir le passage en revue méticuleux des types de données manquantes par [Santos2019]) :

— *Missing completely at random* (MCAR) : la probabilité qu'une donnée soit manquante ne dépend ni des valeurs observées, ni des valeurs non observées (manquantes). En d'autres termes, la probabilité qu'une donnée soit manquante est aléatoire, c'est-à-dire indépendante des valeurs . Par exemple, il peut s'agir d'un arrêt de la mesure pour des raisons techniques imprévisibles entraînant une maintenance, ou à cause de la présence d'intempéries non prédites. Dans les deux cas, la probabilité qu'une valeur manque est indépendante des valeurs mesurées et non mesurées.

— *Missing at random* (MAR) : concept initialement développé par [Rubin1976], il s'agit de dire que la probabilité qu'une donnée soit manquante dépend seulement des données observées. Par exemple, lorsque plusieurs variables sont observées sur un même pixel (comme les images multi-spectrales), la probabilité qu'une observation d'une variable soit manquante dépend d'autres observations sur d'autres variables, et non d'elle-même.

— *Missing not at random* (MNAR) : la probabilité qu'un élément soit manquant dépend seulement des données manquantes. Formulé autrement, cela signifie que la probabilité qu'une valeur soit manquante est reliée à sa propre valeur. Par exemple, on peut décider de supprimer des valeurs aberrantes d'une image : les valeurs sont donc manquantes à cause de leur propre valeur, jugée trop atypique. Notons que dans certains cas, il est théoriquement impossible

de prouver qu'un mécanisme d'incomplétude est MNAR puisqu'on ne connaît simplement pas les valeurs aux points manquants. On se contente donc d'émettre une hypothèse sur ce mécanisme.

Remarque Une définition des mécanismes MCAR, MAR et MNAR, formulée à l'aide des distributions de probabilité des données manquantes, sera fournie au chapitre 4, où l'approche étudiée dépend de l'hypothèse sur le mécanisme responsable de l'incomplétude de données.

1.2 Données manquantes en mesure de déplacement

1.2.1 L'imagerie SAR

Le radar à synthèse d'ouverture (SAR pour *Synthetic Aperture Radar*) est un système d'imagerie actif utilisant les micro-ondes, dont les propriétés permettent l'acquisition d'images (figure 1.3) quelle que soit la météo, de jour comme de nuit. L'imagerie SAR permet notamment de calculer des modèles numériques de terrain (MNT) à précision centimétrique en mesurant les variations du chemin aller-retour de l'onde électromagnétique en fonction du temps d'acquisition et de la position du satellite [Ferretti2007]. L'imagerie SAR trouve des applications dans des domaines extrêmement variés, allant des sciences du climat à la détection de changement, ou de la cartographie 4-D à l'exploration planétaire, à travers des techniques bien connues (polarimétrie, interférométrie) voire plus sophistiquées (interférométrie polarimétrique, tomographie holographique). Le lecteur et la lectrice désireux d'un aperçu global des techniques existantes pourra se référer au tutoriel de [Moreira2013] ou au livre de [Maître2013]. Dans cette thèse, nous nous concentrons sur les données de mesure de déplacement calculées par interférométrie différentielle et corrélation d'amplitude, qui peuvent être sources d'incomplétude de données.

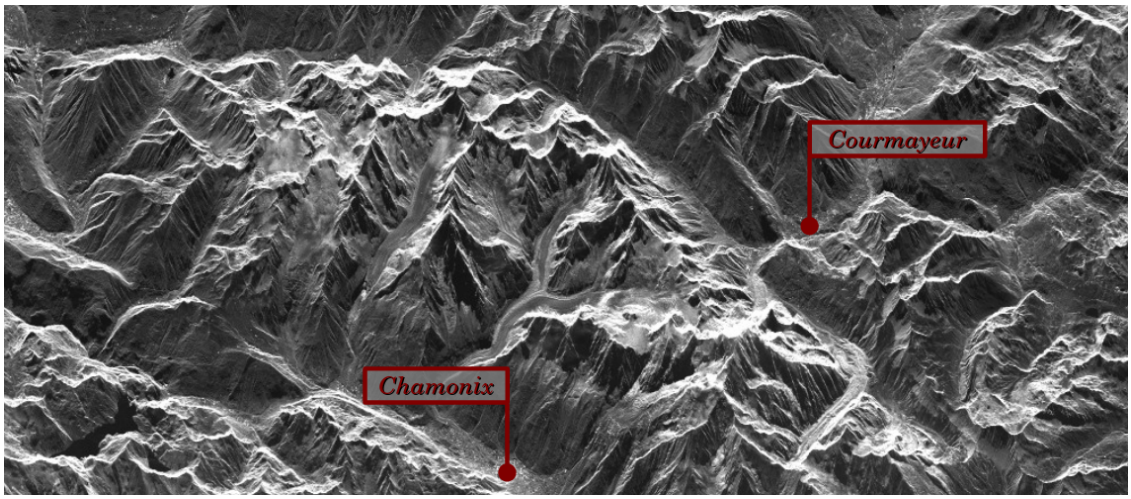


Figure 1.3 – Image SAR acquise par le satellite TerraSAR-X en trajectoire ascendante sur le massif du Mont-Blanc [Fallourd2012]. Certains glaciers du massifs sont clairement visibles, comme la Mer de Glace et le glacier d'Argentière côté Chamonix, ou le glacier de Miage côté Courmayeur. Le nord se situe à gauche.

Interférométrie SAR

Le principe de l'interférométrie, connu depuis longtemps par les opticiens puis importé chez les radaristes au milieu des années 70 [Graham1974], consiste à étudier les interférences entre deux

sources cohérentes. L'interférométrie SAR (InSAR) utilise ce principe pour mesurer la distance entre le satellite et un objet situé sur la surface terrestre avec une grande précision : l'idée consiste à comparer la phase de deux ou plusieurs images SAR complexes ayant été acquises à des temps différents. Comme la phase de chaque pixel de l'image SAR contient une information très précise sur la fauchée (*range*), il est possible de mesurer des différences de chemin aller-retour de l'onde de l'ordre du centimètre voire du millimètre [Moreira2013].

Formation d'un interférogramme La formation d'un interférogramme est le produit hermitien de deux images SAR complexes, appélées images *Single Look Complex* (SLC), que l'on aura préalablement recalées sur une même géométrie. Avant d'extraire les valeurs de la phase, une opération de moyennage local peut être appliquée lors de la formation de l'interférogramme complexe (*multi-looking*) afin de réduire le bruit sur la phase [Goldstein1988]. La phase est dominée par un motif en franges (voir figure 1.4) dans la direction de la fauchée. Ce motif est dû en grande partie à la baisse de la phase interférométrique en fonction de l'augmentation de la distance de fauchée et de l'angle de visée. La qualité de la phase interférométrique est indiquée par la cohérence interférométrique, qui mesure le degré de corrélation entre deux images SAR :

$$\gamma = \frac{\sum_{i,j\in\Omega} z_1(i,j) z_2^H(i,j)}{\sqrt{\sum_{i,j\in\Omega} z_1(i,j) z_1^H(i,j)} \sqrt{\sum_{i,j\in\Omega} z_2(i,j) z_2^H(i,j)}} \tag{1.1}$$

où Ω est la fenêtre d'estimation. La cohérence γ, qui varie entre 0 et 1, peut être utilisée pour mesurer la qualité d'un interférogramme. En pratique, de nombreux facteurs contribuent à une perte de cohérence, comme le niveau de bruit [Zebker1992], la décorrélation temporelle, qui décrit les changements de structure et de permittivité de la scène observée [4] entre deux acquisitions à deux temps différents, ou encore la décorrélation spatiale (géométrique) due à la légère différence entre les deux géométries d'acquisition.

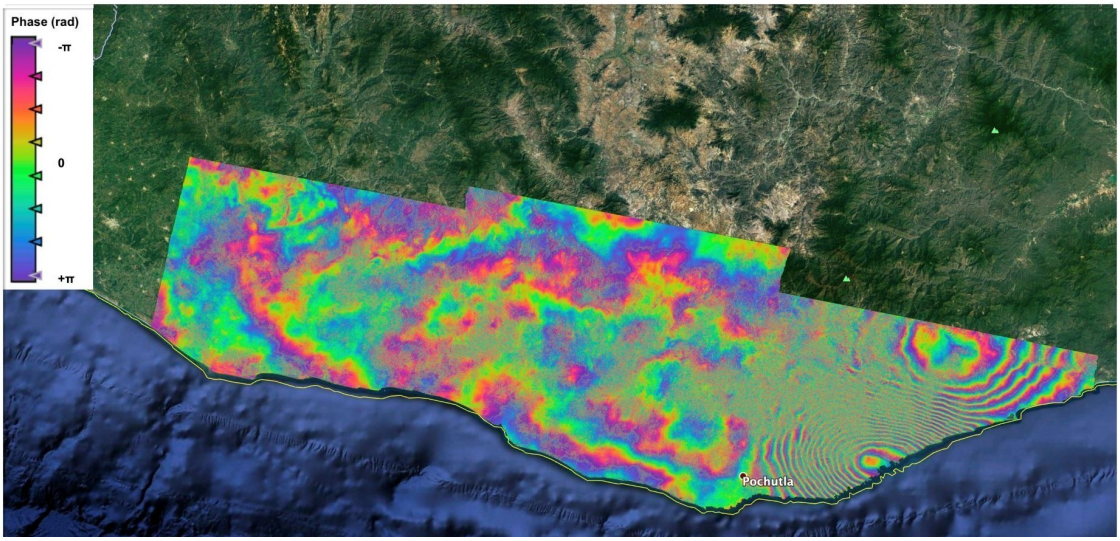


Figure 1.4 – Interférogramme du déplacement cosismique de surface dans la zone d'Oaxaca au Mexique calculé à partir de plusieurs images Sentinel-1 avant et après le séisme du 23 juin 2020 (image ESA). Les franges montrées ici sont dues à la phase du déplacement sismique.

Phase interférométrique Le modèle de la phase interférométrique peut être défini par la somme de ses différentes composantes, soit :

4. Ces changements introduisent une modification des mécanismes de rétrodiffusion de la scène.

$$\varphi = \varphi_{\text{orb}} + \varphi_{\text{topo}} + \varphi_{\text{depl}} + \varphi_{\text{atm}} + \varphi_{\text{bruit}} \tag{1.2}$$

où φ_{orb} est la phase orbitale, φ_{topo} est la composante topographique, φ_{depl} est la phase liée au déplacement, φ_{atm} est la composante due à l'atmosphère, dont la concentration en vapeur d'eau produit des retards non négligeables sur l'onde [Rocca2007], et φ_{bruit} est la phase du bruit. Toutes ces composantes sont susceptibles de créer des franges supplémentaires dans l'interférogramme. De nombreuses techniques de correction de la phase interférométrique existent pour extraire chacune des composantes mentionnées ci-haut. Pour un résumé de ces techniques, ainsi que pour une description plus exhaustive des différentes composantes, nous renvoyons les lecteurs et lectrices intéréssés vers le tutoriel de [Moreira2013] ou la thèse de Y. Yan [Yan2011].

Déroulement de phase interférométrique Le déroulement de phase permet de résoudre l'ambiguïté sur la phase, dont la valeur augmente de 2π lorsque l'onde parcourt une distance égale à la longueur d'onde. Ce problème non-linéaire constitue une étape fondamentale en interférométrie différentielle et nécessite de formuler une hypothèse de départ sur la continuité de la phase. La réalité du terrain, comme la discontinuité des zones cohérentes ou un fort gradient de déplacement, peut faire échouer cette hypothèse. De nombreuses méthodes dédiées au déroulement de phase ont ainsi été développées, dont l'article de [Yu2019] en fait un résumé.

Données manquantes En mesure du déplacement, les données manquantes surgissent principalement pour une raison déjà mentionnée plus haut : la perte de cohérence entre une ou plusieurs acquisitions. Cette perte de cohérence peut entraîner des erreurs de déroulement de phase, telles que des sauts de phase ou des valeurs aberrantes. Dans [Pepe2016], les auteurs retirent les valeurs sur les pixels à faible cohérence temporelle, ce qui engendre une incomplétude de données dans la dimension spatiale (figure 1.5). D'autres exemples existent en milieu urbain, lequel peut être concerné par des déformations terrestres comme des phénomènes de subsidence : on pourra citer les études de [Yan2012] sur la ville de Mexico ou plus récemment de [Aslan2018] sur la ville d'Istanbul, où les déplacements présentent des valeurs manquantes en temps et en espace.

La détection de glissement de terrain par imagerie SAR est également sujette à l'incomplétude de données en espace et en temps [Jakob2012]. De plus, les événements à faible magnitude ne sont souvent pas détectés à cause de la mise en place d'un seuil de détection [Corominas2014] permettant initialement d'exclure les événements non considérés comme des glissements de terrain. En milieu montagneux, la géométrie d'acquisition des images SAR est plus complexe (repliement, zones d'ombre), ce qui rend l'utilisation de la technique InSAR difficile à implémenter et sujette à certaines limitations : décorrélation temporelle due à un changement de surface (végétation, neige), erreurs de MNT, artéfacts dus à l'atmosphère, etc. L'ensemble de ces éléments peut engendrer des erreurs de déroulement de phase interférométrique, et ainsi générer des zones de données manquantes par suppression des valeurs incertaines de la phase. Toujours en étude des glissements de terrain par InSAR, [Aslan2020] suppriment simplement les données acquises durant la saison hivernale, où la neige réduit la capacité de détection en haute altitude [Solari2018]. Dans cette même étude, les données affectées par le bruit atmosphérique corrélé à la topographie de surface sont également supprimées.

En mesure du déplacement des glaciers de montagne, la topographie en relief réduit la visibilité des vallées glaciaires, qui sont rarement visibles sur les deux trajectoires ascendantes et descendantes du satellite [Trouvé2007]. Lorsqu'une seule projection du déplacement dans la ligne de visée SAR est disponible, le calcul du champ de vitesse requiert la formulation d'une hypothèse forte sur le comportement du déplacement de surface, comme sa direction de mouvement [Joughin1998]. De plus, le signal est sensible aux conditions atmosphériques (pression, température, humidité) difficiles à modéliser en milieu montagneux. La perte de cohérence temporelle causée par les changements rapides de la glace en mouvement dans les zones d'ablation ainsi que l'évolu-

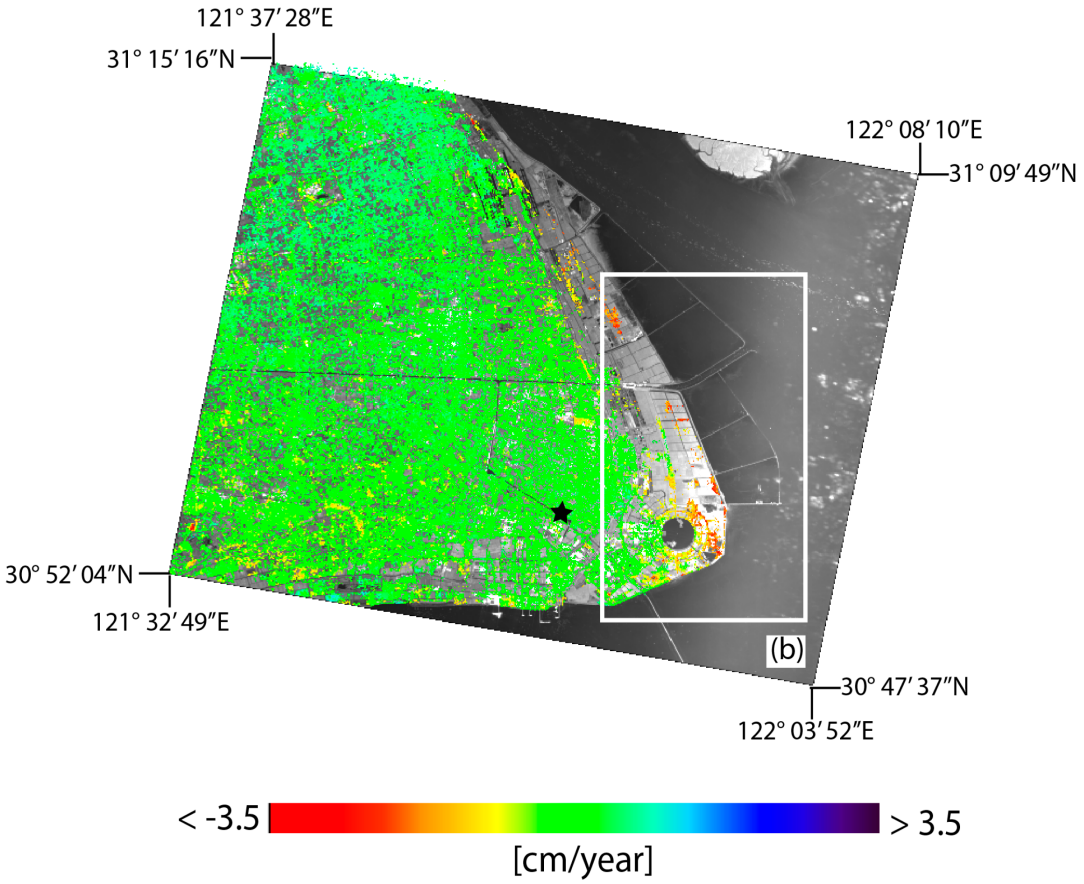


Figure 1.5 – Déplacement DInSAR moyen dans la ligne de visée calculé à partir d'images ENVISAT sur la période 2007-2010 en zone urbaine (Shanghai, Chine) [Pepe2016].

tion de la couverture neigeuse dans la zone d'accumulation [5] exige de traîter des images acquises à un faible intervalle de temps, ce qui limite le nombre de données utilisables [Dehecq2015], en particulier dans les massifs alpins tempérés où la taille moyenne des glaciers est plus petite. La figure 1.6 montre deux interférogrammes déroulés sur les glaciers du Gorner (Suisse) et de Miage (Italie). La plupart des données manquantes sont situées dans des zones à cohérence faible. On constate aussi des sauts de phase qui proviennent d'erreurs lors du déroulement de phase.

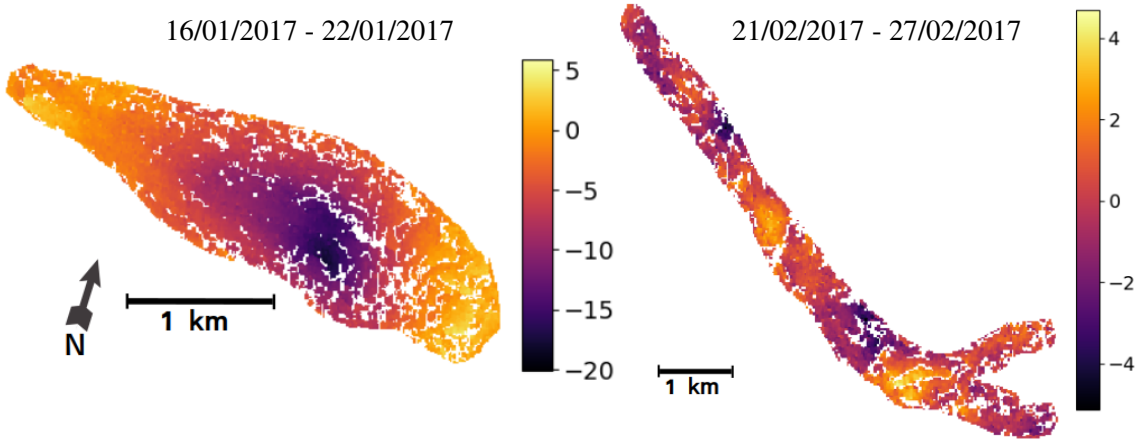


Figure 1.6 – Interférogrammes déroulés (centimètre dans la ligne de visée) calculés à partir de paires (6 jours) d'images SAR Sentinel-1 A/B sur les glaciers du Gorner (gauche) et de Miage (droite). La représentation est en géométrie radar.

La figure 1.7 est un autre exemple d'interférogramme déroulé sur une grande zone située au sud-ouest de la ville de Mexico, dont la topographie en relief provoque occasionnellement des erreurs de déroulement de phase.

Corrélation d'amplitude

La corrélation d'amplitude (*offset tracking*) est une technique de mesure du mouvement entre deux images utilisant la similarité de l'intensité ou l'amplitude des pixels. Cette technique a été utilisée pour estimer les déplacements de surface de forte magnitude issus de séismes, de l'activité volcanique, des glissements de terrain et depuis une vingtaine d'année pour l'estimation de la vitesse de surface des glaciers [Strozzi2002]. Le principe consiste à rechercher le maximum de la valeur de similarité entre les deux images grâce à un système de fenêtres glissantes. Parmi les autres techniques de corrélation ayant été développées depuis, on peut citer la méthode IPS [Serafino2006] qui repose sur l'exploitation du signal de rétrodiffusion de cibles isolées et brillantes, ou encore l'algorithme de [Erten2009], qui propose l'évaluation d'un critère de maximum de vraisemblance à partir de la fonction de probabilité du ratio des chatoiements (*speckle*) des deux images.

Calcul de déplacement par la méthode du maximum de similarité La méthode du maximum de similarité cherche à estimer un déplacement en comparant les points d'un couple d'images (image 1 et image 2) acquises à deux dates différentes. La similarité entre les images 1 et 2 est calculée à l'aide d'un critère de corrélation [Faugeras1993], appelé aussi fonction de similarité. Si l'on note $I_1(i,j)$ et $I_2(i,j)$ les valeurs d'intensités au pixel (i,j) dans l'image 1 et l'image 2 respectivement, la fonction de similarité a pour expression :

5. La zone d'ablation est la partie du glacier concernée par la perte de masse, alors que la zone d'accumulation est la partie où la neige se transforme en glace. Ces deux zones sont séparées par la ligne d'équilibre du glacier, qui sépare la partie du glacier où le bilan de masse est excédentaire (accumulation) et la partie du glacier où le bilan de masse est déficitaire (ablation).

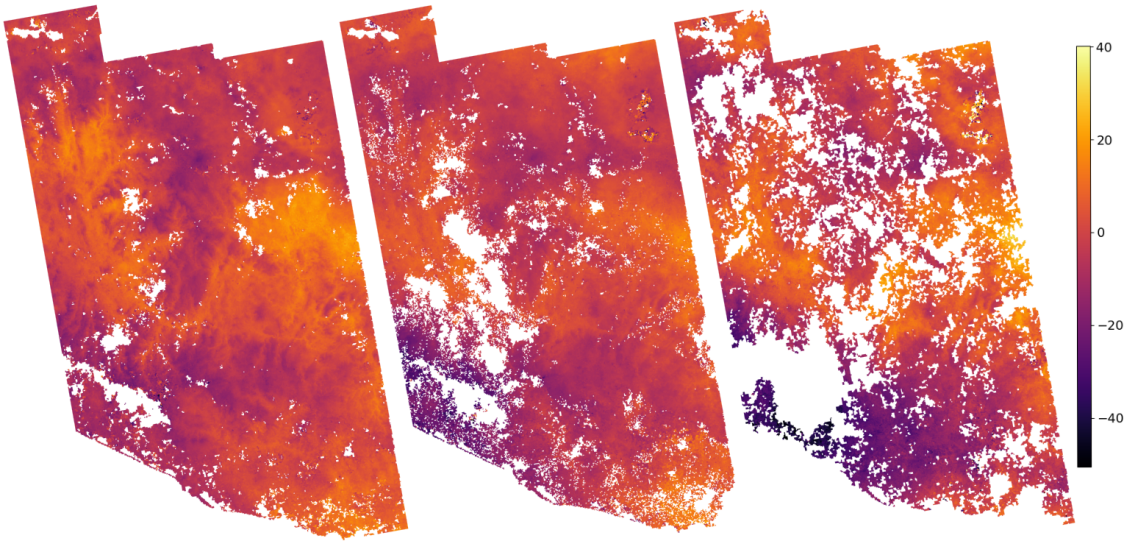


Figure 1.7 – Interférogrammes déroulés non corrigés (centimètres dans la ligne de visée) à trois dates différentes calculés à partir d'image Sentinel-1 A/B sur la région montagneuse de l'ouest et du sud-ouest de Mexico (L. Maubant, ISTerre, communication personnelle).

$$sim(p,q) = \frac{\sum_{i,j\in\Omega_1}\left(I_1(i,j)-\overline{I_1}\right)\times\left(I_2(i+p,j+q)-\overline{I_2}\right)}{\sqrt{\sum_{i,j\in\Omega_1}|I_1(i,j)-\overline{I_1}|^2}\times\sqrt{\sum_{i,j\in\Omega_1}|I_2(i+p,j+q)-\overline{I_2}|^2}} \tag{1.3}$$

où $\overline{I_1}$ et $\overline{I_2}$ désignent respectivement la moyenne des valeurs d'intensités au sein de fenêtres de corrélation Ω_1 et Ω_2 ayant la même taille dans les deux images. Ce critère de corrélation, appelé *Zero-mean Normalized Cross-Correlation* (ZNCC), possède des variantes selon le type de normalisation ou le type de corrélation (pour un récapitulatif, voir la thèse de R. Fallourd [Fallourd2012], et l'étude de [Vernier2011] qui fournit une version optimisée de la ZNCC). Le vecteur de déplacement $\overrightarrow{d}(i,j)$ au pixel (i,j) est obtenu en calculant la ZNCC entre la fenêtre Ω_1 centrée en (i,j) dans l'image I_1 et la fenêtre Ω_2 translatée de (p,q) dans l'image I_2. La recherche est réalisée sur une fenêtre de recherche Δ englobant les fenêtres de corrélation, dont la taille dépend de l'information *a priori* sur le déplacement. Le vecteur de déplacement au pixel (i,j) correspond alors au maximum de la similarité au sein de la fenêtre de recherche Δ :

$$\overrightarrow{d}(i,j) = (p_{\text{opt}}, q_{\text{opt}}) = \underset{(p,q)\in\Delta}{\arg\max}\, sim(p,q) \tag{1.4}$$

où $(p_{\text{opt}}, q_{\text{opt}})$ est le décalage (*offset*) qui maximise la fonction de similarité $sim(p,q)$.

Données manquantes Les fonctions de similarité peuvent être utilisées comme valeur de confiance pour sélectionner les pixels à forte similarité entre deux images. La valeur de la fonction de similarité dépend de plusieurs facteurs liés à la structure de la surface observée. Par exemple, cette valeur sera probablement haute dans une zone crevassée dont la signature apparaît sur les deux images étudiées, à condition que la fenêtre de corrélation soit assez grande pour capturer toute cette zone. À l'inverse, les zones homogènes, comme un sol dépourvu de points isolés à forte intensité ou les zones de couverture neigeuse, résultent en une faible similarité. De plus, si la vitesse de surface est très importante entre deux acquisitions, un point sur la première image peut se retrouver en dehors de la fenêtre de recherche dans la seconde image, ce qui engendrera une valeur faible de la fonction de similarité (c'est-à-dire du pic de corrélation).

Dans la littérature, il est commun que les pixels dont la valeur du maximum de similarité est inférieure à un seuil soient masqués afin de préserver les zones où la confiance en la mesure est

élevée [Nakamura2007,Fallourd2011]. Comme le montre la figure 1.8, les valeurs du maximum de la ZNCC inférieures à 0.2 ont été masquées, ce qui résulte en un champ de déplacement incomplet sur la majeure partie du glacier.

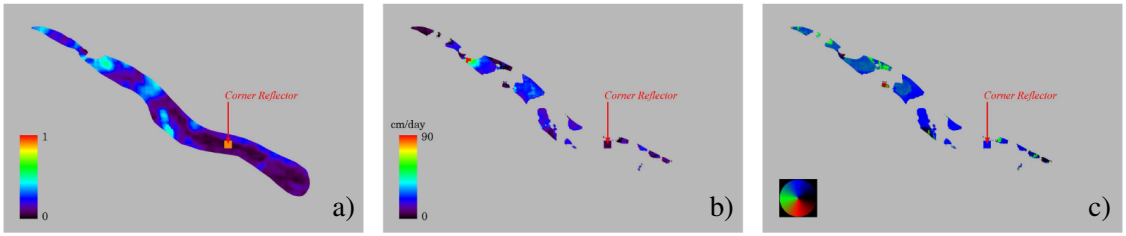


Figure 1.8 – Estimation du déplacement du glacier d'Argentière par corrélation d'une paire d'images TerraSAR-X entre le 29 septembre et le 10 octobre 2008. a) pics de corrélation ZNCC ; b) magnitude du déplacement ; c) estimation de l'orientation du déplacement. Figure tirée de [Fallourd2011] ©2011 IEEE.

Ce phénomène n'est pas réservé au glacier d'Argentière mais à tout type de déplacement, dont la taille, l'orientation par rapport à la ligne de visée, les caractéristiques de la surface observée (présence ou non de structure, zones de saturation) et la vitesse de déplacement (plus ou moins marquée) sont la source d'un phénomène de décorrélation entre deux acquisitions, ce qui ne permet pas toujours d'obtenir une similarité suffisante et par conséquence une valeur observée du déplacement.

1.2.2 L'imagerie optique

Bien avant qu'elle ne soit appliquée à l'imagerie SAR, l'estimation du mouvement par corrélation d'image a été appliquée sur des images optiques numériques [Anuta1970]. En télédétection, [Scambos1992] ont d'abord appliqué une intercorrélation sur des images Landsat pour mesurer les vitesses d'écoulement de glaciers en Antarctique. [Berthier2005] a ensuite appliqué cette technique aux glaciers alpins à partir de séries temporelles d'images moyenne résolution ASTER et haute résolution SPOT, dont la précision de ces dernières se rapproche de celle obtenue par InSAR.

Comme en imagerie SAR, le calcul du déplacement par corrélation d'images optiques est fortement dépendant de la qualité de la corrélation. Lorsque cette technique est automatisée et incorporée dans une chaîne de traitement, elle peut exiger la sélection d'un maximum de paires d'images disponibles. Comme l'écrit [Dehecq2015] dans sa thèse (p. 63),

> [...] une seule paire permet rarement d'avoir une couverture complète de la région imagée, en raison d'ombres, de nuages ou de la saturation dans certaines zones, qui vont induire des trous ou des mauvais appariements dans le résultat final. Mais plusieurs paires peuvent être complémentaires, en raison des différences de conditions de surface, d'éclairement, etc... ce qui permet d'augmenter la couverture spatiale du champs de vitesse.

L'incomplétude de données est donc un problème fréquent : une démarche commune consiste alors à proposer un traitement supplémentaire comme un moyennage temporel des cartes de déplacement, ce qui ne permet pas d'avoir un suivi de l'évolution temporelle du déplacement observé. À titre d'exemple, les figures 1.9 montrent l'estimation des vitesses de surface moyennes sur deux massifs alpins européens calculées à partir d'une archive d'images Landsat 7 entre 1999 et 2003. Dans le massif du Mont-Blanc, la plupart des valeurs manquantes sont situées sur les petits glaciers, ou dans des zones dont la surface est peu structurée, comme la moitié haute du glacier d'Argentière. Dans les alpes bernoises, c'est notamment toute la partie haute du glacier d'Aletsch, le plus grand glacier des Alpes, qui est concernée par l'incomplétude du déplacement.

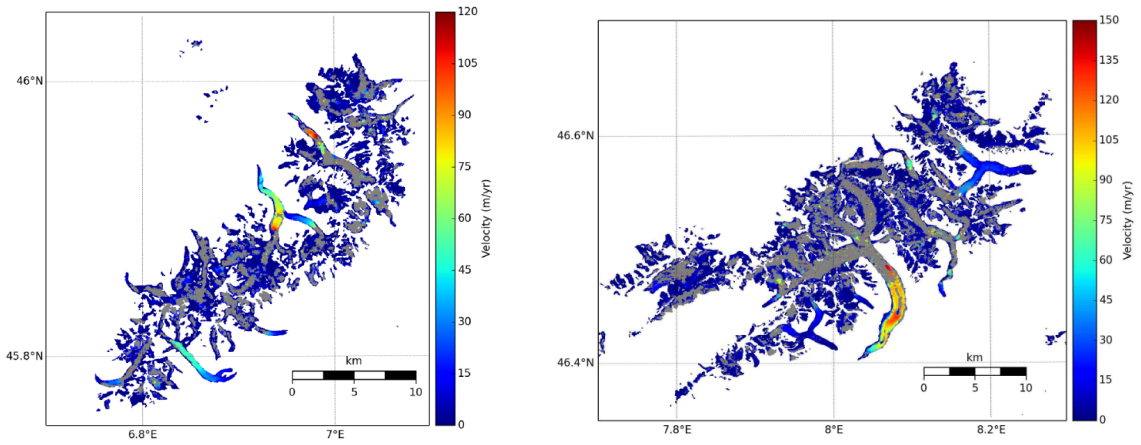


Figure 1.9 – Vitesses de surface (m/an) des glaciers du Mont-Blanc (gauche) et des alpes bernoises (droite) obtenues pour la période 1999-2003 à partir d'images Landsat 7 [Dehecq2015].

La figure 1.10 montre les vitesses de surface du glacier Fox, en Nouvelle-Zélande, estimées par calcul de la corrélation de paires d'images Sentinel-2 à partir d'une version modifiée de l'algorithme `ampcor` du code ROI_PAC (*Repeated Orbit Interferometry Package*) [Rosen2004]. Les déplacements manquants sont issus de valeurs de corrélation trop basses dans des zones à haute saturation, comme en amont du glacier, ou des zones à très fort déplacement comme dans la zone centrale du glacier, réputée rapide [Kääb2016].

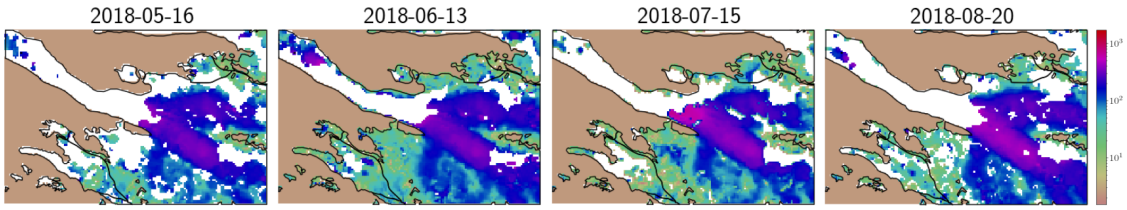


Figure 1.10 – Vitesse de surface (m/an) sur le glacier Fox en Nouvelle-Zélande. Le calcul des vitesses est issu de la chaîne de traitement développée par [Millan2019]. Les contours sont ceux du Randolph Glacier Inventory (RGI) [Consortium2017].

Remarque Notons que les données manquantes semblent, dans la plupart des exemples mentionnés jusqu'ici, posséder une certaine dépendance spatiale et temporelle, voire spatio-temporelle. Bien que nous n'ayons pas encore fait état des différentes méthodes existantes pour l'interpolation de données manquantes, une connaissance de cette dépendance peut s'avérer utile avant de développer une telle méthode, qui pourra exploiter la corrélation spatiale, temporelle ou spatio-temporelle du champ de déplacement pour reconstruire les données.

1.2.3 Le Global Network Satellite System

Le Global Network Satellite System (GNSS) est un terme générique standard pour désigner tous les systèmes de navigation satellite, comme GALILEO, GPS, GLONASS, BeiDou, etc. Le principe du GNSS est le suivant : la distance entre un émetteur et un récepteur d'onde radio est déduite de la mesure des temps de transmission et de réception, en faisant l'hypothèse que l'onde a une vitesse proche de la vitesse de la lumière. Dans notre cas, les émetteurs sont plusieurs satellites situés à 20 200 kilomètres du sol terrestre dont la position exacte est connue [Duquenne2005] et qui transmettent régulièrement un signal en direction de la Terre, et le récepteur est un instrument capable de calculer sa propre position dans un repère géodésique en trois dimensions. Pour détecter

un déplacement d'une surface dans le temps et en déduire une vitesse, on peut alors mesurer l'onde reçue à plusieurs intervalles de temps.

De nombreuses études font état de mesures GNSS incomplètes, dont [Dong2006,Kositsky2010, Xu2016,Gualandi2016,Liu2018a,Khazraei2019,Maubant2020]. Les causes de valeurs manquantes sont multiples : maintenance des capteurs nécessitant une mise hors service, dysfonctionnement dû aux perturbations du milieu naturel (éboulement, zone instable), attenuation ou difficulté de traitement du signal reçu ou suppression de valeurs atypiques. La plupart des données manquantes en mesure GNSS ont une forme corrélée temporellement puisque l'analyse se concentre souvent sur des séries temporelles mesurées par des récepteurs. Certaines zones sont parfois suivies par un ensemble de capteurs, appelé réseau de récepteurs, ce qui est par exemple le cas du Piton de la Fournaise sur l'île de la Réunion. Un arrêt de tout ou partie du réseau de capteurs sur une période peut donc engendrer une incomplétude de données de forme spatio-temporellement distribuée (voir figure 1.2).

1.3 Méthodes pour l'interpolation de données manquantes

1.3.1 Méthodes en télédétection

Il existe en télédétection une quantité considérable de méthodes d'interpolation des données manquantes dont nous n'avons pas la prétention de dresser un portrait exhaustif. Nous voulons simplement faire état des grandes familles de méthodes existantes selon deux types de classification données dans la littérature.

La première, formulée dans le livre *Statistical Analysis with Missing Data* de [Little2002], identifie quatre grandes catégories de méthodes, lesquelles ne sont pas exclusives les unes par rapport aux autres :

1. *Les méthodes basées sur l'omission des données manquantes*. Ces procédures consistent simplement à retirer les données manquantes et à mener l'analyse sur les données observées seulement. On peut citer par exemple l'omission par liste (*listwise deletion*) ou l'omission par paire (*pairwise deletion*). La première consiste à supprimer toutes les variables si une seule observation est manquante. Par exemple, si un sujet dépend des trois variables X, Y, Z et si une observation manque sur Y, on supprime les observations des variables X et Z pour ce même sujet. Ce mode opératoire simple peut se révéler très pratique lorsque la quantité de données manquantes est faible par rapport au nombre de données observées. L'inconvénient repose essentiellement sur la perte d'information que produit l'omission, induisant une perte de précision sur la connaissance des paramètres statistiques régissant les données [Olinsky2003], surtout si ces derniers sont estimés à partir d'une sous-population.

2. *Les méthodes basées sur les poids*. Selon la probabilité d'apparition d'une observation au sein d'un jeu de données, un poids plus ou moins important peut être attribué à cette observation. De manière générale, le poids est inversement proportionnel à la probabilité d'apparition d'une observation.

3. *Les méthodes d'imputations des données manquantes*. [Little2002] définissent l'imputation comme toute procédure visant à remplacer les valeurs manquantes par des valeurs prédites ou observées. On peut citer l'imputation *hot deck* qui consiste à remplacer les valeurs manquantes par des valeurs existantes, l'imputation par la moyenne où les moyennes sont substituées aux données manquantes, ou encore l'imputation multiple [van Buuran2012], qui consiste à remplacer les données manquantes par des valeurs générées m fois à partir d'une distribution donnée, puis à moyenner l'ensemble des m imputations. L'avantage de

cette stratégie est qu'elle est applicable à tout type de données manquantes (MCAR, MAR, MNAR) et est relativement simple à implémenter. [Molnar2008] a cependant souligné qu'un certain type d'imputation, celle consistant à remplacer les données manquantes par la dernière valeur observée, peut introduire des biais non négligeables, voire mener à un résultat catastrophique. Enfin, lorsque les données sont imputées par des valeurs prédites, on parle alors d'interpolation. Beaucoup de méthodes d'interpolation existent dans la littérature, et c'est en partie dans cette catégorie que s'inscrit le travail de cette thèse et notamment les chapitres 2 et 3.

4. *Les méthodes basées sur un modèle statistique*. Une gamme importante de méthodes consiste à définir un modèle régissant les données observées puis à procéder à une inférence reposant sur la distribution *a posteriori* ou la vraisemblance, notamment en estimant les paramètres statistiques régissant ce modèle par des procédures comme le maximum de vraisemblance. Cette approche a plusieurs avantages dont la flexibilité, la possibilité d'évaluer les hypothèses de départ sur le modèle et enfin la garantie de fournir des estimations des paramètres statistiques prenant en compte le type d'incomplétude des données. Ce type de méthode constitue le second socle dans lequel s'inscrit cette thèse.

Remarque À la différence des méthodes d'imputations, les méthodes basées sur un modèle n'ont pas directement pour but de prédire les données manquantes, mais plutôt d'estimer les paramètres statistiques sous contrainte d'un modèle défini. La plupart des méthodes que nous allons présenter ci-après et au cours de ce manuscrit sont soit des méthodes d'imputations des données manquantes, soit des méthodes basées sur un modèle statistique.

La seconde classification, formulée en partie dans l'état de l'art de [Shen2015], est spécifiquement conçue pour catégoriser les méthodes d'interpolation des données manquantes en télédétection. On pourra également se référer à l'étude de [Lepot2017], qui dresse un panorama instructif des méthodes d'interpolation pour l'analyse de séries temporelles. Cette classification repose sur le type d'information utilisée lors de l'interpolation, selon qu'elle est basée sur la corrélation spatiale, temporelle, spectrale ou sur une combinaison d'un ou plusieurs types de corrélation. Nous faisons ici état des approches spatiales, temporelles et spatio-temporelles, avec quelques exemples de méthodes.

1. *Approches spatiales.* La prise en compte de la corrélation spatiale est classique lorsque l'on traîte de champs géophysiques spatiaux. Le krigeage est une méthode très populaire en géostatistique et possède de nombreuses variantes. Leurs principes reposent sur un dénominateur commun : une valeur interpolée est estimée par une combinaison linéaire pondérée des valeurs proches, où les poids et le nombre de valeurs utilisées sont dépendants de la corrélation ou de la covariance des données [Goovaerts1997]. L'avantage de cette famille de méthode repose notamment sur la possibilité de fournir une mesure quantitative de l'incertitude associée à la prédiction des données manquantes, alors que l'inconvénient est principalement le temps de calcul que requiert l'inversion de la covariance spatiale, parfois de grande dimension. [Cressie2008] a cependant proposé un krigeage "à dimension fixe", où la covariance spatiale des données possède une flexibilité qui dépend du type de dépendance spatiale des données, réduisant ainsi la dimension à partir de laquelle est estimé la covariance. D'autres méthodes utilisent la décomposition du signal en fonction empiriques orthogonales (EOFs) spatiales [Beckers2003] qui permettent une prise en compte de la corrélation spatiale du champ étudié ;
2. *Approches temporelles.* Parmi les méthodes les plus classiques, citons l'interpolation aux plus proches voisins (NNI pour *Nearest Neighbor Interpolation*), la pondération inverse à la distance (PID), l'interpolation basée sur un polynôme (interpolation cubique, interpolation par les splines). Ces méthodes ont initialement été développées pour traîter des séries temporelles univariées, mais leur extension aux données multivariées (ensemble de pixels

d'un champ) existe. La méthode CACAO [Verger2013] utilise le motif saisonnier de pixels sur des images d'indice de surface foliaire (LAI, *Leaf Area Index*), ainsi qu'un modèle climatologique [6] comme informations temporelles pour prédire les données manquantes. Le programme TIMESAT [Jönsson2004] repose sur l'ajustement des données par des filtres adaptatifs de Savitzky-Golay et des fonctions gaussiennes asymétriques. Cette méthode nécessite une connaissance préliminaire du phénomène saisonnier et un pré-traitement des valeurs atypiques. Les EOFs temporelles sont également utilisées pour extraire des tendances temporelles à partir de séries incomplètes, technique que l'on peut classer dans l'analyse spectrale singulière (SSA) [Kondrashov2006] et sa version multivariée (M-SSA) ;

3. *Approches spatio-temporelles.* Ces méthodes combinent l'approche spatiale et temporelle. Citons par exemple la méthode *gapfill* [Gerber2018], dont la prédiction des données manquantes repose sur une régression quantile, le krigeage spatio-temporel [Zeng2013b], qui utilise un modèle de variogramme décrivant la corrélation spatio-temporelle des données, la régression spatio-temporelle [De Oliveira2014], qui consiste à prédire les valeurs manquantes au sein d'une fenêtre spatio-temporelle par un modèle de régression linéaire, ou encore la 2D-SSA étendue aux séries temporelles d'images [von Buttlar2014]. Plus récemment, une méthode d'interpolation utilisant des réseaux de neurones profonds a été appliquée à des images MODIS incomplètes [Zhang2018], ce qui nécessite néanmoins un grand volume de données d'entraînement. Ces méthodes, parfois très différentes, ont en commun d'utiliser la corrélation spatiale et temporelle des données observées afin de prédire les données manquantes.

Le diagramme en figure 1.11 offre un aperçu non-exhaustif des méthodes récentes selon cette classification, ainsi que le positionnement de cette thèse.

Remarque D'autres catégories pourraient également être ajoutées, celles des *approches spectrales* ainsi que celles combinant l'approche spectrale et les approches mentionnées ci-dessus. L'approche spectrale fait usage de la diversité spectrale et de la redondance d'information présente dans les images multi-spectrales et hyper-spectrales pour reconstruire les données manquantes sur une bande spécifique.

1.3.2 Méthodes en mesure de déplacement

Dans la continuité de ce qui vient d'être présenté, nous voulons ici passer en revue les quelques moyens mis en oeuvre dans la littérature pour traîter les données manquantes en mesure de déplacement, ce qui est l'objet principal de cette thèse.

En InSAR

Nous avons vu qu'en mesure de déplacement InSAR, la perte de cohérence entre deux images acquises à deux temps différents est génératrice de données incomplètes, transférant ainsi le problème de la cohérence faible à un problème de données manquantes. Les études que nous avons citées en section 1.2.1 omettent les valeurs manquantes, ce qui peut poser problème si l'on veut connaître le comportement (spatial et temporel) local du déplacement ou si l'on veut construire un modèle statistique à partir des données incomplètes. Afin de gérer les valeurs manquantes, [Pepe2016] utilisent un modèle de déformation externe (description analytique de la déformation au cours du temps), ce qui requiert des hypothèses supplémentaires sur le comportement physique de la déformation dans les zones de données manquantes. [Chen2017] utilisent la fusion de données et des observations multi-capteurs pour faire face à l'incomplétude temporelle des données InSAR.

6. Ensemble de mesures décrivant les variations de paramètres climatiques sur une longue période.

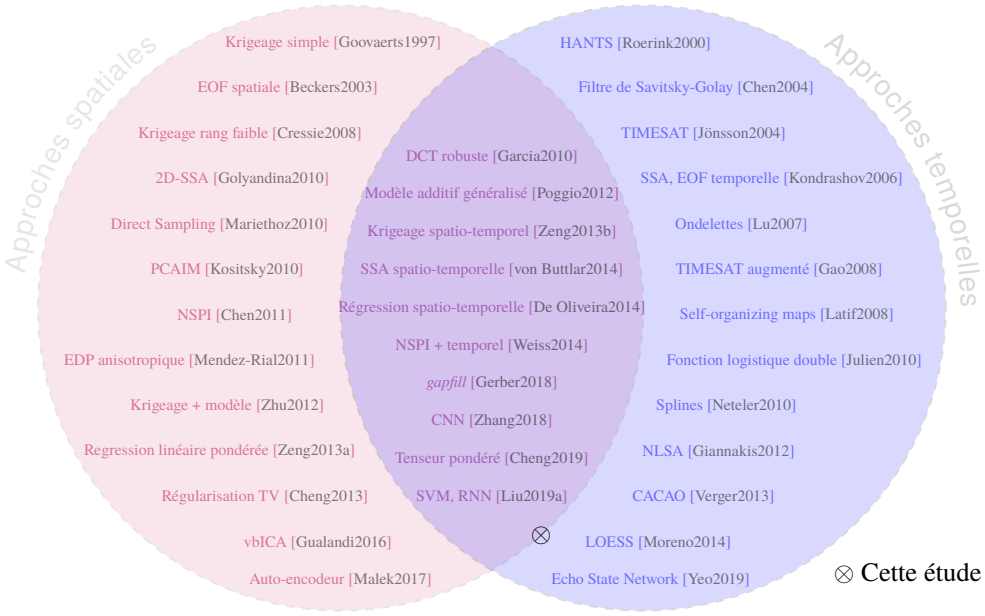


Figure 1.11 – Aperçu de quelques méthodes récentes classées par approche spatiale, temporelle ou spatio-temporelle, et positionnement de notre étude.

Cette approche différente demande un traitement qui dépend de plusieurs types de données dont l'incertitude varie.

Lorsque des méthodes d'interpolation sont mises en oeuvre, celles-ci utilisent assez classiquement l'information spatiale afin d'interpoler les valeurs manquantes. On pourra citer les méthodes comme l'analyse par régression, la NNI, la pondération inverse ou angulaire à la distance, l'interpolation par les splines et/ou bicubique et le krigeage [Gudmundsson2002, Jolivet2011, Wu2013, Chang2018], ou un simple moyennage spatial [Maubant2020]. Ces méthodes font parties des approches spatiales citées précédemment et n'utilisent donc pas l'information temporelle, ce qui peut être problématique lorsque le processus physique étudié évolue dans le temps [Choudhury2015], ce qui est le cas du déplacement terrestre.

Un état de l'art minutieux sur l'interpolation de champs de déplacement InSAR, comme les interférogrammes, ainsi que des champs de déplacement issus de la corrélation d'amplitude, révèle qu'aucune méthode n'y est spécifiquement dédiée, et ce alors que la complétude des champs de déplacement se révèle être un facteur clef pour mieux comprendre les phénomènes observés, surtout lorsqu'il s'agit de cibles à fort taux de décorrélation (glaciers, volcan couvert de végétation, etc.). De plus, il semble nécessaire de considérer la spécificité des données de déplacement InSAR en terme de :

1. complexité du champ de déplacement, dont le comportement dépend à la fois du support de déplacement : glace en mouvement (glaciers alpins, inlandsis), roche ou sédiment (glacier rocheux, plaque terrestre, lave), milieu urbain, et de la nature du déplacement : linéaire, oscillatoire, périodique, non-linéaire (voir [Mora2003]), etc. ;
2. corrélation du bruit à différentes échelles de temps et/ou d'espace, qui peut prendre des formes diverses, telles que des perturbations atmosphériques et/ou des erreurs lors du déroulement de la phase interférométrique.

En effet, les perturbations atmosphériques constituent une source récurrente d'artéfacts affectant la précision de la mesure InSAR [Bürgmann2000, Hanssen2001], bien plus que les artéfacts dus aux erreurs résiduelles orbitales [Fattahi2014]. L'étude de [Doin2009] a par ailleurs montré que la composante troposphérique de la phase peut biaiser la mesure du champ de déplacement InSAR. La contribution atmosphérique au déplacement de surface est souvent modélisée par un

bruit aléatoire corrélé, caractérisé par une fonction d'autocorrélaton ou une fonction de covariance [Tarantola1987, Fukushima2005] définie par :

$$C(r) = \sigma^2 \exp(-r/a) \tag{1.5}$$

où r est la distance entre deux points du champ de déplacement, σ^2 est la variance du bruit et a est la longueur de corrélation.

Positionnement 1. Le développement d'une méthode d'interpolation de champs de déplacement InSAR, pour être pleinement effective, devrait prendre en compte les spécificités des données de déplacement InSAR citées ci-dessus tout en fournissant la possibilité d'utiliser la corrélation temporelle des champs de déplacement. Idéalement, une telle méthode serait indépendante d'un modèle externe afin de pouvoir être appliquée à plusieurs types de champs de déplacement SAR. Ceci est notamment l'objet du chapitre 2 de cette thèse.

En corrélation d'amplitude d'images SAR et optique

En corrélation d'amplitude d'images SAR, une étude a récemment utilisé un réseau de neurones pour interpoler un seul champ de déplacement issu de la corrélation du chatoiement, fournissant ainsi de meilleurs résultats qu'un krigeage [Zhang2019]. Cette technique utilise ainsi strictement l'information spatiale. Une recherche minutieuse n'a pas permis de trouver d'autres techniques existantes. Concernant la corrélation d'images optiques, aucune méthode dédiée à la reconstruction de données manquantes de champs de déplacement n'a, à notre connaissance, été développée jusqu'ici. Dans les deux cas (corrélation d'images SAR et d'images optiques), la logique veut que l'on supprime les valeurs de déplacement associées à une corrélation faible, comme c'est le cas des travaux mentionnés en sections 1.2.1 et 1.2.2. En milieu montagneux, cela peut engendrer des zones de données manquantes particulièrement étendues spatialement et temporellement. Les études mentionnées se contentent d'omettre simplement les valeurs manquantes, ou d'appliquer un moyennage pour calculer des vitesses moyennes sur des longues périodes [Berthier2005, Fallourd2012, Dehecq2015, Millan2019], ce qui peut entraver le suivi continu spatio-temporel de l'évolution des vitesses de surface.

Positionnement 2. La proportion de données manquantes pouvant être assez importante dans les champs de déplacement issus de la corrélation d'images en milieu montagneux, la prise en considération de la corrélation spatiale en plus de la corrélation temporelle peut s'avérer utile pour reconstruire les données manquantes. Le comportement du déplacement en terme de complexité du champ de déplacement (voir sous-section précedénte, item 1.) devrait aussi être pris en compte dans la reconstruction. Comme énoncé lors du positionnement 1, la reconstruction des données manquantes ne devrait pas reposer sur un modèle externe aux données. Ceci est notamment l'objet des chapitres 2 et 3 de cette thèse.

Mesure GNSS

À la différence de l'imagerie optique et SAR, la prédiction de données manquantes dans des séries temporelles de mesure GNSS a fait l'objet de beaucoup d'études. Dans le travail de [Dong2006], les auteurs utilisent une approche basée sur l'analyse en composante principale (ACP) et sur l'expansion du signal en composantes spatiales et temporelles, que l'on peut appeler transformée de Karhunen-Loève, pour reconstruire les données. [Gualandi2016] utilisent également des méthodes similaires afin de prédire les données manquantes dans un problème de séparation de sources. [Xu2016] considère en plus la problématique, toujours d'actualité, d'un bruit dominant sur le signal de déplacement d'origine géophysique, ce qui est assez commun aux données GPS

[Ray2008]. [Khazraei2019] a plus récemment utilisé une technique similaire basée sur la génération de simulations de type Monte Carlo et a montré qu'il était possible d'extraire des variations inter-annuelles ou intra-annuelles à partir de mesures incomplètes. L'amplitude du bruit blanc et la présence d'autres signaux périodiques dans la mesure de déplacement constituent toutefois une limite au succès de l'extraction du signal d'intérêt. Enfin, deux autres méthodes ont également été appliquées : l'interpolation log-normale [Kositsky2010] ainsi que le filtrage de "Kriged" Kalman [Liu2018a] permettant d'interpoler des données manquantes temporellement continues en prenant en compte la corrélation spatiale en plus de la corrélation temporelle des données.

1.4 Approches prédictives

1.4.1 Les fonctions empiriques orthogonales

Karl Pearson, eugéniste anglais de la fin du XIX$^{\text{ème}}$ siècle, fût aussi un promoteur du socialisme allemand, notamment en se proposant à Karl Marx comme traducteur des volumes existants du *Capital* [7]. Si cela peut paraître surprenant, c'est parce que Pearson est plutôt connu aujourd'hui pour être l'un des fondateurs majeurs de la statistique moderne. Parmi ses nombreuses contributions, on peut citer le coefficient de corrélation de Pearson, la méthode des moments, l'analyse en composantes principales (PCA), reprise et développée plus tard par [Hotelling1933, Hotelling1935], dont les fonctions empiriques orthogonales (EOFs) découlent directement. Les EOFs ont d'abord été utilisées en sciences de l'atmosphère par [Obukhov1947], [Fukuoka1951], [Lorenz1956] et [Kutzbach1967], comme méthode d'exploration des données atmosphériques puis météorologiques en temps et en espace, en toute indépendance d'un modèle. Depuis, les EOFs ont été utilisées dans des champs aussi divers que la réduction de dimension de données [Hannachi2001], l'extraction de structures dynamiques et le filtrage [Broomhead1986, Kimoto1991, Plaut1994]. L'analyse en EOF repose sur la recherche d'un ensemble de motifs spatiaux orthogonaux ainsi que de composantes principales (PC) décorrélées, lesquelles peuvent être interprétées indépendamment les unes des autres afin de dégager des variabilités générales et/ou particulières. Pour une description détaillée des techniques issues de l'ACP et des EOFs, le lecteur pourra se référer au livre de [Preisendorfer1988]. L'article de [Hannachi2007] dresse un état de l'art sur les EOFs en sciences de l'atmosphère et du climat, lequel fournit en partie une base à la présente section.

La simplicité d'implémentation, la possibilité d'interprétation physique des EOFs et l'absence d'information *a priori* sur le comportement des données sont quelques uns des avantages pouvant servir à justifier la mise en oeuvre de l'analyse en EOF. Initialement, celle-ci repose sur la décomposition d'un champ spatio-temporel continu $X(t,s)$, où t est le temps et s la position spatiale :

$$X(t,s) = \sum_{i=1}^{R} a_i(t) u_i(s) \tag{1.6}$$

où $u_i(s)$ sont des fonctions de l'espace et $a_i(t)$ des fonctions du temps. On dira ici que $X(t,s)$ est décomposé en R *modes de variabilité* ou *modes EOF*. L'expression (1.6) découle directement du théorème Kosambi-Karhunen-Loève (KKL) [Kosambi1943, Loève1945, Karhunen1946], qui permet une représentation de X par la somme du produit de variables aléatoires non corrélées et de fonctions continues réelles orthogonales sur un intervalle défini. Ces fonctions, désignées plus tard comme fonctions empiriques orthogonales par les atmosphéristes, sont obtenues en diagonalisant

7. https://www.britannica.com/biography/Karl-Pearson

la matrice de covariance ou de corrélation du champ continu X représenté mathématiquement par le théorème KKL. La structure symétrique de ces matrices rend alors la décomposition en valeurs singulières (SVD) [Golub1996] particulièrement adaptée pour en effectuer la diagonalisation. Les EOFs sont dites *temporelles* ou *spatiales* selon sur quelle dimension (temps ou espace) la matrice de covariance est calculée, laquelle est alors appelée matrice de covariance temporelle ou spatiale. Il est d'usage de regrouper les modes EOFs en trois catégories désignant les variabilités qui composent le signal : tendances, formes oscillatoires et bruit [Ghil2002] (voir figure 1.12). On comprend aisément qu'une telle représentation est liée à la répartition en fréquences du signal, où les premiers modes représentent les fréquences les plus basses et les derniers modes les fréquences les plus hautes. La décomposition de Fourier n'est qu'un cas particulier de cette représentation, puisque les fonctions orthogonales y sont imposées (sinusoïdes à fréquences constantes), alors que l'on cherche ici les fonctions décrivant le mieux les données.

Une partie du travail, non la plus aisée, consiste ainsi à sélectionner un nombre réduit de modes pouvant "expliquer" la partie la plus significative du comportement des données en terme de variance, en éliminant ainsi la partie du signal identifiée comme étant du bruit parasite ou inutile.

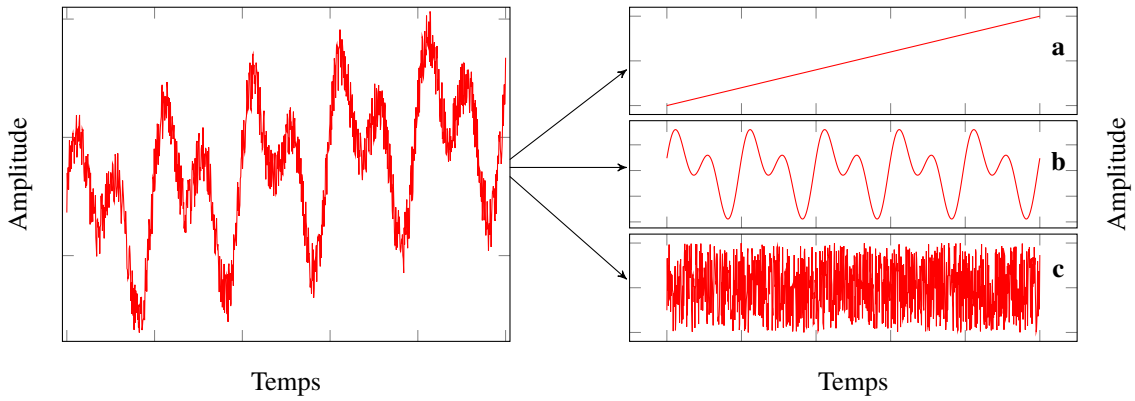


Figure 1.12 – Exemple d'un signal périodique (gauche) et des modes de variabilité qui le composent (droite) : (a) tendance, (b) oscillations et (c) bruit. Ici, la somme des modes de variabilité permet de reconstruire le signal. Spectralement, la tendance correspond à une fréquence basse, tandis que le bruit est (généralement) un événement haute fréquence.

Organisation des données

Dans la littérature de l'analyse en EOF, le champ spatio-temporel $X(t, s)$ est représentée par une grille de données [8], elle-même contenue dans une matrice $\mathbf{X}$ de taille $N \times P$ (ou $P \times N$), où N est le nombre d'observations temporelles et P le nombre de positions spatiales. Nous choisissons d'adopter la seconde représentation, où la matrice $\mathbf{X}$ est de taille $P \times N$. Les valeurs de $\mathbf{X}$ à la position s et au temps t sont notées $(x_{st})_{1 \leq s \leq P, 1 \leq t \leq N}$ et $\mathbf{X}$ admet la définition suivante :

$$\mathbf{X} = \left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\right) = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{P1} & x_{P2} & \cdots & x_{PN} \end{pmatrix} \tag{1.7}$$

Chaque colonne $\mathbf{x}_t = \left(x_{1t}, x_{2t}, \ldots, x_{Pt}\right)^T$ est une observation au temps $t = 1, \ldots, N$, appelée également *vecteur d'état* et chaque ligne est une série temporelle au point s. Une observation $\mathbf{x}_t$

8. Voir par exemple le tutoriel très complet de [Björnsson1997] sur le calcul des EOFs et leur interprétation.

peut être une image ou un champ géophysique, représentée au départ sous la forme d'une matrice 2-D de taille $P_x \times P_y$ puis arrangée en un vecteur colonne de longueur $P = P_x \times P_y$.

Calcul de la covariance

Comme nous l'avons évoqué, les EOFs sont obtenues en diagonalisant la matrice de covariance du champ X contenu dans la matrice $\mathbf{X}$. On définit ainsi la matrice de covariance empirique par :

$$\mathbf{\Sigma} = \frac{1}{P}\mathbf{X}^T\mathbf{X} \tag{1.8}$$

Cette matrice de taille $N \times N$ est appelée matrice de covariance empirique temporelle car elle est construite à partir du produit terme à terme de chaque vecteur d'état, dont la moyenne est supposée nulle. La valeur de $\mathbf{\Sigma}$ à la position (i, j), notée $(\Sigma)_{ij}$, est fournie par :

$$(\Sigma)_{ij} = \frac{1}{P}\sum_{k=1}^{P} x_{ki}x_{kj} \tag{1.9}$$

La matrice de covariance empirique spatiale, de taille $P \times P$, est calculée par $\mathbf{\Sigma} = \frac{1}{N}\mathbf{X}\mathbf{X}^T$. Dans les deux cas (matrice de covariance temporelle et spatiale), $\mathbf{\Sigma}$ est symétrique et définie positive, c'est-à-dire qu'il existe un vecteur $\mathbf{x} \in \mathbb{R}^N \setminus \mathbf{0}$ ou $\mathbb{R}^P \setminus \mathbf{0}$ tel que $\mathbf{x}^T\mathbf{\Sigma}\mathbf{x} > 0$. La diagonalisation de $\mathbf{X}$ est réalisée à l'aide de la décomposition en valeurs singulières (SVD), définie par la factorisation suivante :

$$\mathbf{X} = \mathbf{V}\mathbf{\Lambda}\mathbf{U}^T \tag{1.10}$$

où $\mathbf{V}$ est une matrice unitaire de taille $P \times P$ contenant les vecteurs de base orthonormés à gauche, $\mathbf{\Lambda}$ est une matrice de taille $P \times N$ dont la diagonale est composée de réels positifs et de zéros ailleurs, et $\mathbf{U}^T$ une matrice unitaire de taille $N \times N$ contenant les vecteurs de base orthonormés à droite. La matrice de covariance temporelle peut alors être exprimée en fonction du produit des SVD :

$$\mathbf{\Sigma} = \mathbf{X}^T\mathbf{X} = (\mathbf{V}\mathbf{\Lambda}\mathbf{U}^T)^T(\mathbf{V}\mathbf{\Lambda}\mathbf{U}^T) = \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^T \tag{1.11}$$

Cette dernière définition correspond à la décomposition en valeurs propres (EVD) de $\mathbf{\Sigma}$. On comprend aisément en regardant les deux dernières équations que les valeurs propres de $\mathbf{\Sigma}$ sont égales à la racine carré des valeurs singulières de $\mathbf{X}$. Les vecteurs propres orthogonaux de $\mathbf{U}$, qui sont les EOFs, sont ainsi aisément obtenus par cette procédure.

EOFs et données manquantes

En proposant une implémentation itérative du calcul des EOFs sur des imagettes radiométriques du Advanced Very-High-Resolution Radiometer (AVHRR) contenant des données manquantes (figure 1.13), l'étude de [Beckers2003] est la première à utiliser l'analyse en EOF comme méthode d'interpolation de données manquantes. Parmi les méthodes d'interpolation existantes en analyse de séries temporelles, les EOFs ont depuis été utilisées de nombreuses fois en sciences de l'atmosphère et en océanographie pour interpoler les données manquantes, souvent à des fins d'extraction des caractéristiques spatio-temporelles de signaux géophysiques [Beckers2006, Alvera-Azcarate2007, Taylor2013, Xu2016, Liu2018b]. Le principe de base de l'application des EOFs à l'interpolation est similaire à celle de l'analyse en EOF, sauf que le champ de départ $X(t, s)$ contient à présent des données manquantes. Il n'est donc pas directement possible de calculer la matrice de covariance à partir du champ incomplet ni d'obtenir les EOFs. Une étape d'initialisation des données manquantes est donc nécessaire avant la décomposition. Dans la procédure originale mise en oeuvre dans [Beckers2003], l'initialisation est réalisée par 0 sur un champ dont la moyenne spatio-temporelle a

été préalablement retirée (en sciences de l'atmosphère et du climat, on parle alors de l'*anomalie*). Une SVD est alors appliquée sur le champ initialisé, puis une partie seulement des EOFs est utilisée pour reconstruire le champ. Autrement dit, la SVD tronquée contenant une partie des EOFs permet d'obtenir de nouvelles valeurs là où les données sont manquantes. Une fois les données manquantes "remplies", cette procédure est répétée jusqu'à ce que l'erreur entre le champ initial et le champ reconstruit converge.

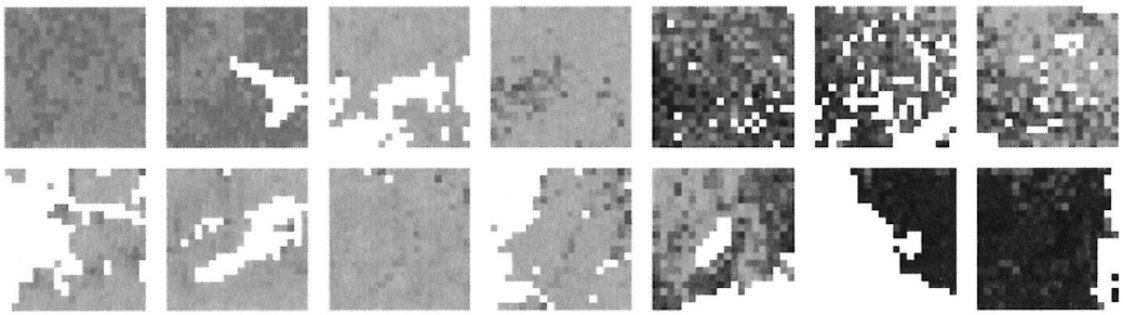


Figure 1.13 – Extrait de la séquence d'images AVHRR (400 pixels) sur la mer Adriatique contenant des nuages utilisées par [Beckers2003]. ©American Meterological Society. Figure utilisée avec permissions.

En mesure de déplacement InSAR, aucune étude ne s'est penchée sur la reconstruction de données manquantes à travers l'analyse en EOF. Dans leur travail récent, [Prébet2019] ont cependant utilisé les EOFs afin d'extraire un signal de déplacement d'une série temporelle d'interférogrammes calculés à partir d'images Sentinel-1 A/B sur le glacier du Gorner en Suisse. En sélectionnant un nombre restreint de modes, la procédure utilisée permet de séparer le signal de déplacement temporellement corrélé d'autres perturbations dans un contexte où le rapport signal sur bruit (SNR) est bas. La figure 1.14 montre notamment que les franges interférométriques peuvent être reconstruites dans un tel environnement chaotique. Afin d'optimiser la capacité de correction des perturbations, les auteurs préconisent d'appliquer la méthode (appelée méthode PM pour *Principal Modes*) deux fois : une fois sur les interférogrammes enroulés, puis une seconde fois après déroulement de la phase interférométrique. Cette stratégie s'avère payante puisqu'elle permet de réduire les artéfacts et d'obtenir ainsi un champ de déplacement plus cohérent. Toutefois, la méthode PM s'avère peu apte à reconstruire la partie du signal de déplacement correspondant à des fréquences plus hautes, ce qui est un des inconvénients des méthodes de type PCA face aux données aberrantes ou atypiques [Serneels2008]. De plus, l'application de la méthode sur les interférogrammes enroulés s'avère délicate, puisqu'il suffit que les valeurs reconstruites diffèrent légèrement des valeurs observées pour que la phase déroulée soit déviée de manière importante par rapport à la vraie valeur. Les auteurs mentionnent la possibilité d'appliquer la méthode PM en présence de données manquantes, sans toutefois s'y engager. En effet, la présence de bruit dans les zones à cohérence faible peut être interprétée comme une zone de données manquantes, transférant ainsi un problème de filtrage en un problème d'interpolation.

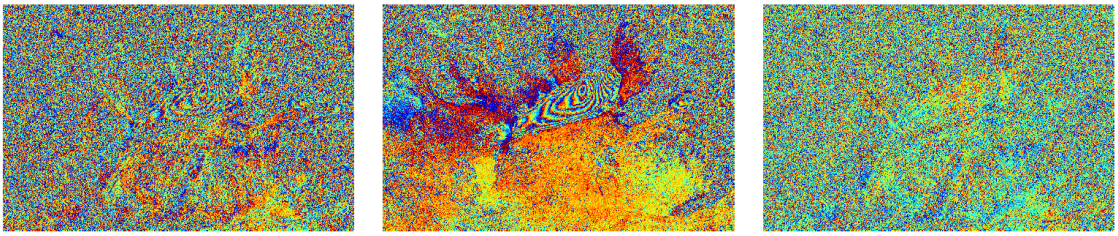


Figure 1.14 – Interférogramme initial (a), reconstruit (b) et résidus (reconstruit-initial) sur le glacier du Gorner. Figure tirée de [Prébet2019]. ©2019 IEEE.

Positionnement 3. Il peut sembler pertinent, dans la continuité des travaux de [Prébet2019], de proposer le développement d'une méthode d'interpolation reposant sur l'analyse en EOF et capable d'utiliser l'information temporelle des champs de déplacement SAR (InSAR, corrélation d'amplitude). Pour cela, il sera nécessaire de considérer la spécificité de ces données en terme de complexité du champ de déplacement et de corrélation du bruit à différentes échelles de temps et/ou d'espace, ce qui n'a pas été réalisé pour le moment. Nous considérons qu'une méthode d'interpolation efficace devra répondre à ces critères spécifiques. Pour ce qui est des déplacements InSAR, une telle méthode pourra être appliquée sur les interférogrammes déroulés afin de minimiser les potentiels biais d'interpolation.

1.4.2 Les fonctions empiriques orthogonales étendues

Les Fonctions Empiriques Orthogonales Étendues (EEOFs), initialement introduite par le travail de [Weare1982], sont une extension des EOFs car elles permettent de capter la corrélation temporelle des données en plus de la corrélation spatiale. Les EEOFs sont numériquement similaires à la version multivariée de la SSA, appelée M-SSA [Broomhead1986, Vautard1992], et les deux méthodes sont souvent prises pour synonyme [Von Storch2001]. La différence d'appellation provient des origines de chacune des méthodes : la première découle de l'analyse de systèmes dynamiques dans des séries temporelles univariées alors que la seconde prend racine dans l'analyse en composantes principales de champs météorologiques [Ghil2002].

En analyse en EEOFs, le vecteur d'état $\mathbf{x}_t = \left(x_{1t}, x_{2t}, \ldots, x_{Pt}\right)^T$ est étendu en alignant dans un seul et même vecteur sa version décalée, ou copiée, M fois dans le temps :

$$\mathbf{y}_t = \big(\underbrace{x_{1t}, \ldots, x_{Pt}}_{\mathbf{x}_t}, \underbrace{x_{1,t+1}, \ldots, x_{P,t+1}}_{\mathbf{x}_{t+1}}, \ldots, \ldots, \underbrace{x_{1,t+M-1}, \ldots, x_{P,t+M-1}}_{\mathbf{x}_{t+M-1}}\big) \tag{1.12}$$

où $t = 1, \ldots N - M + 1$ et M est appelé paramètre de délai ou dimension embarquée. Le choix du délai M, établi au préalable, est conditionné par deux considérations [Ghil2002] : la quantité d'information extraite versus le degré de confiance statistique en cette information. Le premier point exige une grande fenêtre M car plus M est grand, plus la série temporelle $\mathbf{y}_t$ contiendra de versions du vecteur d'état $\mathbf{x}_t$ décalées dans le temps, alors que le second incite à un maximum de répétitions du processus et donc à ce que le ratio N/M soit le plus grand possible. L'inclusion de portions de signal temporellement retardées dans l'expression (1.12) peut être interprétée comme une augmentation de données. La matrice de données admet à présent la forme :

$$\mathcal{X} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_M \\ \mathbf{x}_2 & \mathbf{x}_3 & \cdots & \mathbf{x}_{M+1} \\ \vdots & \vdots & & \vdots \\ \mathbf{x}_{N-M+1} & \mathbf{x}_{N-M+2} & \cdots & \mathbf{x}_N \end{pmatrix} \tag{1.13}$$

Cette matrice est similaire à la forme présentée en équation (1.7), à l'exception du fait que ses éléments sont à présent des vecteurs d'état $\mathbf{x}_t$ et non plus des scalaires x_{ij}. La matrice $\mathcal{X}$ est de taille $(N - M + 1) \times PM$, ce qui est bien plus grand que la matrice initiale $\mathbf{X}$ de taille $P \times N$. Similairement à la décomposition (1.8), la matrice de covariance est obtenue par :

$$\mathbf{C} = \frac{1}{N - M + 1} \mathcal{X}^T \mathcal{X} \tag{1.14}$$

Cette matrice de covariance *augmentée* possède une structure de Toeplitz symétrique, c'est-à-dire constante sur ses diagonales [Vautard1992]. À l'image des EOFs obtenues par décomposition

de $\mathbf{\Sigma}$, les EEOFs sont obtenues en décomposant $\mathbf{C}$. Les signaux périodiques au sein du champ spatio-temporel peuvent être identifiés par l'existence de paires de valeurs propres dégénérées [9] de la matrice de covariance. Cette technique se révèle être apte à identifier et isoler des structures périodiques et propagatives au sein de signaux à partir de données multi-dimentionnelles [Kimoto1991, Plaut1994, Groth2015].

Alors que ces travaux s'inscrivent dans un régime où la dimension temporelle des données N est grande par rapport à la dimension spatiale P, les travaux de [Golyandina2010, Golyandina2015] proposent une implémentation de l'analye spectrale singulière sur une seule image à deux dimensions, c'est-à-dire pour $N = 1$. Les pixels de l'image sont représentés dans une matrice de données spatiales, laquelle est augmentée spatialement à l'aide d'une fenêtre glissante de taille $P_x \times P_y$. Le résultat est une grande matrice, dont la structure particulière est appelée *Hankel-block-Hankel* (HbH). La matrice de covariance est ensuite estimée à partir du jeu de données augmenté. Une approximation en *rang faible* de la covariance est ensuite implémentée par une troncature de sa SVD, processus qui revient à sélectionner un nombre réduit de modes comme dans l'analyse en EOF classique.

EEOFs et données manquantes

L'application des EEOFs (ou M-SSA) à la reconstruction de données manquantes s'est tout naturellement étendue au cas de données géophysiques incomplètes, comme le démontrent les études de [Kondrashov2006, Alvera-Azcarate2007, Xu2016]. Dans la première de ces trois études, une procédure itérative de reconstruction est développée en s'appuyant sur l'algorithme de [Beckers2003]. À la différence de cette dernière étude qui fait usage des EOFs spatiaux, les auteurs font usage des EOFs temporels sur des données multivariées, utilisant donc la corrélation temporelle pour reconstruire les données manquantes. Le travail de [Alvera-Azcarate2007] est une extension directe de l'algorithme de [Beckers2003] aux données multivariées, où chaque variable est un phénomène différent (température de surface de la mer, concentration en chlorophylle, vitesse de vent) dont le vecteur d'état est augmenté. Comme le font remarquer les auteurs, l'utilisation d'une matrice augmentée temporellement pour le calcul des EEOFs par rapport à l'utilisation des EOFs classique présente l'avantage de pouvoir détecter des structures évolutives grâce à la présence d'informations passées et/ou futures, si ces dernières ne sont pas manquantes dans la matrice augmentée [Kim2000, Von Storch2001, Jollife2002]. De plus, la corrélation non-nulle entre les variables physiques utilisées peut aider au processus de reconstruction [Gomis2001]. Enfin, le travail de [Xu2016] porte sur la reconstruction de séries temporelles de déplacement mesurés par GPS, ce qui est en lien avec cette thèse. La sélection du nombre de modes et du délai est effectuée grâce à l'erreur moyenne quadratique entre les données initiales et les données reconstruites. Finalement, [von Buttlar2014] ont étendu les travaux de [Golyandina2010] en présence de données manquantes sur une série temporelle d'images 2-D de divers champs géophysiques (température de l'air, NDVI [10]), alors que l'étude initiale proposait une implémentation sur une seule image.

Positionnement 4. Alors que les champs de déplacement contiennent des données manquantes spatio-temporelles, avec parfois des formes prolongées et corrélées comme dans le cas de champs de déplacement issus d'un calcul de corrélation, il serait intéressant de prendre comme point de départ l'algorithme de [von Buttlar2014] pour augmenter chaque champ de déplacement en espace et ainsi utiliser la corrélation spatiale en plus de la corrélation temporelle pour reconstruire les données manquantes.

9. Des valeurs propres sont dites dégénérées si la différence de leur amplitude respective est inférieure ou équivalente à leur incertitude d'estimation. Ce concept sera étudié en profondeur lors du chapitre 3, notamment pour sélectionner un nombre de modes optimal pour reconstruire des champs de déplacement incomplets.

10. *Normalized difference vegetation index.*

1.4.3 Sélection du nombre de modes

Dans l'analyse en EOF et EEOF, le choix du nombre de modes est souvent un exercice délicat, surtout si les données contiennent des perturbations de type corrélées (bruit corrélé) car ces dernières sont difficiles à séparer du signal d'intérêt. Dans la littérature, ces perturbations sont souvent associées à un "spectre rouge" (*red spectra*), où les modes contaminés par un tel bruit contribuent de manière significative à la variance du signal, voire dominent les basses fréquences du spectre [Kondrashov2006].

Dans la littérature, il est commun que le choix du nombre de modes soit guidé par la mesure de la variance contenue dans les R premiers modes comparée à la variance totale du système [11] [Beckers2003, Hannachi2007], quantité exprimée par :

$$f_i = \frac{\sum_{i=1}^{R} \lambda_i}{\sum_{i=1}^{N} \lambda_i} \tag{1.15}$$

où λ_i sont les valeurs propres du système et N est la dimension temporelle. On pourra exprimer f_i en pourcentage et choisir par exemple les modes dont la variance explique 90% de la variance totale. Lorsque l'incertitude des données est connue, le choix du nombre de mode peut être déterminé de manière à ce que l'erreur entre les données reconstruites et les données initiales bruitées soit en moyenne de l'ordre de l'incertitude [Kositsky2010]. L'étude de [Thacker1996] a également proposé d'inclure l'incertitude de mesure au sein de l'analyse en EOF, notamment en estimant une matrice de covariance d'erreur.

Dans le cas de données fortement corrompues par du bruit corrélé, les modes voisins peuvent être contaminés entre eux, ce qui signifie que leurs valeurs propres correspondantes sont proches entre elles. Une règle empirique proposée par [North1982] permet d'estimer l'incertitude des valeurs propres, et ainsi de fournir une information précise sur les caractéristiques du spectre (ensemble des valeurs propres) [Overland1982], comme les discontinuités, la variation en terme de pente ou les plateaux de valeurs.

Positionnement 5. Les caractéristiques du spectre de valeurs propres citées ci-dessus pourraient être utilisées afin de sélectionner un nombre de modes adéquat pour reconstruire les données manquantes, ce qui implique une estimation de l'incertitude des valeurs propres et une analyse de l'autocorrélation du champ étudié. Ceci est notament l'objet du chapitre 3, où nous reviendrons plus longuement sur les notions d'incertitude et de contamination des valeurs propres.

De plus, lorsque la statistique du bruit (loi statistique, moyenne, covariance) présent dans les données est connue, un critère basé sur une méthode Monte Carlo peut-être appliqué [Overland1982, Björnsson1997]. L'inconvénient réside dans la connaissance d'information *a priori* sur le bruit, qui ne peut être approché que par une modélisation. Dans le travail de [Prébet2019], les auteurs utilisent la racine de l'erreur quadratique moyenne (RMSD ou RMSE) entre les données reconstruites (estimé) et les données bruitées (observation) : à chaque ajout consécutif d'un mode dans la reconstruction, l'erreur est calculée. Ainsi, une valeur de l'erreur est obtenue pour chaque ajout de mode : il suffit en théorie de sélectionner le nombre de modes qui correspond à l'erreur minimale parmi les erreurs calculées ([Beckers2003] avait déjà procédé ainsi). Cette étude montre dans des simulations que cette RMSE est similaire à l'erreur minimale entre l'estimé et les vraies données, ce qui indique la possibilité de s'affranchir d'une connaissance *a priori* lors du traitement de données réelles où la vérité est difficilement accessible.

11. À l'origine, le système désigne l'équation matricielle $\mathbf{A}\mathbf{u} = \lambda\mathbf{u}$ où $\mathbf{A}$ est une transformation linéaire représentée par une matrice carrée, $\mathbf{u} \in \mathbb{R}^N$ est un vecteur propre de $\mathbf{A}$ et λ est la valeur propre de $\mathbf{A}$ correspondante à $\mathbf{u}$.

La validation croisée

Le principe de la validation croisée, formulé initialement par [Wahba1980] puis utilisé en analyse objective par [Brankart1995] pour déterminer les paramètres statistiques optimaux d'un champ océanographique, peut être résumé ainsi : 1) on sélectionne une série de points sur le champ initial X, de manière aléatoire ou non ; 2) ces points sont copiés puis mis de côté ; 3) ces points sont retirés des données, créant ainsi des données manquantes supplémentaires ; 4) la procédure de reconstruction est lancée sur toutes les données ; 5) lors du résultat intermédiaire ou final, les points de validation croisé reconstruits sont comparés aux points mis de côté, souvent par le calcul d'une erreur. Un résumé visuel de ces étapes est fourni en figure 1.15.

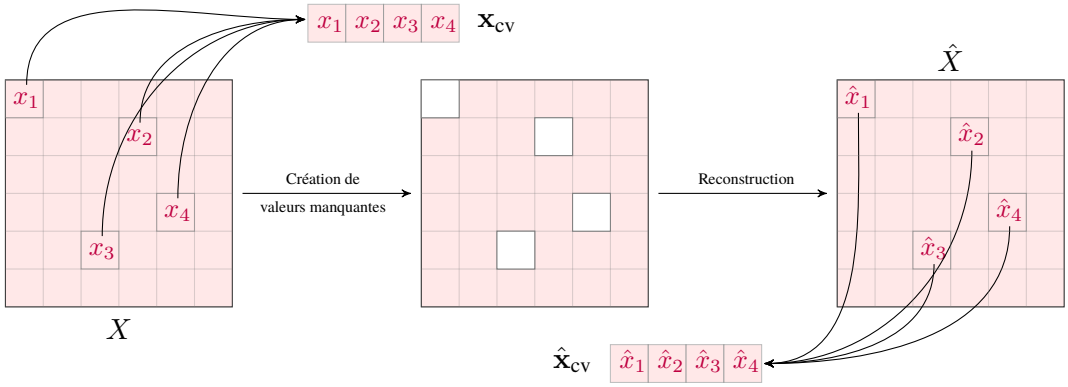


Figure 1.15 – Illustration du principe de validation croisée sur un champ sans données manquantes. $\hat{X}$ désigne le champ reconstruit et $\hat{\mathbf{x}}_{cv}$ les points de validation croisée reconstruits. Après reconstruction, l'erreur est calculée entre $\mathbf{x}_{cv}$ et $\hat{\mathbf{x}}_{cv}$.

La comparaison en 5) peut être réalisée à l'aide de la RMSE, que l'on nomme *cross*-RMSE car celle-ci est calculée seulement sur les points de validation croisée. La quantité de points de validation croisée doit être assez importante afin d'obtenir une estimation statistiquement robuste de l'erreur, sans que cela représente une proportion trop importante des données car cela implique de créer des données manquantes artificielles. On pourra typiquement choisir 1% des données pour mesurer l'erreur de validation croisée.

L'algorithme DINEOF (*Data Interpolating EOF*) [Alvera-Azcárate2005], directement issu des travaux de [Beckers2006], se base sur la cross-RMSE calculée à chaque ajout de modes pour pouvoir sélectionner le nombre optimal de modes. Plus particulièrement, le champ est reconstruit plusieurs fois avec un nombre de modes constant k jusqu'à ce que la cross-RMSE converge. Une fois la convergence atteinte, le champ est reconstruit avec $k+1$ modes, puis la cross-RMSE est de nouveau calculée, et ainsi de suite. Si la cross-RMSE permet de s'affranchir de la vérité terrain, cette erreur est soumise aux perturbations que contiennent les données observées [Ng1997], comme du bruit corrélé représenté par la partie "rouge" du spectre, c'est-à-dire dans les premiers modes. La présence de ce type de perturbation au sein même de l'erreur provoque un phénomène appelé sur-estimation, qui consiste à sélectionner plus de modes que nécessaire. La sous-estimation du nombre de modes est le phénomène inverse, mais dont la cause est due à la présence d'un signal de déplacement haute fréquence représenté dans les derniers modes.

Positionnement 6. Du fait de la difficulté que représente la sélection du nombre de modes, surtout en présence de bruit corrélé, il semble raisonnable de devoir mettre en place un certain nombre de critères robustes et paramétrables pour éviter de sur-estimer le nombre de modes, dans les cas de l'analyse en EOF et en EEOF précédemment décrits. La sélection du nombre optimal de modes pourra ainsi reposer sur un algorithme itératif convergeant vers le minimum de la cross-RMSE, comme proposé dans l'algorithme DINEOF.

1.4.4 Note sur l'initialisation des données manquantes

Avant reconstruction des données manquantes, ces dernières doivent être initialisées dans la matrice de données $\mathbf{X}$. L'étape d'initialisation, qui permet le calcul des EOFs en "complétant" la matrice de covariance, conditionne la vitesse de convergence vers un minimum local (ou global) [Srebro2003] de l'erreur entre le champ initial et le champ reconstruit. [Brankart1995] indiquent qu'à défaut d'avoir une information sur les valeurs probables aux points manquants, on peut initialiser ces dernières par la moyenne de l'observation au temps t. Lorsque la moyenne est retirée avant le traitement, on peut également initialiser les données manquantes par une valeur de 0 en supposant que l'estimation de la moyenne retirée est une estimation non biaisée.

La plupart des travaux que nous avons cités utilisant l'analyse en EOF et EEOF pour l'interpolation de données pratiquent l'initialisation du champ par 0, à condition donc que la moyenne (spatiale, temporelle ou spatio-temporelle) ait été retirée.

Positionnement 7. Alors que l'initialisation par 0 (ou par la moyenne si le champ a une moyenne non nulle) semble faire l'unanimité dans les études existantes, il n'existe pas, à notre connaissance, d'étude comparative empirique sur l'effet de l'initialisation des données manquantes sur la performance de reconstruction, comme l'erreur d'interpolation ou la vitesse de convergence. On s'intéressera donc à mener une telle comparaison, en initialisant par exemple les valeurs manquantes par un bruit se rapprochant du bruit présent dans les données.

1.5 Approches paramétriques

Une autre approche, complémentaire à l'approche prédictive, s'intéresse à l'estimation d'un ou plusieurs paramètres statistiques décrivant le comportement de données incomplètes. En effet, comment estimer de manière précise la moyenne, la variance ou la covariance d'un ensemble de données contenant des observations manquantes ? Ce problème est mitoyen aux sciences sociales et biomédicales [Rubin1976, Walczak2001a, Walczak2001b, Howell2007, Graham2012, Little2014], où l'analyse statistique de données de sondage ou de patients occupe une place notable. Cette approche regroupe des méthodes basées sur un modèle statistique (voir la classification en sous-section 1.3.1), c'est-à-dire qu'elle nécessite de définir un modèle dépendant des paramètres statistiques le plus à même de représenter les données.

1.5.1 L'algorithme Espérance-Maximisation

Une des méthodes les plus utilisées pour l'estimation paramétrique en présence de données manquantes est l'algorithme Espérance-Maximisation (EM) [Dempster1977]. Cette technique ne procède pas directement à l'interpolation des données manquantes, mais propose un calcul itératif des paramètres convergeant vers les paramètres optimaux au sens de l'estimation du maximum de vraisemblance (EMV) lorsque l'équation de vraisemblance ne possède pas de solution analytique. L'avantage de l'algorithme EM est sa simplicité conceptuelle facilitant son implémentation en un programme, ainsi que l'assurance, sous certaines conditions générales, de convergence de la vraisemblance des données vers une valeur stationnaire [Little2002]. L'idée intuitive générale est la suivante : 1) remplacer les valeurs manquantes par les valeurs estimées ; 2) estimer les paramètres ; 3) estimer à nouveau les valeurs manquantes en prenant en compte les nouveaux paramètres estimées ; 4) ré-estimer les paramètres et continuer ce schéma jusqu'à convergence. L'inconvénient principal est la lenteur de convergence lorsque la quantité de données incomplètes est conséquente.

L'estimation des paramètres statistiques peut aussi être utilisée à des fins d'imputations des valeurs manquantes, comme l'a montré l'étude de [Schneider2001] en science du climat. L'EM a aussi été utilisé jointement à l'ACP, et sa version probabiliste (ACPP) [Tipping1999], pour déterminer les composantes principales via l'EMV des paramètres statistiques. Les auteurs de cette étude examinent également l'application de la ACPP à des vecteurs de données contenant au moins une donnée manquante, sans étendre l'étude à plusieurs observations manquantes. La version "robuste" de la ACPP a été par la suite appliquée à la détection de valeurs aberrantes [Chen2009]. La robustesse est ici entendue en la capacité d'un modèle à représenter des données hétérogènes comportant des valeurs aberrantes ou atypiques, lesquelles sont souvent sous-représentées par des modèles plus classiques comme le modèle gaussien.

Positionnement 8. Les analyses en EOF et en EEOF mentionnées en sous-sections 1.4.1 et 1.4.2, embarquées dans un processus itératif d'estimation des données manquantes, peuvent être formalisées en utilisant le concept intuitif de l'EM : estimation des valeurs manquantes à l'étape E puis estimation du ou des paramètres statistiques à l'étape M. Les étapes E et M sont ainsi répétées jusqu'à ce que l'erreur de validation croisée (sous-section 1.4.3) entre les données initiales et les données reconstruites converge.

1.5.2 Initialiser ou ne pas initialiser

Nous avons vu que l'analyse en EOF dans le cas de données manquantes nécessite une initialisation des valeurs manquantes, car l'algorithme repose sur la décompostion du champ spatio-temporel ou de sa matrice de covariance en fonctions spatiales et temporelles continues. D'un point de vue programmatique, ne pas initialiser les valeurs manquantes provoque des erreurs lors de la formation de la covariance et lors de l'application de la SVD. L'approche paramétrique, notamment par l'implémentation de l'EM, ne requiert pas une telle initialisation. La procédure d'application de l'EM est relativement flexible à ce sujet, car elle s'adapte au type de données manquantes (aléatoires, corrélées). L'idée est de "faire avec" les valeurs manquantes, en déduisant leur probabilité en fonction des valeurs observées et de l'estimation des paramètres. En d'autres mots, il s'agit d'inférer leur espérance conditionnelle en sachant les valeurs des données observées et l'estimation des paramètres. Nous reviendrons plus longuement sur cet algorithme, ainsi que sur l'EMV, lors du chapitre 4.

1.5.3 Estimation de la matrice de covariance

En mesure de déplacement, la connaissance des paramètres statistiques peut s'avérer utile pour construire des modèles de déplacement dans un problème d'inversion des données acquises par télédétection, comme les données InSAR ou GNSS. Lorsque des données sont manquantes, ce problème peut être rendu particulièrement difficile. Appuyons-nous sur un exemple d'inversion de données. L'inversion d'un champ de déplacement pour déterminer un modèle optimal de déformation requiert une minimisation de la différence entre les observations et le modèle, dont l'expression est donnée par la métrique suivante [Tarantola2005] :

$$\mathcal{X}^2 = \big(\mathbf{d}_{\text{obs}} - \mathbf{g}(\mathbf{m})\big)^T \mathbf{C}_\mathbf{d} \big(\mathbf{d}_{\text{obs}} - \mathbf{g}(\mathbf{m})\big) \tag{1.16}$$

où $\mathbf{d}_{\text{obs}}$ est le vecteur des déplacements observés, $\mathbf{g}(\mathbf{m})$ est le vecteur des déplacements modélisés avec $\mathbf{m}$ l'ensemble des paramètres physiques du modèles et $\mathbf{g}$ est l'opérateur physique du modèle (qui représente le processus ou le mécanisme sous-jacent). $\mathbf{C}_\mathbf{d}$ est la covariance des données observées par télédétection (InSAR, GNSS). Lorsque les données sont complètes, $\mathbf{C}_\mathbf{d}$ peut être estimée à l'aide d'un estimateur selon l'hypothèse du modèle de distribution (gaussien, gaussien-composé, non-gaussien, etc.), comme la *Sample Covariance Matrix* (SCM) ou un M-estimateur.

Lorsque les données sont incomplètes, il n'est plus possible de calculer directement la covariance ou tout autre paramètre statistique. Nous pouvons identifier trois choix qui s'offrent alors à nous : 1) omettre les données manquantes et procéder à l'estimation des paramètres, ce qui, en plus de poser une difficulté d'implémentation, risque de fournir une estimation biaisée des paramètres ; 2) rendre les données complètes en procédant à une interpolation, ce qui est le but des méthodes de prédiction passées en revue en section 1.4 pour ensuite déterminer la covariance via un estimateur classique ; 3) effectuer une estimation de la covariance en situation de données manquantes, ce qui est le but de l'algorithme EM.

En se plaçant dans l'optique du point 3), il est possible de choisir comme point de départ le modèle centré gaussien, qui permet une première description simple des données, puis d'élargir l'étude au modèle gaussien-composé, qui est un modèle plus flexible et mieux adapté aux données de télédétection, et plus spécifiquement aux données issues de la mesure radar [Mahot2012]. Notons que d'autres modèles ont été étudiés dans le cas de données incomplètes, comme les distributions elliptiques complexes [Frahm2010], qui est une généralisation de la distribution gaussienne, et la distribution Student [Liu2019b].

Positionnement 9. Il semble intéressant de confronter l'approche paramétrique (section 1.4) à l'approche prédictive, notamment en dressant un certain nombre d'hypothèses (gaussianité, non-gaussianité) sur la distribution probabiliste des données. Pour cela, l'utilisation de l'algorithme EM semble adaptée, et nécessitera une définition nouvelle, adaptée à l'approche paramétrique, de la forme des données manquantes (sous-section 1.1.3). Cette étude, plus exploratoire, pourra s'appliquer à l'estimation de la matrice de covariance de données de mesure de déplacement présentant une incomplétude. Pour cela, on pourra utiliser des données de mesure de déplacement issues d'un réseau de stations GNSS, ce qui est l'objet du chapitre 4.

1.6 Synthèse

Cet état de l'art nous a permis de passer en revue un certain nombre de méthodes pour gérer les données manquantes en télédétection. Les méthodes existantes peuvent être condensées en deux classifications. La première, très large, identifie quatre groupes de méthodes : les méthodes basées sur l'omission des données, les méthodes basées sur les poids, les méthodes d'imputation et les méthodes nécessitant un modèle statistique. La seconde classification, plus propre à la télédétection, reconnaît trois types de méthodes selon l'approche : spatiale, temporelle ou spatio-temporelle.

Les quelques positionnements scientifiques et méthodologiques disséminés au cours de ce chapitre sont à la racine des travaux présentés dans les chapitres 2, 3 et 4. Nous en présentons ici une synthèse.

Dans un premier temps, certaines approches prédictives, que l'on peut classer dans les méthodes d'imputation par approches temporelle et spatio-temporelle, ont été introduites. Nous avons ainsi jugé pertinent d'appliquer l'analyse en EOF à la mesure de champs de déplacement incomplets, et ce pour plusieurs raisons. D'une part, l'identification de difficultés liées à la perte de cohérence temporelle des données SAR nous a conduit à considérer l'importance de développer une méthode d'interpolation basée sur l'analyse en EOF, le problème de la cohérence faible étant ainsi entendu comme un problème de données manquantes [Prébet2019]. D'autre part, l'analyse en EOF offre la possibilité de prendre en compte l'information temporelle des champs de déplacements SAR. La reconstruction de données manquantes corrélées devra s'attacher à la spécificité des données en terme de complexité du champ de déplacement et de corrélation du bruit, ce qui constitue, à

notre connaissance, un manque dans les études existantes. De plus, les données peuvent manquer en des proportions parfois notables en espace et en temps, comme en mesure de déplacement par corrélation d'amplitude d'images SAR ou d'images optique. Dans ce cas, il sera intéressant de procéder à une augmentation spatiale des données pour reconstruire le champ incomplet à l'aide d'EOFs étendues [von Buttlar2014] ayant recours à la corrélation spatiale en plus de la corrélation temporelle des données. Ce type de technique n'a, selon nos recherches, pas été appliquée à la mesure de déplacement incomplète. Dans les deux cas (analyse en EOF et EEOF), la présence d'un bruit corrélé perturbant la mesure du déplacement peut entraîner une sur-estimation du nombre de modes. Il semble donc nécessaire de mettre en place des critères basés sur l'erreur entre les données initiales et les données reconstruites, comme la cross-RMSE, afin d'éviter cette sur-estimation. Dans le cas de l'analyse en EEOF, l'augmentation des données engendre une augmentation significative de la dimension de la matrice de covariance, et donc du nombre de vecteurs propres issus de la SVD : l'investigation de l'autocorrélation spatio-temporelle du champ et de l'incertitude des valeurs propres peuvent être d'un intérêt tout particulier pour proposer une sélection robuste du nombre optimal de modes. Pour cela, la règle empirique empirique proposée par [North1982] pourra être étudiée et étendue au cas de données spatialement augmentées.

Dans un second temps, nous avons brièvement introduit certaines approches paramétriques basées sur un modèle statistique : l'algorithme EM a naturellement émergé de cette discussion en tant que technique itérative d'estimation des paramètres statistiques en présence de données manquantes. Nous avons vu que des analyses similaires à l'analyse en EOF, comme l'ACP, ont déjà été intégrées dans un formalisme EM [Tipping1999]. Les analyses en EOF et en EEOF, qui procèdent également par itérations pour estimer les données manquantes, pourront également être intégrées dans un algorithme de *type* EM. Enfin, nous avons vu, à travers un exemple d'inversion de données que la connaissance des paramètres statistiques de données de déplacement, comme la covariance, alimente l'estimation des paramètres d'un modèle de déformation. Dans une étude liminaire, nous chercherons finalement à développer des algorithmes d'estimation de la matrice de covariance en utilisant l'algorithme EM, ce qui, en mesure de déplacement, constitue un objet de recherche original.

2
La méthode EM-EOF

Sommaire

2.1 Introduction 36
2.2 La méthode EM-EOF 36
2.2.1 Organisation des données 37
2.2.2 Décomposition de la covariance 37
2.2.3 Reconstruction des données 38
2.2.4 Estimation du nombre optimal de modes 39
2.2.5 Initialisation des données manquantes 40
2.2.6 L'algorithme EM-EOF 40
2.3 Simulations numériques 41
2.3.1 Type de champ de déplacement 42
2.3.2 Type de perturbation 43
2.3.3 Type de données manquantes 43
2.3.4 Paramètres de simulations 44
2.3.5 Résultats et discussions 45
2.4 Application sur données réelles 54
2.4.1 Glacier du Gorner 55
2.4.2 Glacier de Miage 59
2.4.3 Étude d'un cas limite : le glacier d'Argentière 59
2.4.4 Comparaison avec les méthodes NNI et krigeage 62
2.5 Conclusion 62

2.1 Introduction

Dans cette première étude, nous proposons de combiner l'analyse en EOF avec un algorithme de *type* EM, méthode que l'on désignera par l'acronyme EM-EOF pour *Expectation-Maximization Empirical Orthogonal Functions*. Les valeurs manquantes sont, avant traitement, initialisées : en prenant l'hypothèse de départ que la valeur d'initialisation est proche des valeurs espérées aux points manquants, un algorithme itératif convergeant vers les valeurs les plus vraisemblables peut ainsi être construit.

La méthode EM-EOF est itérative et adaptative aux données. Elle permet d'interpoler les valeurs manquantes au sein de séries temporelles de champs de déplacement issus de l'imagerie en télédétection (e.g. SAR, optique). La complexité des données en terme de comportement du signal de déplacement et du bruit est systématiquement prise en compte.

Comme la vérité terrain n'est pas ou peu disponible en mesure de déplacement par télédétection, une technique de *validation croisée* [Brankart1995] est mise en oeuvre afin de calculer la distance de reconstruction. La reconstruction finale minimise ainsi la distance entre le champ reconstruit et les données utilisées comme validation (voir figure 1.15)

L'originalité de cette étude réside principalement dans la mise en oeuvre d'une méthode itérative prenant en compte la corrélation temporelle du champ de déplacement puis dans son application aux séries temporelles de mesures de déplacement issues d'images de télédétection.

Ce chapitre suit la trame suivante. La section 2.2 décrit la méthode EM-EOF, ce qui comprend, dans l'ordre : un rappel sur l'organisation des données, le calcul et la décomposition de la matrice de covariance temporelle, la reconstruction avec un nombre de modes appropriés, l'initialisation des données manquantes, la validation croisée puis la description de l'algorithme. Les résultats d'application de la méthode sur des jeux de données synthétiques sont exposés en section 2.3, l'objectif étant d'étudier l'impact de paramètres externes tels que le niveau de bruit et la quantité de données manquantes sur la méthode. L'avant dernière section (2.4) concerne l'application de la méthode EM-EOF à des données réelles : trois séries temporelles de champs de déplacement calculés par InSAR et corrélation d'amplitude d'images Sentinel-1 A/B acquises entre septembre 2016 et décembre 2017 sur les glaciers du Gorner (Suisse), Miage (Italie) et Argentière (France). Enfin, la section 2.5 sert de conclusion quant aux avantages et inconvénients de la méthode EM-EOF, lesquels nous permettront de tirer des perspectives pour la suite de ce manuscrit.

2.2 La méthode EM-EOF

La méthode EM-EOF exploite essentiellement la corrélation temporelle du déplacement : elle s'appuie ainsi sur l'analyse en EOF de la matrice de covariance temporelle de la série temporelle de mesures de déplacement pour reconstruire les données manquantes.

Le principe général de la méthode EM-EOF se résume à deux étapes distinctes (figure 2.1). La première étape consiste à estimer le nombre optimal de modes EOF à partir d'une initialisation des données manquantes. Pour cela, la covariance temporelle du jeu de données est décomposée en modes distincts. Le nombre optimal de modes (noté R) pour reconstruire la série temporelle est ensuite estimé en minimisant l'erreur entre les données initiales de validation et les données reconstruites. La deuxième étape est une mise à jour itérative des valeurs manquantes utilisant l'estimation du nombre de modes de l'étape 1 comme l'entrée de l'algorithme de type EM : à l'étape E, les données manquantes sont estimées et comblées par leur valeurs espérées. À l'étape M, la covariance temporelle de la série est à son tour estimée à partir des données complétées lors de l'étape E. L'algorithme prend fin lorsque l'erreur de validation croisée converge vers un seuil pré-défini.

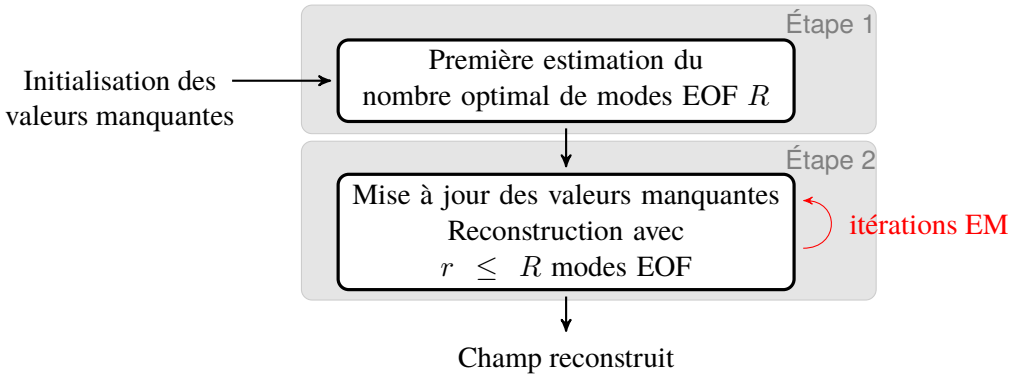


Figure 2.1 – Diagramme simplifié de la méthode EM-EOF.

2.2.1 Organisation des données

On rappelle ici quelques notations de l'analyse en EOF introduites en sous-section 1.4.1. Supposons que $\mathbf{X}$ désigne une matrice de données de taille $P \times N$, où P est le nombre de *variables* et N le nombre d'*observations*. Plus spécifiquement, P peut désigner un nombre de pixels, ou points, observés N fois au cours du temps. On appellera donc $\mathbf{X}$ *champ spatio-temporel*, et ses valeurs au point s et au temps t seront notées $(x_{st})_{1\leq s\leq P, 1\leq t\leq N}$. En forme matricielle, le champ $\mathbf{X}$ peut s'écrire :

$$\mathbf{X} = \Big(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\Big) = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{P1} & x_{P2} & \cdots & x_{PN} \end{pmatrix} \tag{2.1}$$

où chaque vecteur colonne $\mathbf{x}_t = (x_{1t}, x_{2t}, \ldots x_{Pt})^T$ est une observation de P points à un temps t. Inversement, chaque ligne de $\mathbf{X}$ est une série temporelle de taille N à un point s. Chaque vecteur $\mathbf{x}_t$ peut être un champ de déplacement incomplet, initialement représenté par une matrice 2D puis ordonné en vecteur colonne de taille p. Avant tout traitement ultérieur, la moyenne spatiale [1] du champ à chaque temps lui est soustraite afin d'obtenir l'*anomalie spatiale* $\mathbf{X}'$:

$$\mathbf{X}' = \mathbf{X} - \mathbf{1}_P \bar{\mathbf{x}} \tag{2.2}$$

où $\mathbf{1}_P = (1, \ldots, 1)^T$ est le vecteur unité de taille P et $\bar{\mathbf{x}} = (\mu_1, \mu_2, \ldots \mu_N)$ est un vecteur ligne contenant les moyennes empiriques spatiales de chaque observation $\mathbf{x}_t$ définies par :

$$\mu_t = \frac{1}{P} \sum_{s=1}^{P} x_{st} \tag{2.3}$$

2.2.2 Décomposition de la covariance

On définit tout d'abord la matrice de covariance temporelle empirique par :

$$\mathbf{\Sigma} = \frac{1}{P} \mathbf{X}'^T \mathbf{X} \tag{2.4}$$

1. Comme l'a fait remarquer [Björnsson1997], ne pas soustraire la moyenne n'empêche pas le calcul ultérieur des vecteurs propres, mais permet d'interpréter correctement la matrice de covariance des données.

$\mathbf{\Sigma}$ est une matrice de taille $N \times N$, symétrique, réelle et définie positive. Par conséquent, pour tout $\mathbf{z} \in \mathbb{R}^N$, $\mathbf{z}^T\mathbf{\Sigma}\mathbf{z} \geq 0$, et les valeurs propres de $\mathbf{\Sigma}$ sont réelles et positives, rangées par ordre décroissant $\lambda_1 > \lambda_2 > \cdots > \lambda_N$.

Les vecteurs propres, appelés *fonctions empiriques orthogonales* (EOF), satisfont l'équation linéaire suivante, communément appelée équation des valeurs propres :

$$\mathbf{\Sigma U} = \mathbf{U\Lambda} \tag{2.5}$$

où $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_N)$ est une matrice orthogonale de taille $N \times N$ contenant les vecteurs propres $\mathbf{u}_i = (u_{1i}, \ldots, u_{Ni})^T$ de $\mathbf{\Sigma}$ et $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots \lambda_N)$ contient les valeurs propres de $\mathbf{\Sigma}$ sur sa diagonale. Chaque vecteur $\mathbf{u}_i$ propre correspond à la valeur propre λ_i. L'équation (2.5) résulte du problème de maximisation suivant :

$$\max(\mathbf{u}^T\mathbf{\Sigma u}) \tag{2.6}$$

soumis à la contrainte $\mathbf{u}^T\mathbf{u} = 1$. Cela revient à trouver un vecteur unité $\mathbf{u}$ tel que $\mathbf{X'u}$ ait une variabilité maximale. Notons que la propriété d'orthogonalité [2] de $\mathbf{U}$, dont tient le nom de fonctions orthogonales, permet d'écrire l'équation (2.5) sous la forme $\mathbf{\Sigma} = \mathbf{U\Lambda U}^T$, ce qui correspond à la *décomposition en valeur propres* (EVD) de la matrice $\mathbf{\Sigma}$. Cette dernière peut être décomposée sous la forme d'une somme, écrite comme suit :

$$\mathbf{\Sigma} = \lambda_1\mathbf{u}_1\mathbf{u}_1^T + \lambda_2\mathbf{u}_2\mathbf{u}_2^T + \cdots + \lambda_N\mathbf{u}_N\mathbf{u}_N^T \tag{2.7}$$

Cette écriture n'est autre que la *décomposition spectrale* de $\mathbf{\Sigma}$. Chaque terme $\lambda_i\mathbf{u}_i\mathbf{u}_i^T$ de cette équation est appelé *mode*, ou *mode EOF*. Chaque mode permet de décrire une variabilité temporelle de l'anomalie spatiale $\mathbf{X'}$ [Hannachi2007] et chaque valeur propre permet de mesurer la fraction de variance totale expliquée par le mode correspondant. De manière générale, les premiers modes représentent la variabilité majeure du signal, ce qui signifie qu'une grande partie de la variance, qui peut aussi se traduire en terme de puissance spectrale, est contenue dans les premiers modes. Cela signifie également que le comportement du champ $\mathbf{X'}$ peut être expliqué par quelques modes dits *dominants*, c'est-à-dire ceux correspondants aux plus grandes valeurs propres de $\mathbf{\Sigma}$.

2.2.3 Reconstruction des données

La reconstruction de $\mathbf{X'}$, notée $\hat{\mathbf{X}}'$, est obtenue en sommant le produit des *composantes principales* (PC) $\mathbf{a}_i$ et des vecteur propres :

$$\hat{\mathbf{X}}' = \sum_{i=1}^{N} \mathbf{a}_i\mathbf{u}_i^T \tag{2.8}$$

où la $i$ème composante $\mathbf{a}_i = \mathbf{X'u}_i$ est la projection de $\mathbf{X'}$ sur le $i$ème vecteur propre, et dont chaque élément a_{ki}, pour $k = 1, \ldots, P$, s'exprime par :

$$a_{ki} = \sum_{j=1}^{N} x'_{kj}u_{ji} \tag{2.9}$$

On peut se représenter chaque $\mathbf{a}_i$ comme un champ spatial associé à chaque vecteur propre $\mathbf{u}_i$. On peut alors se référer aux PC comme aux "modes" de variabilité spatiale de la série temporelle, alors que les vecteur propres permettent de visualiser comment ces modes évoluent dans le temps.

Afin de retrouver le champ reconstruit $\hat{\mathbf{X}}$, on ajoute finalement la moyenne spatiale à l'anomalie reconstruite :

2. Propriété vérifiable par $\mathbf{U}^T\mathbf{U} = \mathbf{UU}^T = \mathbf{I}$ où $\mathbf{I}$ est la matrice identité.

$$\hat{\mathbf{X}} = \hat{\mathbf{X}}' + \mathbf{1}_P \bar{\mathbf{x}} \tag{2.10}$$

2.2.4 Estimation du nombre optimal de modes

La troncature de l'expression (2.8) à $R \ll N$ termes permet de ne retenir que les modes EOFs qui correspondent aux premières (plus grandes) valeurs propres. On notera dès lors $\hat{\mathbf{X}}'_R$ le champ reconstruit avec R modes, défini par :

$$\hat{\mathbf{X}}'_R = \sum_{i=1}^{R \ll N} \mathbf{a}_i \mathbf{u}_i^T \tag{2.11}$$

En procédant ainsi, il est possible d'extraire les principales caractéristiques du signal car les premiers modes capturent la plus grande part de la dynamique temporelle du signal alors que les modes suivants représentent diverses perturbations [Prébet2019], souvent associées à du bruit.

Le choix du nombre optimal de modes pour reconstruire le signal est donc crucial, l'idéal étant de laisser de côté les modes *secondaires* comme le ferait un filtre passe-bas avec les hautes fréquences.

Nous proposons ci-après deux techniques afin de sélectionner le nombre optimal de modes : le calcul d'une erreur de validation croisée, puis la construction d'un critère de convergence basé sur cette même erreur.

Validation croisée

La sélection du nombre optimal de modes se base notamment sur une erreur de reconstruction, dite erreur de validation croisée. Cette erreur n'est autre que la racine de l'erreur quadratique moyenne (cross-RMSE) [Brankart1995] exprimée par :

$$\delta_k = \left[\frac{1}{Q} \sum_{i=1}^{Q} (\hat{x}_i - x_i)^2 \right]^{1/2} = \frac{1}{\sqrt{Q}} ||\hat{\mathbf{x}}_{\text{cv},k} - \mathbf{x}_{\text{cv}}||_2 \tag{2.12}$$

où $\mathbf{x}_{\text{cv}} = \{x_i\}_{1 \leq i \leq Q} \subseteq \mathbf{X}_{\text{obs}}$ est un vecteur contenant Q points choisis aléatoirement parmi les données existantes $\mathbf{X}_{\text{obs}}$, et $\hat{\mathbf{x}}_{\text{cv},k} = \{\hat{x}_i\}_{1 \leq i \leq Q}$ est sa reconstruction avec k modes. Après tirage aléatoire des Q points de validation croisée, ceux-ci sont artificiellement convertis en données manquantes, une copie de leur valeur étant stockée dans $\mathbf{x}_{\text{cv}}$. Après chaque reconstruction du signal avec k modes, les valeurs de $\mathbf{x}_{\text{cv}}$ sont comparées à l'estimation $\hat{\mathbf{x}}_{\text{cv},k}$. Le choix du nombre Q de points doit se faire de manière à représenter statistiquement les données tout en sachant qu'augmenter Q signifie également augmenter la quantité de données manquantes, ce qui pourrait affecter la qualité de la reconstruction. La cross-RMSE est particulièrement adaptée lorsqu'aucune vérité terrain n'est disponible pour valider les résultats, ce qui est souvent le cas en mesure de déplacement. L'usage d'une telle erreur permet également de s'affranchir de toute connaissance *a priori* sur l'évolution spatio-temporelle du champ.

Biais de surestimation

Lorsqu'un signal de déplacement est fortement perturbé par un bruit corrélé, le nombre optimal de modes peut être surestimé. Afin de parer à ce problème, on examine la quantité suivante :

$$\Lambda = 1 - \frac{\delta_{k+1}}{\delta_k} \tag{2.13}$$

Cette métrique permet de mesurer la variation de la cross-RMSE lors de l'ajout d'un mode supplémentaire. Une petite variation, disons inférieure à un seuil β, implique que peu d'information est apportée au nouveau champ reconstruit par l'ajout d'un nouveau mode : dans ce cas, ce mode

n'est pas pris en compte. Si l'incertitude des données est connue, β peut être déterminé de façon à ce que l'incertitude de reconstruction soit dans l'intervalle de l'incertitude des données. A défaut de connaître l'incertitude, β est déterminé empiriquement. Dans la plupart des cas, une valeur de 0.1 (ce qui signifie que l'algorithme prend fin si la variation de la cross-RMSE entre k et $k+1$ modes tombe en-dessous de 10%) est suffisante pour éviter ce biais.

2.2.5 Initialisation des données manquantes

L'initialisation des données manquantes constitue une étape clef de la méthode EM-EOF. En effet, la valeur initiale peut impacter l'estimation de la covariance temporelle $\mathbf{\Sigma}$ lors des itérations, et donc avoir un impact sur le calcul des EOFs. Par ailleurs, l'initialisation étant considérée comme la première estimation des données manquantes, il sera pertinent de choisir une valeur en accord avec la distribution statistique des données observées. Ce choix implique qu'une telle initialisation devrait perturber le moins possible la répartition de la variance des différents modes qui composent le champ de déplacement. Afin d'éviter tout biais au sein de l'anomalie, les données manquantes peuvent être initialisées par la moyenne spatiale, ce qui revient à initialiser l'anomalie par zéro [Schneider2001, Alvera-Azcarate2007]. L'étude de [Beckers2003] a montré qu'une telle initialisation tend à diminuer la variance des modes non dominants et inversement, à augmenter celle des modes dominants. D'autre part, l'information à petite échelle peut se voir effacée car un effet de filtrage se met en place sur le voisinage proche des données manquantes. Nous choisissons, dans un premier temps, d'initialiser les valeurs manquantes de l'anomalie par 0. Nous nous attacherons, lors des simulations, à comparer différentes valeurs d'initialisation et à analyser brièvement leur impact sur l'estimation des valeurs manquantes.

2.2.6 L'algorithme EM-EOF

Dans cette section, nous décrivons en détail le déroulé de l'algorithme EM-EOF présenté succintement en figure 2.1.

Étape 1 : première estimation du nombre optimal de modes

L'étape 1 est décrite en détail en pseudo-code par l'algorithme 1. Après initialisation des données manquantes, les équations (2.4) et (2.5) sont calculées puis l'anomalie est reconstruite par l'expression (2.8), en y ajoutant un mode supplémentaire à la fois. A chaque ajout d'un mode supplémentaire k, la cross-RMSE δ_k est calculée. Le nombre optimal de modes R est celui qui minimise le set de cross-RMSE :

$$R = \underset{R\in[1,N]}{\arg\min}\, \delta_k, \qquad k = 1, \ldots, N \tag{2.14}$$

où N est le nombre maximal de modes, soit la dimension temporelle ici.

Étape 2 : mise à jour des valeurs manquantes

L'étape 2 de la méthode EM-EOF a pour but d'affiner d'une part le nombre de mode estimé à l'étape 1, puis la valeur même des valeurs manquantes, d'autre part. Une description est fournie par le pseudo-code de l'algorithme 2. Nous décrivons ci-dessous cette étape dans le cadre formel de l'algorithme EM. Si $\hat{\mathbf{X}}_{\text{mis}}^{(m)}$ désigne l'estimation de $\mathbf{X}_{\text{mis}}$ et $\delta_k^{(m)} = ||\hat{\mathbf{x}}_{\text{cv},k}^{(m)} - \mathbf{x}_{\text{cv}}||_2$ la cross-RMSE à l'itération m de l'algorithme, les étapes de l'EM sont décrites comme suit :

Initialisation. $\mathbf{X}_{\text{mis}}^{(m=0)} = 0$.

Étape E. Calcul des valeurs manquantes espérées sachant les données observées, la covariance estimée et les EOFs :

Algorithme 1 Étape 1 : première estimation de R

Entrée: $\mathbf{X}$, init_value
Sortie: R

$\mathbf{x}_{\text{cv}} \leftarrow$ init_value
pour $k \leftarrow 1, N$ **faire**
    Calculer (2.4), (2.5) pour estimer $\hat{\mathbf{\Sigma}}$, $\hat{\mathbf{U}}$
    Calculer (2.8) avec k modes pour estimer $\hat{\mathbf{x}}_{\text{cv},k} \subseteq \hat{\mathbf{X}}_k$
    Calculer δ_k
fin pour
return $\underset{R\in[1,N]}{\arg\min}\, \delta_k$

$$\hat{\mathbf{X}}_{\text{mis}}^{(m)} = \mathbb{E}\big[\mathbf{X}_{\text{mis}} | \mathbf{X}_{\text{obs}}, \hat{\mathbf{\Sigma}}^{(m)}, \hat{\mathbf{u}}^{(m)}\big]$$

Cette étape est réalisée en calculant les équations (2.4), (2.5) et (2.8).

Étape M. Calcul de la cross-RMSE et examen de la condition

$$\delta_k^{(m)} < \delta_k^{(m-1)}$$

Si cette condition est remplie, les valeurs manquantes estimées à l'étape précédente sont remplacées par les valeurs estimées à l'itération courante : $\hat{\mathbf{X}}_{\text{mis}}^{(m)} = \hat{\mathbf{X}}_{\text{mis}}^{(m-1)}$. L'algorithme consiste alors en une répétition des étapes E et M jusqu'à ce que l'erreur converge, c'est-à-dire lorsque la différence $\Delta_k = \delta_k^{(m)} - \delta_k^{(m-1)}$ atteint un seuil de tolérance prédéfini α. Une fois le critère de convergence atteint, un mode supplémentaire est ajouté à la reconstruction. L'EM est alors parcouru une nouvelle fois avec $k+1$ modes, et ainsi de suite. A tout moment, si l'erreur augmente, ou si le critère Λ (équation (2.13)) dépasse un seuil β, la procédure prend fin. Sinon, celle-ci continue jusqu'à ce que le nombre de mode R estimé à l'étape 1 soit atteint. Notons que le nombre R n'est pas nécessairement atteint durant cette étape. Cela dépend essentiellement du rapport signal-sur-bruit (SNR) et de la quantité de données manquantes. Durant le processus itératif de l'étape 2, la mise à jour des valeurs manquantes améliore la cohérence globale du champ en modifiant les valeurs reconstruites vers leur valeur espérée.

A la fin de l'étape 2, la série temporelle est reconstruite avec le nombre optimal de modes en faisant appel à l'équation (2.8). La moyenne est finalement ajoutée à l'anomalie pour retrouver le champ reconstruit final.

2.3 Simulations numériques

Dans cette partie, nous allons simuler des champs de déplacement physique d'ordres différents. L'intérêt de la simulation réside dans la disponibilité de la vérité terrain, avec laquelle il est possible d'effectuer une validation absolue. Le but est d'une part de démontrer la capacité de la méthode EM-EOF à interpoler des valeurs manquantes au sein de série temporelles de champs à complexités variables puis, d'autre part, de montrer l'impact de paramètres clefs sur la qualité de la reconstruction, comme le type de bruit présent au sein des données, le type de données manquantes et la valeur d'initialisation.

Algorithme 2 Étape 2 : mise à jour des valeurs manquantes

Entrée: $\mathbf{X}$, R, α, β
Sortie: $\hat{\mathbf{X}}_k$

pour $k \leftarrow 1, R$ **faire**
  tant que $|\Delta_k| < \alpha$ **faire**
    Calculer (2.4), (2.5), (2.8) pour estimer $\hat{\mathbf{\Sigma}}^{(m)}$, $\hat{\mathbf{U}}^{(m)}$, $\hat{\mathbf{x}}_{\text{cv},k}^{(m)} \subseteq \hat{\mathbf{X}}_k^{(m)}$
    $\hat{\mathbf{x}}_{\text{cv},k}^{(m-1)} \leftarrow \hat{\mathbf{x}}_{\text{cv},k}^{(m)}$
    Calculer $\delta_k^{(m)}$
    si $\delta_k^{(m)} > \delta_{k-1}^{(m)}$ or $\Lambda < \beta$ **alors**
      return $\hat{\mathbf{X}}_{k-1}$
    fin si
    $m \leftarrow m + 1$
  fin tant que
fin pour
return $\hat{\mathbf{X}}_k$

2.3.1 Type de champ de déplacement

Afin de simuler les champs de déplacement, divers modèles d'ordres variables sont générés (le tableau 2.1 récapitule tous les champs simulés). Les modèles vont d'un simple champ linéaire d'ordre 1 g_1 à des champs d'ordre n (2, 3, 4, ≥ 4), appelés respectivement g_2, g_3, g_4 et g_5 (voir la figure 2.2 pour visualiser les champs synthétisés). Ces champs correspondent à des sommes de produits de sinusoïdes afin de reproduire un déplacement oscillatoire à fréquences multiples. Le champ g_5 correspond quant à lui à un champ multi-forme, c'est-à-dire composé de plusieurs cibles physiques à variabilités distinctes. Un sixième champ g_6 correspondant à un déplacement post-sismique exponentiellement décroissant est également généré à l'aide de l'outils Pyrocko [Heimann2017].

Nom	$g(r,t)$	Ordre
g_1	$(1-0.5r_1)t$	1
g_2	$g_1 + \sin(2\pi f_1 t)\cos(2\pi f_1 r_1)$	2
g_3	$g_2 + 0.5\cos(2\pi f_2 t)\cos(2\pi f_3 r_1)$	3
g_4	$g_3 + 0.1\sin(2\pi f_4 t)\cos(2\pi f_5 r_1)$	4
g_5	$g_1(r_1) + g_3(r_2) + g_3(r_3) + g_4(r_1)$	≥ 4
g_6	$A - b\exp(-t/\tau_e) + 10^{-4}t$	-

Tableau 2.1 – Récapitulatif des champs déterministes synthétisés. g_1 est linéaire en temps et en espace alors que les chamsp g_2 à g_5 sont non-linéaires et incluent diverses oscillations à fréquences variables. g_6 est un modèle de déformation post-sismique avec un temps de décroissance $\tau_e = 1.5$. Les constantes A et b sont fixées à 0 et 1 respectivement. $r_1 = \sqrt{x^2+y^2}$, $r_2 = \sqrt{(x-1)^2+(y-1)^2}$ et $r_3 = \exp\left(-(x+y)^2\right) + xy + \tan(x)$ sont les distances à l'origine. Les coordonnées (x, y) varient dans l'intervalle compacte $[-1, 1]^2$ et t est la variable temps. Les valeurs des fréquences sont fixées à : $f_1 = 0.25$, $f_2 = 0.75$, $f_3 = 2.5$, $f_4 = 1.25$, $f_5 = 5$.

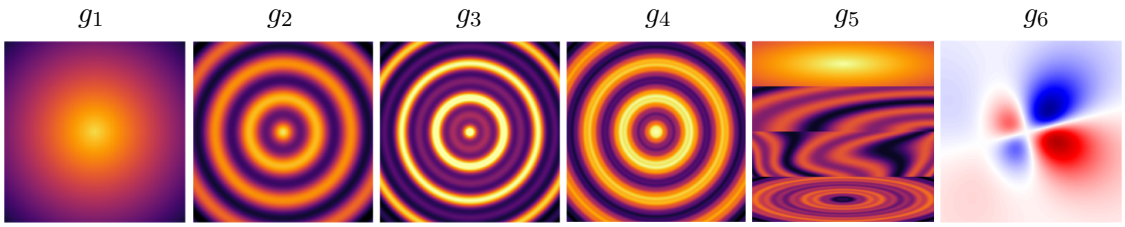


Figure 2.2 – Exemples de champs de déplacement synthétisés dans cette étude. On notera que le champ g_5 est une composition spatiale de champs d'ordre 1, 3 et 4, avec diverses formes de distances à l'origine (voir tableau 2.1).

2.3.2 Type de perturbation

Les mesures de déplacement dérivées de l'imagerie en télédétection sont souvent sujettes à des perturbations d'origines diverses (voir sous-section 1.3.2 et notamment l'équation 1.5) : perturbation atmosphérique, erreurs introduites lors du traitement (e.g. erreur de déroulement de phase en InSAR), bruit thermique, etc. Ces perturbations, tout comme le champ de déplacement, peuvent être corrélées en temps et/ou en espace. Afin de représenter au mieux (sans toutefois prétendre à l'exhaustivité) la nature de ces perturbations au sein des champs simulés, nous synthétisons trois types de perturbations :

— un bruit spatialement corrélé (SCN)

— un bruit spatio-temporellement corrélé (STCN)

— des erreurs localisées issues du traitement, simulées par des sauts artificiels des valeurs déplacement (e.g. sauts de phase)

Un exemple des bruits SCN et STCN est présenté en figure 2.3. Le SCN est synthétisé en utilisant une fonction d'auto-corrélation de la forme $c(r) = r^{-\gamma}$, où la variation du paramètre $\gamma \geq 0$ permet de régler le degré de corrélation entre deux points distants de r (figure 2.4). Cette fonction est ensuite utilisée afin de filtrer puis moduler un bruit blanc gaussien dans le domaine fréquentiel. Le STCN est la somme d'un SCN et d'un bruit temporellement corrélé, et utilise de manière connexe un coefficient $\rho \geq 0$ afin de régler le degré de corrélation temporelle. Le lecteur est invité, pour un descriptif détaillé de la synthétisation du STCN, à consulter l'Annexe A.

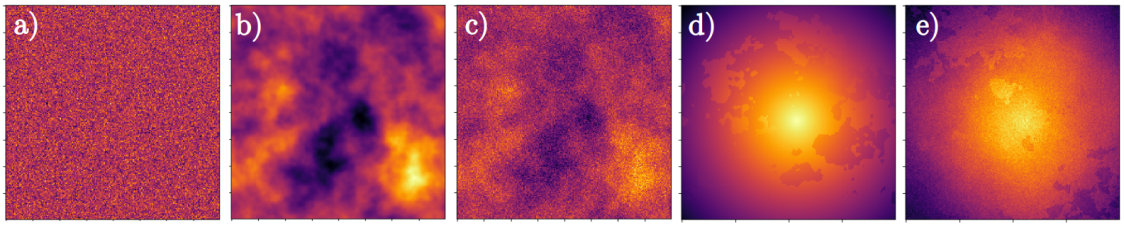


Figure 2.3 – Exemples de perturbations : a) bruit blanc gaussien (non simulé dans l'étude) ; b) bruit spatialement corrélé (SCN) ; c) bruit spatio-temporellement corrélé (STCN) ; d) erreur localisée issue du traitement sur champ d'ordre 1 ; e) erreur localisée issue du traitement sur champ d'ordre 1 perturbé par un bruit STCN.

2.3.3 Type de données manquantes

Selon leur origine, les données manquantes peuvent être distribuées de manière très variée. Parmi les formes de données manquantes présentées au chapitre 1 (section 1.1.3), on distingue deux grands cas : distribution aléatoire indépendante du temps et de l'espace et distribution corrélée en temps et/ou espace. Par exemple, on observe fréquemment des données manquantes spatialement corrélées en mesure de déplacement par corrélation d'images, puisque de fortes chutes de neige

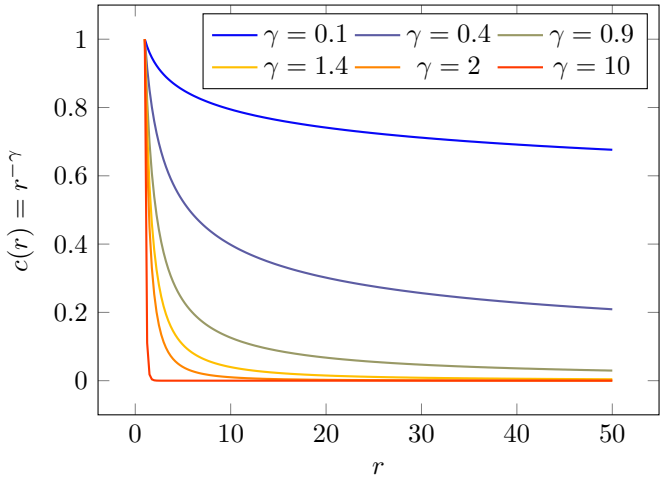


Figure 2.4 – Evolution du degré de corrélation $c(r)$ en fonction de la distance r. Plus l'exposant γ est grand, moins le bruit est corrélé. A l'inverse, plus γ est petit et plus $c(r)$ est invariant à la distance r, ce qui correspond à un bruit plus corrélé.

ou un écoulement très rapide de la surface du glacier peuvent être le signe d'une corrélation basse. Il est alors souvent préférable de supprimer ces valeurs à forte incertitude que de les conserver. En InSAR, des phénomènes saisonniers, comme les chutes de neige ou la densification du couvert végétal peuvent induire une perte de cohérence. Nous considérons donc les types de données manquantes suivants (figure 2.5) : 1) aléatoires dans les deux dimensions espace et temps, puis 2) spatio-temporellement corrélés, c'est-à-dire possédant une forme évoluant en temps et en espace. Dans ce deuxième cas, les données manquantes sont synthétisées sur dix champs de déplacement consécutifs afin de simuler une origine saisonnière.

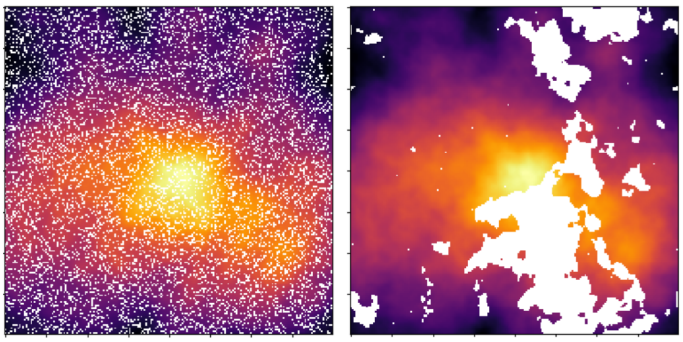


Figure 2.5 – Types de données manquantes superposées à un champ de déplacement du premier ordre contenant un bruit SCN. A gauche : données manquantes aléatoires ; à droite : données manquantes corrélées.

2.3.4 Paramètres de simulations

Pour les besoins de l'expérience, six séries temporelles de 40 champs de déplacement sont synthétisées. Les champs g_1 à g_4 sont de taille 200×200, alors que g_5 et g_6 sont de taille 4000×4000. Les champs sont ensuite artificiellement corrompus par des données manquantes aléatoires et corrélées ainsi que des bruits SCN et STCN. Dans un premier temps, la quantité de données manquantes est fixée à 30% et le rapport signal-sur-bruit (SNR) autour de 1.4 (tableau 2.2). Le SNR est ici définit comme le ratio suivant :

$$\mathrm{SNR} = \frac{\mu^2_{\text{dépl}}}{\sigma^2_{\text{bruit}}} \tag{2.15}$$

où $\mu_{\text{dépl}}$ est la moyenne spatio-temporelle du signal de déplacement et σ_{bruit} est l'écart-type du bruit. On considère que le signal de déplacement est totalement submergé par le bruit lorsque le SNR est inférieur à 1. Inversement, plus le SNR est grand, moins l'amplitude du bruit est importante.

Dans un second temps, une analyse plus fine est réalisée : la quantité de données manquantes est modulée entre 0 et 80%, alors que l'on fait varier le SNR dans un intervalle allant de 0.5 à 4.5, soit d'un signal fortement à moyennement bruité.

Cas	Ordre	Type de données manquantes	Type de bruit	SNR
1	1	Aléatoire	SCN	1.44
		Corrélé	STCN	1.47
	2	Aléatoire	SCN	1.45
		Corrélé	STCN	1.24
2	3	Aléatoire	SCN	1.61
		Corrélé	STCN	1.43
	4	Aléatoire	SCN	1.46
		Corrélé	STCN	1.24
	-	Aléatoire	SCN	1.7
3	multi-forme	Aléatoire	SCN	1.4
		Corrélé	STCN	1.4

Tableau 2.2 – Paramètres des expériences des cas 1, 2 et 3 (voir section 2.3.5 ci-après). Tous les champs de déplacement contiennent 30% de données manquantes.

2.3.5 Résultats et discussions

Dans les sous-parties qui suivent, nous présentons les échantillons de résultats de reconstruction sur les champs simulés décrits en section 2.3.1. Les résultats sont classés comme suit :

— Comparaison de différentes valeurs d'initialisation ;

— **Cas 1** : Champs du premier et second ordre g_1 et g_2 ;

— **Cas 2** : Champs du troisième, quatrième ordre et déplacement post-sismique : g_3, g_4, g_6 ;

— **Cas 3** : Champs multi-formes g_5.

Comparaison des différentes initialisations

Afin d'évaluer l'effet de la valeur d'initialisation sur la qualité de la reconstruction des données, les simulations sont reproduites avec trois valeurs d'initialisation :

1. Moyenne spatiale ;
2. Moyenne spatiale + bruit blanc Gaussien ;
3. Moyenne spatiale + SCN.

Le résultat des expériences sur un nombre suffisant de répétitions (500) permet de conclure qu'il n'existe pas, selon la valeur d'initialisation considérée, de différence notable sur la valeur de l'erreur finale ni sur l'estimation du nombre optimal de modes. Cependant, lorsque le champ est initialisé par une des deux dernières initialisations, la durée de convergence se voit systématiquement augmentée durant l'étape 2. Ces deux initialisations pourraient s'avérer efficaces si le bruit présent dans les données était connu, ce qui n'est le cas ici. Par conséquent, dans toutes les expériences qui suivent, nous choisissons d'initialiser les valeurs manquantes par la moyenne spatiale.

Cas 1 : champs du premier et second ordre

Un exemple de reconstruction des champs g_1 et g_2 est présenté en figure 2.6. Le nombre optimal de modes estimé est de 1 pour le champ du premier ordre et de 2 pour le champ du second ordre. On observe que les champs de déplacement reconstruits sont en accord avec le vrai déplacement simulé, sans dégradation visible dans les zones de données manquantes (aléatoires et corrélées).

Il semble néanmoins que la qualité de reconstruction soit plus affectée par du bruit STCN, dont la granularité est encore faiblement visible au sein du champ reconstruit (figure 2.6 (b)(d)). La figure 2.7 montre également l'évolution d'un point du champ g_2 au cours de la série temporelle ainsi que sa reconstruction. Ces résultats démontrent la capacité de la méthode EM-EOF à reconstruire des tendances temporelles en plus des variations spatiales, et ce quel que soit le type de données manquantes et bruit.

Nous simulons, en plus des bruits SCN et STCN, des sauts locaux des valeurs de déplacement sur le champ du second ordre (figures 2.7 (a) et 2.8). En interférométrie radar, ces sauts sont souvent le résultats d'erreurs lors du processus de déroulement de la phase interférométrique (d'où l'expression saut de phase). Les valeurs reconstruites aux points concernés par des sauts de phase montrent une certaine robustesse de la méthode par rapport à ce type d'événement haute fréquence. L'observation des valeurs résiduelles permet de confirmer que les sauts de phase y ont été absorbés. En effet, ces perturbations sont représentées par des modes EOF secondaires qui ne sont pas sélectionnés dans la reconstruction.

L'analyse sous forme de carte d'erreur de reconstruction en fonction du pourcentage de données manquantes et du SNR (figure 2.9) permet également de mieux connaître la sensibilité de la méthode à de tels paramètres, et ainsi de poser plus clairement les conditions d'applicabilité de la méthode.

Ces cartes d'erreur permettent ici de dégager deux observations, confirmées par l'expérience. Premièrement, le niveau de bruit joue un rôle majeur dans la performance de reconstruction, et ce comparé à la quantité de données manquantes. Cette dernière n'affecte sensiblement la reconstruction que lorsqu'elle est supérieure à 60% des points existants. Ce constat vaut pour des données manquantes aléatoires. On observe que les données manquantes corrélées affectent moins la qualité de reconstruction. Cette différence de sensibilité entre données manquantes aléatoires et corrélées s'explique par la qualité épisodique de ces dernières. Un calcul simple peut nous permettre de comprendre qu'une plus grande quantité de données manquantes aléatoires signifie également une plus grande probabilité qu'un point soit manquant sur N déplacements consécutifs. Si on désigne par $\mathcal{P}$ cette probabilité, on peut l'exprimer par $\mathcal{P} = (q/100)^n$ où q est le pourcentage de données manquantes. Si $n = 40$, $\mathcal{P}$ est presque négligeable pour $q < 60\%$, ce qui est cohérent avec une augmentation de l'erreur au delà de 60%.

Cas 2 : champs du troisième et quatrième ordre

Les champs considérées ici étant plus complexes, on s'attend à ce que le nombre de mode sélectionné soit plus élevé. En effet, le nombre de mode sélectionné est de 3 dans le cas d'un champ du troisième ordre, 5 pour un champ du quatrième ordre et 2 pour un déplacement post-sismique (figure 2.10).

L'observation des champs reconstruits permet de dégager une consistance globale avec les champs réels, tant dans la variabilité spatiale que temporelle, même en présence d'erreurs de

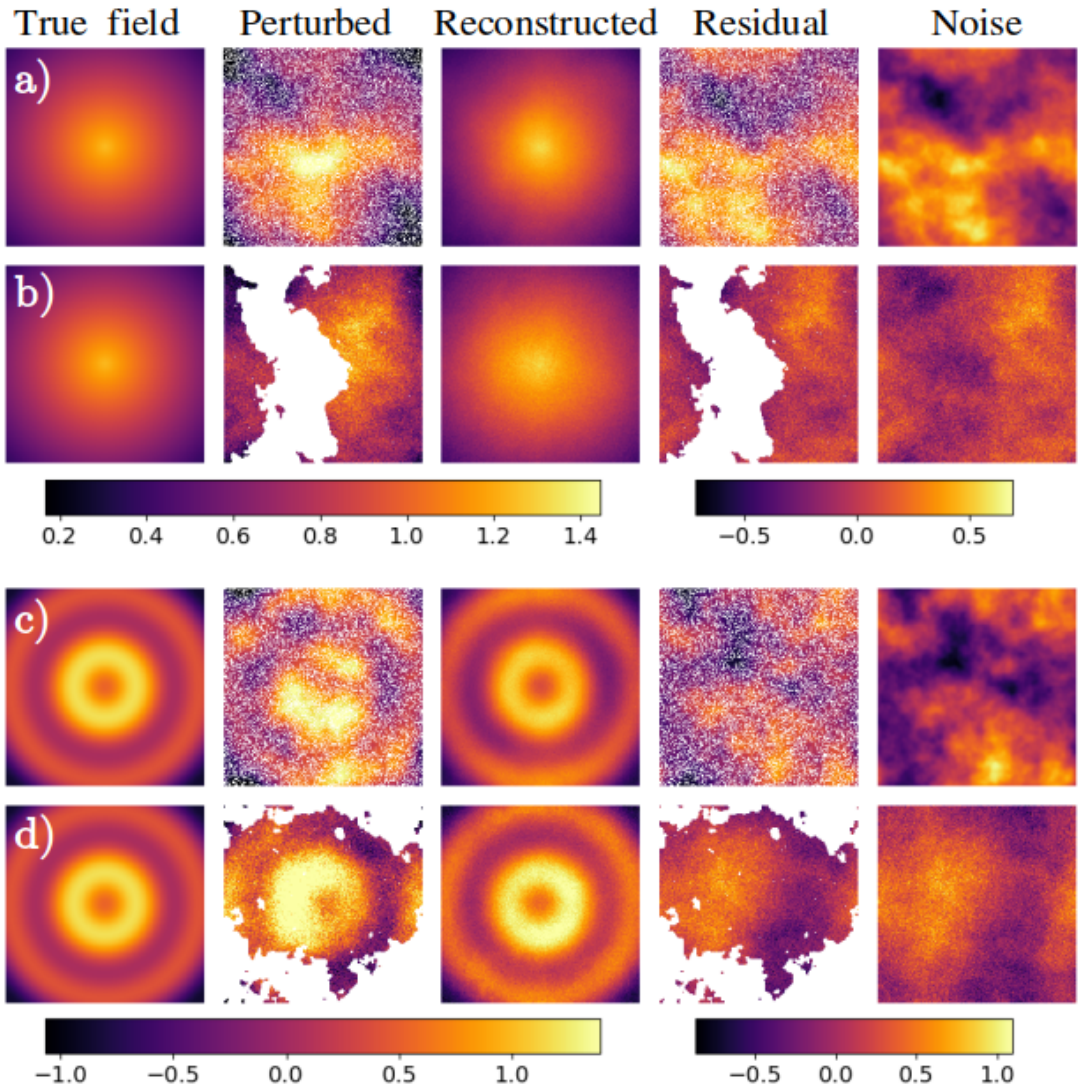


Figure 2.6 – Reconstruction des champs de déplacement [cm] du premier ordre (a)(b) et second ordre (c)(d). Le SNR varie entre 1.24 et 1.44 et la quantité de données manquantes est de 30%. (a)(c) : données manquantes aléatoires et SCN ; (b)(d) : données manquantes corrélées et STCN. Les résidus sont la différence entre le champ reconstruit et le champ perturbé.

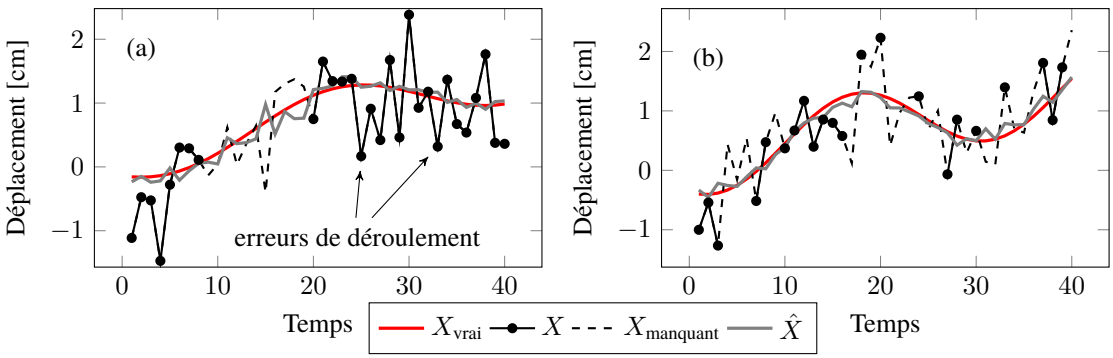


Figure 2.7 – Série temporelle d'un champ du second ordre perturbé par (a) données manquantes corrélées sur 10 dates consécutives et (b) données manquantes aléatoires. Rouge : déplacement vrai ; cercles noirs : déplacement bruité avec données manquantes ; courbe hachée noire : données manquantes ; ligne grise : série temporelle reconstruite.

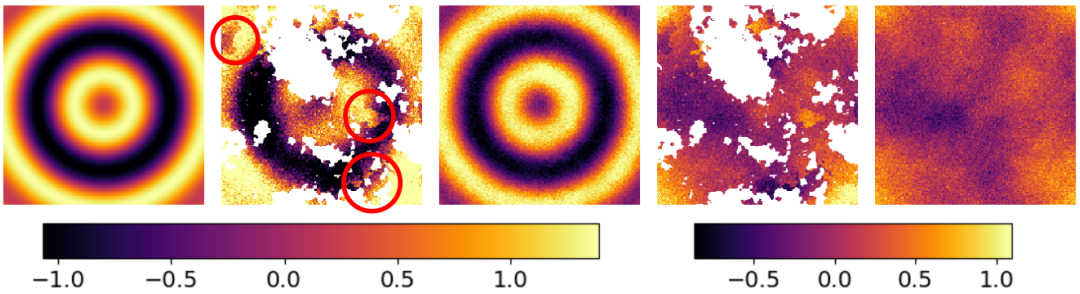


Figure 2.8 – Reconstruction d'un champ du second ordre perturbé par 30% de données manquantes, un bruit STCN et plusieurs erreurs de déroulement de phase (cercles rouges).

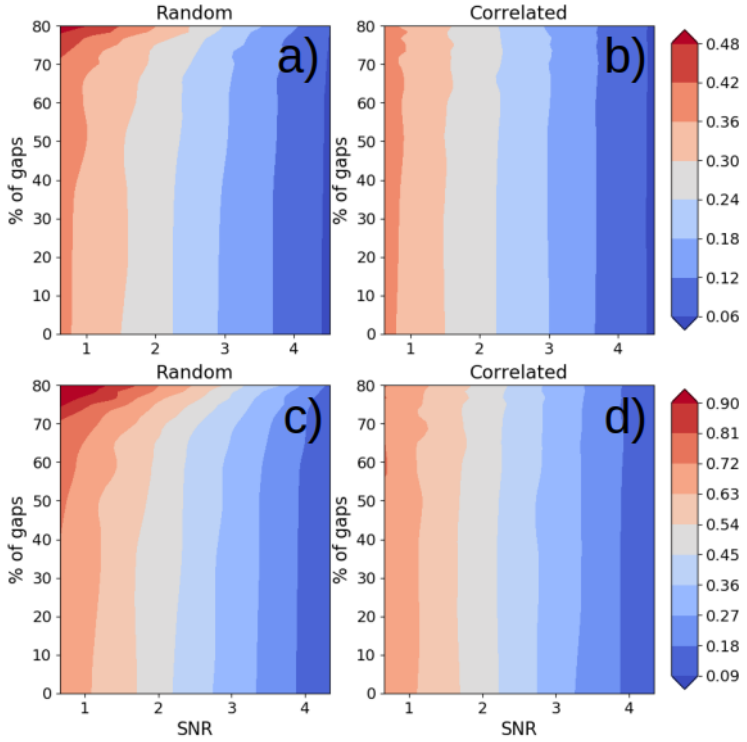


Figure 2.9 – Cartes d'erreur [cm] en fonction du % de données manquantes et du SNR dans le cas d'un champ du premier ordre (a)(b) et du second ordre (c)(d) perturbé par des trous aléatoires (a)(c) et corrélés (b)(d). Tous les déplacement sont également perturbés par un bruit spatio-temporellement corrélé (STCN).

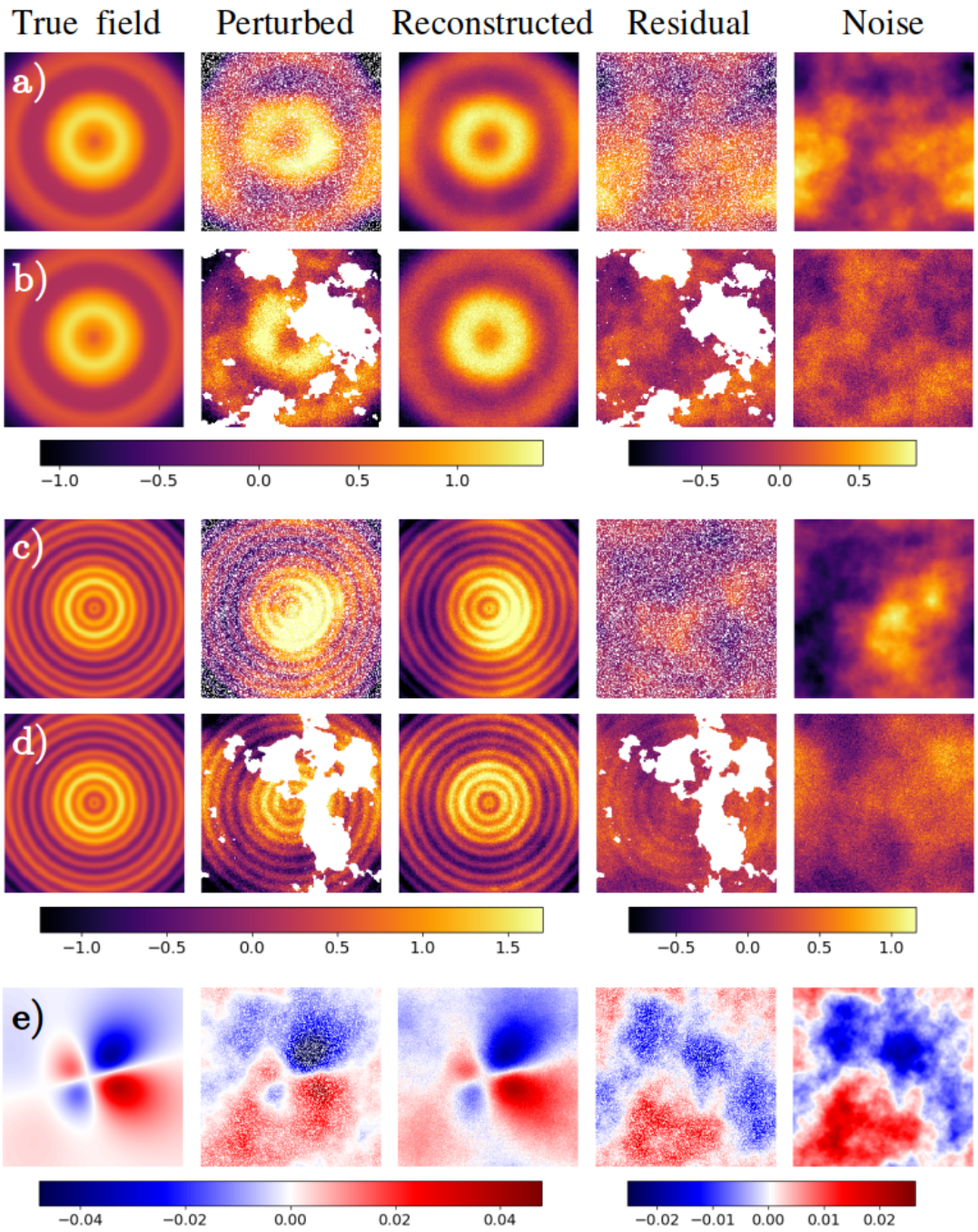


Figure 2.10 – Reconstruction des champs de déplacement [cm] du troisième ordre (a)(b), quatrième ordre (c)(d) et post-sismique (e). Le SNR varie entre 1.24 et 1.61 et la quantité de données manquantes est de 30%. (a)(c)(e) : données manquantes aléatoires et SCN ; (b)(d) : données manquantes corrélées et STCN. Les résidus sont la différence entre le champ reconstruit et le champ perturbé.

déroulement de phase comme le montre la figure 2.14. Dans le cas du champ du quatrième ordre (figure 2.10 (c)(d)), une partie du signal de déplacement se distingue dans les résidus, même si l'amplitude moyenne de ces derniers est faible. De plus, le champ reconstruit contient encore du bruit. En effet, le choix du nombre optimal de modes est ici plus délicat car le champ de déplacement est complexe et fortement perturbé par un bruit dont le comportement peut être proche de celui du déplacement. La présence de ce bruit corrélé en espace et en temps peut conduire à une sur-estimation du nombre de modes, notamment parce que la similarité entre ce bruit et le déplacement rend leur séparation difficile. En effet, comme le montre la figure 2.11, la variance d'un bruit SCN/SCTN à forte corrélation s'explique par seulement quelques modes dominants, tout comme celle d'un champ de déplacement, ce qui introduit un certain mélange spectral.

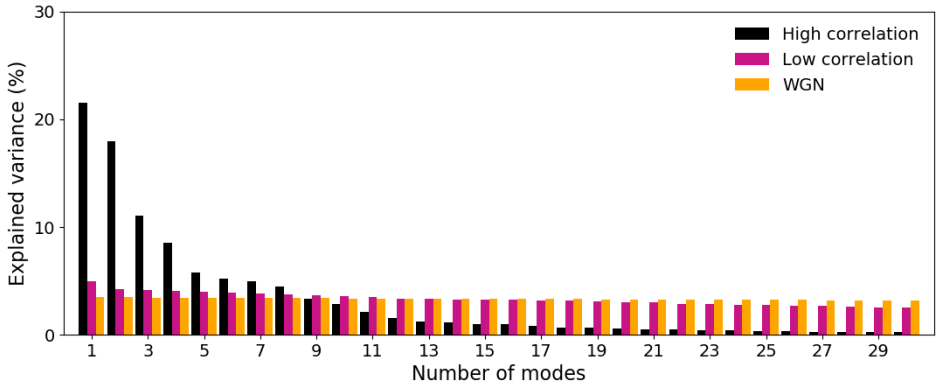


Figure 2.11 – Pourcentage de variance expliquée pour trois bruits : bruit blanc gaussien (WGN, couleur orange) et bruit fortement (noir) et faiblement (magenta) corrélé spatialement (respectivement $\gamma = 0.9$ et $\gamma = 0.2$).

Dans ce cas spécifique, la métrique Λ (équation 2.13) peut être utilisée et on pourra augmenter le seuil β afin de limiter cette sur-estimation. Notons qu'une valeur trop haute de β conduira à l'effet inverse, c'est-à-dire à une sous-estimation du nombre de modes.

Les cartes d'erreurs des champs du troisième et quatrième ordre montrent, comme dans le cas 1, que de grandes quantités de données manquantes affectent plus la reconstruction lorsqu'elles sont aléatoirement distribuées (figure 2.12 (a)(c)), et ce particulièrement lorsque le déplacement est plus complexe (figure 2.12 (c)). Dans le cas du déplacement post-sismique, la méthode est moins sensible à de grandes quantités de trous aléatoires comparé aux trous corrélés (figure 2.13).

Cas 3 : champs multiformes

Dans ce dernier cas d'étude, chaque champ de déplacement est une composition spatiale de quatre champs appelés *blocs* : un champ linéaire (g_1), deux champs d'ordre 3 (g_3) et un champ d'ordre 4 (g_4).

On effectue tout d'abord une seule reconstruction de l'ensemble des blocs. Comme dans les expériences précédentes, les valeurs manquantes sont initialisées par la moyenne spatiale. Le nombre optimal de modes estimé s'élève à 4 pour un champ perturbé par un bruit SCN et avec données manquantes aléatoires (figure 2.15 (a)) et à 6 pour un champ perturbé par un bruit STCN et des données manquantes corrélés (figure 2.15 (b)). On remarque au sein des champs reconstruits que les transitions entre chaque bloc sont conservées, même lorsque les trous corrélés s'étendent sur plusieurs blocs. Les résidus montrent une faible amplitude de déplacement (bloc g_4), ce qui correspond à une légère sous-estimation du nombre de mode. A l'inverse, le bloc g_1 reconstruit contient du bruit, ce qui correspond à une sur-estimation.

Afin de comparer ces résultats, les quatre blocs sont reconstruits séparement les uns des autres. Les résultats de cette reconstruction bloc par bloc sont exposés en figure 2.16. Les caractéristiques

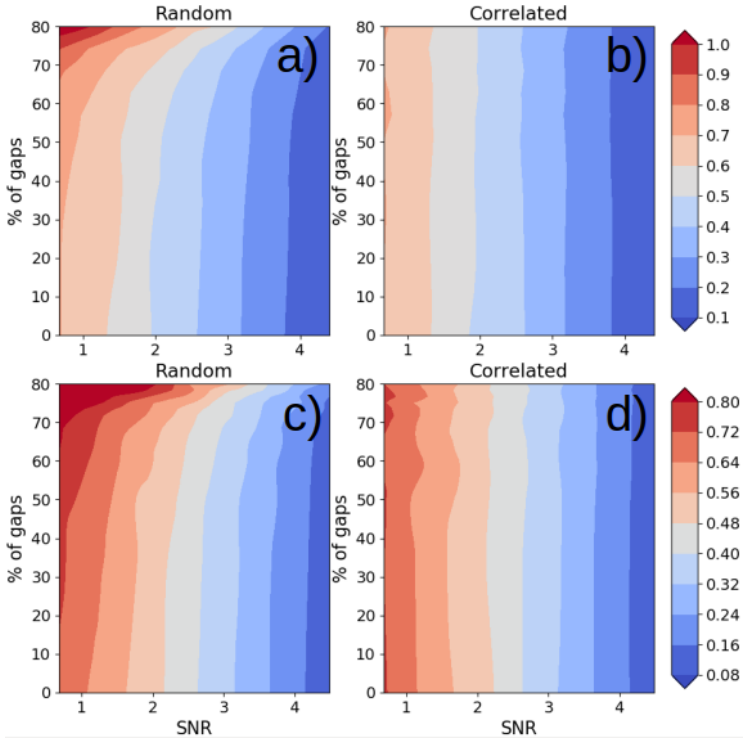


Figure 2.12 – Cartes d'erreur [cm] en fonction du % de données manquantes et du SNR dans le cas d'un champ du premier ordre (a)(b) et du second ordre (c)(d) perturbé par des trous aléatoires (a)(c) et corrélés (b)(d). Tous les déplacements sont également perturbés par un bruit STCN.

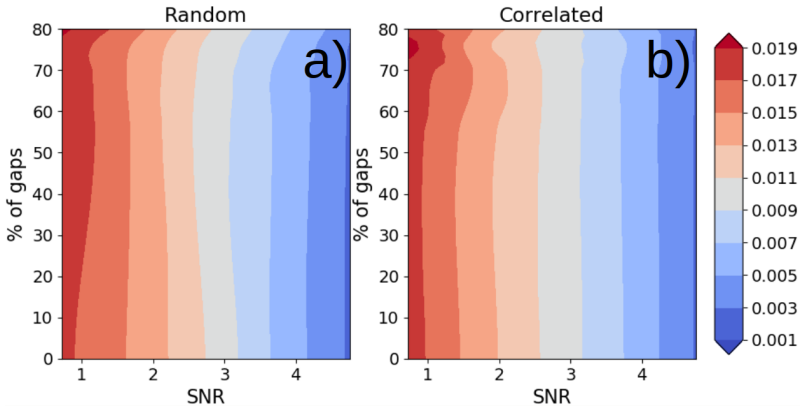


Figure 2.13 – Cartes d'erreur [cm] en fonction du % de données manquantes et du SNR dans le cas d'un déplacement post-sismique perturbé par des trous aléatoires (a)(c) et corrélés (b)(d). Tous les déplacements sont également perturbés par un bruit STCN.

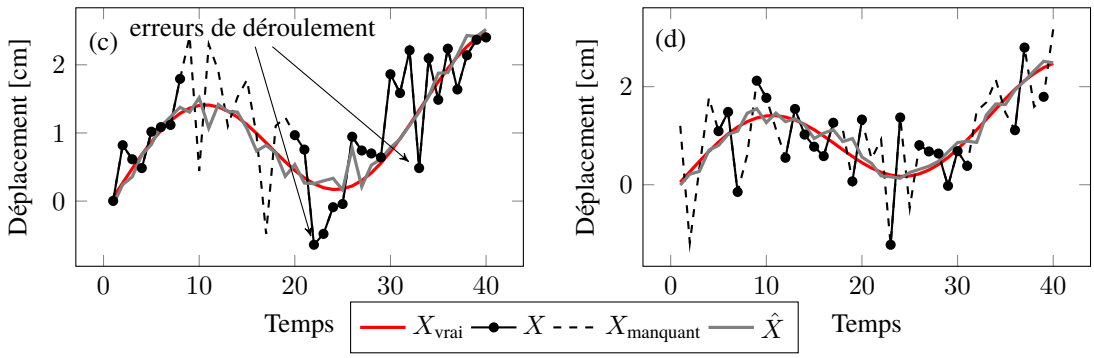


Figure 2.14 – Série temporelle d'un champ du troisième ordre perturbé par (a) données manquantes corrélées sur 10 date consécutives et (b) données manquantes aléatoires. Rouge : déplacement vrai ; cercles noirs : déplacement bruité avec données manquantes ; courbe hachée noire : données manquantes ; ligne grise : série temporelle reconstruite.

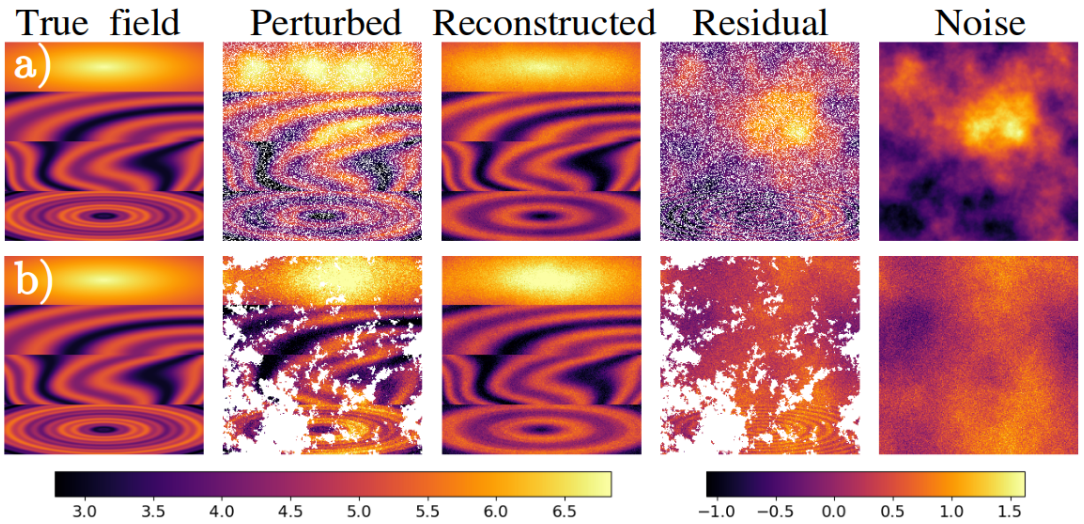


Figure 2.15 – Reconstruction d'un champ de déplacement [cm] multiformes avec 30% de données manquantes et SNR = 1.6. Le champ est composé de plusieurs blocs, de haut en bas : champ linéaire (g_1), deux champs d'ordre 3 (g_3) puis champ d'ordre 4 (g_4) (voir tableau 2.1). Les champs de déplacement sont perturbés par des données manquantes aléatoire et un SCN (a), puis des données manquantes corrélées et un STCN (b).

des déplacements reconstruits sont cette fois-ci mieux conservées, et aucune sur- ou sous-estimation n'est observée. Cela est notamment confirmé par le cas du champ d'ordre 4, où aucun signal de déplacement n'est observé dans les résidus.

On préférera donc, lorsque les zones de transition entre champs de déplacement sont bien délimitées (donc connues), une reconstruction bloc par bloc à une seule et même reconstruction. Il sera d'autant plus intéressant d'utiliser cette technique si la nature du déplacement diffère entre chaque bloc, ce qui impliquera potentiellement un nombre optimal de modes différent par bloc. A l'inverse, si les zones de transition ne sont pas clairement identifiables (comme dans beaucoup de champs de déplacement de surface), une seule reconstruction sera préférée afin d'éviter la création de toute discontinuité entre champs de déplacement.

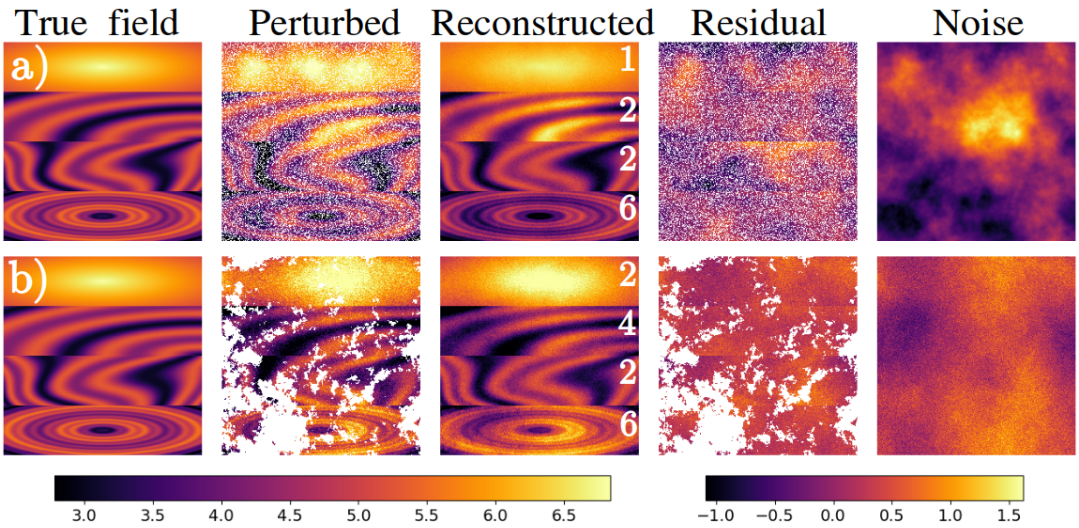


Figure 2.16 – Reconstruction bloc par bloc du champ de déplacement [cm] multiformes présenté en figure 2.15 (même perturbations). Les chiffres correspondent au nombre optimal de modes par bloc.

On pourra conclure cette étude de trois cas de la manière suivante : la qualité de reconstruction est dans l'ensemble plus dépendente du niveau de bruit (SNR) que de la quantité de données manquantes. Il a cependant été observé qu'en moyenne cette même qualité est, au-dessus d'un certain seuil de données manquantes (>60%), autant sensible au SNR qu'à la quantité de données manquantes. De plus, plus le déplacement est complexe (ordre 4 et plus), plus la méthode EM-EOF est sensible au bruit spatio-temporellement corrélé (STCN) qu'au bruit spatialement corrélé (SCN). Lors du premier cas, un problème de sur-estimation du nombre optimal de modes peut advenir : cela peut être régulé par l'utilisation de la métrique Λ. Enfin, lorsque le champ de déplacement est composé de plusieurs cibles physiques, chacune constituant un bloc indépendant de l'autre, une stratégie de reconstruction multiple peut être considérée, et ce en fonction des différences de dynamique entre chaque bloc et du nombre de bloc.

Temps de calcul

La méthode EM-EOF opère sur la covariance temporelle, de dimension $N \times N$. Le facteur limitant le temps de calcul est donc la dimension temporelle. Si R est le nombre optimal de modes et c une constante, l'algorithme EM-EOF a une complexité algorithmique de l'ordre de $\mathcal{O}(R^2N + c)$ puisque ce dernier repose essentiellement sur une EVD tronquée à $R < N$. Il est donc possible de traiter de grandes grilles spatiales dans des temps de calculs raisonnables. Quelques exemples de temps de calcul moyens sont indiqués en tableau 2.3, sur un Intel Xeon E5-2650 v3 à 2.3GHz (processeur standard d'un ordinateur portable).

Taille de l'image (pixels)	Distance en données Sentinel-1 (km)	Temps de calcul (s)
100×100	0.35×2.2	0.07
1000×1000	3.5×22.2	17.1
2000×2000	7×44.4	82.4 (1 min 22.4 s)
4000×4000	14×88.8	295.5 (4 min 55.5 s)
5000×5000	17.5×111	499 (8 min 19 s)

Tableau 2.3 – Temps de calculs moyens de l'algorithme EM-EOF pour une série temporelle de 40 images synthétiques avec 30% de données manquantes.

Comparaison avec d'autres méthodes d'interpolations

Afin d'évaluer plus largement les performances de la méthode EM-EOF, nous présentons une comparaison de l'erreur de reconstruction avec des méthodes d'interpolation classiques : l'interpolation au plus proche voisin (Nearest-Neighbor Interpolation, abrégé NNI) [Sibson1980] et le krigeage [Jones2001]. Ces deux méthodes d'estimation linéaire sont largement utilisées en géostatistique dans le traitement de champs spatiaux. Due au temps de calcul important de l'algorithme de krigeage, l'expérience est menée sur des petites grilles spatiales de 50×50. Le résultat de cette comparaison est présenté en figure 2.17. Dans tous les cas considérés, la méthode EM-EOF montre de meilleures performances en termes d'erreur, et ce d'autant plus que le SNR est bas (gain supérieur). On remarque également que le krigeage est moins sensible à de grandes quantités de données manquantes aléatoires que EM-EOF et NNI (figure 2.17 (a)), notamment parce que cette méthode repose sur la semi-variance et non la distance géométrique [Gentile2012], comme c'est le cas pour NNI. La hausse de l'erreur concernant EM-EOF se traduit par l'indisponibilité croissante de points temporels, l'occurrence d'un même point au sein de la série temporelle étant un gage de performance. Dans le cas de données manquantes corrélées (figure 2.17 (b)), les performances du krigeage sont similaires à celles de EM-EOF alors que l'algorithme NNI est inopérant.

2.4 Application sur données réelles

Il est naturel, après les simulations synthétiques, de se pencher sur des cas de données réelles. Le but de notre méthode, est, en partie, de fournir des champs de déplacement complets aux spécialistes qui étudient de plus près les phénomènes physiques qui dépendent directement du déplacement de surface. Dans le cas des glaciers alpins, les champs de vitesse sont particulièrement utiles aux modélisateurs afin de mieux contraindre le frottement basale.

La méthode EM-EOF est donc appliquée ici sur trois séries temporelles de mesures de déplacement dérivées d'images SAR Sentinel-1 A/B, constellations de deux satellites équipés chacun d'un radar à ouverture synthétique opérant en bande C (4-8 GHz). Ces jeux de données sont constitués de deux séries temporelles d'interférogrammes calculés à six jours sur les glaciers du Gorner (Suisse) et de Miage (Italie), et d'une série temporelle de mesures de déplacements calculés par corrélation d'amplitude sur des images SAR à douze jours d'intervalle sur le glacier d'Argentière en France (figure 2.18).

On choisit l'option d'une seule et même reconstruction par jeu de données, puisque chacun

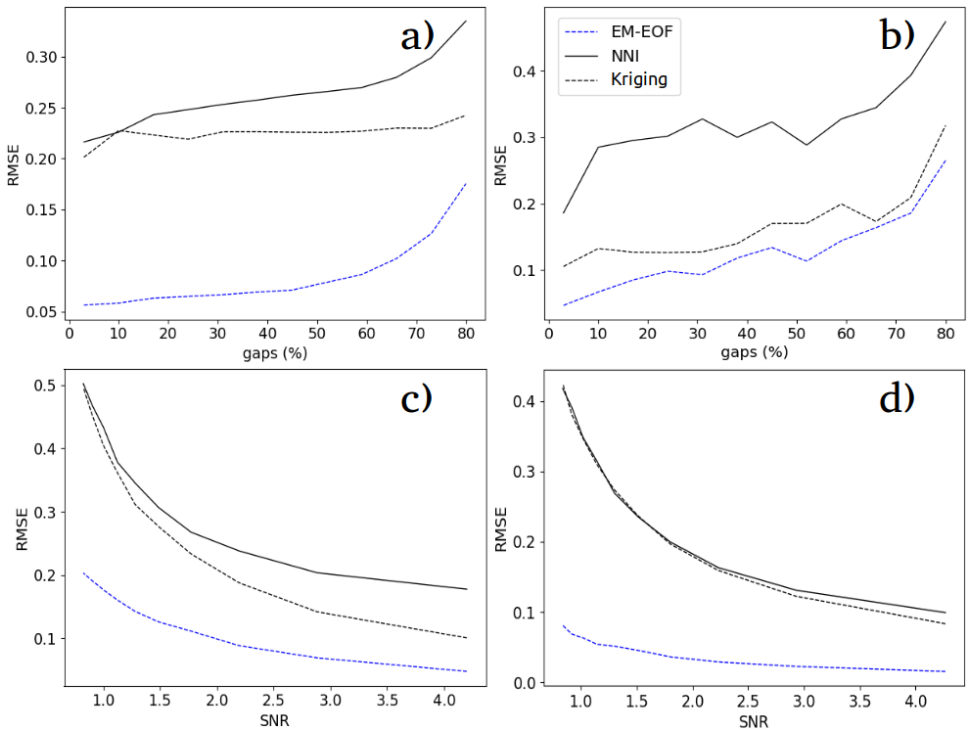


Figure 2.17 – Erreur moyenne (cross-RMSE) des méthodes EM-EOF, NNI et krigeage en fonction de la quantité de données manquantes et du SNR (100 simulations). Légende : (a) Données manquantes aléatoires, SNR = 2, SCN ; (b) données manquantes corrélées, SNR = 2, SCN ; (c) SCN, 30% de données manquantes aléatoires ; (d) STCN, 30% de données manquantes aléatoires.

d'entre eux contient une seule cible (équivalent au bloc des simulations synthétiques). Tous les jeux de données sont sujets à une incomplétude de données, elle-même dûe à un déplacement rapide de surface ou à des chutes importantes de neige durant l'acquisition des images. Comme lors des simulations, nous initialisons les données manquantes par la moyenne spatiale. On choisit également de fixer le nombre de points de validation croisée à 1% des points observés par champ de déplacement.

2.4.1 Glacier du Gorner

Ce jeu de données est constitué de seize interférogrammes traîtés puis calculés [3] à partir d'images Sentinel-1 A/B acquises entre novembre 2016 et mars 2017. Les données manquantes sont spatialement corrélées et varient entre 11,8 et 27,4% par interférogramme. La série temporelle contient également quatre interférogrammes manquants. La qualité du jeu de données varie d'un interférogramme à l'autre : on peut estimer, en se basant sur la cohérence, que douze des seize interférogrammes sont de bonne qualité (franges interférométriques correctement déroulées, peu de sauts de phase, etc.).

Le nombre optimal de modes EOF pour reconstruire cette série temporelle est de 3. Quelques exemples représentatifs des champs reconstruits sont présentés en figure 2.19 : le premier cas (première ligne) contient 14,6% de données manquantes, le second cas est un interférogramme manquant et le troisième cas contient 27,4% de données manquantes.

Concernant le premier cas, les valeurs reconstruites montrent un motif de déplacement en accord avec la tendance globale du champ non reconstruit. Pour ce qui est des valeurs observées, l'effet de filtrage est visible sans qu'il y ait de dégradation des valeurs de déplacement (perte/gain

3. Le traitement a été réalisé lors du stage de recherche de Master 2 de Rémi Prébet, voir [Prébet2017].

Figure 2.18 – Emplacement géographique des glaciers du Gorner (massif du Mont Rose), Miage et Argentière (massif du Mont-Blanc).

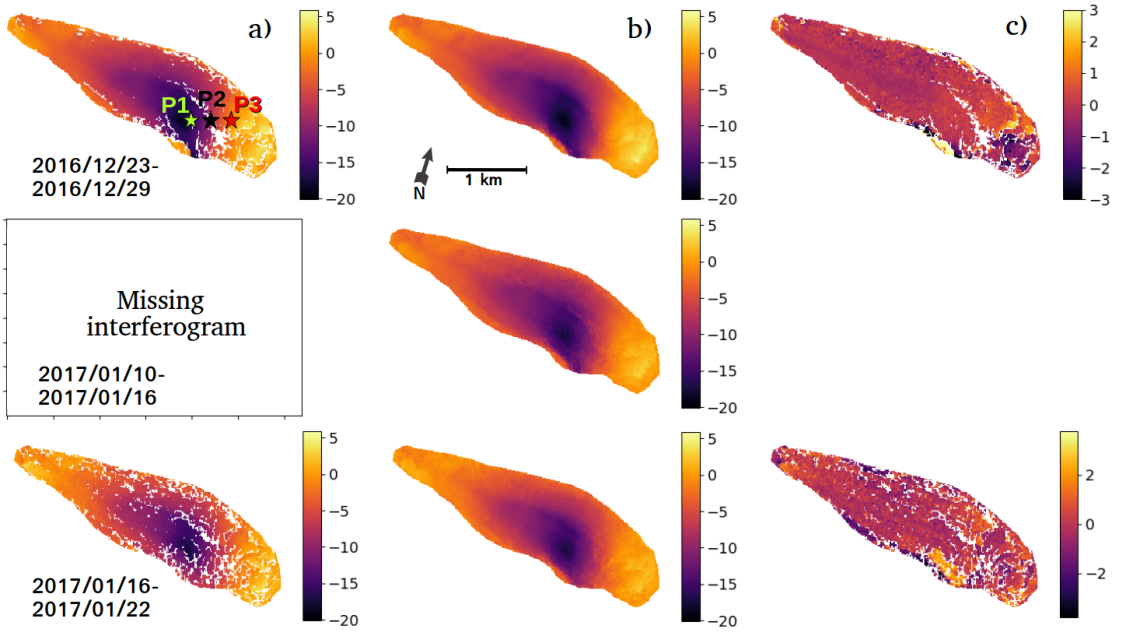


Figure 2.19 – Interférogramme initial (a) et reconstruit (b), et résidus (reconstruit-initial) (c) en géométrie radar sur le glacier du Gorner à trois intervalles de temps (2016/12/23-2016/12/29, 2017/01/10-2017/01/16, 2017/01/16-2017/01/22, format YYYY/MM/JJ). Les séries temporelles aux points P_1, P_2 et P_3 sont présentées en figure 2.20. Les valeurs de déplacement sont en centimètres dans la ligne de visée (line-of-sight, abrégé LOS) du radar.

d'amplitude, saut de valeur, etc.). Aucun signal relatif à un champ de déplacement n'est observé au sein des résidus, que l'on peut par ailleurs décrire comme homogènes et centrés en zéro sur une large partie du glacier. De plus grandes valeurs résiduelles sont observées près de la rive gauche du glacier, précisément là où résident des erreurs de déroulement de phase dûes à la transition abrupte entre la roche statique et la glace en mouvement. Cette zone est également sujette à des discontinuités dûes à une perte de cohérence.

L'évolution temporelle de points situés en P_1, P_2 et P_3 (voir figure 2.19) montre que la reconstruction suit les valeurs observées, même lors de fluctuations abruptes comme à la mi-novembre 2016 et à la mi-février 2017 (figure 2.20). A défaut de vérité terrain, la reconstruction au point P_2, qui contient peu d'observations temporelles, peut être validée par le point P_1 situé à proximité et contenant plus d'observation au cours du temps. On notera également la légère différence observée entre quelques valeurs observées et reconstruites : celle-ci est dûe à la propriété de filtrage de la méthode EM-EOF.

Afin de reconstruire l'interférogramme manquant (deuxième cas), la moyenne temporelle (au lieu de la moyenne spatiale qui ne peut être calculée) est ajoutée à l'anomalie spatiale (voir équation 2.2). Le motif de déplacement reconstruit semble en continuité avec les autres interférogrammes et est cohérent avec les interférogrammes obtenus récemment dans l'étude de [Prébet2019] (même jeu de données).

La reconstruction du cas numéro 3 montre également un résultat satisfaisant. Au sein de l'interférogramme initial, les données manquantes (dûes à une cohérence faible) ont induit des erreurs de déroulement de phase dans quelques zones localisées, notamment à l'amont du glacier. Il en résulte des discontinuités dans les valeurs résiduelles, où de grandes amplitudes sont notamment observées. Cependant, aucune valeur résiduelle n'est identifiable à un signal de déplacement.

Afin de garantir également une analyse quantitative des résidus, on calcule les moyennes et écart-types de ces derniers sur les points observés et sur les points de validation croisée. Les valeurs sont exposées dans le tableau 2.4. Ce tableau nous permet d'observer que les moyennes des résidus sont faibles : moins de 0.05 cm pour les points observés et moins de 0.2 cm pour les points de validation croisée. Cela confirme l'hypothèse selon laquelle la méthode EM-EOF ne dégrade pas les points observés de bonne qualité. La différence de valeur entre les deux catégories de points s'explique par la représentation statistique moindre des points de validation croisée (1% des valeurs observées). Concernant les mesures d'écart-types, celles-ci sont généralement inférieures à 1 cm sur les points observés et 1.8 cm sur les points de validation croisée. Les quelques hautes valeurs observées sont principalement dûes aux erreurs de déroulement de phase au sein des interférogrammes initiaux.

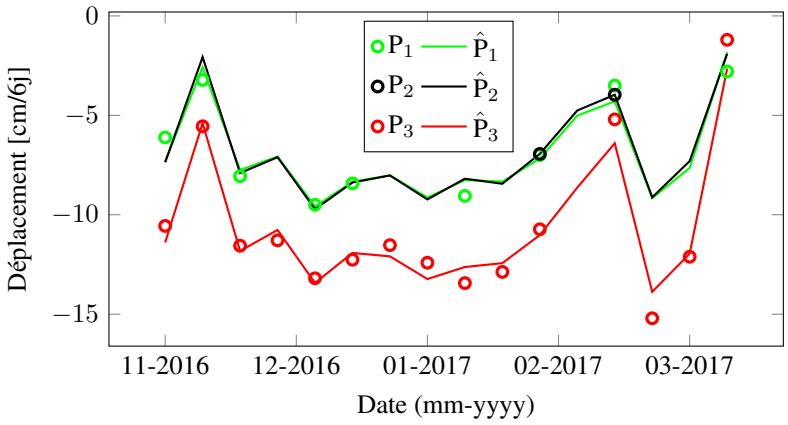


Figure 2.20 – Série temporelle de mesures de déplacement sur différentes zones P_1, P_2 et P_3 (figure 2.19) situées sur le glacier du Gorner ainsi que leur reconstruction $\hat{P}_1$, $\hat{P}_2$ et $\hat{P}_3$ par la méthode EM-EOF. Les cercles représentent les valeurs existantes alors que les lignes pleines correspondent aux valeurs reconstruites.

	Gorner				Miage			
	Observé		**CV**		**Observé**		**CV**	
Date	Moyenne	σ	Moyenne	σ	Moyenne	σ	Moyenne	σ
1	0.028	0.741	0.114	1.043				
2	x	x	x	x				
3	0.089	0.951	0.046	1.724				
4	0.008	0.814	0.088	1.369	0.015	0.984	-0.134	0.974
5	0.048	0.808	-0.158	1.273	0.0008	1.068	-0.029	0.920
6	-0.005	0.662	0.057	1.191	0.002	0.720	-0.058	0.632
7	0.045	0.859	0.024	1.342	0.002	0.885	0.173	0.904
8	0.022	0.921	-0.072	1.419	-0.004	1.029	0.051	1.186
9	-0.048	0.721	-0.05	0.856	0.011	1.137	-0.056	0.773
10	0.0006	0.863	-0.022	1.108	0.001	0.837	0.016	0.825
11	x	x	x	x	x	x	x	x
12	-0.031	0.667	0.015	0.593	-0.0003	0.935	0.015	0.959
13	x	x	x	x	-0.002	1.396	-0.103	1.329
14	-0.034	0.869	0.09	0.968	-0.008	1.030	-0.083	1.039
15	-0.001	0.906	-0.083	1.518	x	x	x	x
16	-0.085	1.032	-0.219	1.782	0.003	1.091	-0.046	0.804
17	-0.011	0.884	-0.118	1.462	-0.009	1.042	0.232	1.029
18	-0.034	0.749	0.002	1.038	-0.032	1.070	-0.074	1.208
19	x	x	x	x	x	x	x	x
20	0.004	1.228	0.026	1.374				

Tableau 2.4 – Moyenne (cm) et écart-type σ (cm) des champs de résidus des points observés et des points de validation croisée (abrégés CV) sur les glaciers du Gorner et de Miage. Le symbol 'x' indique les interférogrammes manquants. Les numéros de date s'étendent entre novembre 2016 et mars 2017.

2.4.2 Glacier de Miage

Treize interférogrammes ont été calculés [4] à partir de quatorze acquisitions Sentinel-1 consécutives de décembre 2016 à mars 2017. La quantité de données manquantes varie entre 11.4 et 23.1%. Beaucoup d'interférogrammes sont sujets à des données manquantes dans la partie centrale du glacier. Il y a également trois interférogrammes manquants dans la série temporelle. On notera également que la forme longiligne et étroite du glacier, en plus des discontinuités dûes à une perte de cohérence, rendent le déroulement de la phase interférométrique difficile. Ainsi, cinq interférogrammes sur treize ont été sujets à des sauts de phase lors du déroulement de phase. La correction de tels sauts de phase est rendue complexe par l'absence d'un autre jeu de données dont l'étendue spatiale et la précision soient similaires.

Des exemples représentatifs de la reconstruction sont présentés en figure 2.21. Le cas 1 (première ligne) contient 12.3% de données manquantes, le cas 2 (seconde ligne) en contient 18.6% alors que le cas 3 (troisième ligne) présente 23.1% de données manquantes. L'estimation du nombre optimal de modes est de 2. On observe globalement une similarité correcte entre les interférogrammes reconstruits et originaux. Dans le cas 1, le champ résiduel est homogène mais présente parfois de claires discontinuités. Cela est dû aux sauts de phase dans l'interférogramme initial, eux-même causés par une perte de cohérence dans la partie centrale du glacier. La série temporelle d'un point situé dans la zone de discontinuité (P_2, figure 2.21) ainsi que sa version reconstruite $\hat{P}_2$, sont tracées en figure 2.22. Ces deux séries, dont l'erreur est située en deçà de la limite de précision nominale de l'InSAR (<1cm), montrent des variations similaires. On notera que la plupart des décalages observés entre P_2 et $\hat{P}_2$ se justifient par les erreurs de déroulement de phase dans la zone de discontinuités à ces dates (Fig 2.21).

Dans le cas 2, la forme de déplacement est reconstruite avec succès, mais présente quelques discontinuités dans la zone de séparation des deux lobes terminaux (partie basse du glacier), qui est une zone perturbée par des données manquantes fortement corrélées temporellement. Une autre série temporelle est également tracée au point P_1 situé dans cette zone (figure 2.22). On observe ici que la méthode EM-EOF permet de reconstruire les interférogrammes en corrigeant notamment les sauts de phase. Le cas numéro 3 présente un interférogramme fortement dégradé, avec de grandes quantités de données manquantes et des sauts de phase à répétition. L'interférogramme reconstruit présente une forme de déplacement en accord et en continuité avec le reste de la série temporelle, même si l'amplitude des valeurs de déplacement dans la zone de séparation semble amoindrie par rapport à l'interférogramme initial. Les moyennes et écart-types des résidus sont exposés dans le tableau 2.4. L'observation est sensiblement la même que dans le cas du glacier du Gorner. Dans ce cas, les résidus sont centrés en zéro dans la plupart des cas (moins de 0.01 cm sur les parties observées et moins de 0.3 cm sur les points de validation croisée). De plus, les écart-types sont plus grands, ce qui est essentiellement dû aux sauts de phase et aux erreurs de déroulement localisées. En plus de la possibilité d'interpoler et de filtrer, la méthode EM-EOF peut détecter et corriger les inconsistences au sein du signal de déplacement sur chaque interférogramme de la série temporelle.

2.4.3 Étude d'un cas limite : le glacier d'Argentière

À partir d'une base de 66 images radar Sentinel-1 A/B acquises entre octobre 2016 et décembre 2017, nous avons généré 65 champs de déplacement par corrélation d'amplitude (voir sous-section 1.2.1) à un intervalle de douze jours [5]. Les images Sentinel-1 A/B en trajectoire ascendante ont été téléchargées depuis le portail Alaska Vertex [6] sur une zone de 250 km de large s'étendant entre le massif du Mont Blanc jusqu'au Valais suisse. L'intégralité de ces images sont préalablement

4. Le calcul des interférogrammes a été réalisé par Y. Yan.
5. Le recalage des images, ainsi que le calcul de la corrélation, ont été effectués à l'aide du logiciel de télédétection Gamma (www.gamma-rs.ch).
6. vertex.daac.asf.alaska.edu.

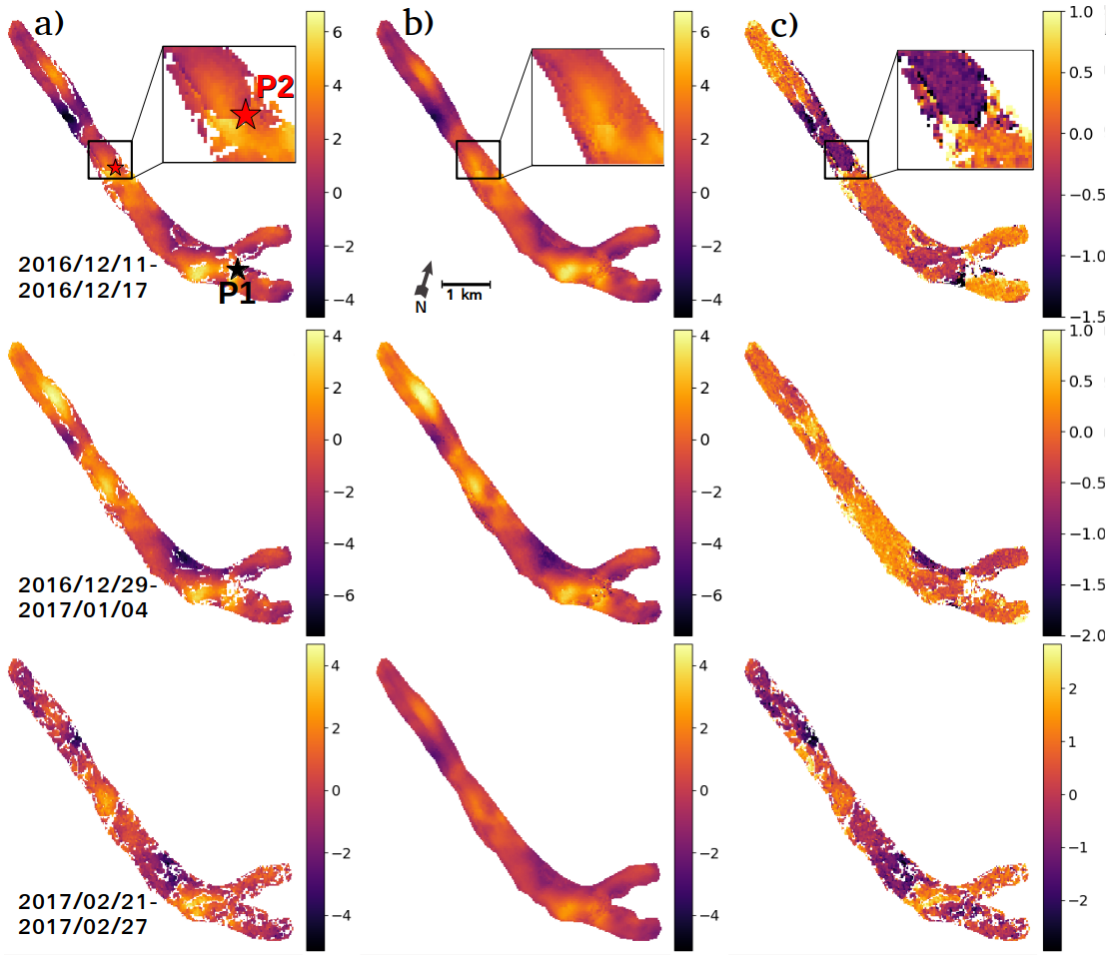


Figure 2.21 – Interférogramme initial (a) et reconstruit (b), et résidus (reconstruit-initial) (c) en géométrie radar sur le glacier de Miage à trois intervalles de temps (2016/12/11-2016/12/17, 2016/12/29-2017/01/04 et 2017/02/21-2017/02/27, format YYYY/MM/JJ). Les séries temporelles aux points P_1 et P_2 sont présentées en figure 2.22. Les valeurs de déplacement sont en centimètres dans la ligne de visée (LOS) du radar.

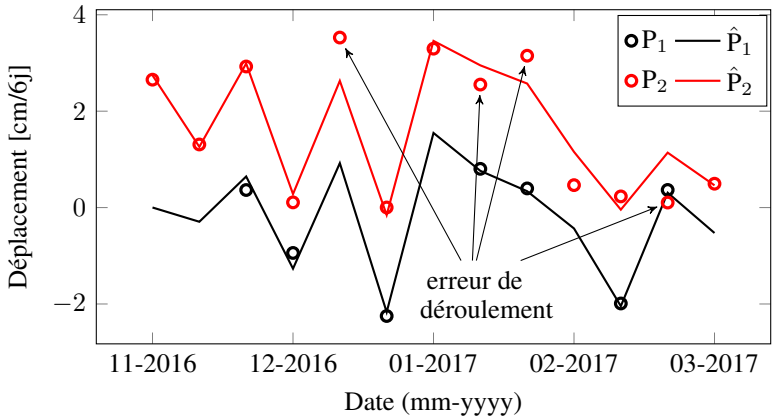


Figure 2.22 – Séries temporelles de mesures de déplacement sur différentes zones P_1 et P_2 (figure 2.21) situées sur le glacier de Miage, ainsi que leur reconstruction $\hat{P}_1$ et $\hat{P}_2$ par la méthode EM-EOF. Les cercles représentent les valeurs existantes alors que les lignes pleines correspondent aux valeurs reconstruites.

recalées sur la géométrie d'une image de référence afin de pouvoir comparer leurs informations respectives en les superposant dans un seul et même repère géométrique.

La corrélation d'amplitude s'effectue à l'aide d'une fenêtre de corrélation carrée de taille variable suivant la période d'acquisition (figure 2.23). Un seuillage sur les valeurs de la fonction de corrélation est par la suite appliqué afin de ne retenir que les valeurs de confiance. Les seuils sur la valeur de la fonction de corrélation sont déterminés empiriquement : il s'agit de minimiser la perte d'information tout en gardant un niveau de confiance acceptable. La valeur de seuil la plus souvent choisie est de 0.2. La série temporelle finale de 65 champs de déplacement contient des valeurs manquantes dues aux valeurs trop faibles de la corrélation qui n'ont pas atteint le seuil défini.

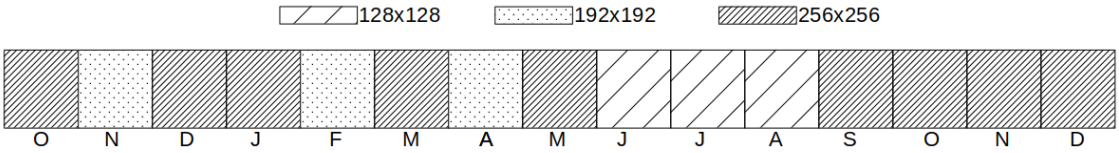
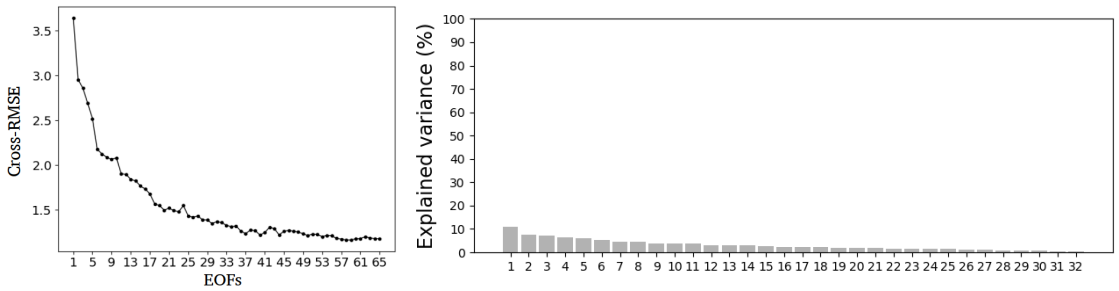


Figure 2.23 – Taille de la fenêtre de corrélation utilisée selon la période d'acquisition des images entre octobre 2016 et décembre 2017 (J : janvier, F : février, etc.).

Remarque Sur les 65 images que comporte la série temporelle, beaucoup sont concernées par des données manquantes localisées sur une partie du glacier, ce qui est identifiable à une corrélation spatio-temporelle des valeurs manquantes.

Ce cas constitue, de notre point de vue, un cas limite du fait de la qualité globale dégradée des données, et ce comparé aux jeux de données précédemment utilisés. Les deux facteurs dégradants sont ici le bruit (SNR faible) et la forte corrélation spatio-temporelle des données manquantes. En raison de l'orientation du glacier d'Argentière, l'amplitude de déplacement est supérieure dans la direction azimuthale (direction du mouvement de la plateforme satellite) que dans la direction de la ligne de visée (direction perpendiculaire au mouvement du satellite). Pour cette raison, nous privilégions l'analyse du déplacement en azimuth. La quantité de données manquantes par champ de déplacement varie entre 2% et presque 50%. Afin d'éclairer le lecteur sur la difficulté que représente ce cas d'étude, l'erreur est calculée en fonction du nombre de modes utilisés dans la reconstruction (figure 2.24 (a)). On y observe que le minimum de l'erreur correspond au mode 58, c'est-à-dire proche de la dimension de la covariance temporelle. Cela est essentillement dû au fort mélange entre le signal de déplacement et bruit. En effet, une grande proportion de modes étant dominés par un bruit spatialement corrélé, ceux-ci sont interprétés comme un potentiel signal de déplacement, faisant ainsi baisser l'erreur à chaque ajout d'un nouveau mode (figure 2.24 (b)). Trois exemples de champs reconstruits sont présentés en figure 2.25, avec un nombre de mode de 20 correspondant au résultat après l'étape 2 (minimum local dans l'erreur). Les cas 1, 2 et 3 (première, seconde et troisième ligne) contiennent respectivement 7.39%, 7.42% et 4.12% de données manquantes. La reconstruction montre, comme dans les cas d'étude précédents, une cohérence globale avec les champs initiaux. Malgré cela, l'exactitude des valeurs reconstruites dans les zones de données manquantes peut être remise en question : dans les cas 2 et 3, les valeurs reconstruites au sein des zones basses du champ de déplacement (partie haute du glacier) peuvent différer des valeurs voisines et ainsi marquer de franches discontinuités. Comme l'ont montré les simulations, la qualité de reconstruction dépend plus fortement du niveau de bruit que de la quantité de données manquantes. Nous suggérons donc que ces discontinuités sont l'effet du fort niveau de bruit, la quantité de données manquantes étant faible dans ce cas (<10%).



(a) (b)

Figure 2.24 – (a) Cross-RMSE $E(k)$ versus nombre de modes EOF k utilisés pour la reconstruction du jeu de données du glacier d'Argentière. Le minimum de $E(k)$ est atteint pour $k = 58$ (étape 1). Après l'étape 2, le nombre optimal de modes est réduit à 20 modes, ce qui correspond à un minimum local de $E(k)$. (b) Pourcentage de variance expliquée par chaque mode (jusqu'au numéro 32). L'énergie du système est bien répartie sur un grand nombre de modes, rendant ainsi difficile la tâche de sélection du nombre optimal de modes.

2.4.4 Comparaison avec les méthodes NNI et krigeage

La performance de la méthode EM-EOF dans le cas d'applications réelles est analysée au travers de la comparaison, entamée lors des simulations, avec les méthodes d'interpolation NNI et krigeage. La comparaison est menée sur le glacier du Gorner car ce jeu de données présente peu de sauts de phase. En effet, cela pourrait être préjudiciable quant à l'objectivité de la comparaison avec des méthodes d'interpolation n'utilisant que l'information spatiale. La figure 2.26 montre ainsi un exemple de reconstruction de l'interférogramme 2017/01/16-2017/01/22. Dû au temps de calcul important que nécessite le krigeage, la comparaison n'est menée que sur la partie haute du glacier, plus sujette à des données manquantes et à des sauts de phase. L'observation des résultats indique des conclusions similaires à celles des simulations : la méthode EM-EOF permet d'interpoler plus finement comparé aux méthodes NNI et krigeage, et ce plus spécifiquement dans les zones concernées par des données manquantes corrélées et des sauts de phase. L'analyse visuelle suggère dans ces cas une reconstruction plus cohérente avec les valeurs voisines et plus lissée par effet de filtrage des perturbations haute fréquence.

2.5 Conclusion

EM-EOF est une méthode itérative, sans information a priori sur le comportement spatio-temporel des données, pour l'interpolation de valeurs manquantes au sein de séries temporelles de champs de déplacement SAR. Le principe est le suivant. Tout d'abord, les données manquantes sont initialisées par la moyenne spatiale, puis la covariance temporelle est décomposée en modes EOF. Le nombre optimal de modes pour reconstruire les données incomplètes est ensuite estimé grâce au calcul d'une erreur sur une petite proportion des données, adaptée et proportionnelle au nombre de données observées (étape 1). Enfin, les valeurs manquantes estimées sont mises à jour jusqu'à ce que l'erreur converge (étape 2).

Les simulations, munies d'une analyse rigoureuse de l'erreur de reconstruction en fonction des facteurs potentiellement limitants (type et niveau de bruit, type et quantité de données manquantes), ont montré l'effacité de la méthode à l'égard de champs à complexité variable et comportant divers types de bruit et de données manquantes. Nous avons pu montrer empiriquement que la méthode EM-EOF est plus sensible au bruit qu'aux données manquantes, excepté dans le cas où ces

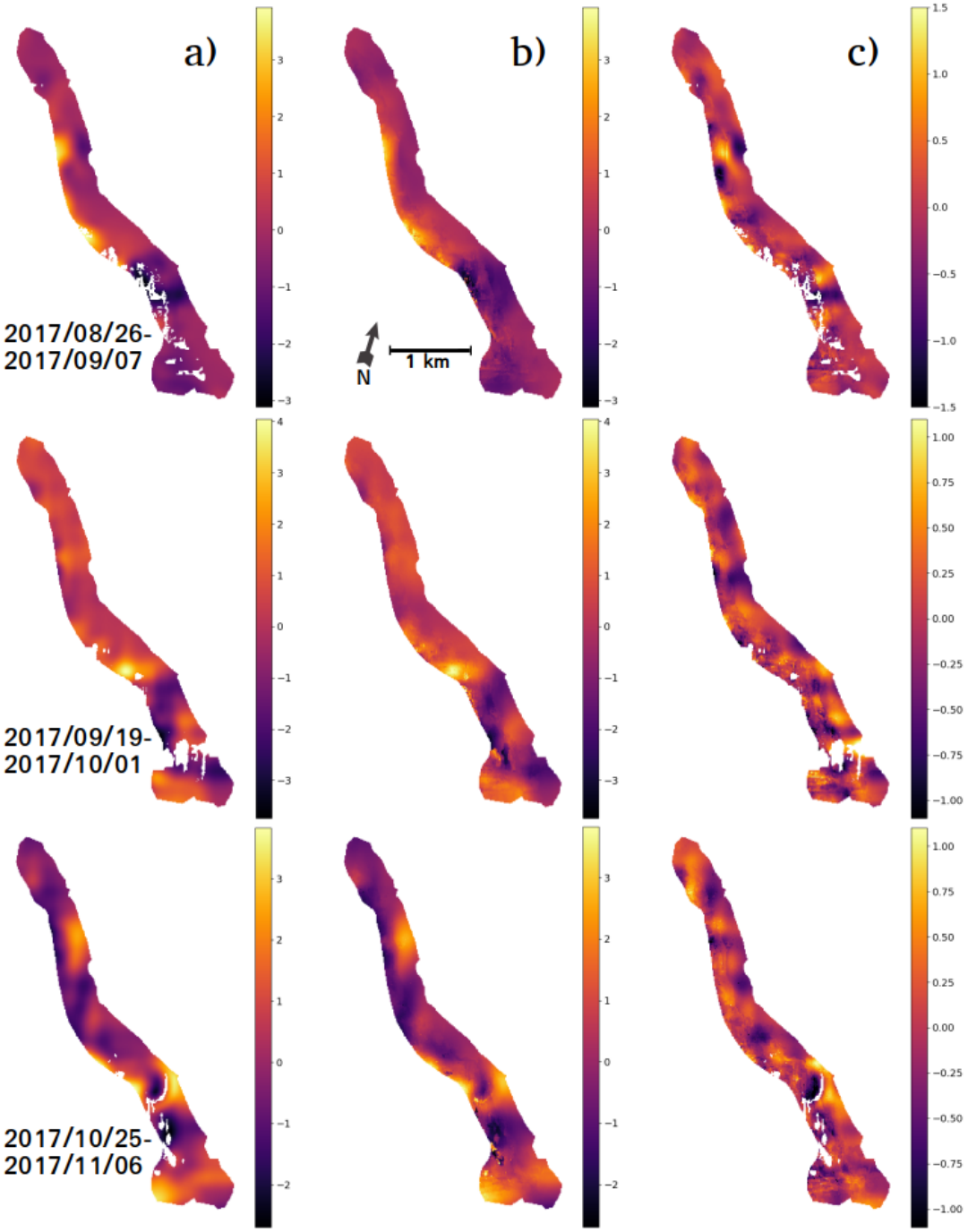


Figure 2.25 – Champ de déplacement (corrélation d'amplitude) initial (a), reconstruit (b) et résidus (reconstruit-initial) (c) en géométrie radar sur le glacier d'Argentière à trois intervalles de temps (2017/08/26-2017/09/07, 2017/09/19-2017/10/01 et 2017/10/25-2017/11/06, format YYYY/MM/JJ). Les valeurs de déplacement sont en mètres dans la direction azimuthale.

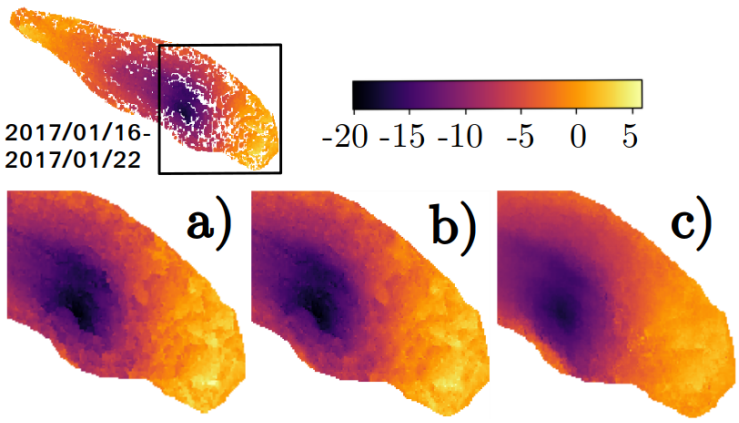


Figure 2.26 – Interférogramme reconstruit [cm] (2017/01/16-2017/01/22) sur la partie haute du glacier du Gorner par les méthodes NNI (a), krigeage (b) et EM-EOF (c).

dernières sont aléatoires et en quantité suffisamment importante, cas qui peut affecter la qualité de reconstruction. Lorsque le bruit est spatio-temporellement correlé, le nombre optimal de modes EOF peut être sur-estimé à cause de la difficulté à séparer le signal de déplacement du bruit, qui peuvent se comporter de manière similaire, tant spectralement que corrélativement.

L'application de la méthode EM-EOF sur un jeu de données réelles confirme également le potentiel de prise en charge de cas complexes combinant de multiples écueils : données manquantes à forte corrélation, interférogrammes manquants, rapport signal sur bruit bas, sauts de phase à répétition. En ce sens, cette méthode est tout à fait apte à augmenter la taille effective d'un jeu de données, et ce d'autant plus que les trous de données constituent un problème fréquent dans l'analyse de séries temporelles de champs de déplacement. Ainsi, cela pourra constituer une aide au géophysicien (modélisateur, glaciologue) pour mieux contraindre les paramètres rhéologiques (contrainte de frottement basal, hydrologie sous-glaciaire), dont la connaissance dépend directement de vitesses de surface continues en espace et en temps.

La mise en évidence d'un cas limite lors du dernier cas d'étude sur le glacier d'Argentière permet de constater la limite d'opérativité de la méthode. Cela s'exprime notamment par un signal de déplacement fortement contaminé par un bruit de type corrélé, ce qui est appuyé par une forte incertitude due à la basse résolution spatiale (22 m dans la direction azimutale en mode interférométrique), et un déplacement insuffisamment grand au regard de la capacité (précision) de la technique de corrélation d'amplitude.

Ces différentes conclusions nous amènent à continuer la réflexion sur la caractérisation et la reconstruction de structures fortement corrélées spatialement et corrompues par des données manquantes parfois persistantes (corrélées) sur des zones localisées. Le problème de sur-estimation, qui malgré la mise en oeuvre de la métrique Λ seuillée par le paramètre β, ne prend pas en compte le groupement des valeurs propres, appelé *multiplet*, et qui, comme souligné par [Hannachi2007], correspond à la description d'une variabilité particulière du signal. Dans la suite de ce manuscrit, nous proposons d'inclure de l'information supplémentaire dans le champ spatio-temporel $\mathbf{X}$. Cette *augmentation* de données conduit à une estimation d'une matrice de covariance spatio-temporelle. Ces travaux, comme nous allons le voir, font directement référence à l'analyse spectrale singulière (SSA) et sa version multivariée [Vautard1992,Ghil2002,Golyandina2010] introduite lors du chapitre 1 sous le nom de fonctions empiriques orthogonales étendues (section 1.4.2).

3

La méthode EM-EOF étendue

Sommaire

3.1 Introduction 66
3.2 La méthode EM-EOF étendue 67
3.2.1 Organisation et augmentation des données 67
3.2.2 Estimation et décomposition de la covariance spatio-temporelle 68
3.2.3 Reconstruction de la covariance spatio-temporelle 69
3.2.4 Sélection du nombre optimal de modes 70
3.2.5 Détermination du décalage spatial 73
3.2.6 Synthèse de la méthode EM-EOF étendue 74
3.3 Simulations numériques 75
3.3.1 Type de champ de déplacement 76
3.3.2 Type de perturbation et type de données manquantes 76
3.3.3 Paramètres de simulations 77
3.3.4 Résultats et discussion 78
3.3.5 Bilan de l'étude synthétique 91
3.4 Application sur données optiques : le cas du glacier Fox 93
3.5 Conclusion et perspectives 98

3.1 Introduction

Le but de ce chapitre est d'introduire une extension de la méthode EM-EOF (appelée EM-EOF étendue ci-après). Comme présentée au chapitre précédent, la méthode EM-EOF se base sur l'analyse en EOF de la covariance temporelle pour la reconstruction de données manquantes au sein de séries temporelles de déplacement. Cette technique, comme nous l'avons vu, décompose la covariance temporelle $\mathbf{\Sigma}$ puis estime, de manière itérative, un nombre optimal de modes pour extraire des tendances significatives à partir de champs de déplacement bruités. Des résultats prometteurs ont ainsi pu être présentés sur des séries temporelles de champs de déplacement de surface de glacier alpins calculés par interférométrie différentielle de couples d'images Sentinel-1.

Cependant, la méthode EM-EOF peut présenter certaines limites lorsque :

— la corrélation spatiale du champ de déplacement domine la corrélation temporelle ;

— le champ de déplacement présente des caractéristiques spatiales hétérogènes et locales ;

— la série temporelle est courte, renforçant ainsi la probabilité qu'une zone ou un pixel soit peu observé au cours du temps, ce qui est la conséquence de données manquantes fortement corrélées.

En particulier, comme la méthode EM-EOF utilise la décomposition de la covariance temporelle, des biais de reconstruction peuvent apparaître lorsqu'un pixel n'est jamais observé dans la série temporelle. Cela constitue une motivation supplémentaire à la prolongation de l'étude, notamment en faisant l'usage d'une covariance non pas temporelle mais spatio-temporelle pour prendre en compte les corrélations spatiale et temporelle du champ de déplacement.

À la manière des travaux dérivés de l'analyse spectrale singulière [Vautard1992, Ghil2002, Kondrashov2006, Golyandina2010], qui procèdent par une augmentation des données par fenêtrage (ou décalage) de dimension finie, nous utilisons l'information spatiale décalée et redondante afin d'augmenter une série temporelle de champs de déplacement. En cela, le principe de la 2D-SSA [Golyandina2010, Golyandina2015], qui se concentre sur une seule et même image, puis sur une série temporelle d'images [von Buttlar2014], est directement étendu pour le traitement de séries temporelles de champs de déplacement incomplets.

Du fait de la structure de la matrice de covariance spatio-temporelle augmentée, le lien direct entre les modes EOF et leur interprétation physique peut être rendu difficile. Un nouveau critère basé sur l'incertitude des valeurs propres du système est ainsi proposé afin de sélectionner le nombre optimal de modes, notamment afin de traîter et d'analyser finement le problème de sous et/ou sur-estimation du nombre de modes. Une approximation du décalage spatial utilisé pour augmenter les données spatialement est également fournie par une approche simple liant corrélation des données et théorie de l'estimation de covariance.

Les différentes sections de ce chapitre s'organisent ainsi : la méthode EM-EOF étendue est décrite en section 3.2 : organisation des données (section 3.2.1), estimation puis décomposition de la covariance spatio-temporelle (sections 3.2.2 et 3.2.3), sélection du nombre optimal de modes (section 3.2.4), détermination du décalage spatial (section 3.2.5) et enfin description de l'algorithme de type EM (section 3.2.6). L'application sur trois champs de déplacement synthétiques est ensuite exposée et discutée en section 3.3. La section 3.4 propose une application de la méthode sur des données réelles : des champs de vitesse de surface obtenus par corrélation d'images Sentinel-2 sur le glacier de Fox en Nouvelle-Zélande [Millan2019]. Enfin, un bilan de cette étude ainsi que quelques perspectives sont dressés en section 3.5.

3.2 La méthode EM-EOF étendue

Le principe général de la méthode EM-EOF étendue (figure 3.1) reprend celui de la méthode EM-EOF sous la forme de deux étapes. Après initialisation des valeurs manquantes, la première étape effectue une première estimation du nombre optimal de modes appelée R : cette estimation correspond au minimum de l'erreur de validation croisée (cross-RMSE) entre le champ reconstruit et le champ initial. La seconde étape est une mise à jour des valeurs manquantes et de l'estimation de nombre de modes par un algorithme de type Expectation Maximization (EM). À l'étape E, les valeurs manquantes sont estimées par leurs valeurs espérées sachant l'estimation de la covariance spatio-temporelle. À l'étape M, l'erreur de validation croisée est calculée : si cette erreur baisse par rapport à l'itération précédente, les valeurs manquantes sont mises à jour par les valeurs manquantes estimées lors de l'étape E. On obtient alors en sortie un nombre optimal de modes mis à jour $r \leq R$. Après les étapes 1 et 2 (figure 3.1), une mesure de confiance basée sur l'incertitude d'estimation des valeurs propres est calculée puis associée à r : en fonction de la valeur et des caractéristiques de cette mesure de confiance, qui seront détaillées plus loin, le nombre de modes est mis à jour ou non en retournant à l'étape 2.

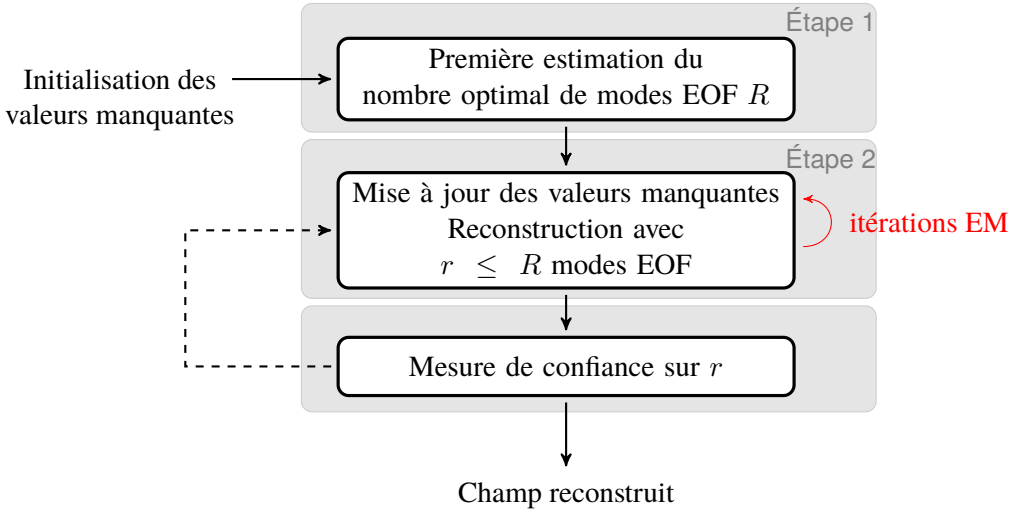


Figure 3.1 – Diagramme simplifié de la méthode EM-EOF étendue.

3.2.1 Organisation et augmentation des données

Soit $\mathbf{X}_t$ un champ spatial [1] de taille $P_x \times P_y = P$ observé aux temps discrets $t = 1, \ldots, N$. On note $x_{ij}(t), 1 \leq i \leq P_x, 1 \leq j \leq P_y$ chaque élément de $\mathbf{X}_t$ à la position (i, j). Chaque champ $\mathbf{X}_t$ peut être ordonné au sein d'une matrice de données spatio-temporelle :

$$\mathbf{Y} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N) \tag{3.1}$$

En pratique, $\mathbf{Y}$ a une moyenne nulle, ce qui signifie que sa moyenne spatiale lui a été retirée au préalable. Chaque $\mathbf{X}_t$ est ensuite augmenté en une matrice *Hankel-block Hankel* (HbH), c'est-à-dire une matrice de Hankel par bloc de sous-matrices [2]. Cette matrice, que l'on note $\mathbf{D}_t$, est de taille $K_x K_y \times M_x M_y$, avec $K_x = (P_x - M_x + 1)$, $K_y = (P_y - M_y + 1)$, où $M_x \times M_y$ est la taille d'une fenêtre glissante à deux dimensions :

1. $\mathbf{X}_t$ peut correspondre, par exemple, au vecteur colonne $\mathbf{x}_t$ défini au chapitre 2, avant ordonnancement.
2. On rappelle qu'une matrice de Hankel est une matrice carrée dont les anti-diagonales sont constantes. Une matrice Hankel-block Hankel est constituée de blocs identiques sur ses anti-diagonales.

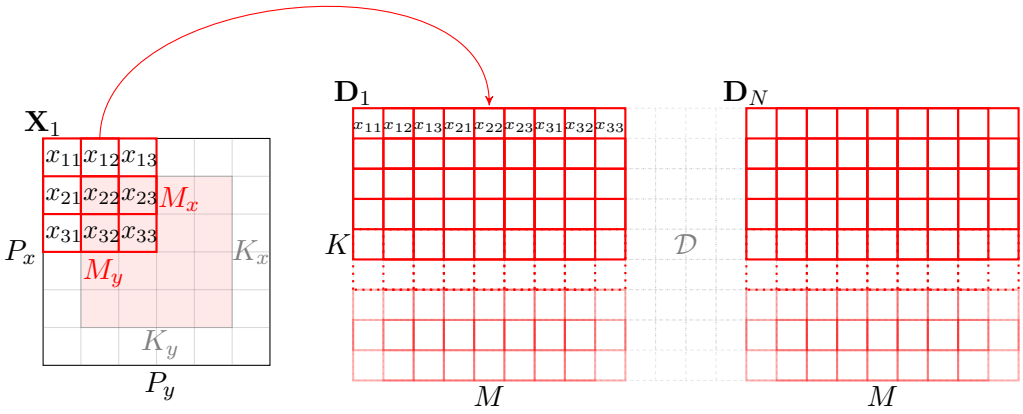


Figure 3.2 – Illustration de l'augmentation spatiale des champs $(\mathbf{X}_t)_{1\leq t\leq N}$ à l'aide d'une fenêtre glissante de taille $M_x \times M_y$. Le champ $\mathbf{X}_1$ est ici augmenté en une matrice $\mathbf{D}_1$ de taille $K_xK_y \times M_xM_y$, laquelle est stockée dans une grande matrice spatio-temporelle $\mathcal{D}$. Chaque $\mathbf{D}_t$ correspondant à $\mathbf{X}_t$ est ensuite ordonné en ligne, ce qui résulte en une matrice de taille $(K \times NM)$.

$$\mathbf{D}_t = \begin{pmatrix} \mathbf{H}_{1,t} & \mathbf{H}_{2,t} & \dots & \mathbf{H}_{M_x,t} \\ \mathbf{H}_{2,t} & \mathbf{H}_{3,t} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{H}_{K_x,t} & \dots & \dots & \mathbf{H}_{P_x,t} \end{pmatrix} \tag{3.2}$$

Chaque sous-matrice $\mathbf{H}_{i,t}$ est une matrice de Hankel de taille $K_y \times M_y$ définie par :

$$\mathbf{H}_{i,t} = \begin{pmatrix} x_{i1}(t) & x_{i2}(t) & \dots & x_{i,M_y}(t) \\ x_{i2}(t) & x_{i3}(t) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ x_{i,K_y}(t) & \dots & \dots & x_{i,P_y}(t) \end{pmatrix} \tag{3.3}$$

Dans un souci d'allègement de la notation, nous noterons $K = K_xK_y$, $M = M_xM_y$ et $P = P_xP_y$ par la suite. De manière similaire à $\mathbf{X}_t$ et $\mathbf{Y}$, chaque matrice $\mathbf{D}_t$ est à son tour ordonnée au sein d'une grande matrice spatio-temporelle de taille $(K \times NM)$ (voir figure 3.2) :

$$\mathcal{D} = (\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_N) \tag{3.4}$$

En référence à la littérature en analyse spectrale singulière multi-variée (M-SSA) [Broomhead1986, Vautard1992, Ghil2002], $\mathcal{D}$ est appelée matrice de donnée augmentée. La différence réside ici dans le fait que chaque $\mathbf{X}_t$ est augmenté spatialement et non temporellement. La fenêtre uni-dimensionnelle de taille M utilisée en M-SSA pour augmenter une série temporelle est maintenant une fenêtre à deux dimensions de taille $M_x \times M_y$ [Golyandina2010]. La matrice augmentée est donc spatio-temporelle dans le sens où sa structure est une alternation de blocs temporels au sein desquels sont imbriqués des blocs spatiaux augmentés.

3.2.2 Estimation et décomposition de la covariance spatio-temporelle

A partir de la matrice augmentée $\mathcal{D}$, on peut calculer la covariance empirique augmentée par :

$$\mathbf{\Sigma} = \frac{1}{K}\mathcal{D}^T\mathcal{D} \tag{3.5}$$

$\mathbf{\Sigma}$ est une matrice de taille $NM \times NM$, symétrique, réelle et définie positive. Ses valeurs propres sont par conséquent réelles et positives, rangées par ordre décroissant : $\lambda_1 > \lambda_2 > \cdots > \lambda_{NM}$. Chaque entrée de $\mathbf{\Sigma}$ à la position (i, j), notée σ_{ij}, peut être exprimée en fonction de chaque entrée de $\mathcal{D}$:

$$\sigma_{ij} = \frac{1}{K}\sum_{k=1}^{K}\mathcal{D}_{ki}\mathcal{D}_{kj}, \qquad i, j = 1, \ldots, NM \tag{3.6}$$

La décomposition en valeurs propres de la matrice $\mathbf{\Sigma}$ s'écrit de manière commune :

$$\mathbf{\Sigma} \overset{\text{EVD}}{=} \sum_{i=1}^{NM}\lambda_i\mathbf{u}_i\mathbf{u}_i^T \tag{3.7}$$

Les vecteurs $\mathbf{u}_i$ sont les NM fonctions empiriques orthogonales étendus (EEOF) de la matrice $\mathcal{D}$. Chaque terme de l'équation (3.7), qui provient du théorème de représentation spectral, est un mode EOF [Hippert-Ferrer2020] (voir également chapitre 2). La fraction de la variance totale expliquée par un mode i est indiquée par sa valeur propre λ_i. Cette fraction peut être facilement calculée par $\lambda_i / \sum_i \lambda_i$. De manière générale, les premiers modes représentent la plus grande part de variabilité du signal dans $\mathcal{D}$. Les valeurs propres λ_i fournissent également une information importante quant à la distribution de la puissance spectrale du signal, formule que l'on peut englober sous le terme d'énergie.

3.2.3 Reconstruction de la covariance spatio-temporelle

Afin de reconstruire la série temporelle de champs 2D, il faut tout d'abord définir les composantes principales (PC) $\{\mathbf{a}_k\}_{1\leq k\leq NM}$, qui sont, comme dans l'analyse en EOF, la projection de la matrice de données (augmentée) sur chaque (E)EOF, et dont chaque élément est défini par :

$$a_{ik} = \sum_{j=1}^{NM}\mathcal{D}_{ij}u_{jk}, \qquad i = 1, \ldots, K \tag{3.8}$$

où u_{jk} On notera que chaque PC a une taille K. La matrice augmentée peut être partiellement ou totalement reconstruite en projetant les PC sur les vecteurs propres étendus. Si $\mathbf{A}$ est la matrice $K \times NM$ contenant tous les PC sur ses colonnes et si $\mathbf{U}$ est la matrice contenant les vecteurs propres étendus, obtenus par EVD de $\mathbf{\Sigma}$ (équation (3.7)), alors la reconstruction prend la forme matricielle $\hat{\mathcal{D}} = \mathbf{A}\mathbf{U}^T$, où chaque élément de $\hat{\mathcal{D}}$ est donné par :

$$\hat{\mathcal{D}}_{ij} = \sum_{k=1}^{NM}a_{ik}u_{jk}^T \tag{3.9}$$

Avant reconstruction de $\mathbf{Y}$, un moyennage sur les diagonales [3] est appliqué sur chaque matrice $\mathbf{H}_{i,t}$ et $\mathbf{D}_t$, que l'on exprime de la manière suivante :

$$x_{ik}(t) = \frac{1}{\#\mathcal{A}_k}\sum_{(l,l')\in\mathcal{A}_k}x_{ll'}(t) \tag{3.10}$$

$$\mathbf{H}_{k,t} = \frac{1}{\#\mathcal{B}_k}\sum_{(l,l')\in\mathcal{B}_k}\mathbf{H}_{ll',t} \tag{3.11}$$

3. Cette opération est appelée *hankelization*.

où $\mathcal{A}_k = \{(l,l') : 1 \leq l \leq K_y, 1 \leq l' \leq M_y, l+l' = k+1\}$ et $\mathcal{B}_k = \{(l,l') : 1 \leq l \leq K_x, 1 \leq l' \leq M_x, l+l' = k+1\}$. Cela signifie que le moyennage a tout d'abord lieu sur chaque bloc de $\mathbf{H}_{i,t}$, puis entièrement sur $\mathbf{H}_{i,t}$ [Golyandina2010], ce qui est cohérent avec la structure de Hankel par bloc.

Le champ reconstruit, noté $\hat{\mathbf{Y}}$, consiste finalement à inverser le processus d'ordonnancement décrit en figure 3.2 en procédant par correspondance directe entre chaque $\hat{\mathbf{D}}_t$ et chaque $\hat{\mathbf{X}}_t$.

Il est important de mentionner que la troncature de l'expression (3.9) par un nombre $R \ll MN$ de modes permet de filtrer la partie indésirable et/ou non significative du signal (par exemple le bruit). Le choix d'une troncature appropriée, qui revient à détecter la structure rang faible de la matrice de covariance $\mathbf{\Sigma}$, est l'objet de la prochaine sous-section.

3.2.4 Sélection du nombre optimal de modes

Plusieurs techniques sont utilisées afin de sélectionner le nombre optimal de modes. Premièrement, on utilise une erreur de validation croisée entre le champ reconstruit et le champ initial, ce qui permet de ne pas dépendre de la vérité terrain et d'obtenir une estimation du nombre optimal de modes. Ensuite, la métrique Λ (chapitre 2, section 2.2.4) permettant de traiter le problème de sur-estimation du nombre de modes est utilisée. Enfin, une mesure de confiance est construite à partir de l'incertitude des valeurs propres, lesquelles procurent à la fois une information sur la variation des valeurs propres et sur la répartition en énergie du système. Cette mesure permet ainsi d'affiner la sélection du nombre optimal de modes puisque ces derniers sont directement liés aux valeurs propres.

Validation croisée

Afin d'estimer le nombre optimal de modes, on utilise une erreur sur des données de validation croisée (cross-RMSE) parmi les données existantes $\mathbf{Y}_{\text{obs}}$. Ces données de validation, notées $\mathbf{y}_{\text{cv}} \in \mathbf{Y}_{\text{obs}}$, sont choisies aléatoirement à hauteur de 1% du nombre total de points observés, puis sont retirées des données et gardées en copie. La cross-RMSE est la norme ℓ_2 de la différence entre $\mathbf{y}_{\text{cv}}$ et son estimation basée sur k modes EEOF $\hat{\mathbf{y}}_{\text{cv},k}$, puis normalisée par la taille de $\mathbf{y}_{\text{cv}}$. Si δ_k désigne cette erreur, alors celle-ci s'exprime par :

$$\delta_k = \frac{1}{\sqrt{Q}} ||\hat{\mathbf{y}}_{\text{cv},k} - \mathbf{y}_{\text{cv}}||_2 \tag{3.12}$$

L'ensemble des cross-RMSE $\{\delta_k\}_{1 \leq k \leq NM}$ est calculé, le nombre optimal de modes étant celui qui correspond à la valeur minimale dans cet ensemble.

Biais de surestimation

Comme évoqué au chapitre 2, un mélange fort entre bruit corrélé et signal brut (par exemple un signal de déplacement) peut conduire à une sur-estimation du nombre de modes. Cela est plus précisément dû à la contamination des données de validation croisée par du bruit [Ng1997], lesquelles sont utilisées comme base de l'estimation du nombre optimal de mode. Afin de se prémunir contre ce problème, le critère suivant a été proposé dans le chapitre 2, section 2.2.4 :

$$\Lambda = 1 - \frac{\delta_{k+1}}{\delta_k} \tag{3.13}$$

Cette quantité permet de mesurer la variation négative de la cross-RMSE lors de l'ajout d'un mode supplémentaire dans la reconstruction (3.9). Une petite variation de Λ implique que peu d'information est apportée au nouveau champ reconstruit par l'ajout d'un nouveau mode : dans ce cas, ce mode n'est pas pris en compte. On pourra alors définir un seuil sur Λ en-dessous

duquel l'algorithme s'arrête. Cela signifie que l'erreur δ_{k+1} aura peu varié par rapport à δ_k : on sélectionnera alors k modes et non $k+1$ modes pour reconstruire les données.

Mesure de confiance

Incertitude des valeurs propres Si le critère Λ est sensible aux variations de la cross-RMSE, ce dernier ne permet pas de prendre en compte la structure du spectre de valeurs propres dans le choix du nombre de modes. En effet, la manière dont varient les valeurs propres peut fournir une information utile quant à la répartition en fréquences du signal de déplacement.

Par exemple, les valeurs propres sont dites *dégénérées* si elle possèdent une amplitude similaire. La dégénéréscence des valeurs propres rend l'interprétation des EEOF difficile puisque toute combinaison linéaire de ces EEOF est également un EEOF, ce qui conduit à un mélange des EEOF [Hannachi2007]. Le phénomène inverse à la dégénéréscence est la séparation des valeurs propres, laquelle peut être synonyme d'un changement de variabilité spatio-temporelle entre une valeur propre et la suivante, ou un groupe de valeurs propres et le suivant. Deux ou plusieurs valeurs propres consécutives, groupement appelé *multiplet*, sont dégénérées lorsque l'incertitude d'une valeur propre est du même ordre ou plus grande que la différence entre cette valeur propre et sa plus proche voisine. Par conséquent, toute étude de la dégénéréscence des multiplets devra tout d'abord s'intéresser à l'estimation de l'incertitude des valeurs propres du système. La première étude à ce sujet [North1982] a ainsi proposé une règle empirique afin de fournir une approximation de l'incertitude des valeurs propres. Cette règle peut être résumée par les équations suivantes :

$$\Delta\lambda_k \approx \sqrt{\frac{2}{L^*}}\lambda_k \tag{3.14}$$

$$\Delta\mathbf{u}_k \approx \frac{\Delta\lambda_k}{\lambda_j - \lambda_k}\mathbf{u}_j \tag{3.15}$$

où λ_j est la valeur propre la plus proche de λ_k, $\mathbf{u}_j$, $\mathbf{u}_k$ sont les vecteurs propres correspondants (EEOF) et L^* est le nombre d'observations indépendantes au sein de l'échantillon spatio-temporel, aussi appelé *taille effective d'échantillon* (effective sample size, abrégée ESS ci-après). $\Delta\lambda_k$ désigne l'incertitude de la valeur propre λ_k et $\Delta\mathbf{u}_k$ celle du vecteur propre $\mathbf{u}_k$. L'intervalle d'incertitude de λ_k est ainsi donné par $\lambda_k \pm \Delta\lambda_k$. L'interprétation de l'équation (3.15) est la suivante : si l'incertitude de la valeur propre λ_k est proche de la différence entre cette valeur propre et sa plus proche voisine, alors les vecteurs propres correspondant ont de fortes chances d'être contaminés l'un par l'autre. Cette contamination existe par exemple lorsque deux EEOF décrivent ensemble la même variabilité spatio-temporelle (comme une tendance ou une oscillation) ou si le signal est perturbé par un bruit corrélé, ce qui aurait pour effet de disperser la variance du système sur l'ensemble du spectre.

Estimation de l'ESS Afin de fournir une estimation de L^*, celle-ci est séparée en deux termes multiplicatifs de sorte que $L^* = N^*M^*$. N^* correspond à l'ESS temporelle alors que M^* correspond à l'ESS spatiale. [Thiébaux1984] ont formulé une estimation de N^* pour l'analyse de séries temporelle, définie par :

$$N^* = N\left[1 + 2\sum_{k=1}^{N-1}(1 - \frac{k}{N})\rho(k)\right]^{-1} \tag{3.16}$$

où $\rho(k)$ est l'autocorrélation temporelle de la série temporelle étudiée et N est le nombre d'observations temporelles. Cette définition est valable pour une série temporelle univariée composée de N observations, comme un pixel d'une image dont l'amplitude varie au cours du temps. A partir de cette définition, nous estimons l'ESS spatiale M^* au sein de chaque fenêtre spatiale de taille

$M = M_x \times M_y$. Si ν désigne l'autocorrélation spatiale moyenne de la série temporelle, M^* peut être estimé par :

$$M^* = M\left[1 + 2\nu \sum_{k=1}^{M}(1 - \frac{k}{M})\right]^{-1} \tag{3.17}$$

Cette forme est analogue à la forme de l'ESS temporelle (3.16). Une simplification permet de réduire l'ESS spatiale à l'expression suivante :

$$M^* = \frac{M}{1 + \nu(M-1)} \tag{3.18}$$

L'autocorrélation spatiale moyenne ν est la moyenne des indices I de Moran du champ $\mathbf{X}_t$, noté I_t, qui permet de caractériser la corrélation entre des points spatialement proches :

$$\nu = \frac{1}{N}\sum_{t=1}^{N} I_t = \frac{1}{N}\sum_{t=1}^{N} \frac{N}{W} \frac{\sum_i \sum_j w_{ij}\, \mathrm{vec}(\mathbf{X}_t)_i\, \mathrm{vec}(\mathbf{X}_t)_j}{\sum_i \mathrm{vec}(\mathbf{X}_t)_i^2} \qquad i,j = 1,\ldots,P \tag{3.19}$$

où $\mathrm{vec}(.)$ est l'opérateur vectorisation [4], w_{ij} sont les poids de toutes les paires de points définis par l'inverse de la distance au carré entre les points aux positions i et j :

$$w_{ij} = \begin{cases} \frac{1}{|i-j|^2} & i \neq j \\ 0 & \text{sinon,} \end{cases}$$

et où W est la somme de tous les poids pour toutes les paires de points, soit $W = \sum_i \sum_j w_{ij}$. La définition des poids w_{ij} varie selon une hypothèse sur la corrélation spatiale du champ en question. L'hypothèse prise en compte ici est que le champ possède bien une autocorrélation spatiale et que les structures spatiales qui le composent ont une dépendance décroissante à d'autres structures en fonction de la distance entre ces structures, et ce dans un rayon restreint (longueur de corrélation). Au-delà de cette longueur de corrélation, la dépendance entre structures est nulle. Cela justifie donc le choix de poids indexés sur l'inverse de la distance au carré.

Mesure de confiance À partir de l'estimation des incertitudes des valeurs propres (équation (3.15)), un indice de confiance $\mathcal{C}_k$ associé à chaque valeur propre λ_k peut être calculé dans l'intervalle $[0, 1]$, prenant ainsi la forme :

$$\mathcal{C}_k = \frac{\max(\Gamma_k) - \Gamma_k}{\max(\Gamma_k) - \min(\Gamma_k)} \qquad k = 1,\ldots,NM \tag{3.20}$$

où Γ_k est donné par :

$$\Gamma_k = \log\left(\frac{\Delta\lambda_k}{\lambda_j - \lambda_k}\right) \tag{3.21}$$

$\mathcal{C}_k$ permet de détecter la dégénérescence et/ou la séparation des valeurs propres dans le spectre de la matrice augmentée $\mathcal{D}$, qui correspondent respectivement à des valeurs basses et hautes de $\mathcal{C}_k$. Ainsi, chaque pic au sein de $\mathcal{C}_k$ coïncide avec une séparation plus ou moins marquée entre deux multiplets. A l'inverse, les valeurs situées de part et d'autre des pics (les creux) correspondent à une dégénérescence de multiplet, ou plus largement à des valeurs propres dont l'amplitude est proche.

4. La vectorisation d'une matrice est la transformation linéaire d'une matrice en un vecteur colonne. Si $\mathbf{A}$ désigne une matrice de taille $N \times P$, $\mathrm{vec}(\mathbf{A})$ est le vecteur de taille $NP \times 1$ contenant les colonnes de $\mathbf{A}$ apposées les unes aux autres.

Afin d'ajuster le nombre optimal de modes r issu de l'étape 2 (figure 3.1), l'indice $\mathcal{C}_k$ est calculé pour $k = 1, \ldots, MN$. On note $\{\mathcal{C}_k\}$ le set des indices calculés. Les pics $\{\mathcal{C}_k^p\} \in \mathcal{C}_k$, correspondant à des séparations dans le spectre de valeurs propres, sont alors détectés. Si un des pics détectés se trouve à un indice $1 \leq k \leq MN$ correspondant à r, l'algorithme s'arrête. En effet, dans ce cas idéal, le nombre optimal de mode r estimé correspond à une séparation dans le spectre. Si tel n'est pas le cas, l'indice k le plus proche de r correspondant à un pic au-dessus d'un certain seuil est détecté. Le nombre optimal de modes est alors mis à jour par $\hat{r} = k$ en retournant à l'étape 2 de la méthode. L'ajustement est décrit par le pseudo-code ci-après (algorithme 3).

Algorithme 3 Ajustement du nombre de modes par calcul de $\mathcal{C}_k$

Entrée: λ_k, r, $seuil < 1$
Sortie: $\hat{r}$

```
Calculer {C_k}_{1≤k≤MN}
Détecter les pics {C_k^p} ∈ {C_k}
Extraire les indices {k} des pics {C_k^p}
si r ∈ {k} alors
    r̂ = r
sinon
    Extraire l'indice k du pic C_k^p ≥ seuil le plus proche de r
    r̂ = k
fin si
```

Afin d'illustrer l'ajustement du nombre de modes par le calcul de $\{\mathcal{C}_k\}$, un exemple simplifié est fourni en figure 3.3. Le spectre est constitué des séquences de valeurs propres suivantes : une valeur propre dominante λ_1, un multiplet $\{\lambda_2, \lambda_3, \lambda_4\}$, une valeur propre isolée λ_5, un multiplet $\{\lambda_6, \lambda_7\}$, puis un "plateau" de valeurs propres. On voudrait sélectionner 7 modes car la dernière séparation significative a lieu entre λ_7 et λ_8. À l'issue de l'étape 2, le nombre optimal de modes r est ici de 6 (ligne verticale verte), ce qui est dû à une variation insuffisante de la cross-RMSE lors de l'ajout du mode 7 (voir critère Λ, section 3.2.4). Or, les principaux pics de $\mathcal{C}_k$ au-dessus d'un seuil, fixé ici à 0.9, sont situés aux indices $k = 1, k = 4$ et $k = 7$, correspondant aux principales séparations. Le nombre de modes r est alors ajusté de sorte que $\hat{r} = k = 7$, indice le plus proche de $r = 6$ correspondant à un pic de $\mathcal{C}_k$ (ligne verticale rouge). Cet ajustement fait appel à l'étape 2 qui met à jour les valeurs manquantes jusqu'à l'obtention d'un champ reconstruit avec $\hat{r} = 7$ modes.

Dans ce cas, l'indice le plus proche de r correpondant à un pic est supérieur à r, ce qui justifie l'appel à l'étape 2. Si cet indice est inférieur à r, les modes correspondant à la différence entre l'indice et r sont retirés au champ reconstruit.

Notons que lorsque le spectre ne comporte pas de séparation dominante, $\mathcal{C}_k$ ne comporte alors pas de pic dominant. Dans ce cas, le seuil choisi pourra être abaissé afin d'élargir le choix de pics et éviter une sous-estimation du nombre de modes. L'ajustement, moins significatif dans ce cas, consiste alors à choisir un nombre de modes correspondant au pic non dominant le plus proche.

3.2.5 Détermination du décalage spatial

Le choix du décalage spatial pour augmenter les données, appelé aussi fenêtre spatiale, est généralement dicté par le compromis entre la quantité d'information extraite au sein de la fenêtre (M) et le nombre de répétitions (K) de la fenêtre spatiale au sein de chaque image [Groth2015]. La première logique requiert une taille de fenêtre la plus grande possible alors que la seconde appelle à réduire la taille de la fenêtre. Plutôt que de chercher une seule valeur de M, il est admis qu'un intervalle peut fournir des résultats satisfaisants [Ghil2002]. La variation du décalage spatial modifie par ailleurs la dimension de la covariance spatio-temporelle de taille $K \times MN$:

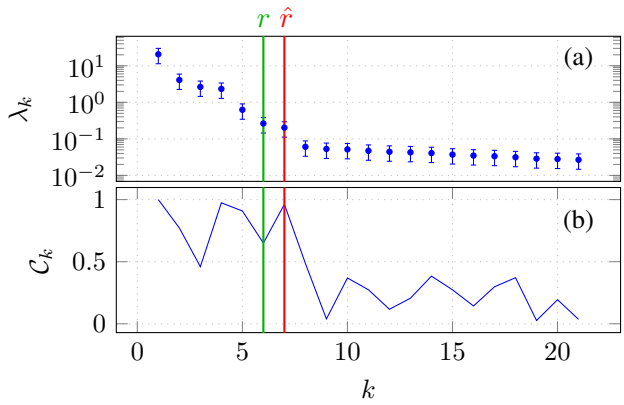


Figure 3.3 – (a) Spectre de valeurs propres λ_k et (b) mesure de confiance associée $\mathcal{C}_k$. La ligne verte correspond à l'estimation du nombre de modes après l'étape 2. La ligne rouge correspond à l'ajustement du nombre de modes par le calcul de $\mathcal{C}_k$. Les barres verticales sont les intervalles d'incertitude de chaque valeur propre issus de la règle empirique (3.15).

la répartition de l'énergie par valeur propre change aussi, et avec elle la variation de la cross-RMSE en fonction du nombre de modes (figure 3.4). En effet, l'augmentation du décalage spatial provoque une translation du minimum de la cross-RMSE vers de plus hautes valeurs. L'estimation du nombre de modes augmentera aussi en moyenne car celle-ci repose en partie sur le minimum de la cross-RMSE.

Nous proposons ici deux mesures correspondant aux limites inférieures et supérieures de l'intervalle de M. La première mesure se base sur la théorie de l'estimation de la covariance : le nombre d'échantillons indépendants doit être au moins supérieur à deux fois le nombre de variables [Reed1974] (ici les M points au sein de la fenêtre spatiale). La valeur maximale de M peut être estimée en résolvant l'inégalité $K > 2M$, ce qui, par un calcul simple, conduit à l'approximation $M < P/6$.

La seconde mesure est basée sur l'autocorrélation spatiale du champ de déplacement. Soit τ la distance de décorrélation spatiale définie par :

$$\tau = -\frac{\Delta P}{\log \tilde{r}} \tag{3.22}$$

où $\tilde{r}$ est l'auto-corrélation à la distance unité et ΔP est le pas d'échantillonnage spatial, ici 1 pixel. Si l'on suit l'étude de [Ghil2002], M peut être généralement approximé par $M \simeq P/\tau$. Dans la plupart des cas, $\tilde{r}$ est supposé être inférieur à 0.95, ce qui donne finalement $M > P/20$.

3.2.6 Synthèse de la méthode EM-EOF étendue

Nous donnons ici un résumé succinct, étape par étape, du déroulé de la méthode EM-EOF étendue en reprenant le principe général déjà illustré en figure 3.1.

Étape 1

Après initialisation des données manquantes, la matrice de données spatio-temporelle $\mathbf{Y}$ est augmentée (équations (3.2 et (3.4)). La covariance spatio-temporelle est ensuite estimée à partir de $\mathcal{D}$ (équation 3.5), puis décomposée en modes EEOF (équation 3.7). La matrice augmentée est ensuite reconstruite (équations (3.8) et (3.9)), puis moyennée (équation (3.10)). La matrice augmentée est ensuite ré-ordonnée de manière à retrouver $\hat{\mathbf{Y}}$. Finalement, la moyenne spatiale de $\mathbf{Y}$ est ajoutée à $\hat{\mathbf{Y}}$ afin de retrouver un champ à moyenne non nulle. À chaque ajout d'un mode dans la reconstruction, la cross-RMSE δ_k est calculée. Le nombre optimal de modes R correspond au minimum de l'ensemble des cross-RMSE calculées à chaque ajout de mode supplémentaire.

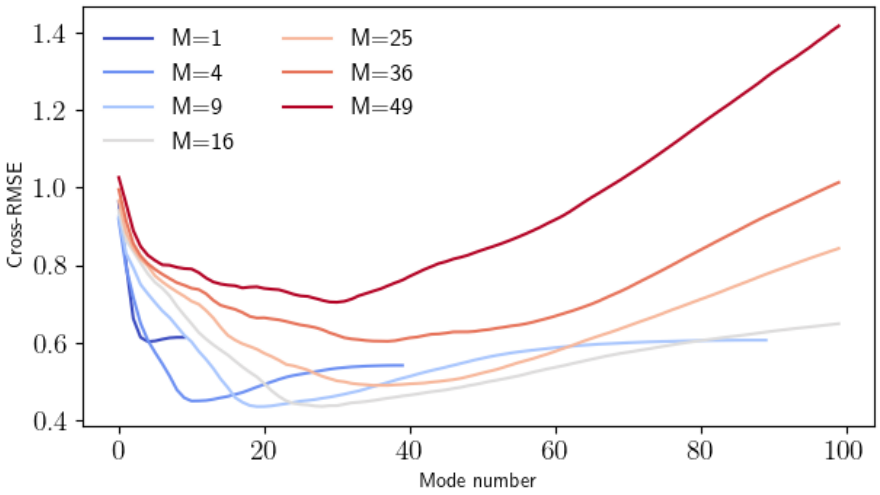


Figure 3.4 – Variation de la cross-RMSE en fonction du nombre de modes pour différent décalages spatiaux sur des données synthétiques incomplètes (N=10, P=144). Le minimum de la cross-RMSE évolue avec la variation du décalage spatial : plus ce dernier est grand, plus le nombre de modes sélectionnés sera grand.

Étape 2

L'étape 2 est une mise à jour des données manquantes et du nombre de modes R estimé à l'étape 1. Cette étape repose sur la forme de l'algorithme EM, lequel a déjà été décrit au chapitre précédent, en section 2.2.6. La différence est que l'on calcule les équations (3.5), (3.7), (3.8), (3.9) et (3.10) à chaque itération. La convergence de l'étape 2 se base sur la cross-RMSE, et le traitement du problème de sur-estimation sur le critère Λ. Ces deux éléments ont été traités en section 2.2.6 du chapitre précédent. La sortie de l'étape 2 est le champ reconstruit avec r modes, $\hat{\mathbf{Y}}_r$.

Mesure de confiance associée à r

Le set de mesures de confiance $\{\mathcal{C}_k\}_{1\leq k\leq MN}$ est calculé et, comme détaillé en sous-section 3.2.4, les pics dominants en sont extraits puis leurs rangs comparés à r. Si ce dernier nécessite un ajustement, un retour à l'étape 2 est opéré. La sortie finale est le champ reconstruit avec $\hat{r}$ modes, $\hat{\mathbf{Y}}_{\hat{r}}$.

3.3 Simulations numériques

Dans cette partie, des simulations sur champs de déplacement synthétiques sont présentées. En l'absence de vérité terrain dans les données de télédétection, de telles simulations s'avèrent utiles afin de s'assurer que les valeurs manquantes interpolées sont le plus proche de la vérité simulée. D'une part, il s'agit de confirmer la capacité d'interpolation de la méthode EM-EOF étendue en opérant là où la méthode EM-EOF peut présenter certaines limites, comme énoncé en introduction : séries temporelles courtes de champs de déplacement pertubés par de forts taux de données manquantes, où la corrélation spatiale prévaut sur la corrélation temporelle et présentant des caractéristiques spatiales hétérogènes. D'autre part, il s'agit de déceler, à travers une comparaison systématique avec la méthode EM-EOF, les avantages et inconvénients des deux méthodes et ce en fonction du type de champ de déplacement, du type de données manquantes et du type de bruit présent dans les données.

3.3.1 Type de champ de déplacement

Dans cette étude, nous nous intéressons à quatre types de champ de déplacement (figure 3.5) dont les modèles sont décrits dans le tableau 3.1 ci-après.

Le champ g_0 est un champ linéaire croissant. Le champ g_1 est constitué d'une partie linéaire et d'une partie oscillatoire à trois fréquences, ce qui en fait un champ d'ordre 3. Le champ g_2 contient de multiples fréquences et possède une distance à l'origine non-linéaire : nous l'appelons champ d'ordre n. Enfin, le champ g_3 est un champ d'ordre 5 entouré spatialement d'une zone stable dont les valeurs sont fixées à 0.

Nom	$g(r,t)$	Ordre
$g_0(r_1,t)$	$(1+0.5r_1)t$	1
$g_1(r_1,t)$	$(1+0.5r_1)t + \sin(w_1 t)\cos(w_1 r_1) + 0.5\cos(w_2 t)\cos(w_3 r_1)$	3
$g_2(r_2,t)$	$\sin(w_1 t)\cos(w_1 r_2) + 0.5\cos(w_2 t)\cos(w_3 r_2) + 0.1\sin(w_4 t)\cos(w_5 r_2)$ $+0.3\sin(w_6 r_2)\sin(w_7 t) + 0.1\sin(w_8 r_2)\sin(w_8 t)$	n
$g_3(r_2,t)$	$\sin(w_1 t)\cos(w_1 r_2) + 0.5\cos(w_2 t)\cos(w_3 r_2)$ $+0.1\sin(w_4 t)\cos(w_5 r_2)+$ zone stable	5

Tableau 3.1 – Modèles des champs déterministes synthétisés. t est la variable temps, les coordonnées (x,y) discrétisent le compact $[-1,1]^2$, et $r_1 = \sqrt{(x+0.1)^2+(y+0.3)^2}$ et $r_2 = \exp(-(x+y)^2)+xy+\tan(x)$ sont les distances à l'origine. $w_1,\ldots,w_8 = 2\pi f_1,\ldots,2\pi f_8$ sont les vitesses angulaires du signal dont les fréquences sont fixées à $\{f_1,\ldots,f_8\} = \{0.25, 0.75, 2.5, 1.25, 5, 7.5, 1.75, 0.5\}$.

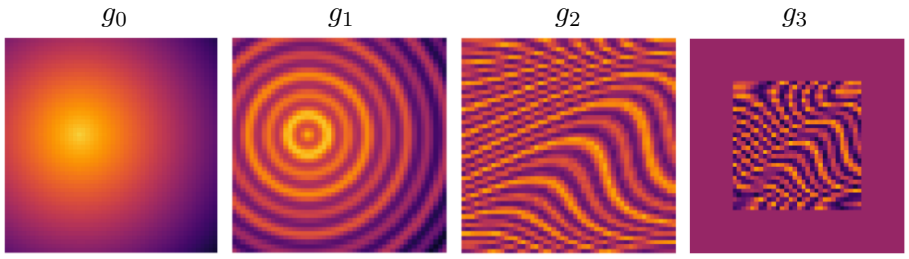


Figure 3.5 – Champs de déplacement considérés dans cette étude. Une description des modèles mathématiques des champs g_0 à g_3 est fournie en tableau 3.1

.

3.3.2 Type de perturbation et type de données manquantes

Afin de simuler différents types de perturbations, nous reprenons deux des trois types de bruits simulés lors du chapitre précédent. Le but est de mettre en évidence l'intérêt pratique de la méthode EM-EOF étendue dans différents cas de bruit corrélés. Nous simulons un bruit spatialement corrélé (SCN) et un bruit spatio-temporellement corrélé (STCN). Le degré de corrélation de ces deux types de bruit est contrôlé au moyen de coefficients de corrélation spatiale et temporelle $\gamma \in \mathbb{R}^+$ et $\rho \in [0,1]$. Le bruit SCN est ainsi synthétisé par le biais d'une fonction d'auto-corrélation $c(r) = r^{-\gamma}$, où la modification de γ permet de jouer sur le degré de corrélation entre deux points distants de r, et donc sur la longueur de corrélation. Le bruit STCN est la somme d'un bruit SCN et d'un bruit corrélé temporellement. Plus de détails concernant sa synthèse sont fournis en annexe A.

Concernant le type de données manquantes, deux cas sont également examinés : données manquantes aléatoires indépendantes de l'espace et du temps et données manquantes spatio-temporellement corrélées. Dans ce second cas, les données manquantes sont générées sur quatre dates consécutives et sur une zone particulière afin de reproduire l'effet de phénomènes saisonniers

et localisés (chute de neige, densification du couvert végétal) créateurs de données manquantes en mesure de déplacement par télédétection.

3.3.3 Paramètres de simulations

Quatre séries temporelles de 10 champs de déplacement correspondant aux modèles g_0, g_1, g_2 et g_3 sont générées sur une grille régulière de taille 50×50 [5]. Les séries temporelles sont ensuite perturbées par des bruits de type SCN et STCN, ainsi que par des données manquantes aléatoires et corrélées (voir tableau 3.2 ci-après). Les données manquantes aléatoires sont simulées sur l'ensemble des champs qui composent la série temporelle, alors que les données manquantes corrélées sont simulées sur quatre dates consécutives. Les coefficients de corrélation γ, ρ des bruits SCN et STCN sont fixés à 0.5 : si l'on considère l'intervalle des valeurs possibles de γ et ρ, cela correspond à un bruit SCN fortement corrélé (voir figure 2.4) et un bruit STCN à corrélation moyenne. Dans les cas g_0, g_1 et g_2, la taille du décalage spatial est fixée à $M = 121$, soit une fenêtre spatiale de taille $M_x \times M_y = 11 \times 11$. Cette valeur de M correspond à la limite inférieure de l'intervalle approximé en section 3.2.5, c'est-à-dire à $\approx P/20$. Dans le cas g_3, étant donné la taille réduite de la cible physique (30×30), M est abaissé à 64, ce qui correspond à $\approx P/15$. De plus, le seuil sur la détection des pics de $\mathcal{C}_k$ est fixé à 0.8. Enfin, les données manquantes sont initialisées par la moyenne spatiale de chaque champ $\mathbf{X}_t$.

Champ	**Type de données manquantes**	**Type de bruit**	**SNR**	**% de données manquantes**	M	**Initialisation**
g_0	Aléatoire	SCN ; $\gamma = 0.5$	2	30%	121	Moyenne spatiale
	Corrélée	STCN ; $\gamma, \rho = 0.5$	1.8			
g_1	Aléatoire	SCN ; $\gamma = 0.5$	2	30%		
	Corrélée	STCN ; $\gamma, \rho = 0.5$	1.8			
g_2	Aléatoire	SCN ; $\gamma = 0.5$	1.8	50%		
	Corrélée	STCN ; $\gamma, \rho = 0.5$	1.5			
g_3	Aléatoire	SCN ; $\gamma = 0.5$	1.8	30%	64	
	Corrélée	STCN ; $\gamma, \rho = 0.5$				

Tableau 3.2 – Principaux paramètres de simulations des quatre cas d'étude (g_0, g_1, g_2, g_3).

5. Le choix de cette grille relativement petite est conditionné par le temps de calcul (qui est plus important que celui de la méthode EM-EOF) qu'engendre l'augmentation de la dimension de la matrice de covariance.

3.3.4 Résultats et discussion

Impact du bruit et des données manquantes sur l'estimation de $\hat{r}$

L'estimation du nombre optimal de modes, qui dépend directement de la forme du spectre de valeurs propres, peut être directement impactée selon les caractéristiques du bruit (corrélation, SNR) présent dans les données ainsi que selon la quantité de données manquantes. Afin de mieux appréhender cet impact, on peut analyser la structure spectrale d'un tel bruit. Par exemple, le spectre des valeurs propres d'un bruit SCN est présenté en figure 3.6. On remarque que plus le bruit est corrélé (γ petit), plus le spectre est structuré. À l'inverse, un bruit très peu corrélé (γ grand) possède un spectre homogène, à l'image d'un bruit blanc. La reconstruction d'un signal de déplacement perturbé par ce type de bruit consistera dès lors à sélectionner un nombre de modes qui minimise la part de variabilité due au bruit sans toutefois décalagesser celle correspondant au signal de déplacement. La figure 3.6 montre que cela peut poser un problème car, tout comme pour un signal de déplacement, la variabilité d'un bruit corrélé s'exprime dans les premiers modes.

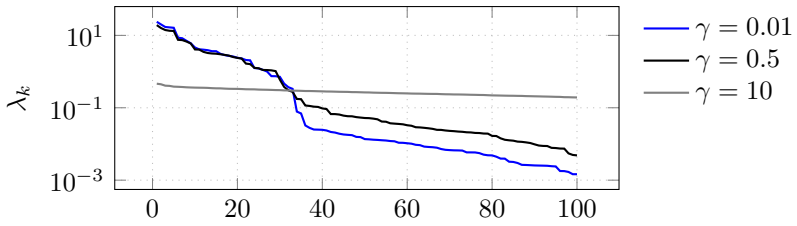


Figure 3.6 – Valeurs propres d'un bruit SCN pour différentes corrélations.

En plus de la corrélation du bruit, la variation du SNR et de la quantité de données manquantes peuvent impacter l'estimation du nombre optimal de modes. La figure 3.7 permet d'illustrer ce phénomène. On remarque que la diminution du SNR provoque un réhaussement de l'ensemble des valeurs propres (augmentation de l'énergie du système). À l'inverse, on observe que l'augmentation de la quantité de données manquantes tend à diminuer l'énergie du système, ce qui est dû à l'initialisation des données manquantes [Beckers2003]. Dans les deux cas, l'augmentation des données manquantes et/ou du niveau de bruit se traduit par moins de valeurs propres dominantes. La détection du nombre optimal de modes est alors plus sujette à une sous-estimation car une partie des valeurs propres décrivant la variabilité du signal de déplacement est soit noyée dans celle du bruit, soit effacée par l'initialisation des données manquantes. Ces cas (faible SNR et/ou grande quantité de données manquantes) constituent donc une limite à l'utilisation pratique de la mesure $\mathcal{C}_k$.

Analyse quantitative : reconstruction des champs g_0 à g_3

Cas 0 : champ g_0 Le champ linéaire g_0 est perturbé par des données manquantes aléatoires et un bruit SCN, puis par des données manquantes corrélées et un bruit STCN (tableau 3.2). Le nombre de modes estimé par la méthode EM-EOF étendue est $\hat{r} = 3$ dans les deux configurations. Dans la première configuration, la mesure $\mathcal{C}_k$ (figure 3.8) est utilisée pour ajuster le nombre de modes estimé à l'issue de l'étape 2 ($r = 4$) au nombre de modes correspondant à la séparation dominante la plus proche dans le spectre, située à $k = 3$. Dans le second cas, la mesure $\mathcal{C}_k$ n'est pas utilisée car le nombre de modes sélectionnés à l'étape 2 correspond déjà à une séparation dans le spectre.

Des exemples de reconstruction sont présentés en figure 3.9 (champs synthétiques à la date $t = 5$) et en figure 3.10 (séries temporelles d'un point choisi aléatoirement) pour les méthodes EM-EOF et EM-EOF étendue. Dans la première configuration (données manquantes aléatoires et bruit SCN), la méthode EM-EOF étendue fournit une reconstruction spatialement plus lisse que la méthode EM-EOF (1 mode sélectionné dans les deux configurations). De plus, les résidus montrent que les deux méthodes filtrent correctement le bruit SCN. L'évolution temporelle (figure 3.10 (a))

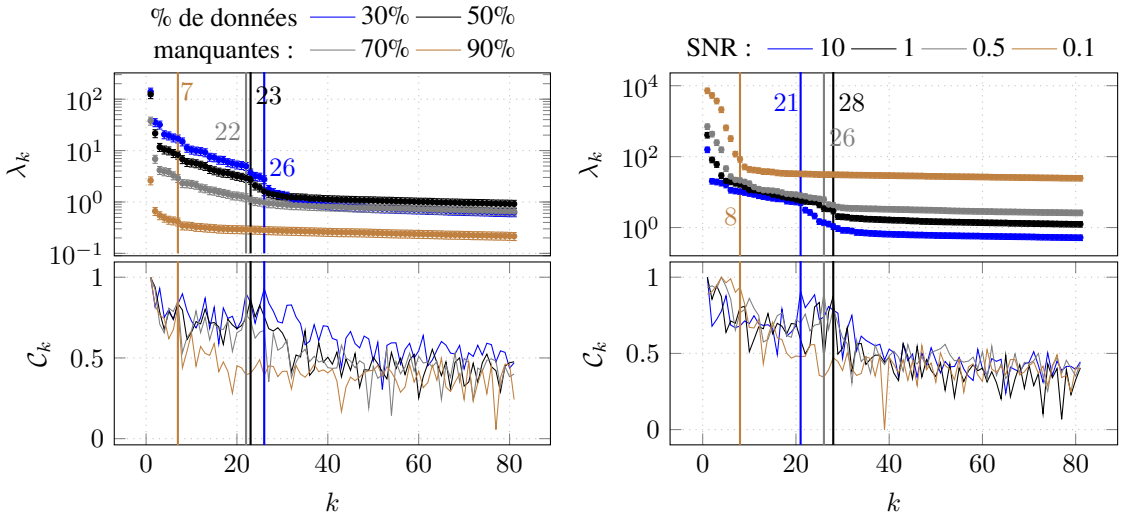


Figure 3.7 – Spectres de valeurs propres λ_k d'un champ de déplacement synthétique augmenté et mesure $\mathcal{C}_k$ pour différentes quantités de données manquantes (gauche) et différents SNR (droite). Lignes verticales et nombres en couleur : estimation du nombre optimal de modes $\hat{r}$.

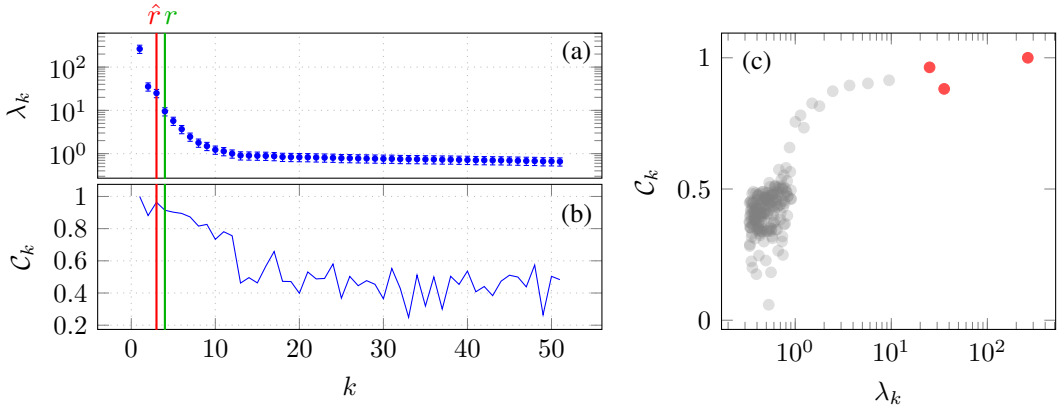


Figure 3.8 – (a) Valeurs propres λ_k de la matrice de données augmentées $\mathcal{D}$ (50 premières) du champ g_0 perturbé par des données manquantes aléatoires et un bruit SCN ; (b) mesure de confiance associée $\mathcal{C}_k$ et estimation du nombre de modes à l'issue de l'étape 2 (ligne verte) puis après ajustement (ligne rouge) ; (c) $\mathcal{C}_k$ versus λ_k. Les cercles rouges correspondent au nombre de modes sélectionnés.

suggère que la méthode EM-EOF est plus proche de la vérité terrain. Dans la seconde configuration (données manquantes corrélées et bruit STCN), la méthode EM-EOF étendue fournit également une reconstruction spatialement plus homogène. Le champ reconstruit par la méthode EM-EOF est spatialement plus hétérogène, mais est en réalité plus proche de la vérité terrain sur la série temporelle (figure 3.10 (b)). Dans les deux configurations, la méthode EM-EOF est donc plus proche de la vérité terrain. La méthode EM-EOF étendue fournit un champ reconstruit plus proche des données réelles et spatialement plus homogène, ce qui est notamment dû au moyennage spatial sur le champ augmenté (équation (3.10)). Toutefois, on remarque que le champ reconstruit est faiblement décentré car il conserve une partie du bruit SCN (figure 3.9 (a)) : le moyennage ne permet donc pas de réduire totalement le bruit SCN.

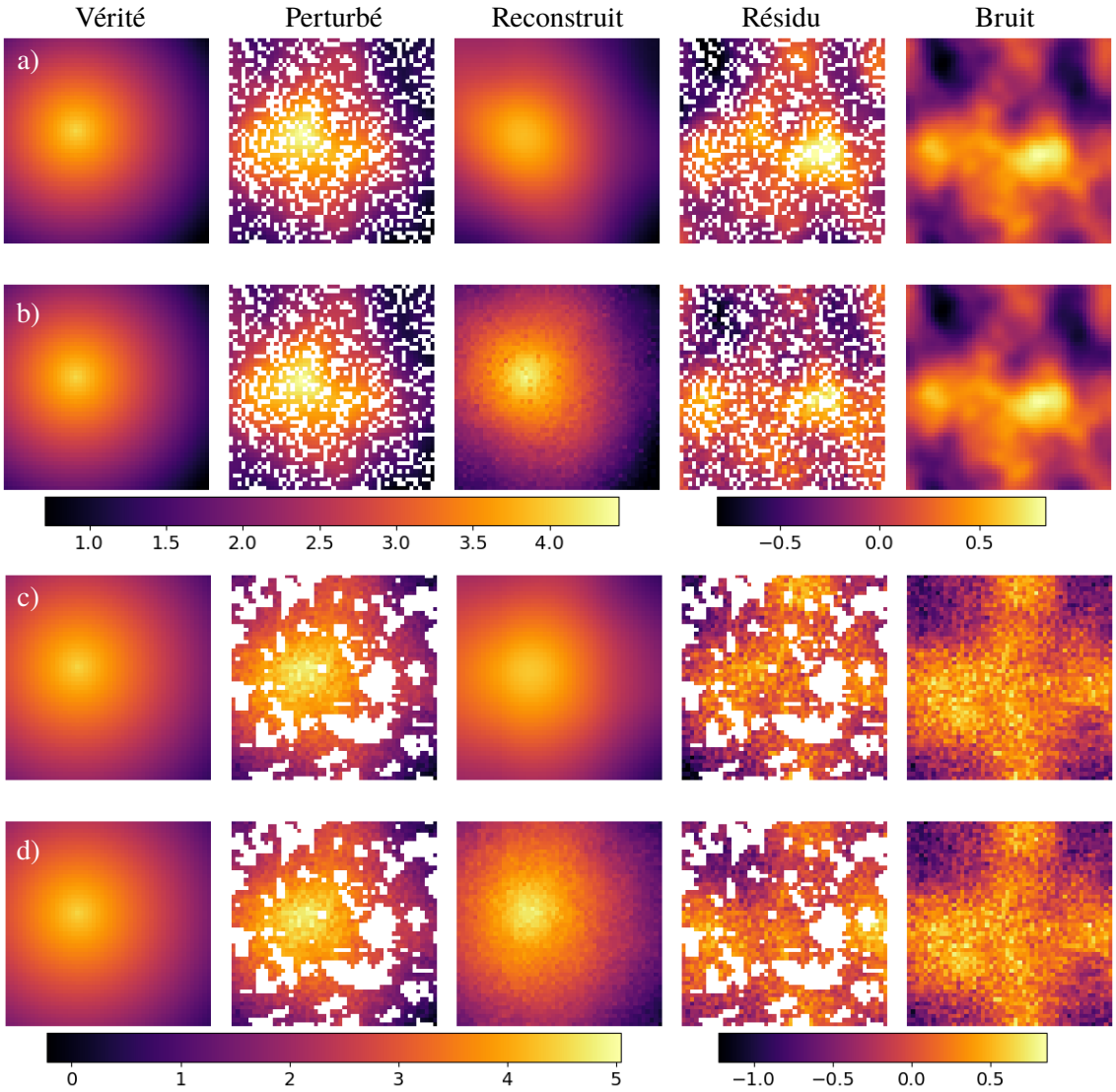


Figure 3.9 – Reconstruction [cm] d'un champ d'ordre 1 perturbé par des données manquantes aléatoires et un bruit SCN ((a), (b)) puis des données manquantes corrélées et un bruit STCN ((c), (d)), par les méthodes EM-EOF étendue ((a), (c)) et EM-EOF ((b), (d)). La quantité de données manquantes est fixée à 30% et le SNR à 1.8 ((a)(b)) et 2 ((c)(d)). Le résidu est la différence entre le champ reconstruit et le champ perturbé.

Cas 1 : champ g_1 Dans un premier temps, le champ g_1 est perturbé par des données manquantes aléatoires ainsi que par du bruit SCN. Le nombre de modes estimé est $\hat{r} = 37$. Le spectre de valeurs propres ainsi que les mesures $\mathcal{C}_k$ sont illustrées en figure 3.11. Ici, l'ajustement n'est pas

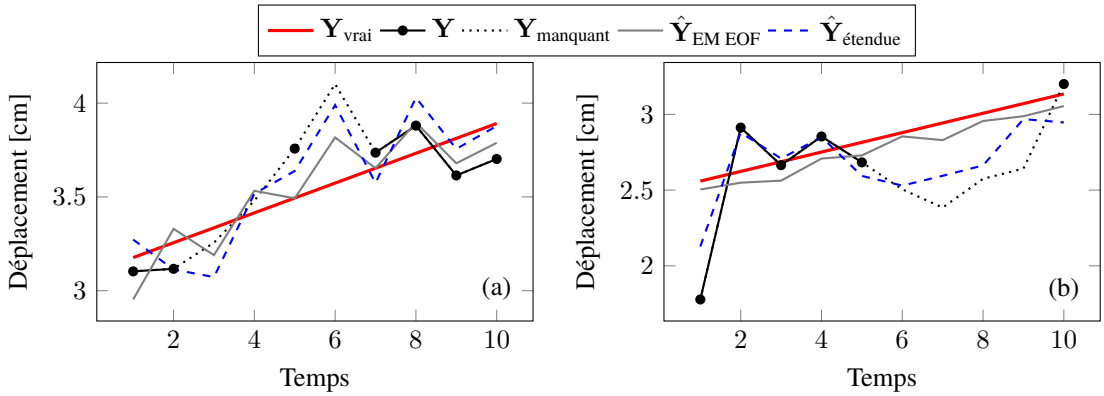


Figure 3.10 – Série temporelle d'un champ d'ordre 3 perturbé par (a) données manquantes aléatoires et bruit SCN ; (b) données manquantes corrélées et bruit STCN. Rouge : déplacement vrai ; cercles noirs : déplacement perturbé observé ; courbe en pointillés noire : déplacement perturbé non observé (données manquantes) ; ligne grise : reconstruction par la méthode EM-EOF ; courbe en pointillés bleue : reconstruction par la méthode EM-EOF étendue.

nécessaire car l'estimation du nombre de modes à l'issue de l'étape 2 correspond à un pic de $\mathcal{C}_k$ situé à $k = 37$ marquant une séparation du spectre. Les modes correspondant aux valeurs propres situées en deçà de cette séparation décrivent à la fois la variabilité du signal de déplacement et du bruit SCN.

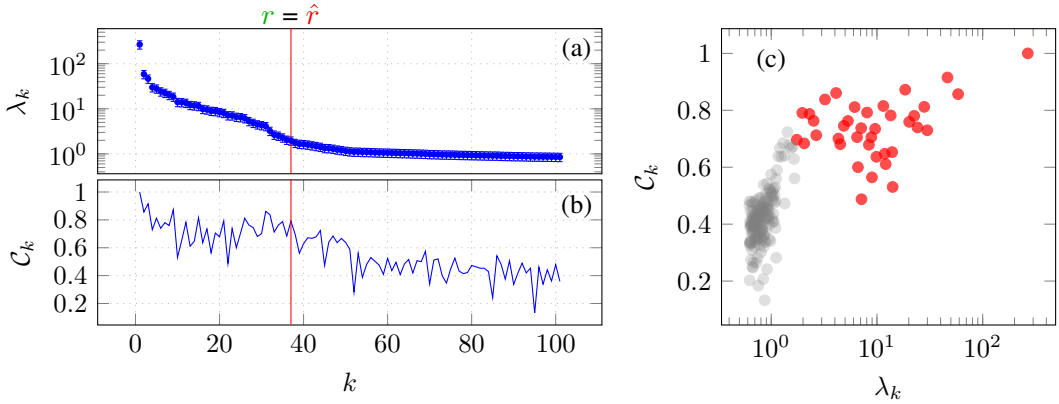


Figure 3.11 – (a) Valeurs propres λ_k de la matrice de données augmentées $\mathcal{D}$ (100 premières) du champ g_1 perturbé par des données manquantes aléatoires et un bruit SCN ; (b) mesure de confiance associée $\mathcal{C}_k$ et estimation du nombre de modes (ligne rouge) ; (c) $\mathcal{C}_k$ versus λ_k. Les cercles rouges correspondent au nombre de modes sélectionnés.

Le résultat de reconstruction est présenté en figure 3.12 (date $t = 5$), pour les méthodes EM-EOF étendue et EM-EOF. Le nombre de modes estimé par la méthode EM-EOF est de 3. Les motifs de déplacement présentent, pour les deux méthodes, une cohérence globale entre champ reconstruit et vérité terrain, bien qu'une comparaison visuelle permette de constater que la méthode EM-EOF étendue fournit un champ reconstruit plus homogène. La plupart des composantes du signal (tendance, oscillations) sont reconstruites, comme l'illustre la décomposition mode par mode d'un champ d'ordre 3 (figure 3.13), dont la répartition est la suivante : le premier mode est lié à la partie linéaire du signal ; les modes 2 à 8 concernent la partie du signal de déplacement à fréquence basse ; les modes supérieurs à 8 voient une oscillation supplémentaire s'ajouter au signal, ce qui s'apparente aux fréquences plus hautes qui le composent. Le bruit SCN est présent dès le second mode et s'étend sur les modes suivants. La méthode EM-EOF étendue reconstruit donc une partie du bruit SCN qui, avec les paramètres considérés (SNR=2, $\gamma = 0.5$), possède une

structure spectrale mélangée à celle du signal de déplacement dès les premiers modes. La méthode EM-EOF filtre davantage le bruit, ce qui est visible sur le résidu qui est proche du bruit SCN.

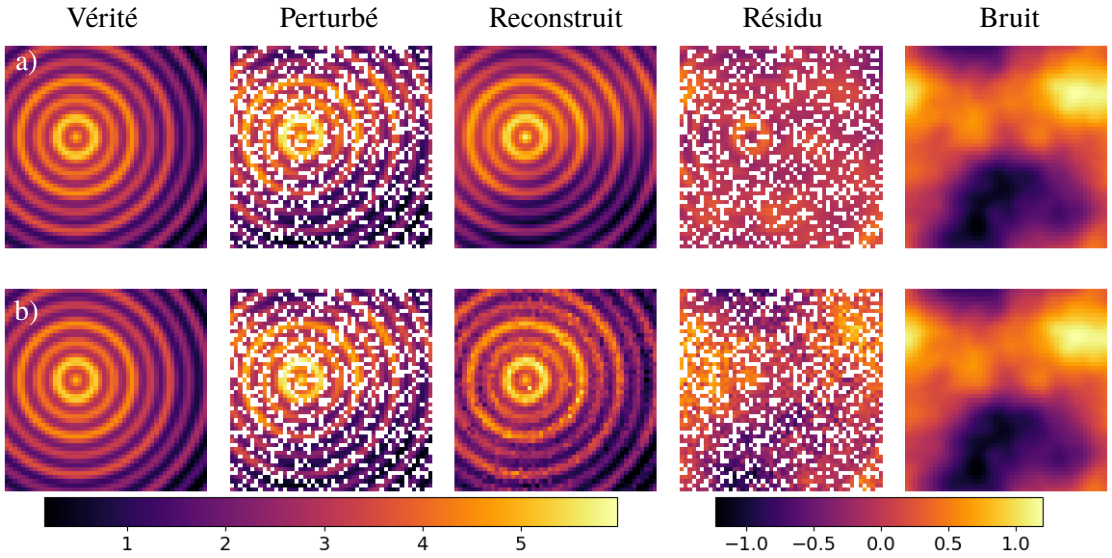


Figure 3.12 – Reconstruction [cm] d'un champ d'ordre 3 perturbé par des données manquantes de type aléatoire et un bruit SCN, par les méthodes EM-EOF étendue (a) et EM-EOF (b). La quantité de données manquantes est fixée à 30% et le SNR à 2. Le résidu est la différence entre le champ reconstruit et le champ perturbé.

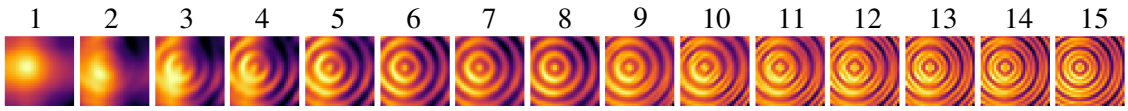


Figure 3.13 – Reconstruction par la méthode EM-EOF étendue d'un champ d'ordre 3 par ajout de modes successifs (jusqu'à 15).

La figure 3.14 (a) illustre la reconstruction d'une série temporelle d'un point choisi aléatoirement. La reconstruction par la méthode EM-EOF étendue fournit une interpolation plus proche des données perturbées (réelles) mais contient une partie du bruit SCN. La reconstruction par la méthode EM-EOF fournit un résultat similaire dont la variation dépend moins des données réelles.

Le champ g_1 est ensuite perturbé par des données manquantes corrélées et un bruit STCN. Le nombre optimal de modes estimé est ici de 39 pour EM-EOF étendue et 3 pour EM-EOF. Dans cette configuration, l'estimation du nombre optimal de modes repose essentiellement sur la cross-RMSE et sur le critère Λ car le spectre ne comporte pas de séparation dominante. La figure 3.15 présente un exemple de champ reconstruit. La reconstruction par la méthode EM-EOF étendue s'avère spatialement plus conforme à la vérité terrain, alors que la reconstruction par la méthode EM-EOF est plus hétérogène, avec quelques discontinuités visibles dans la partie basse au sein des zones de données manquantes. Comme déjà souligné auparavant, l'augmentation spatiale des données permet de prendre en compte la corrélation spatiale du champ en créant de multiples copies des données. Lors de la reconstruction, le moyennage de ces données permet alors de corriger les potentielles discontinuités dans la reconstruction, effet directement visible ici. La série temporelle d'un point choisi aléatoirement (figure 3.14 (b)) permet de constater que les deux méthodes produisent une reconstruction temporelle similaire, c'est-à-dire proche des données réelles.

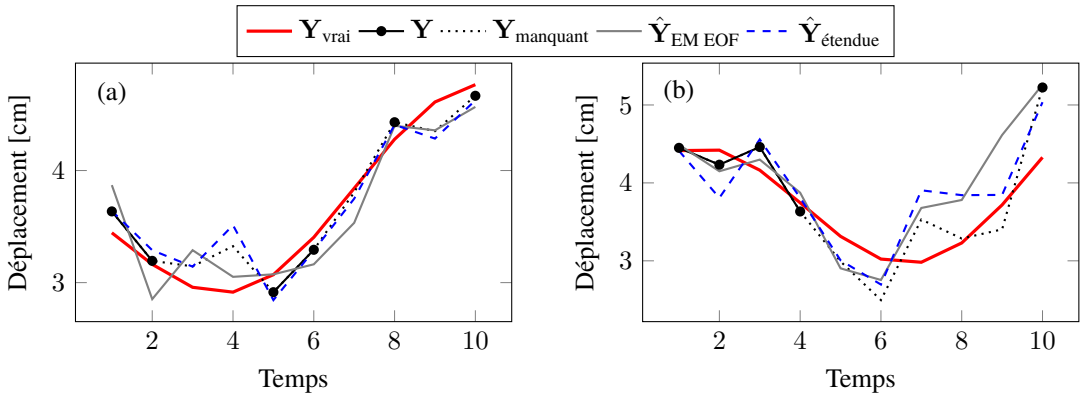


Figure 3.14 – Série temporelle d'un champ d'ordre 3 perturbé par (a) données manquantes aléatoires et bruit SCN ; (b) données manquantes corrélées et bruit STCN. Rouge : déplacement vrai ; cercles noirs : déplacement perturbé observé ; courbe en pointillés noire : déplacement perturbé non observé (données manquantes) ; ligne grise : reconstruction par la méthode EM-EOF ; courbe en pointillés bleue : reconstruction par la méthode EM-EOF étendue.

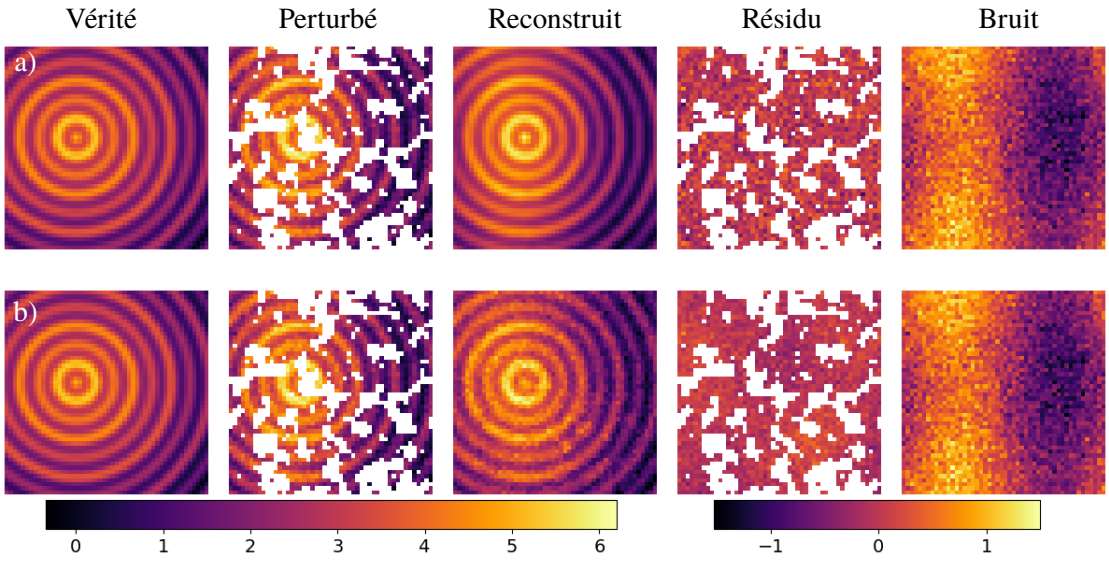


Figure 3.15 – Reconstruction d'un champ de déplacement synthétique d'ordre 3 [cm] perturbé par des données manquantes corrélées et un bruit STCN, pour les méthodes EM-EOF étendue (a) et EM-EOF (b). La quantité de données manquantes est fixée à 30% et le SNR à 1.8.

Cas 2 : champ g_2 Les résultats de la reconstruction du champ d'ordre n g_2 sont, à leur tour, analysés. Comme dans le cas précédent, nous examinons les résultats sur un champ perturbé par des données manquantes aléatoires et un bruit SCN, puis sur un champ perturbé par des données manquantes corrélées et un bruit STCN. Les champs sont ici perturbés par 50% de données manquantes.

Dans le premier cas, les estimations du nombre de modes sont les suivantes : 61 pour EM-EOF étendue et 4 pour EM-EOF. La mesure $\mathcal{C}_k$ (figure 3.16) permet d'ajuster l'estimation de $\hat{r}$ à une séparation dans le spectre par rapport à l'étape 2 ($r = 59$). Une troncature en amont de cette séparation entraînerait une sous-estimation du nombre de modes : la figure 3.17 montre notamment qu'il est difficile de filtrer le bruit (présent dès les premiers modes) sans filtrer le signal de déplacement. Cela est notamment dû à la structure du bruit SCN, qui possède une grande longueur de corrélation et s'apparente spectralement à un signal basse fréquence, comme déjà souligné en sous-section 3.3.4 et dans le cas d'étude du champ g_1.

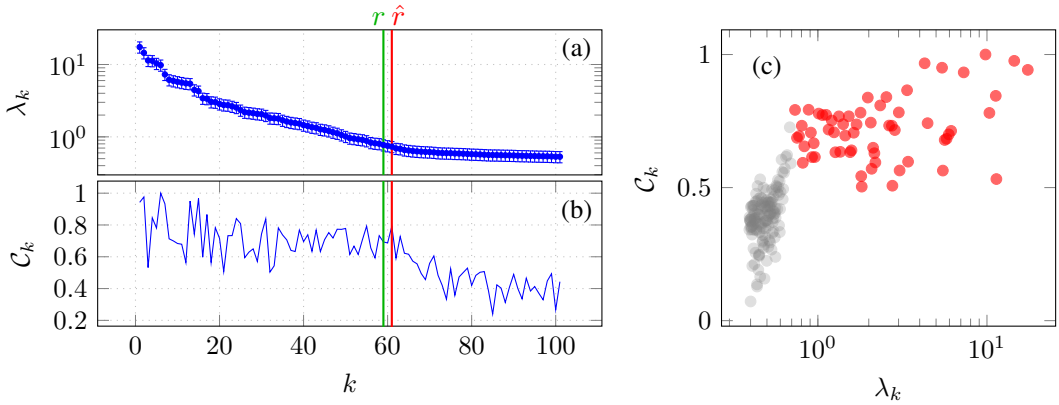


Figure 3.16 – (a) Valeurs propres λ_k de la matrice de données augmentées $\mathcal{D}$ (100 premières) du champ g_2 perturbé par des données manquantes aléatoires et un bruit SCN ; (b) mesure de confiance associée $\mathcal{C}_k$ et estimation du nombre de modes à l'issue de l'étape 2 (ligne verte) puis après ajustement (ligne rouge) ; (c) $\mathcal{C}_k$ versus λ_k. Les cercles rouges correspondent au nombre de modes sélectionnés.

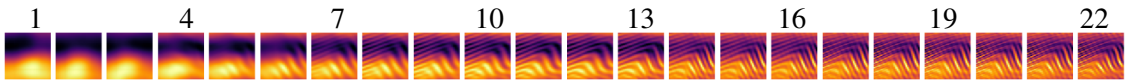


Figure 3.17 – Reconstruction par la méthode EM-EOF étendue d'un champ d'ordre n par ajout modes successifs (jusqu'à 22).

Le champ reconstruit par la méthode EM-EOF étendue (figure 3.18) montre un résultat satisfaisant en comparaison à la vérité terrain. La variabilité spatiale du champ de déplacement synthétique est globalement bien conservée. La structure est moins affectée par la présence de données manquantes aléatoires (50%) dans le cas d'EM-EOF étendue, ce qui est dû à deux mécanismes. Le premier est la prise en compte de la corrélation spatiale du champ dans la recontruction alors que l'incomplétude des données rend l'information temporelle peu disponible. Le second est le moyennage dans la reconstruction, qui permet de réduire les discontinuités spatiales induites par l'incomplétude des données et/ou le bruit. Le champ résiduel est proche de zéro dans le cas d'EM-EOF étendue, alors que l'on retrouve une partie du bruit SCN dans les résidus d'EM-EOF. La méthode EM-EOF étendue fournit donc un champ plus homogène et mieux interpolé, mais ne permet pas de filtrer totalement le bruit SCN, contrairement à la méthode EM-EOF qui permet de mieux filtrer le bruit mais fournit un champ plus perturbé par la quantité importante de données manquantes, et ce compte tenu de la petite taille de la série temporelle.

Les séries temporelles reconstruites pour ce cas sont présentées en figure 3.19 (a). Malgré la

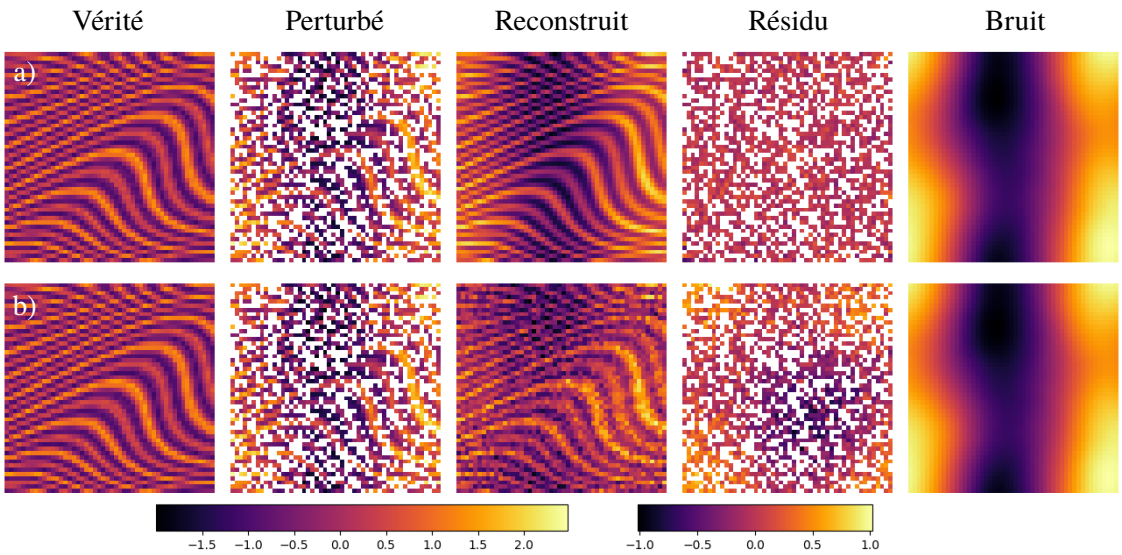


Figure 3.18 – Reconstruction d'un champ d'ordre n perturbé par des données manquantes aléatoires et un bruit spatialement corrélé, pour les méthodes EM-EOF étendue (haut) et EM-EOF (bas). La quantité de données manquantes est fixée à 50% et le SNR à 1.8.

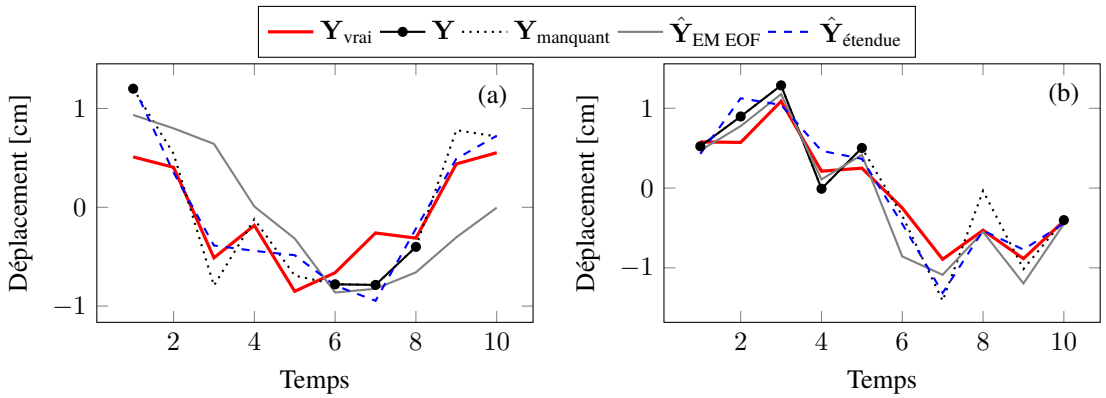


Figure 3.19 – Série temporelle d'un champ de déplacement synthétique d'ordre n perturbé par (a) données manquantes aléatoires et bruit SCN ; (b) données manquantes corrélées et bruit STCN. Rouge : déplacement vrai ; cercles noirs : déplacement bruité avec données manquantes ; courbe noire en pointillés : données manquantes ; ligne grise : série temporelle reconstruite par la méthode EM-EOF ; courbe bleue en pointillés : reconstruction par la méthode EM-EOF étendue.

complexité spatio-temporelle du champ considéré, la cohérence entre reconstruction et vérité terrain est assurée. La série reconstruite par la méthode EM-EOF étendue est proche de la vérité terrain et du champ perturbé, alors que celle reconstruite par la méthode EM-EOF est temporellement plus lisse mais "n'accroche pas" à la vérité terrain. Cela pourrait être dû notamment à une estimation biaisée de la covariance temporelle (équation (2.4)), puisque peu de points sont observés au cours de la série, soulignant au passage l'intérêt de l'augmentation spatiale des données pour estimer une matrice de covariance spatio-temporelle et ainsi réduire les biais potentiellement émergents.

Concernant le cas d'un champ perturbé par des données manquantes corrélées et un bruit STCN, le nombre optimal de modes estimé est de 71 pour EM-EOF étendue et 4 pour EM-EOF. Le spectre des valeurs propres ne possédant pas de séparation dominante, la mesure $\mathcal{C}_k$ n'est pas illustrée car l'ajustement du nombre est réalisé sur un pic non dominant.

Un exemple de champ reconstruit est affiché en figure 3.20. La reconstruction par la méthode EM-EOF étendue fournit, comme dans les exemples précédents, un champ interpolé plus homogène mais contenant une partie du bruit STCN car les modes correspondant aux valeurs propres significatives du bruit ont été pris en compte dans la reconstruction. Concernant la méthode EM-EOF, une partie du bruit est filtré (voir résidu) mais le champ reconstruit contient plus de discontinuités spatiales. L'évolution temporelle de la reconstruction montre que les deux méthodes fournissent un résultat similaire dans la zone de données manquantes.

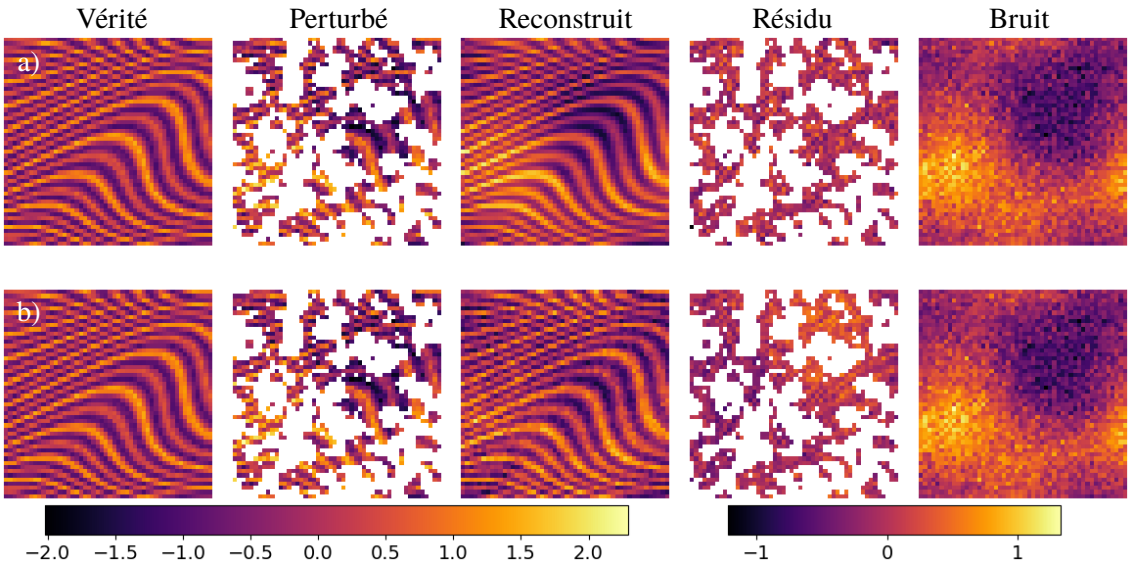


Figure 3.20 – Reconstruction d'un champ de déplacement synthétique [cm] d'ordre n perturbé par des données manquantes corrélées et un bruit STCN, pour les méthodes EM-EOF étendue (a) et EM-EOF (b). La quantité de données manquantes est fixée à 50% et le SNR à 1.5.

En définitive, la méthode EM-EOF étendue fournit un champ spatial généralement mieux interpolé, quel que soit le bruit considéré (SCN ou STCN), mais reconstruit également une partie de ce bruit corrélé. La méthode EM-EOF fournit une interpolation spatiale dont la structure spatiale est plus hétérogène du fait de la quantité importante de données manquantes et de la faible dimension temporelle des données, ce qui peut se traduire par des biais de reconstruction.

Cas 3 : champ g_3 Le champ g_3 est perturbé par des données manquantes aléatoires et un bruit SCN, puis des données manquantes corrélées et un bruit STCN. Le nombre optimal de modes est estimé à $\hat{r} = 79$ et $\hat{r} = 62$ respectivement. Dans la première configuration, $\mathcal{C}_k$ n'est pas utilisé car le nombre de mode estimé à l'étape 2 correspond déjà à une séparation dans le spectre. Dans la second configuration, l'illustration de $\mathcal{C}_k$ (figure 3.21) montre un ajustement d'un mode par rapport au nombre de mode estimé à l'étape 2.

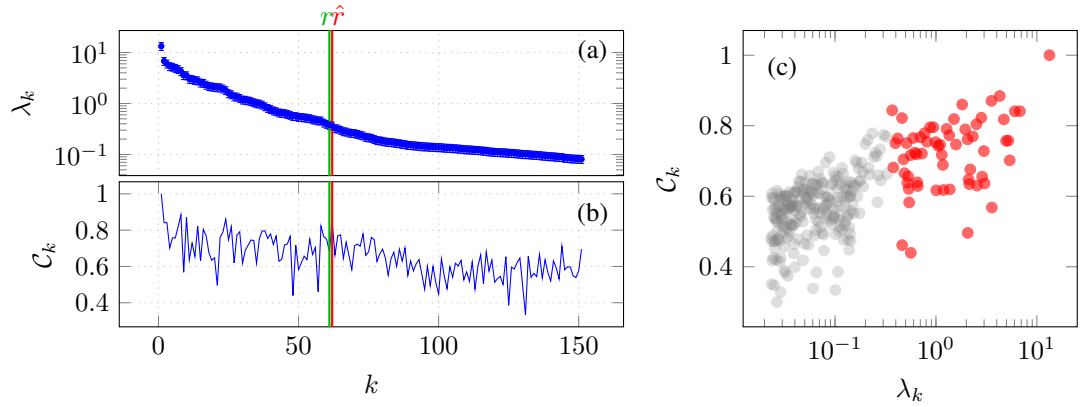


Figure 3.21 – (a) Valeurs propres λ_k de la matrice de données augmentées $\mathcal{D}$ (150 premières) du champ g_3 perturbé par des données manquantes corrélées et un bruit STCN ; (b) mesure de confiance associée $\mathcal{C}_k$ et estimation du nombre de modes à l'issue de l'étape 2 (ligne verte) et après ajustement (ligne rouge) ; (c) $\mathcal{C}_k$ versus λ_k. Les cercles rouges correspondent au nombre de modes sélectionnés.

Un exemple de résultat pour les deux configurations est présenté en figure 3.22. L'interpolation est comparable entre les méthodes EM-EOF étendue et EM-EOF (nombre de modes : 4). Concernant la première configuration (figure 3.22 (a)(b)), le bruit SCN est présent dans la reconstruction des deux méthodes. Concernant les zones stables, ces dernières sont moyennées par la méthode EM-EOF étendue. On remarque également la présence d'effets de bord dans la zone de transition entre l'objet physique et la zone stable (à gauche et à droite), effet qui est dû au fenêtrage spatial. Cela se traduit par un léger résidu dans les zones de transition du champ reconstruit. Cet effet peut être corrigé si l'on sait détecter les bords en amont du traitement afin de masquer la zone stable. Ce même effet est visible dans la reconstruction du champ perturbé par des données manquantes corrélées et un bruit STCN (figure 3.22 (c)(d)). Dans cette configuration, la méthode EM-EOF (nombre de modes : 4) permet de filtrer le bruit STCN (voir résidu) et comme dans le cas précédent, préserve aussi la zone de transition.

Enfin, la reconstruction de séries temporelles de points choisis aléatoirement est présentée en figure 3.23. Dans les deux configurations considérées, la méthode EM-EOF étendue permet d'obtenir une interpolation plus fidèle au champ perturbé par rapport à la méthode EM-EOF, alors que cette dernière fournit une reconstruction plus proche de la vérité terrain, surtout lorsque le champ est perturbé par des données manquantes corrélées et du bruit STCN (figure 3.23 (b)).

Analyse des erreurs de reconstruction

Afin de fournir des indicateurs quantitatifs de la performance des reconstructions illustrées précédemment, on analyse et compare trois erreurs de reconstruction en fonction de la quantité de données manquantes et du SNR :

— La cross-RMSE correspondant au nombre de modes sélectionnés $\hat{r}$ (équation (3.12)) :

$$\delta^{\text{CV}} = \delta_{k=\hat{r}} \tag{3.23}$$

— La RMSE entre les données reconstruites et les données réelles :

$$\delta = \frac{1}{\sqrt{NP}} ||\hat{\mathbf{Y}}_{\hat{r}} - \mathbf{Y}||_F \tag{3.24}$$

— La RMSE entre les données reconstruites et la vérité terrain :

$$\delta^{\text{vrai}} = \frac{1}{\sqrt{NP}} ||\hat{\mathbf{Y}}_{\hat{r}} - \mathbf{Y}_{\text{vrai}}||_F \tag{3.25}$$

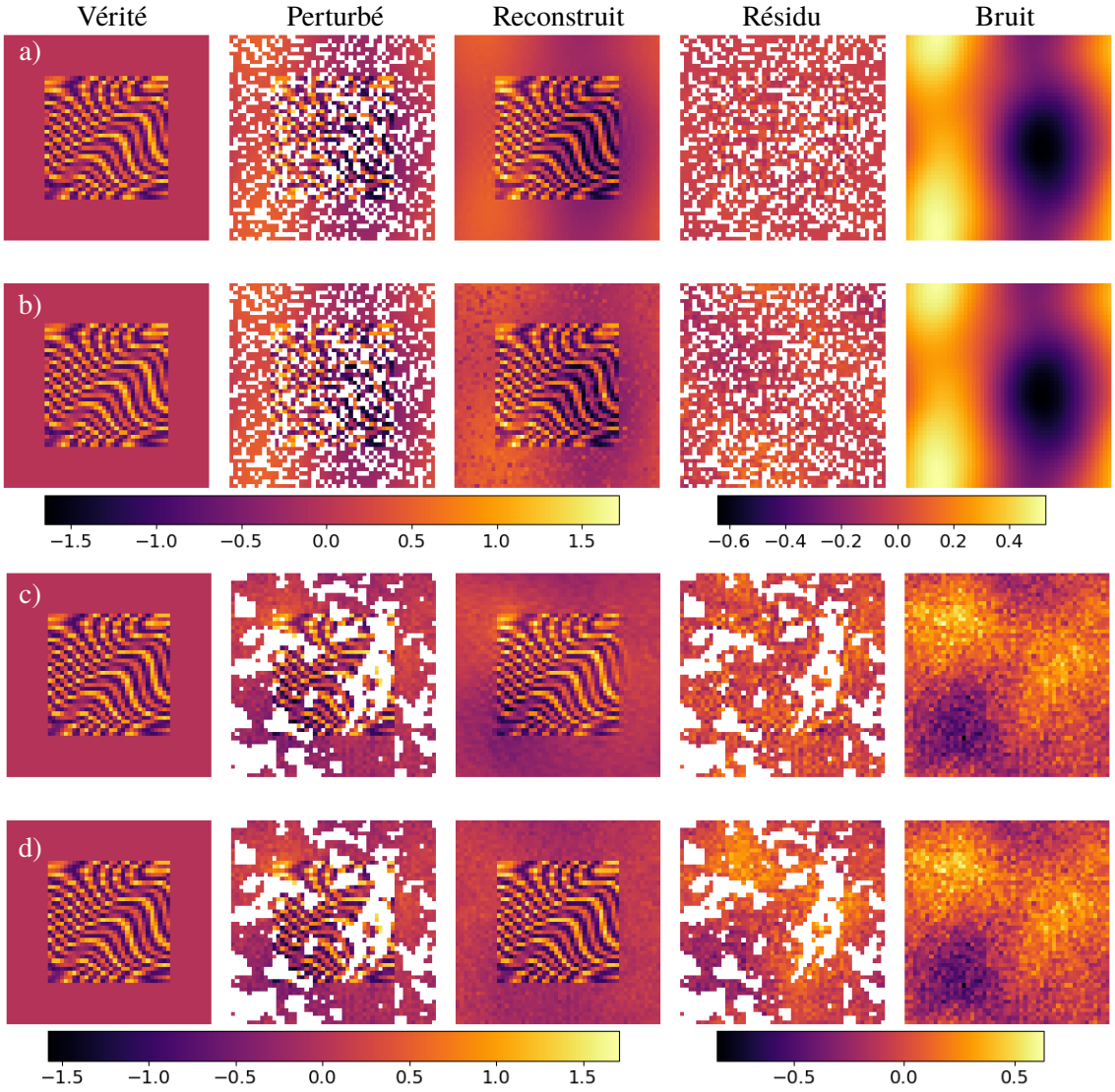


Figure 3.22 – Reconstruction d'un champ de déplacement synthétique [cm] d'ordre n perturbé par des données manquantes aléatoires et un bruit SCN ((a), (b)), puis par des données manquantes corrélées et un bruit STCN ((c), (d)), pour les méthodes EM-EOF étendue ((a), (c)) et EM-EOF ((b), (d)). La quantité de données manquantes est fixée à 30% et le SNR à 1.8.

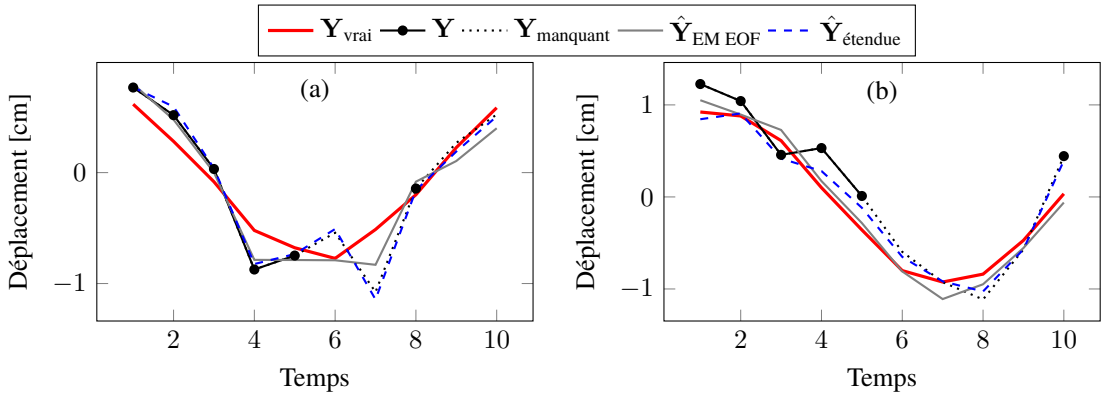


Figure 3.23 – Série temporelle d'un champ de déplacement synthétique d'ordre n perturbé par (a) données manquantes aléatoires et bruit SCN ; (b) données manquantes corrélées et bruit STCN. Rouge : déplacement vrai ; cercles noirs : déplacement bruité avec données manquantes ; courbe noire en pointillés : données manquantes ; ligne grise : série temporelle reconstruite par la méthode EM-EOF ; courbe bleue en pointillés : reconstruction par la méthode EM-EOF étendue.

Contrairement à la cross-RMSE qui utilise une norme vectorielle, les autres erreurs sont les normes matricielles (norme de Frobenius $||.||_F$) [Golub1996] de la différence des matrices spatio-temporelles $\hat{\mathbf{Y}}$ et $\mathbf{Y}$ (équation (3.1)) ordonnées en matrices $N \times P$. Notons que la RMSE n'est pas indépendante de l'amplitude des données ni de leur variance. Par conséquent, les erreurs présentées ci-après (moyenne sur 100 simulations) ne sont pas dans l'ordre de grandeur des champs reconstruits précédemment.

L'évolution des erreurs en fonction de la quantité de données manquantes est présentée en figure 3.24. De manière générale, on remarque une tendance similaire entre la cross-RMSE δ^{CV} et la RMSE δ.

Quel que soit le type de données manquantes, on remarque que les erreurs de reconstruction sur les champs peu complexes (g_0 et g_1) sont plus basses pour la méthode EM-EOF étendue que pour la méthode EM-EOF. Pour ces mêmes champs, lorsque les données manquantes sont aléatoires, on remarque que l'écart des erreurs entre EM-EOF et EM-EOF étendue augmente en même temps que la quantité de données manquantes, pour atteindre son maximum lorsque la quantité de données manquantes est importante (90%). La méthode EM-EOF étendue est donc plus robuste à une augmentation de la quantité de données manquantes aléatoires pour des champs synthétiques relativement simples. Concernant le champ g_2, la méthode EM-EOF étendue fournit également des erreurs plus faibles, dont l'écart avec la méthode EM-EOF diminue lorsque la quantité de données manquantes aléatoires augmente. Lorsque les données manquantes sont corrélées, la méthode EM-EOF étendue est plus proche des données réelles que la méthode EM-EOF sur les champs g_0, g_1 et g_2, ce qui a en partie été observé précédemment. Concernant le champ g_3, la méthode EM-EOF fournit des erreurs plus basses quel que soit le type de données manquantes. Cela peut s'expliquer, comme nous l'avons souligné auparavant, par les biais de reconstruction engendrés par effet de bord sur la zone de transition entre le champ de déplacement et la zone stable.

Les erreurs en fonction du SNR pour les deux types de bruits SCN et STCN sont présentées en figure 3.25. Lorsque le bruit est important (SNR < 1), on constate que la méthode EM-EOF étendue est plus performante que la méthode EM-EOF, et ce quel que soit le type de bruit. Quel que soit le type de champ de déplacement, la méthode EM-EOF étendue est peu sensible à la variation du bruit SCN, ce qui n'est pas le cas pour le bruit STCN. Cette différence est due à la variation de la partie temporelle du bruit : en effet, le moyennage prend effet sur des données augmentées spatialement (et non temporellement), d'où l'augmentation de l'erreur lorsque le bruit STCN est important. La méthode EM-EOF est plus performante sur le champ g_3 lorsque le SNR augmente alors que la méthode EM-EOF étendue est plus robuste à une diminution de ce dernier, surtout

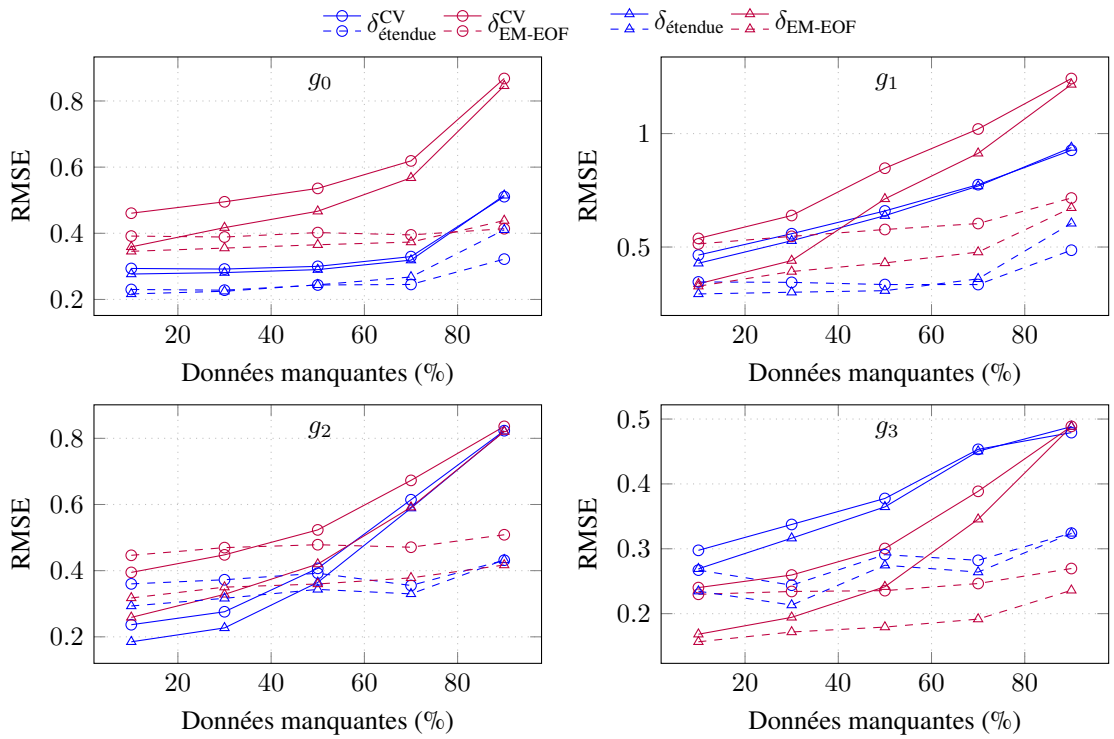


Figure 3.24 – RMSE en fonction de la quantité de données manquantes (%). Trait plein : données manquantes aléatoires et bruit SCN ; trait en pointillés : données manquantes corrélées et bruit STCN. SNR = 2.

si le bruit est spatialement corrélé, confirmant ainsi la possibilité de reconstruire des signaux à corrélation spatiale plus importante.

Enfin, on analyse les erreurs de reconstruction des champs de déplacement synthétiques par date (figure 3.26). De manière générale, on remarque que les erreurs varient au cours de la série temporelle. Cette variation est dûe à la dépendance de la RMSE à l'échelle d'amplitude des données ainsi qu'à la variance des données interpolées [Willmott2006] [6]. Ainsi, les erreurs du champ g_0 ont tendance à augmenter au cours de la série car l'amplitude moyenne du déplacement augmente linéairement. Concernant les autres champs, l'erreur varie de manière non-linéaire car l'amplitude moyenne des champs de déplacement est de nature oscillatoire. De plus, on observe sur tous les cas présentés une augmentation de l'erreur entre les dates 5 et 8, ce qui correspond aux données manquantes corrélées simulées sur une période seulement. On remarque que les RMSE δ augmente toutefois moins dans le cas de la méthode EM-EOF étendue, confirmant une nouvelle fois la performance de cette méthode vis-à-vis des données manquantes corrélées. Notons que la cross-RMSE est peu affectée par cette variation, puisque celle-ci est calculée sur des points choisis aléatoirement sur l'ensemble de la série temporelle.

Concernant les trois premiers cas de champs de déplacement g_0, g_1 et g_2, les erreurs δ^{CV} et δ sont plus faibles pour la méthode EM-EOF étendue que pour la méthode EM-EOF, indiquant que la reconstruction par la méthode EM-EOF étendue est plus proche des données perturbées (réelles), ce qui corrobore les observations précédentes. Dans un seul de ces cas (champ du premier ordre g_0), la reconstruction par la méthode EM-EOF est plus proche de la vérité ($\delta^{vrai}_{EM\text{-}EOF} < \delta^{vrai}_{étendue}$). Concernant le champ g_3, la méthode EM-EOF étendue fournit un champ recontruit plus proche de la vérité mais plus éloigné des données réelles dans le cas de données manquantes aléatoires.

6. À la différence de l'erreur moyenne absolue (MAE), dont la variabilité ne dépend pas de la variance des données interpolées.

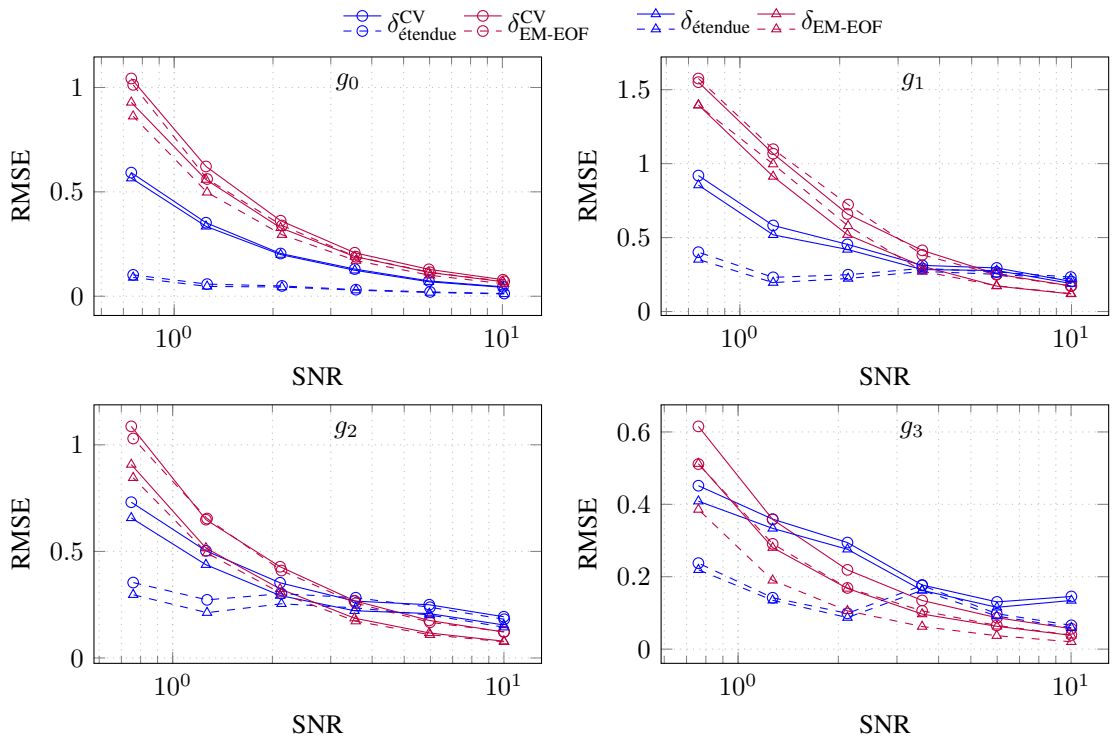


Figure 3.25 – RMSE en fonction du SNR. Trait plein : bruit STCN et données manquantes corrélées ; trait en pointillés : bruit SCN et données manquantes aléatoires. La quantité de données manquantes est fixée à 30%.

3.3.5 Bilan de l'étude synthétique

Si l'on récapitule ce qui a été énoncé lors de l'introduction, la méthode EM-EOF peut présenter certaines limites lorsque : 1) la corrélation spatiale du champ de déplacement prévaut sur la corrélation temporelle, 2) le champ de déplacement présente des caractéristiques spatiales hétérogènes et locales et 3) la série temporelle est courte avec plus de chance que des points soient peu observés au cours du temps.

Les résultats des simulations numériques nous permettent d'aborder directement ces limites.

Concernant le point 1), la prise en compte de la corrélation spatiale (par l'augmentation des données) permet de ne pas dépendre exclusivement de la corrélation temporelle lorsque peu d'observations temporelles sont disponibles, ce qui rejoint alors le point 3). Pour ce dernier, l'utilisation d'information spatiale redondante pour estimer une covariance augmentée permet de minimiser de potentielles discontinuités dans la reconstruction lorsque peu d'observation temporelles sont disponibles, effet notamment visible sur les séries temporelles reconstruites entre les résultats des méthodes EM-EOF étendue et EM-EOF. La méthode EM-EOF étendue permet d'interpoler plus fidèlement des séries temporelles courtes avec de grandes quantités de données manquantes. Ce cas constitue une limite à la méthode EM-EOF qui est plus performante lorsque la dimension temporelle N est grande. Le fenêtrage spatial des données permet de répondre au point 2) en traîtant des sous-ensembles de pixels relativement plus homogènes que l'image entière, réduisant ainsi de potentiels biais d'estimation de la matrice de covariance. Dans le cas d'un champ avec zone stable (champ g_3) dont le contour n'est pas connu, le fenêtrage peut provoquer des effets de bords mais permet d'éliminer les discontinuités dans les zones stables (méthode EM-EOF), elles-mêmes dues à une estimation biaisée de la matrice de covariance. Dans ce cas de figure, l'utilisation de la méthode EM-EOF est recommandée si le champ est peu bruité, alors qu'on préferera l'utilisation de la méthode EM-EOF étendue si le champ est fortement bruité.

En complément à ces points, le moyennage spatial permet de réduire la plupart des discon-

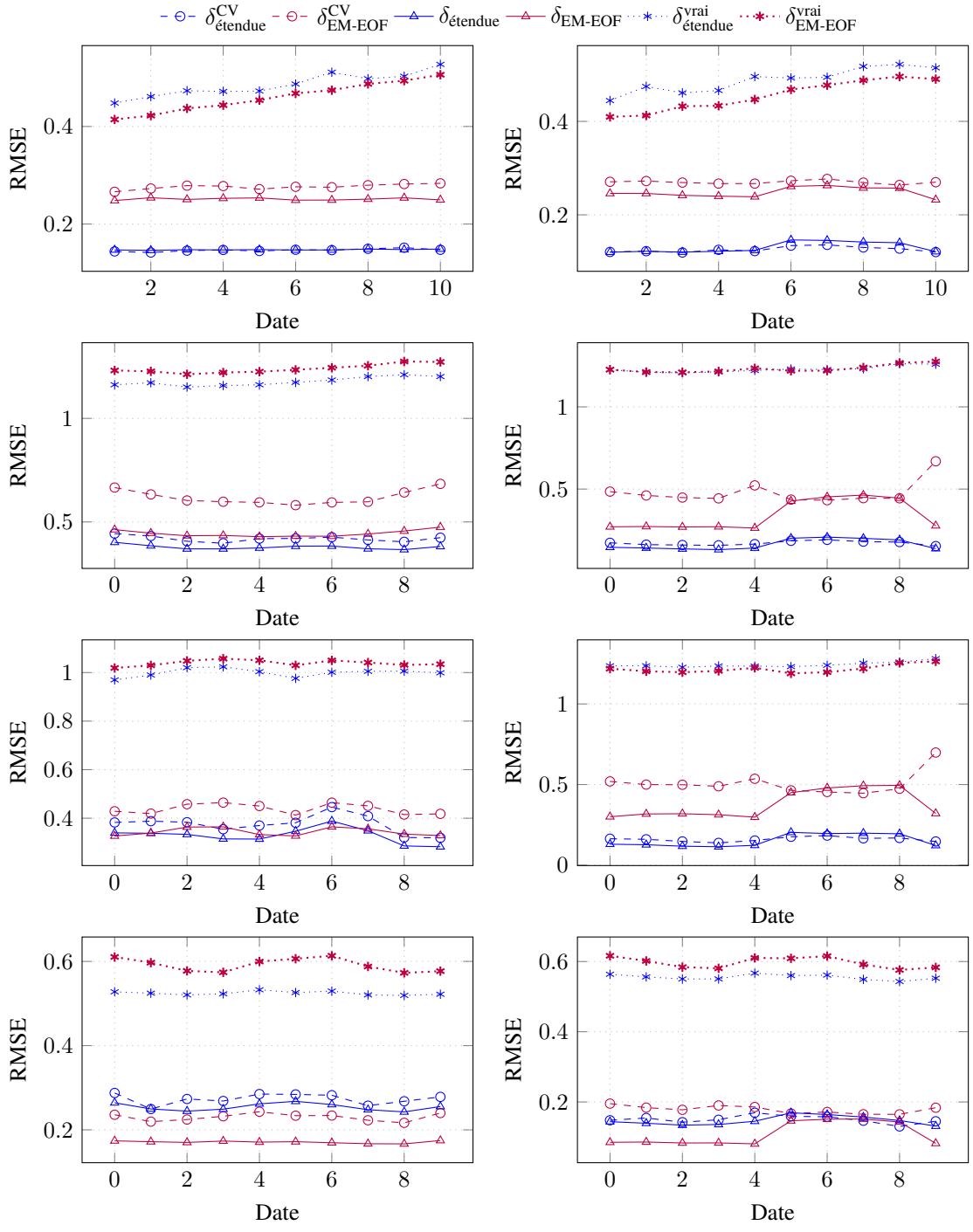


Figure 3.26 – Évolution de la RMSE dans les séries temporelles des champs g_0 (première ligne) à g_3 (dernière ligne). Colonne de gauche : données manquantes aléatoires ; colonne de droite : données manquantes corrélées. La quantité de données manquantes est fixée à 30% et le SNR à 2.

tinuités provoquées par les données manquantes, et ce d'autant plus que leur quantité augmente, comme l'a notamment souligné le second cas d'étude ainsi que l'analyse des erreurs en fonction de la quantité de données manquantes sur les champs g_0 et g_1. Ce moyennage a l'avantage de pouvoir fournir un résultat plus homogène lorsque le champ de déplacement est constitué de structures fortement corrélées spatialement (bruit SCN, données manquantes corrélées). À ce titre, les erreurs ont notamment montré une plus grande robustesse de la méthode EM-EOF étendue face aux variations du niveau de bruit SCN.

En ce qui concerne l'utilisation de la mesure de confiance $\mathcal{C}_k$, les simulations ont montré que celle-ci peut constituer une aide à la sélection du nombre de modes, elle-même rendue plus complexe du fait de la grande dimension des données, de la présence de bruit fortement corrélé et de grandes quantités de données manquantes (section 3.3.4). Lorsque le spectre ne contient pas de séparation dominante, le critère $\mathcal{C}_k$ est moins discriminant mais permet tout de même d'ajuster le nombre de modes, dont l'estimation repose essentiellement sur la cross-RMSE et le critère Λ.

Enfin, il paraît important de mentionner que le temps de calcul de la méthode EM-EOF étendue est en moyenne 100 fois supérieur à celui de la méthode EM-EOF. En effet, le fait d'opérer sur une covariance augmentée par échantillonnage spatial de taille M provoque une augmentation de la complexité algorithmique, passant de $\mathcal{O}(\hat{r}N + c)$ à $\mathcal{O}(\hat{r}MN + c)$. Cet élément doit donc être pris en compte dans la décision de mettre en oeuvre la méthode EM-EOF étendue sur de grands champs spatiaux nécessitant un fenêtrage spatial plus large. Il existe toutefois la possibilité de réduire la complexité algorithmique par des techniques d'implémentation efficace de l'EVD [Korobeynikov2010], ce qui constitue une suite logique de ce travail. Un bilan comparatif général des avantages et inconvénients des méthodes EM-EOF et EM-EOF étendue est présenté ci-après (tableau 3.3).

Méthode	**Taille de la série N**		**Type de bruit**		**SNR**		**Type de données manquantes**		**% données manquants**		**Temps de calcul**
	↗	↘	SCN	STCN	↗	↘	Aléa.	Corr.	↗	↘	
EM-EOF	$+++$	$-$	$+$	$+$	$++$	$+$	$+$	$+$	$+$	$++$	$+$
EM-EOF étendue	$+$	$++$	$++$	$+$	$+$	$++$	$++$	$+++$	$++$	$+$	$-$

Tableau 3.3 – Comparaison des méthodes EM-EOF étendue et EM-EOF selon la grille suivante : $+++$ très adaptée ; $++$ adaptée ; $+$ moyennement adaptée ; $-$ peu adaptée. Abbréviations - ↗ : Grand ; ↘ : Petit ; Aléa. : Aléatoire ; Corr. : Corrélé.

3.4 Application sur données optiques : le cas du glacier Fox

La méthode EM-EOF étendue est appliquée à un jeu de données de vitesse de surface obtenues par corrélation croisée d'images Sentinel-2 sur le glacier Fox dans les Alpes néo-zélandaises (figure 3.27). La chaîne de traitement utilisée pour obtenir ce produit de vitesse a été mise au point dans l'étude de [Millan2019].

Le tableau 3.4 récapitule les caractéristiques des données utilisées. Ce jeu de données consiste en une série temporelle de douze champs de vitesse de surface s'étendant entre février et septembre 2018 (figure 3.28). L'intervalle temporel (baseline) entre chaque image varie entre dix et quarante jours. La taille de chaque grille spatiale est $P = 100 \times 150 = 15000$ pixels. Chaque champ de vitesse contient des valeurs manquantes qui correspondent à des valeurs initialement retirées

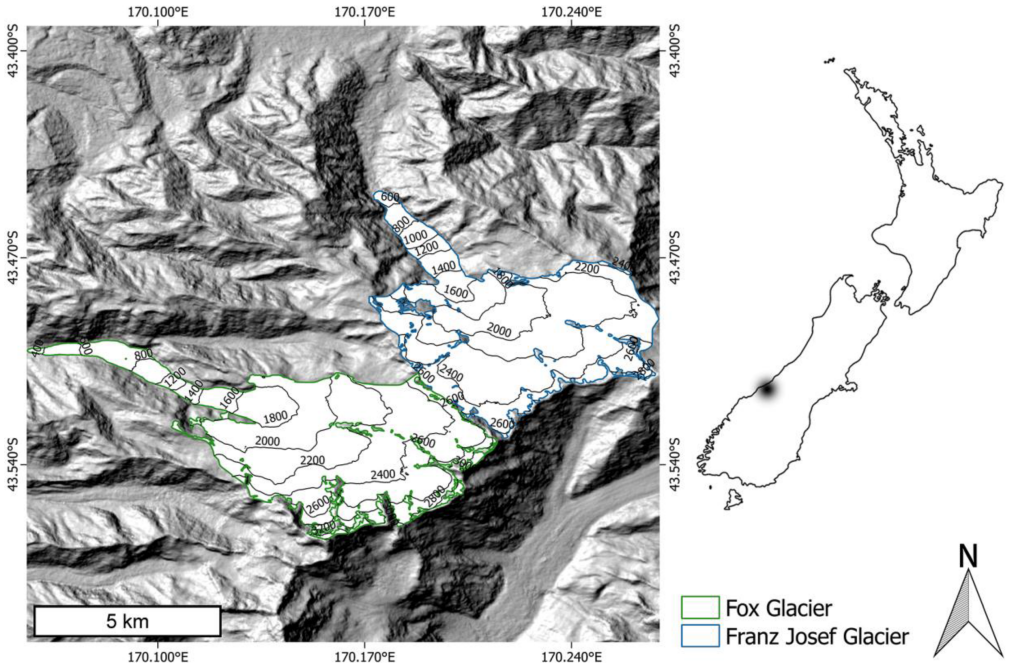


Figure 3.27 – Emplacement géographique du glacier Fox (contours vert et point noir en gradient) dans les Alpes du sud en Nouvelle-Zélande. Les contours sont ceux du Randolph Glacier Inventory (RGI) [Consortium2017]. Figure tirée de [Wang2015].

Période	**Plateforme**	**Type de données**	**Taille de la série**	**[min, max]% manquants**
02/2018-09/2018	Sentinel-2	Corrélation d'amplitude	12	[10, 60]%

Tableau 3.4 – Principales caractéristiques du jeu de données de vitesse de surface sur le glacier Fox.

à cause d'un calcul de corrélation défectueux, d'un pic de corrélation trop faible ou de valeurs aberrantes. La quantité de valeurs manquantes varie ainsi entre 10% et 60% du nombre de points par champ de vitesse. Les données manquantes sont fortement corrélées, notamment dans la partie basse du glacier (plus étroite) où le calcul de corrélation est plus délicat. Ainsi, certaines zones ne sont jamais observées au cours de la série temporelle. Avant lancement de la méthode, les valeurs manquantes sont initialisées par la moyenne spatiale. Les points de validation croisée sont choisis aléatoirement et leur nombre est fixé à 1% du total de points observés (non-manquants) par champ de vitesse. Enfin, le décalage spatial est fixé à $M = 225$ (fenêtre carrée de taille 15×15), ce qui, au regard de la quantité de points situés sur le glacier, correspond approximativement à la limite inférieure de l'intervalle de valeurs du décalage spatial proposé en sous-section 3.2.5.

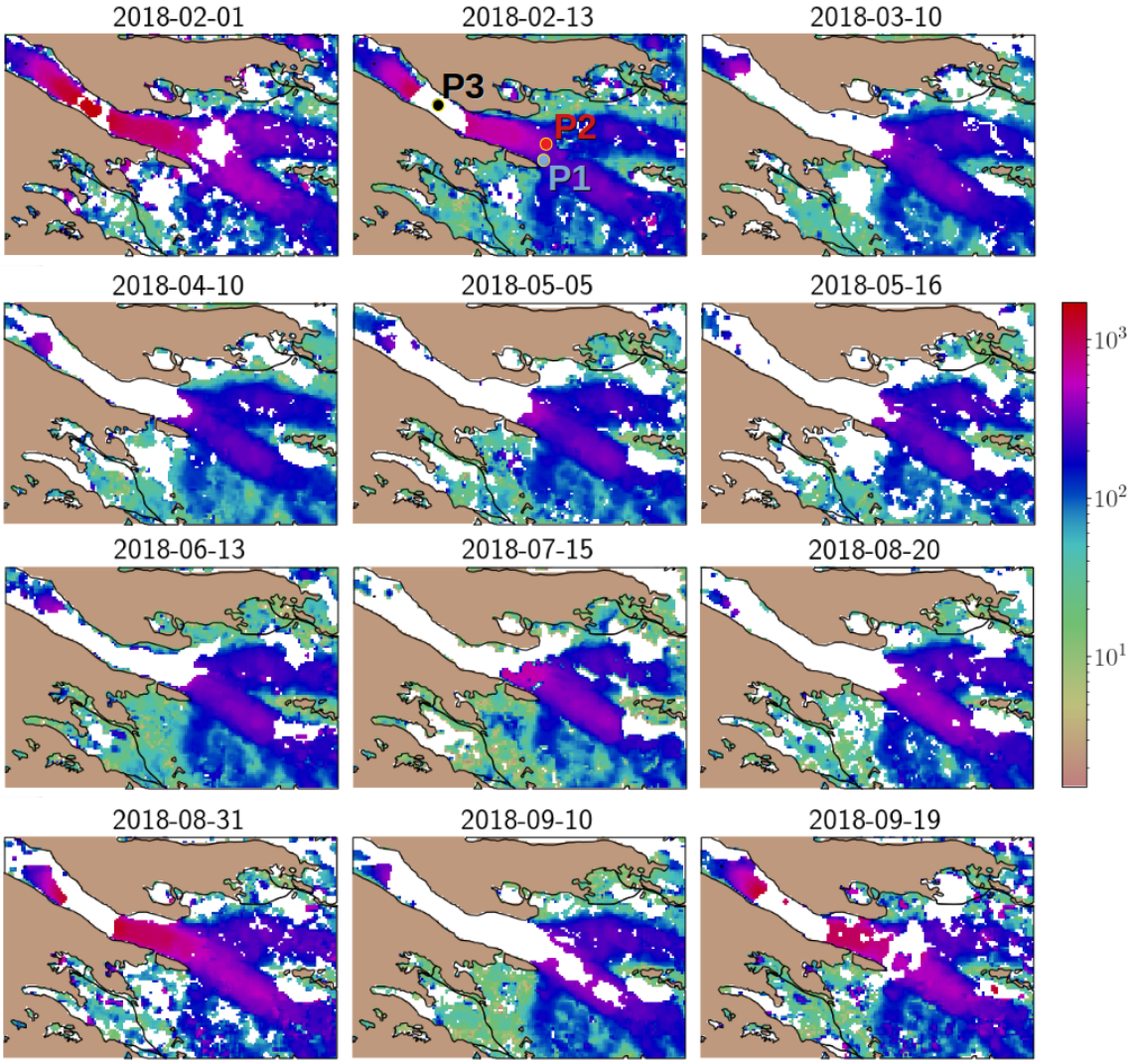


Figure 3.28 – Série temporelle de champs de vitesse de surface (mètres/an) obtenus par corrélation d'amplitude d'images optiques Sentinel-2 sur le glacier Fox entre février et mi-septembre 2018. L'emplacement des zones étudiées P1, P2 et P3 est également illustré à la date 2018-02-13. Les contours proviennent du Randolph Glacier Inventory (RGI) [Consortium2017].

Le nombre optimal de modes estimé est de 13. La figure 3.29 montre le spectre de valeurs propres du jeu de données, ainsi que la valeur de l'indice $\mathcal{C}_k$ associé à chaque valeur propre. On observe ici que le nombre de modes estimé correspond à un pic dans $\mathcal{C}_k$, ce qui coïncide avec une séparation au sein du spectre de valeurs propres. On peut également identifier trois multiplets qui correspondent aux valeurs propres $\{\lambda_5 - \lambda_7\}, \{\lambda_8 - \lambda_{11}\}$ et $\{\lambda_{12}, \lambda_{13}\}$, lesquels sont retenus au sein des données reconstruites. Les champs de vitesse reconstruits (figure 3.30) montrent que

la variation saisonnière (correspondant ici à un demi-cycle saisonnier) est reconstituée, avec des amplitudes de vitesse similaires à celles de [Millan2019]. On remarque que les vitesses de surface peuvent atteindre 1500 m/an dans la partie basse du glacier, laquelle est une zone étroite et pentue, voyant ainsi ses vitesses de surface accélérer. Ces valeurs hautes sont cohérentes avec la valeur maximale de vitesse de 4.5 m/jour en-dessous du principal mur de glace situé en aval du glacier [Herman2011, Kääb2016]. Notons que l'effet du moyennage sur les données reconstruites est visible puisque certaines aspérités spatiales sont systématiquement corrigées, notamment sur en amont du glacier (dates 2018-02-01, 2018-08-31 et 2018-09-19).

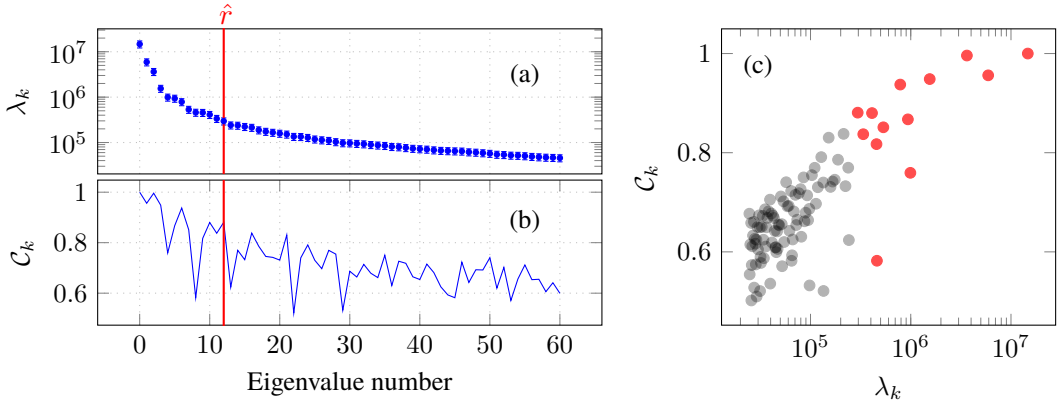


Figure 3.29 – (a) Valeurs propres λ_k (60 premières) du jeu de données augmenté sur le glacier Fox, (b) mesure de confiance associée $\mathcal{C}_k$ et (c) $\mathcal{C}_k$ versus λ_k. Les cercles et ligne rouges représentent les valeurs propres correspondant aux modes sélectionnés.

Afin d'évaluer la capacité de la méthode EM-EOF étendue à reconstruire la variabilité temporelle du champ de déplacement, des séries temporelles de vitesse de surface sur trois zones P1, P2 et P3 sont présentées en figure 3.31. Ces résultats sont également comparés à ceux de la méthode EM-EOF. P1 est la même zone que celle choisie dans l'étude de [Millan2019] et ne contient aucune valeur manquante au cours de la série temporelle. P2 est plus proche de la ligne d'écoulement centrale du glacier et contient cinq valeurs manquantes sur les douze dates d'observation. P3 est située dans la partie basse du glacier et ne contient qu'une seule valeur de vitesse observée à la date 2018-02-01 à cause d'une grande vélocité de surface dans cette zone. Du fait de la proximité des zones P1 et P2, nous considérons que l'évolution saisonnière des vitesses de surface ne varie qu'à un facteur d'échelle près entre ces deux points. Par conséquent, la tendance de l'évolution temporelle en P1 peut être utilisée à des fins de validation des valeurs de vitesse de déplacement reconstruites en P2.

On observe que les valeurs reconstruites en P2 sont globalement cohérentes avec les valeurs observées dans la série temporelle, c'est-à-dire dans l'intervalle d'erreur [7] dans la plupart des cas. Notons également que la reconstruction de la date 2018-07-15 par les deux méthodes est plus faible que la valeur observée. Une observation détaillée du champ de vitesse à cette date suggère la présence de valeurs aberrantes dans les zones limitrophes aux trous de données manquantes. La reconstruction en P3 montre que la tendance saisonnière est reconstituée et ce malgré la quasi absence de données observées. En comparaison avec la méthode EM-EOF, l'extension de celle-ci permet d'obtenir une amélioration des résultats de reconstruction avec un gain de $\simeq$15 m/an en moyenne, surtout dans la période d'avril à août où les trous de données manquantes sont importants. Cette observation souligne la contribution de l'exploitation de la corrélation spatiale, en plus de la corrélation temporelle, dans la reconstruction, ainsi que du moyennage des données augmentées

7. On notera que les erreurs ne sont pas issues de la reconstruction mais sont celles issues de la chaîne de traitement pour le calcul des vitesses de surface développée dans [Millan2019].

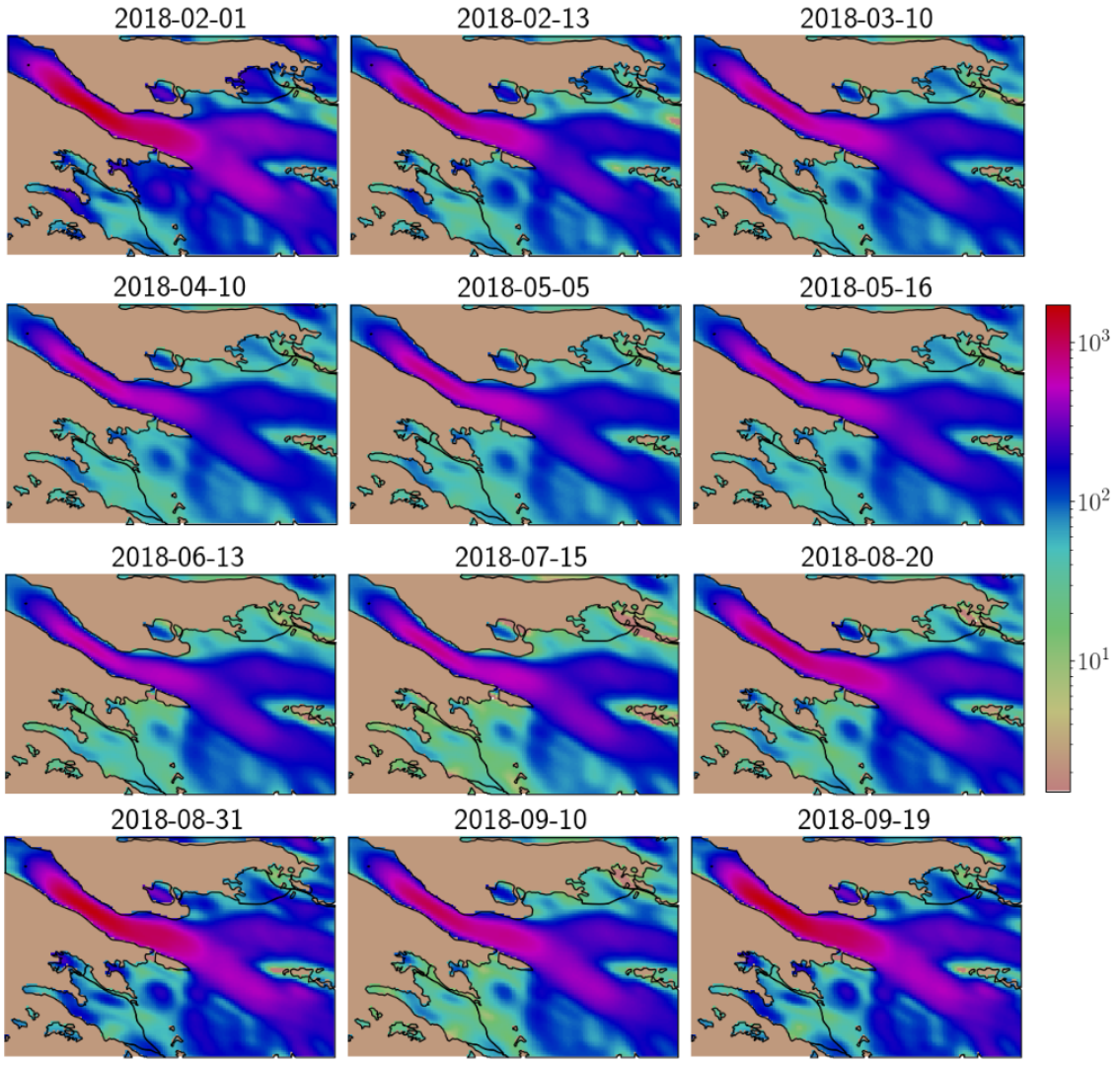


Figure 3.30 – Reconstruction de la série temporelle de champs de vitesse de surface (mètres/an) sur le glacier Fox entre février et mi-septembre 2018. Les contours proviennent du RGI.

reconstruites. Une comparaison entre l'évolution de P1, P2, P3 permet de remarquer que la même tendance saisonnière est reconstruite, et ce quelle que soit la zone considérée, y compris sur le bas de la langue glaciaire (P3) où très peu de données observées existent. En cela, la méthode EM-EOF étendue peut fournir une aide importante au glaciologue modélisateur afin de mieux contraindre des paramètres sous-glaciaires, telle que la contrainte de frottement basal, à partir de vitesses de surface spatio-temporellement résolues.

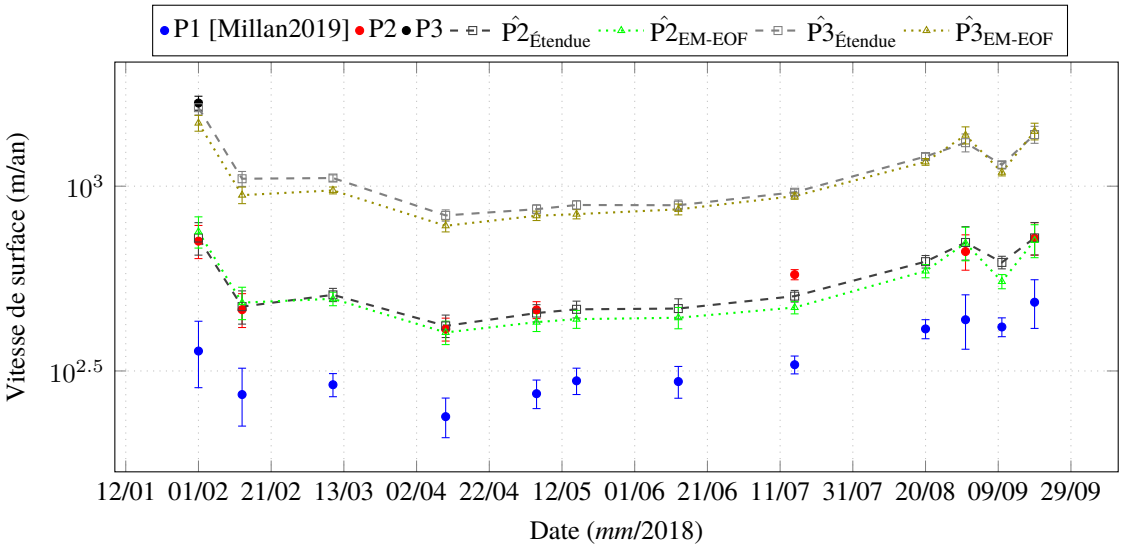


Figure 3.31 – Évolution des vitesses de surface entre février et mi-septembre 2018 sur les zones P1 (cercles bleus), P2 (cercles rouges), P3 (cercles noirs) et séries temporelles reconstruites en P2 et P3. Les intervalles d'erreur sont ceux de l'étude de [Millan2019].

3.5 Conclusion et perspectives

Dans ce chapitre, une extension de la méthode EM-EOF développée au chapitre 2 a été présentée pour la reconstruction de données manquantes au sein de séries temporelles de champs de déplacement. Par une augmentation spatiale des données, la matrice de covariance temporelle est étendue à une matrice de covariance spatio-temporelle. De plus, une sélection robuste du nombre de modes a été développée, se basant sur 1) l'erreur de validation croisée, 2) un critère permettant de prévenir la sur-estimation du nombre de modes et 3) l'analyse de l'incertitude des valeurs propres. Concernant ce dernier point, le problème de dégénérescence des valeurs propres est traîté en étendant la règle empirique proposée par [North1982] avec échantillonnage spatio-temporel. À partir de cette règle, une mesure de confiance associée à chaque valeur propre est construite afin d'aider à la détection du nombre optimal de modes. De plus, un intervalle sur le décalage spatial est proposé en utilisant des règles simples tirées des propriétés de l'estimation de la covariance et de la décorrélation spatiale.

Les simulations, associées à des comparaisons quantatives et qualitatives à la méthode EM-EOF, ont permis de mieux saisir les avantages et inconvénients de l'extension proposée, notamment lorsque la structure spatiale du champ de déplacement est perturbée par un bruit corrélé, mais aussi lorsque le champ de déplacement dispose de peu d'observations temporelles en certaines zones. Dans ce cas, la prise en compte de la corrélation spatiale permet de notamment réduire les biais d'estimation de la covariance et ainsi de fournir une interpolation temporelle adéquate. De plus, le moyennage spatio-temporel fournit un champ interpolé spatialement mieux résolu et dépourvu

de toute hétérogénéité. L'estimation du nombre de modes a lieu à travers trois critères cités ci-dessus : l'erreur de validation croisée, le critère Λ et la mesure de confiance $\mathcal{C}_k$. Ces trois techniques permettent d'estimer, de manière semi-automatique, le nombre optimal de modes pour reconstruire des données incomplètes. La mesure de confiance $\mathcal{C}_k$, dont les limites d'application existent et ont été dressées, apporte une connaissance supplémentaire sur la structure spectrale et permet d'ajuster l'estimation du nombre de modes en présence de bruit corrélé contribuant significativement à la variance du signal (voir section 1.4.3).

À travers l'application à une série temporelle de champs de vitesse de surface sur le glacier Fox, en Nouvelle Zélande, la capacité de la méthode EM-EOF étendue à interpoler une série temporelle courte de champs de vitesses de surface disposant de grandes quantités de données manquantes spatialement corrélées a été montrée. La méthode EM-EOF étendue permet ainsi d'interpoler un signal de déplacement complexe même lorsque peu d'observations temporelles sont disponibles. Cette conclusion s'appuie notamment sur la reconstitution du cycle saisonnier des vitesses de surface sur plusieurs zones du le glacier, y compris des régions où très peu de données sont disponibles à cause des difficultés que représentent leur calcul par des méthodes d'estimation du déplacement. Une comparaison avec la méthode EM-EOF montre également un gain de précision de l'ordre de 15 mètres par an par rapport aux données observées, confirmant l'intérêt de cette extension.

De manière générale, cet outil peut constituer une aide pour augmenter la taille effective de séries temporelles courtes et incomplètes pour des applications diverses en mesure de déplacement de surface. En particulier, concernant l'application à l'étude des glaciers alpins, il peut s'agir d'une étape supplémentaire vers l'obtention de vitesses de surface complètes et continues, ce qui permettrait in fine d'approfondir la connaissance des paramètres rhéologiques qui contrôlent le déplacement de surface (e.g. contrainte de cisaillement basal, viscosité de la glace, hydrologie sous glaciaire) [Rabatel2018]. En effet, l'étude de l'effet de l'hydrologie sur le glissement des glaciers nécessite des produits de vitesses complets et résolus, ce qui est peu souvent le cas. Cette méthode peut aussi être considérée comme complémentaire à la méthode EM-EOF, puisqu'on préfère utiliser cette dernière lorsque la dimension temporelle est grande.

Une perspective possible de ce travail est d'estimer une matrice de covariance spatio-temporelle à l'aide d'une fenêtre spatiale adaptative, c'est-à-dire de travailler directement avec un masque sur les contours de la cible observée. De tels travaux pourraient notamment s'inspirer de la *Shaped* 2D-SSA [Golyandina2015]. Cela permettrait en théorie d'éviter ou de diminuer les effets de bords potentiels créés sur les zones de transition entre la zone de déplacement et les zones fixes qui l'entourent, comme le glacier et ses rives.

4

Vers une estimation robuste de la matrice de covariance de données incomplètes

Sommaire

4.1 Introduction . . . 102
4.2 Modèles statistiques et type de données manquantes . . . 102
4.2.1 Modélisation statistique . . . 103
4.2.2 Type de données manquantes . . . 106
4.3 Principe de l'algorithme EM . . . 107
4.3.1 Estimation du maximum de vraisemblance . . . 107
4.3.2 L'algorithme EM . . . 108
4.4 Estimation de la matrice de covariance en modèle Gaussien . . . 108
4.4.1 Estimation avec données manquantes de forme générale . . . 109
4.4.2 Estimation en rang faible . . . 111
4.4.3 Simulations numériques . . . 112
4.5 Estimation robuste de la matrice de covariance . . . 113
4.5.1 Estimation avec données manquantes en bloc . . . 114
4.5.2 Estimation en rang faible . . . 117
4.5.3 Simulations numériques . . . 117
4.6 Comparatif avec la méthode EM-EOF dans le cas Gaussien . . . 122
4.6.1 Estimation de la matrice de covariance sur données synthétiques . . . 122
4.6.2 Reconstruction de données manquantes sur données réelles . . . 123
4.6.3 Discussion . . . 130
4.7 Synthèse . . . 131

4.1 Introduction

Ce chapitre a pour objet d'élargir le cadre d'analyse des données manquantes en mesure de déplacement par télédétection et plus généralement à tout type de signal détecté sujet à une incomplétude de données. Ce cadre a pour ambition d'aborder le problème des données manquantes au sens large, c'est-à-dire d'un point de vue statistique, et ce quelle que soit l'application qui en découle. Ce paradigme nécessite le plus souvent de formuler des hypothèses sur le modèle statistique paramétrique s'exprimant à l'aide d'une distribution de probabilité décrivant le comportement des données reçues. Dès lors que ces hypothèses sont posées, on peut s'intéresser à l'estimation de paramètres statistiques dont dépend cette distribution. Ces paramètres sont l'objet d'une attention particulière de la part du chercheur confronté à tout type de données car ils permettent de décrire, d'analyser et de prédire la variabilité des données, c'est-à-dire de mieux connaître les données.

Jusqu'à présent, lors des chapites 2 et 3, nous n'avons formulé aucune hypothèse particulière sur la distribution des données de mesure de déplacement étudiées. Cela est notamment dû au fait que le problème a été posé sous la forme d'un *problème d'interpolation* des données manquantes (approche prédictive) et non sous celle d'un *problème d'estimation paramétrique* à partir de données incomplètes (approche paramétrique). La matrice de covariance calculée jusqu'à présent prenait la forme :

$$\mathbf{\Sigma} = \frac{1}{N}\mathbf{X}\mathbf{X}^T \tag{4.1}$$

Ce calcul est rendu possible car les données manquantes de $\mathbf{X}$ sont, rappelons-le, initialisées. En effet, le calcul des fonctions empiriques orthogonales (EOFs), ou vecteurs propres, ne peut être obtenu directement à partir de données incomplètes non initialisées [Beckers2003]. Il faut alors rendre les données complètes pour calculer (4.1). Par ailleurs, (4.1) est un *estimateur* connu de la matrice de covariance, appelée matrice de covariance empirique (SCM pour *Sample Covariance Matrix*). Cet estimateur est adapté pour des données suivant une distribution gaussienne, ce qui n'est pas toujours le cas.

Une autre approche consiste à ne pas initialiser les données manquantes : la matrice de covariance est alors estimée à partir des données observées. L'estimateur, qui dépend de la distribution des données, peut être approché par des méthodes itératives. Plus spécifiquement, le but est de trouver des *statistiques exhaustives*, c'est-à-dire à même de décrire statistiquement les données manquantes sachant les données observées. Ceci est en partie l'objet de ce chapitre. D'autre part, on se posera la question de la robustesse de l'estimation paramétrique, c'est-à-dire comment estimer des paramètres statistiques robustes à des données reçues contenant des valeurs aberrantes ou fortement hétérogènes, ce qui est très souvent le cas en mesure de déplacement terrestre. Par exemple, le déroulement de la phase interférométrique peut être sujet à des erreurs ou sauts de phase, produisant ainsi des valeurs aberrantes et localisées de déplacement de surface. On s'intéressera donc, en dernier ressort, à l'estimation robuste de paramètres décrivant des données incomplètes.

4.2 Modèles statistiques et type de données manquantes

Dans la plupart des applications de traitement du signal en télédétection, on s'intéresse à modéliser la distribution des données reçues. Cette distribution dépend des paramètres statistiques des données, comme la moyenne [1] (statistique du premier ordre) et/ou la matrice de covariance

1. Dans ce chapitre, la moyenne est considérée nulle.

(statistique du second ordre). Ces paramètres étant la plupart du temps indisponibles, l'enjeu consiste à émettre une ou plusieurs hypothèses sur la distribution des données. Cela permet de se référer à un modèle probabiliste qui dépend de paramètres dont l'estimation est rendue possible, directement ou indirectement.

Lorsque les données sont incomplètes, l'estimation des paramètres nécessite souvent la mise en oeuvre d'une stratégie différente car il s'agit d'estimer les paramètres des données manquantes à partir des données observées. Cette stratégie dépend de la forme des données manquantes. Prenons l'exemple d'un capteur extérieur recevant un signal continu depuis l'espace. Si la réception du signal est sensible aux intempéries, un certain nombre de données reçues devront potentiellement être retirées lors du traitement car trop bruitées ou atypiques. Dans ce cas, la forme des données manquantes est, sur le long terme, quasi-aléatoire, car elle dépend de la météo qui est un système chaotique. À l'inverse, si le capteur subit une vérification mensuelle entraînant un arrêt temporaire de la réception, la forme des données est alors périodisée.

4.2.1 Modélisation statistique

Dans cette section, nous introduisons les modèles statistiques utilisés dans ce chapitre. Chaque modèle est défini par une loi statistique associée à une fonction de densité de probabilité (on utilisera simplement le raccourci $p.d.f$ pour désigner une telle fonction). Notons que de nombreux éléments de cette section, notamment les définitions et références citées, s'appuient sur les thèses de Mélanie Mahot [Mahot2012] et d'Ammar Mian [Mian2019], ainsi que sur l'Habilitation à diriger des Recherches d'Arnaud Breloy, dont la préparation est en cours.

Dans ce qui suit, on représente par $\{\mathbf{y}_i\}_{i\in[1,N]} \in \mathbb{R}^P$ l'échantillon de taille N de vecteurs réels indépendants et identiquement distribués (vecteurs iid) de dimension P.

La distribution gaussienne

Définition 4.2.1. *Loi gaussienne*

Le vecteur réel $\mathbf{y}$ *de taille* P *suit une distribution gaussienne centrée (ou normale) si sa p.d.f. s'écrit :*

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{P/2}|\mathbf{\Sigma}|^{1/2}} exp\left(-\frac{1}{2}\mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y} \right) \tag{4.2}$$

où $\mathbf{\Sigma} = \mathbb{E}[\mathbf{y}\mathbf{y}^T]$ *est la matrice de covariance de* $\mathbf{y}$.

Dans toute notre étude, la moyenne statistique est considérée comme nulle, c'est-à-dire qu'on l'exprime sous la forme d'un vecteur de taille P composé de 0, noté $\mathbf{0}$. Cette distribution est ainsi notée $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. La distribution gaussienne est populaire dans de nombreuses applications de traitement de signal et en télédétection. Il est commun de formuler une hypothèse de gaussianité sur la distribution des données, notamment du fait de l'absence de connaissance particulière sur les caractéristiques du signal reçu comme sa phase. De plus, la distribution gaussienne admet un estimateur optimal de $\mathbf{\Sigma}$ au sens du maximum de vraisemblance (EMV), dont les calculs constituent des résultats standards en analyse multivariée [2] [Anderson1965, Rao1972]. Dans le cas gaussien, $\mathbf{\Sigma}$ est la SCM. Si les vecteurs $\mathbf{y}_i$ suivent une loi gaussienne, la SCM s'écrit alors :

$$\hat{\mathbf{\Sigma}}_{\text{SCM}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i\mathbf{y}_i^T \tag{4.3}$$

2. L'analyse de données multivariées examine le comportement d'observations issues de plusieurs variables à la fois dont dépendent des modèles statistiques.

Cet estimateur est simple d'utilisation lorsqu'on ne dispose pas d'information sur la distribution des données. C'est d'ailleurs cette forme que nous avons utilisée dans les méthodes EM-EOF et EM-EOF étendue pour calculer respectivement les covariances empiriques temporelles (2.4) et spatio-temporelles (3.5). Cependant, de nombreuses études ont mis en évidence l'existence d'applications qui traîtent des données dont la distribution est non-gaussienne. En traitement du signal radar haute résolution (HR) et en analyse de données hyperspectrales, on pourra notamment citer les travaux de [Jakeman1980, Gini2000, Manolakis2001, Theiler2005, Shnidman2005]. En mesure de déplacement terrestre, l'hypothèse gaussienne peut également s'avérer insuffisante [Dehecq2015, Mantovani2019], et ce alors que la connaissance de la distribution du déplacement est nécessaire dans de nombreux cas de modélisation utilisant la covariance des données [Hergert2010, Smittarello2019b]. Dans ces exemples, et plus généralement lorsque les données contiennent des valeurs aberrantes, l'estimateur (4.3) est peu *robuste*, terme dont nous dresserons les contours dans ce qui suit. Il est donc naturel de chercher un modèle robuste face à des données parfois hétérogènes, comme le modèle elliptique, ce qui est l'objet de la section suivante.

Les distributions elliptiques symétriques

Lorsque l'on traite des images ou des signaux potentiellement corrompus par diverses perturbations, l'hypothèse de gaussianité peut se révéler insuffisante. Le but est alors de trouver un modèle robuste qui puisse englober un grand nombre de distributions tout en étant capable de fournir une estimation précise. C'est le cas des distributions elliptiques symétriques [Ollila2012], qui généralisent notamment les distributions gaussiennes et de nombreuses distributions à queue lourde (*heavy-tail*), qui sont particulièrement adéquantes pour prendre en compte des données présentant des valeurs aberrantes [Greco2007, Gao2010].

Définition 4.2.2. *Modèle elliptique (distribution elliptique symétrique)*
Le vecteur complexe $\mathbf{y}$ *suit une distribution centrée elliptique symétrique (ES) si sa p.d.f s'écrit :*

$$f(\mathbf{y}) = C|\mathbf{\Sigma}|^{-1} g_{\mathbf{y}}\left(\mathbf{y}^T \mathbf{\Sigma}^{-1} \mathbf{y}\right) \tag{4.4}$$

où $\mathbf{\Sigma}$ *est la matrice de covariance de* $\mathbf{y}$*,* C *désigne une constante de normalisation et* $g_{\mathbf{y}} : [0, \infty) \to [0, \infty)$ *est toute fonction, appelée génératrice de densité, garantissant que (4.4) soit une p.d.f. Cette distribution est notée* $\mathcal{ES}(\mathbf{0}, \mathbf{\Sigma}, g_{\mathbf{y}})$*. De plus, ce vecteur admet la représentation stochastique suivante :*

$$\mathbf{y} = \sqrt{\mathcal{Q}}\mathbf{A}\mathbf{v} \tag{4.5}$$

où $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}^T$*,* $\mathbf{v}$ *est uniformément distribuée sur la sphère complexe* $\mathcal{U}_1^P$ *et* $\mathcal{Q}$ *est une variable aléatoire réelle non-négative, appelée variable modulaire, indépendante de* $\mathbf{v}$ *avec une p.d.f dépendante seulement de* $g_{\mathbf{y}}$*.*

Soit $\mathbf{y}_i \sim \mathcal{ES}(\mathbf{0}, \mathbf{\Sigma}, g_{\mathbf{y}})$. Un M-estimateur, noté $\hat{\mathbf{\Sigma}}$, est alors défini par solution de l'équation de point-fixe suivante :

$$\hat{\mathbf{\Sigma}} = \frac{1}{N} \sum_{i=1}^{N} u(\mathbf{y}_i^T \hat{\mathbf{\Sigma}}^{-1} \mathbf{y}_i) \mathbf{y}_i \mathbf{y}_i^T \tag{4.6}$$

où u est toute fonction de pondération réelle définie sur $[0, \infty)$ respectant les conditions de Maronna [Maronna1976] et garantissant l'existence et l'unicité de (4.6). Lorsque ces conditions sont respectées, cet estimateur peut être calculé par un algorithme de point-fixe (PF) défini par :

$$\mathbf{\Sigma}^{(m+1)} = \mathcal{H}(\mathbf{\Sigma}^{(m)}) \tag{4.7}$$

où m désigne l'indice d'itération de l'algorithme et $\mathcal{H} : \mathbb{R}^{P \times P} \to \mathbb{R}^{P \times P}$ une fonction différentiable définie par (4.6). Lorsque $u(t) = -g'_{\mathbf{y}}(t)/g_{\mathbf{y}}(t)$, l'estimateur (4.6) correspond à

l'EMV de la matrice de covariance de $\mathbf{y} \sim \mathcal{ES}(\mathbf{0}, \mathbf{\Sigma}, g_{\mathbf{y}})$. Cependant, certains M-estimateurs sont construits avec une fonction u qui n'est pas reliée à $g_{\mathbf{z}}$, cette dernière étant souvent inconnue en pratique. C'est notamment le cas de l'estimateur de Tyler, auquel on s'intéresse dans la sous-section suivante.

Robustesse : de la distribution $\mathcal{ES}$ au gaussien composé

Afin de définir le modèle gaussien composé, reprenons la forme (4.5) du modèle CES défini précédemment. Soit un vecteur $\mathbf{y}$ suivant une distribution elliptique symétrique. Il est possible de représenter ce vecteur par le modèle :

$$\mathbf{y} = \frac{\sqrt{\mathcal{Q}}}{||\mathbf{n}||}\mathbf{A}\mathbf{n} \tag{4.8}$$

où $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Le vecteur $\mathbf{x} = \mathbf{A}\mathbf{n}$ est appelé coeur gaussien de $\mathbf{y}$ [Draškovic2018]. Si l'on pose $\sqrt{\tau} = \sqrt{\mathcal{Q}}/||\mathbf{n}||$, où τ est un scalaire positif et indépendant de $\mathbf{n}$, alors le modèle (4.8) appartient à la famille des modèles gaussiens composés. Un vecteur $\mathbf{y} \in \mathbb{R}^P$ suit une distribution gaussienne composée s'il admet la représentation stochastique suivante :

$$\mathbf{y} = \sqrt{\tau}\mathbf{x} \tag{4.9}$$

où $\mathbf{x}$ est un vecteur complexe gaussien de dimension P dont la moyenne est un vecteur nul et la matrice de covariance est $\mathbb{E}\{\mathbf{y}\mathbf{y}^T\} = \mathbf{\Sigma}$, où $(\mathbf{\Sigma})_{ii} = 1$ pour $i = 1, \ldots, P$. En écriture compacte, on note donc $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. Le paramètre τ est généralement appelé *texture*. La distribution de (4.9) est notée $\mathcal{N}(\mathbf{0}, \tau_i\mathbf{\Sigma}, f_\tau)$, où f_τ est la p.d.f de τ. De manière générale, f_τ n'est pas connue : on peut alors relaxer le modèle gaussien composé de sorte que τ soit une variable déterministe inconnue. L'EMV de ce modèle prend alors la forme :

$$\hat{\mathbf{\Sigma}} = \frac{P}{N}\sum_{i=1}^{N}\frac{\mathbf{y}_i\mathbf{y}_i^T}{\mathbf{y}_i^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{y}_i}, \qquad \hat{\tau}_i = \frac{\mathbf{y}_i^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{y}_i}{P} \tag{4.10}$$

Cet estimateur est connu sous le nom d'estimateur de Tyler [Tyler1987,Pascal2008]. En ce sens, l'estimation est "robuste" car la distribution de (4.10) ne dépend pas de la distribution initiale des données [Maronna2006, Zoubir2018]. De plus, le modèle gaussien composé possède une certaine flexibilité puisqu'il est possible d'assigner des poids différents (τ_i) aux observations qui pourraient s'apparenter à des données aberrantes ou atypiques.

Modèle par facteur

Dans de nombreuses applications en télédétection et notamment en traitement du signal radar, il est assez commun que les signaux étudiés vivent dans un sous-espace à faible dimension. Pour faire un parallèle avec l'analyse en EOF ou l'ACP, on dira dans ces cas qu'il est possible de représenter le signal à partir d'un nombre réduit R de fonctions orthogonales ou de composantes principales. Dans le cas présent, cette répresentation se traduit par une structure particulière de la matrice de covariance, que l'on nomme structure rang faible :

$$\mathbf{\Sigma} = \mathbf{\Sigma}_R + \sigma^2\mathbf{I} \tag{4.11}$$

où $\mathbf{\Sigma}_R$ est une matrice positive semi-définie de rang R plus faible que la dimension des données P et définie par :

$$\mathbf{\Sigma}_R \overset{\text{EVD}}{=} \sum_{i=1}^{R}\lambda_i\mathbf{u}_i\mathbf{u}_i^T \tag{4.12}$$

$\mathbf{u}_i$ sont les vecteurs orthogonaux issus de la décomposition en valeur propres (EVD) de $\mathbf{\Sigma}_R$ et λ_i sont les valeurs propres associées à chaque $\mathbf{u}_i$. Le modèle auquel se réfère l'équation (4.11) est plus communément appelé modèle par facteur et est directement lié à l'analyse en EOF. Afin de bénéficier de l'avantage du modèle par facteur, celui-ci peut être combiné aux distributions gaussienne et gaussienne composée, que l'on désigne alors respectivement par $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_R + \sigma^2\mathbf{I})$ et $\mathcal{N}(\mathbf{0}, \tau_i(\mathbf{\Sigma}_R + \sigma^2\mathbf{I}))$.

4.2.2 Type de données manquantes

Dans cette section, nous apportons des détails supplémentaires aux mécanismes liés à l'incomplétude de données et aux formes de données manquantes déjà définis lors du chapitre 1. Ces éléments sont ici définis suivant le formalisme en vigueur en analyse statistique de données manquantes.

Forme des données manquantes

Il est utile, au sein d'un jeu de données, de distinguer les formes que prennent les données manquantes parmi les données observées. Suivant l'origine des données manquantes, celles-ci peuvent ainsi être distribuées aléatoirement, ou suivre des formes bien définies.

Soit $\mathbf{Y}$ un jeu de données rectangulaire contenant des données manquantes, que l'on représente par une matrice de taille $(P \times N)$. $\mathbf{Y}$ est représentée sous la forme d'une partition $\mathbf{Y} = \{\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}\}$, où $\mathbf{Y}_{\text{obs}}$ est la partie observée des données et $\mathbf{Y}_{\text{mis}}$ est la partie manquante. Chaque colonne de $\mathbf{Y}$ est un vecteur aléatoire de dimension P observé N fois de sorte que :

$$\mathbf{Y} = \{\mathbf{y}_i = (y_{1,i}, y_{2,i}, \ldots, y_{P,i})^T\}, \qquad i = 1, \ldots, N \tag{4.13}$$

Ainsi, la variable à l'indice i observée à l'indice j est noté y_{ij}. Trois représentations formes de données manquantes sont illustrées en figure 4.1. Certaines méthodes s'appliquent à toutes les formes, alors que d'autres ne s'appliquent qu'à certains types de formes en particulier.

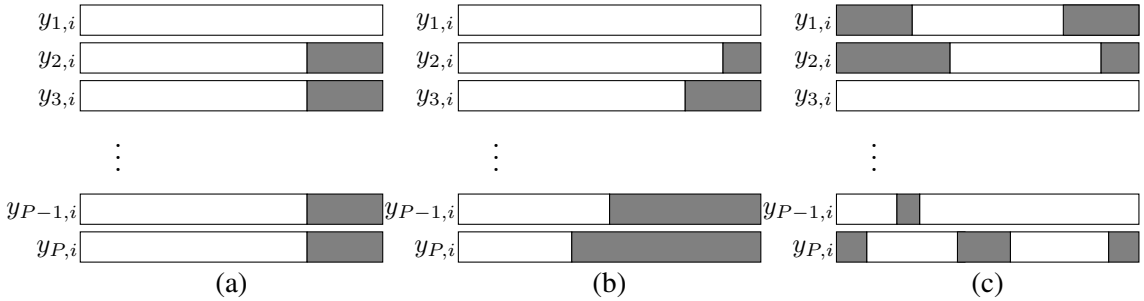


Figure 4.1 – Forme de données manquantes : (a) bloc, (b) monotone et (c) générale. En blanc : valeurs observées ; en gris : valeurs manquantes.

Dans cette étude, nous traîterons des formes en bloc et générale. Si l'on reprend les termes utilisés lors des chapitres 2 et 3, ces formes correspondent respectivement à des données manquantes corrélées et aléatoires au sein des données. Dans certains cas particuliers, la forme en bloc peut être obtenue par une permutation en ligne et en colonne de la matrice $\mathbf{Y}$. Cela consiste alors à trouver des matrices de permutation en ligne $\mathbf{L}$ et colonne $\mathbf{C}$ tel que $\mathbf{Y}_{\text{bloc}} = \mathbf{LYC}$. Notons que cette opération n'est pas toujours possible et dépend essentiellement de la configuration initiale des données manquantes $\mathbf{Y}$.

Dans la suite du manuscrit, quelle que soit la forme des données manquantes considérée, la partie observée de $\mathbf{Y}$ est représentée comme suit :

$$\mathbf{Y}_{\text{obs}} = (\mathbf{y}_{1,\text{obs}}, \ldots, \mathbf{y}_{N,\text{obs}}) \tag{4.14}$$

où $\{\mathbf{y}_{i,\text{obs}}\}_{i\in[1,N]} \in \mathbb{R}^{P_i\times 1}$ est l'ensemble de P_i variables observées dans la $i$ème observation, c'est-à-dire la $i$ème colonne de $\mathbf{Y}$. De manière équivalente, on notera $\{\mathbf{y}_{i,\text{mis}}\}$ l'ensemble de $P-P_i$ variables manquantes dans la $i$ème observation de $\mathbf{Y}$.

Mécanismes liés à l'incomplétude de données

Soit $\mathbf{M}$ une matrice dont les éléments m_{ij} valent 1 si l'élément y_{ij} est manquant et 0 si y_{ij} est observé. Cette matrice est appelée matrice indicatrice des données manquantes. Le mécanisme des données manquantes est caractérisé par la distribution conditionnelle de $\mathbf{M}$ sachant $\mathbf{Y}$, notée $p(\mathbf{M}|\mathbf{Y},\boldsymbol{\theta})$, où $\boldsymbol{\theta}$ est le vecteur des paramètres inconnus du modèle. On rappelle que les trois mécanismes liés à l'incomplétude de données sont MCAR, MAR et MNAR. Lorsque les données manquantes sont MCAR, la probabilité d'occurence des données manquantes est aléatoire, soit, en terme de probabilité conditionnelle :

$$p(\mathbf{M}|\mathbf{Y},\boldsymbol{\theta}) = p(\mathbf{M}|\boldsymbol{\theta}) \tag{4.15}$$

Ici, l'incomplétude ne dépend donc pas des données complètes $\mathbf{Y}$. Lorsque les données manquantes sont MAR, la probabilité que les données soient manquantes ne dépend que des données observées $\mathbf{Y}_{\text{obs}}$, soit :

$$p(\mathbf{M}|\mathbf{Y},\boldsymbol{\theta}) = p(\mathbf{M}|\mathbf{Y}_{\text{obs}},\boldsymbol{\theta}) \tag{4.16}$$

Finalement, le cas NMAR consiste à remplacer $\mathbf{Y}_{\text{obs}}$ par $\mathbf{Y}_{\text{mis}}$ dans l'expression ci-dessus, puisque dans ce cas la probabilité que les données soient manquantes ne dépend que de la valeur des données manquantes. Dans l'ensemble de cette étude, les données seront supposées MAR.

4.3 Principe de l'algorithme EM

Afin d'estimer la covariance du signal représenté par $\mathbf{Y}$, nous faisons appel à l'estimation itérative du maximum de vraisemblance. Pour cela, l'algorithme espérance maximisation (EM), dont nous avons introduit quelques éléments lors du chapitre 1 (section 1.5.1), propose un schéma de calcul itératif des paramètres statistiques convergeant vers les paramètres optimaux au sens de l'estimation du maximum de vraisemblance (EMV). Rappelons que $\mathbf{Y} = \{\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}\}$, où $\mathbf{Y}_{\text{obs}}$ est la partie observée des données et $\mathbf{Y}_{\text{mis}}$ est la partie manquante.

4.3.1 Estimation du maximum de vraisemblance

Afin d'introduire l'algorithme EM, on rappelle ici quelques définitions. Soit $\boldsymbol{\theta} \in \boldsymbol{\Omega}_\theta$ le vecteur de paramètres appartenant à l'espace des paramètres $\boldsymbol{\Omega}_\theta$. L'*estimation du maximum de vraisemblance* (EMV) d'un paramètre $\boldsymbol{\theta}$ sont les valeurs de $\boldsymbol{\theta}$ qui maximisent la fonction de vraisemblance $\mathcal{L}(\boldsymbol{\theta}|\mathbf{Y}_{\text{obs}})$:

$$\hat{\boldsymbol{\theta}}_{\text{EMV}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Omega}_\theta}{\arg\max}\, \mathcal{L}(\boldsymbol{\theta}|\mathbf{Y}_{\text{obs}}) \tag{4.17}$$

Le logarithme étant une fonction monotone croissante, le maximum de la fonction de vraisemblance intervient aux mêmes valeurs de $\boldsymbol{\theta}$ que le maximum du logarithme de la fonction de vraisemblance, appelé *fonction du log-vraisemblance* et définie par $\ell(\boldsymbol{\theta}|\mathbf{Y}_{\text{obs}}) = \log \mathcal{L}(\boldsymbol{\theta}|\mathbf{Y}_{\text{obs}})$. Si la fonction du log-vraisemblance est différentiable et admet une limite supérieure, l'EMV peut être calculé en dérivant la vraisemblance par rapport à $\boldsymbol{\theta}$ et en posant l'équation suivante :

$$D_\ell(\boldsymbol{\theta}, \mathbf{Y}_{\text{obs}}) \equiv \frac{\partial \ell(\boldsymbol{\theta}|\mathbf{Y}_{\text{obs}})}{\partial \boldsymbol{\theta}} = 0 \tag{4.18}$$

Cette équation est appelée *équation de vraisemblance*. En réalité, cette équation est un ensemble d'équations définies par la dérivée partielle de $\ell(\boldsymbol{\theta}|\mathbf{Y}_{\text{obs}})$ par rapport à tous les paramètres qui composent $\boldsymbol{\theta}$. Dans de nombreux modèles, cette équation peut être résolue explicitement mais le problème de maximisation (4.17) n'admet pas de solution analytique et n'est pas directement implémentable. Dans ce cas, des méthodes itératives qui convergent vers le EMV peuvent être appliquées. De nombreuses méthodes existent, dont les plus populaires sont les méthodes basées sur le gradient [3] (descente de gradient, sous-gradient, gradient conjugé, gradient conditionnel) et sur l'algorithme de Newton-Raphson. Ces méthodes, dont certaines sont à présent perçues comme obsolètes (comme la descente de gradient), sont aussi difficilement implémentables, en particulier car elles nécessitent de calculer la dérivée seconde de la fonction du log-vraisemblance.

4.3.2 L'algorithme EM

Lorsque les données contiennent des données manquantes, une solution alternative est d'utiliser l'algorithme Espérance-Maximisation (EM), qui a notamment l'avantage de ne pas avoir recours au calcul des dérivées secondes et possède aujourd'hui des applications variées (on pourra consulter l'ouvrage de [Little2002] ou la synthèse de [Gupta2011]). L'algorithme EM, initialement développé dans l'étude fondatrice de [Dempster1977], permet de relier directement les EMV de $\boldsymbol{\theta}$ obtenues à partir de $\ell(\boldsymbol{\theta}|\mathbf{Y}_{\text{obs}})$ aux EMV obtenus à partir de $\ell(\boldsymbol{\theta}|\mathbf{Y})$. Après initialisation des paramètres $\boldsymbol{\theta}$, l'algorithme EM consiste à alterner, de manière itérative, l'estimation de l'espérance conditionnelle des "données manquantes" à partir des données observées et de l'estimation courante des paramètres (étape E), puis à rechercher $\boldsymbol{\theta}$ qui maximise la fonction du log-vraisemblance (étape M). L'utilisation du terme "données manquantes" est ici un abus de langage, car il ne s'agit pas à proprement dit de remplacer les données manquantes mais plutôt de remplacer les fonctions de $\mathbf{Y}_{\text{mis}}$ apparaissant dans le log-vraisemblance des données complètes $\ell(\theta|\mathbf{Y})$. Cette stratégie permet ainsi d'assurer la convergence de $\ell(\hat{\boldsymbol{\theta}}_{\text{EM}}|\mathbf{Y})$ vers $\ell(\hat{\boldsymbol{\theta}}_{\text{EMV}}|\mathbf{Y})$.

Plus spécifiquement, soit $\boldsymbol{\theta}^{(m)}$ l'estimé courant à l'itération m des paramètres $\boldsymbol{\theta}$. L'étape E de l'algorithme EM consiste à trouver l'espérance du log-vraisemblance des données complètes :

$$\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}) = \int \ell(\boldsymbol{\theta}|\mathbf{Y}) f(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}, \theta^{(m)}) d\boldsymbol{Y}_{mis} \tag{4.19}$$

L'étape M est une mise à jour des paramètres $\boldsymbol{\theta}$ à partir des données remplies à l'étape E en maximisant la fonction surégatoire $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)})$:

$$\mathcal{Q}(\boldsymbol{\theta}^{(m+1)}|\boldsymbol{\theta}^{(m)}) \geq \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}) \tag{4.20}$$

Le but est alors de répéter les étapes E et M en incrémentant m jusqu'à ce qu'un critère d'arrêt soit atteint, par exemple en calculant successivement l'écart $||\boldsymbol{\theta}^{(m+1)} - \boldsymbol{\theta}^{(m)}||$. L'algorithme EM possède de nombreuses variations [Liu1999, Little2002] à convergence plus rapide. Nous nous contenterons, dans cette étude, d'utiliser la version "classique" de l'EM.

4.4 Estimation de la matrice de covariance en modèle Gaussien

Dans un premier temps, nous faisons l'hypothèse de données centrées gaussiennes contenant des données manquantes de forme générale (Fig. 4.1 (b)). L'hypothèse de gaussianité constitue

3. Dont l'origine remonte à Augustin-Louis Cauchy, mais dont les propriétés de convergence seront démontrées presque 100 ans plus tard par Haskell Curry [Curry1944].

une étape vers la distribution gaussienne composée, dont nous avons vu que la flexibilité est plus adaptée aux mesures contenant des données aberrantes ou atypiques.

4.4.1 Estimation avec données manquantes de forme générale

Soit $\mathbf{Y} = \{\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}\} = \{\mathbf{y}_i\}_{i\in[1,N]} \in \mathbb{R}^P$ une matrice contenant des données incomplètes (voir (4.13)) décrivant une forme générale. On note $\mathbf{Y}_{\text{obs}} = \{\mathbf{y}_{i,\text{obs}}\}_{i\in[1,N]} \in \mathbb{R}^{P_i}$ la partie observée de $\mathbf{Y}$, où chaque observation $\mathbf{y}_{i,\text{obs}}$ contient P_i variables. De manière équivalente, on note $\{\mathbf{y}_{i,\text{mis}}\}$ l'ensemble de $P - P_i$ variables manquantes dans la $i$ème observation de $\mathbf{Y}$. On suppose ici que tous les $\mathbf{y}_i$ suivent une distribution gaussienne centrée multivariée :

$$\{\mathbf{y}_i\} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{4.21}$$

où $\boldsymbol{\Sigma} = (\sigma_{ij})$ est la matrice de covariance de taille $(P \times P)$. Le but est ici d'estimer la matrice de covariance de la distribution (4.21) qui admet pour densité de probabilité (4.2). Considérons le vecteur des paramètres $\boldsymbol{\zeta}$ de taille $P^2/2$ contenant la diagonale et les entrées de la matrice triangulaire inférieure de $\boldsymbol{\Sigma}$. Si l'on note $\boldsymbol{\theta} = \boldsymbol{\zeta}^T$, le modèle ci-dessus admet la fonction de vraisemblance suivante :

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{Y}) = -N \log|\boldsymbol{\Sigma}| - \sum_{i=1}^{N} \mathbf{y}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{y}_i \tag{4.22}$$

Pour estimer $\boldsymbol{\theta}$, on a recours à une estimation paramétrique à l'aide du maximum de vraisemblance (voir section 4.3.1), lequel permet d'estimer les paramètres de la distribution au sens de l'EMV :

$$\hat{\boldsymbol{\theta}}_{\text{EMV}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}|\mathbf{Y}) \tag{4.23}$$

En réalité, la fonction de vraisemblance ne dépend que des données observées $\mathbf{Y}_{\text{obs}}$. Ainsi, maximiser $\mathcal{L}(\boldsymbol{\theta}|\mathbf{Y})$ revient à maximiser $\mathcal{L}(\boldsymbol{\theta}|\mathbf{Y}_{\text{obs}})$, soit à poser le problème suivant :

$$\hat{\boldsymbol{\theta}}_{\text{EMV}} = \arg\max_{\boldsymbol{\theta}} - \sum_{i=1}^{N} \log|\boldsymbol{\Sigma}_{i,\text{obs}}| - \sum_{i=1}^{N} \mathbf{y}_{i,\text{obs}}^T \boldsymbol{\Sigma}_{i,\text{obs}}^{-1} \mathbf{y}_{i,\text{obs}} \tag{4.24}$$

où $\boldsymbol{\Sigma}_{i,\text{obs}}$ est la matrice de covariance de $\mathbf{y}_{i,\text{obs}}$. La maximisation de $\mathcal{L}(\boldsymbol{\theta}|\mathbf{Y}_{\text{obs}})$ revient à résoudre l'équation du maximum de vraisemblance (4.18). Dans ce cas, celle-ci n'admet pas de solution analytique : on ne peut donc pas maximiser l'expression (4.22) de manière directe. C'est ce qui justifie la mise en place d'un algorithme itératif de type EM dont la solution converge vers $\hat{\boldsymbol{\theta}}_{\text{EMV}}$. Plutôt que de procéder à un remplacement (*imputation*) direct des données manquantes qui ne serait pas optimal, on estime les statistiques exhaustives [4] des données manquantes, lesquelles admettent une relation linéaire à la fonction de vraisemblance (4.24). Pour notre problème, les statistiques exhaustives sont définies par :

$$s_j = \sum_{i=1}^{N} y_{ij}, \quad j = 1, \ldots, P; \qquad\qquad s_{jk} = \sum_{i=1}^{N} y_{ij} y_{ik}, \quad j, k = 1, \ldots, P \tag{4.25}$$

L'EM consiste alors à maximiser l'expression (4.24) en calculant à chaque étape l'espérance conditionnelle des quantités (4.25) sachant les données observées $\mathbf{Y}_{\text{obs}}$ et les paramètres $\boldsymbol{\theta}$. Si $\boldsymbol{\theta}^{(m)} = \boldsymbol{\zeta}^{(m)^T}$ désigne l'estimée courante des paramètres à l'itération m de l'algorithme EM, les étapes E et M sont alors décrites par :

4. Les paramètres statistiques d'une famille de distribution sont dits exhaustifs (on parle de statistique exhaustive) si l'échantillon à partir duquel les paramètres sont calculés ne donne aucune autre information que ces paramètres statistiques [Fisher1922].

Initialisation - À l'itération 0, les paramètres initialisés sont arrangés dans une matrice $\boldsymbol{\Theta}$ de taille $(P+1)\times(P+1)$:

$$\boldsymbol{\Theta}^{(0)} = \begin{bmatrix} -1 & \mathbf{0} \\ \mathbf{0}^T & \hat{\boldsymbol{\Sigma}} \end{bmatrix}^{(0)} \tag{4.26}$$

où $\hat{\boldsymbol{\Sigma}}$ est obtenue à partir des données observées.

Étape E - Calcul de l'espérance des statistiques exhaustives (équation 4.25) à partir de $\mathbf{Y}_{\text{obs}}$ et $\boldsymbol{\theta}^{(m)}$:

$$\mathbb{E}[s_j|\mathbf{Y}_{\text{obs}},\boldsymbol{\theta}^{(m)}] = \mathbb{E}\left[\sum_{i=1}^{N} y_{ij}|\mathbf{Y}_{\text{obs}},\boldsymbol{\theta}^{(m)}\right] \tag{4.27}$$

$$\mathbb{E}[s_{jk}|\mathbf{Y}_{\text{obs}},\boldsymbol{\theta}^{(m)}] = \mathbb{E}\left[\sum_{i=1}^{N} y_{ij}y_{ik}|\mathbf{Y}_{\text{obs}},\boldsymbol{\theta}^{(m)}\right] \tag{4.28}$$

Cela revient à calculer $\mathbb{E}[y_{ij}|\mathbf{Y}_{\text{obs}},\boldsymbol{\theta}^{(m)}]$ et $\mathbb{E}[y_{ij}y_{ik}|\mathbf{Y}_{\text{obs}},\boldsymbol{\theta}^{(m)}]$. Pour cela, on se munit de l'opérateur sweep [Goodnight1979], outil fournissant un moyen simple et pratique de faire les calculs de maximum de vraisemblance pour des données incomplètes, dont une description détaillée est fournie en Annexe B.1. Soit un vecteur $\mathbf{y}_i$, où au moins une valeur manquante existe. On suppose que les indices $1 \leq j_1,\ldots,j_s \leq P$ correspondent aux positions des variables observées de $\mathbf{y}_i$, c'est-à-dire $\mathbf{y}_{i,\text{obs}} = (y_{j_1,i},\ldots,y_{j_s,i})^T$. La distribution conditionnelle $\mathbf{y}_{i,\text{mis}}|\mathbf{Y}_{\text{obs}}$ est équivalente à $\mathbf{y}_{i,\text{mis}}|\mathbf{y}_{i,\text{obs}}$ car les variables sont considérées indépendantes. Cette distribution peut être obtenue en appliquant l'opérateur sweep sur la matrice $\boldsymbol{\Theta}$ successivement sur les positions des variables observées $j_1,\ldots,j_s$ de $\mathbf{y}_i$:

$$\mathbf{B} = \text{SWP}[j_1,\ldots,j_s]\boldsymbol{\Theta}^{(m)} \tag{4.29}$$

$\mathbf{B} = (b_{ij})_{i,j\in[0,P]}$ contient les estimés du maximum de vraisemblance de la régression linéaire multivariée des données manquantes $\{\mathbf{y}_{i,\text{mis}}\}$ sur les données observées $\{\mathbf{y}_{i,\text{obs}}\}$. Ainsi, les quantités (4.27) et (4.28) sont calculées comme suit :

$$\mathbb{E}[y_{ij}|\mathbf{Y}_{\text{obs}},\boldsymbol{\theta}^{(m)}] = \begin{cases} b_{0j} + \sum_{k\in\{j_1,\ldots,j_s\}} b_{kj}y_{ik} & \text{si } y_{ij} \text{ est manquant} \\ y_{ij} & \text{si } y_{ij} \text{ est observé} \end{cases} \tag{4.30}$$

$$\mathbb{E}[y_{ij}y_{ik}|\mathbf{Y}_{\text{obs}},\boldsymbol{\theta}^{(m)}] = \mathbb{E}[y_{ij}|\mathbf{Y}_{\text{obs}},\boldsymbol{\theta}^{(m)}]\mathbb{E}[y_{ik}|\mathbf{Y}_{\text{obs}},\boldsymbol{\theta}^{(m)}] + Cov[y_{ij},y_{ik}|\mathbf{Y}_{\text{obs}},\boldsymbol{\theta}^{(m)}] \tag{4.31}$$

où $Cov[.]$ désigne la covariance conditionnelle exprimée par :

$$Cov[y_{ij},y_{ik}|\mathbf{Y}_{\text{obs}},\boldsymbol{\theta}^{(m)}] = \begin{cases} b_{jk} & \text{si } y_{ij} \text{ et } y_{ik} \text{ sont manquants} \\ 0 & \text{si } y_{ij} \text{ ou } y_{ik} \text{ sont manquants} \end{cases}$$

Étape M - Mise à jour des paramètres $\boldsymbol{\theta}^{(m+1)}$ à partir des statistiques exhaustives nouvellement estimées $\mathbb{E}[s_j|\mathbf{y}_{\text{obs}},\boldsymbol{\theta}^{(m)}]$ et $\mathbb{E}[s_{jk}|\mathbf{y}_{\text{obs}},\boldsymbol{\theta}^{(m)}]$ à l'étape E :

$$\boldsymbol{\Theta}^{(m+1)} = \begin{bmatrix} -1 & \mathbf{0} \\ \mathbf{0}^T & \boldsymbol{\Sigma}^{(m+1)} \end{bmatrix} = \text{SWP}[0]N^{-1}\mathbf{S} \tag{4.32}$$

où $\mathbf{S}$ contient les espérances calculées ci-dessus :

$$\mathbf{S} = \begin{bmatrix} n & \mathbb{E}[y_{i1}] & \dots & \mathbb{E}[y_{iP}] \\ \mathbb{E}[y_{i1}] & \mathbb{E}[y_{i1}y_{i1}] & \dots & \mathbb{E}[y_{i1}y_{iP}] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[y_{iP}] & \mathbb{E}[y_{iP}y_{i1}] & \dots & \mathbb{E}[y_{iP}y_{iP}] \end{bmatrix} \tag{4.33}$$

Notons que l'étape M revient à calculer l'EMV de la matrice de covariance du modèle (4.21), qui est la SCM déjà définie en (4.3) :

$$\hat{\mathbf{\Sigma}} = \mathbf{C}/N \tag{4.34}$$

où $\mathbf{C} = (c_{jk})$ est la somme des produits croisés dont les éléments sont données par $c_{ij} = \sum_{i=1}^{N} y_{ij}y_{ik}$.

Algorithme L'algorithme EM est décrit par le pseudo-code ci-après (algorithme 4). La convergence de l'algorithme est évaluée à partir de l'erreur quadratique moyenne normalisée (NMSE) entre l'estimé à l'itération m et $m+1$ définie par :

$$\text{NMSE} = \frac{||\hat{\mathbf{\Sigma}}^{(m+1)} - \hat{\mathbf{\Sigma}}^{(m)}||_{\text{F}}^2}{||\hat{\mathbf{\Sigma}}^{(m)}||_{\text{F}}^2} \tag{4.35}$$

Lorsque la NMSE passe sous un certain seuil de tolérance tol prédéfini, l'algorithme prend fin.

Algorithme 4 EM pour l'estimation de la covariance en présence de données manquantes.

Entrée: -
Sortie: $\hat{\mathbf{\Sigma}}$

1: Initialiser $\mathbf{\Sigma}^{(0)} = \mathbf{I}$
2: **tant que** NMSE < tol **faire**
3: Calculer $\mathbb{E}[y_{ij}|\mathbf{Y}_{\text{obs}}, \mathbf{\Sigma}^{(m)}]$, $\mathbb{E}[y_{ij}y_{ik}|\mathbf{Y}_{\text{obs}}, \mathbf{\Sigma}^{(m)}]$ ▷ Étape E
4: Calculer $\mathbf{\Sigma}^{(m+1)} = \text{SWP}[0]N^{-1}\mathbf{S}$ ▷ Étape M
5: Calculer NMSE
6: $\hat{\mathbf{\Sigma}}^{(m)} \leftarrow \hat{\mathbf{\Sigma}}^{(m+1)}$
7: **fin tant que**

4.4.2 Estimation en rang faible

Afin de procéder à une réduction de la dimension des données, nous prenons le modèle de covariance par facteur (4.11) qui est adapté aux données dont le comportement physique peut être exprimé par un nombre réduit de composantes spectrales, en utilisant des outils comme la transformée de Fourier, l'ACP ou les EOFs.

Comme souligné lors de la présentation du modèle par facteur, la distribution gaussienne peut être associée à une covariance de structure rang faible, où les vecteurs $\{\mathbf{y}_i\}$ suivent la distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_R+\sigma^2\mathbf{I})$ où $\mathbf{\Sigma}_R = \sum_{i=1}^{R} \lambda_i \mathbf{u}_i\mathbf{u}_i^T$ et $R \leq P$ est le rang de $\mathbf{\Sigma}_R$. Le problème de maximisation (4.24) est ainsi soumis à la contrainte de structure sur la covariance $\mathbf{\Sigma} = \mathbf{\Sigma}_R + \sigma^2\mathbf{I}$. Une solution globale à ce problème [Tipping1999] existe et prend la forme suivante :

$$\mathbf{R} = \sum_{i=1}^{R} \hat{\lambda}_i \mathbf{u}_i\mathbf{u}_i^T + \hat{\sigma}^2\mathbf{I} \tag{4.36}$$

où

$$\hat{\sigma}^2 = \frac{1}{P-R}\sum_{i=R+1}^{P}\lambda_i \tag{4.37}$$

$$\hat{\lambda}_i = \lambda_i - \hat{\sigma}^2, \quad \text{pour } i = 1,\dots,R \tag{4.38}$$

et $\mathbf{\Sigma} \overset{\text{EVD}}{=} \sum_{i=1}^{P}\lambda_i\mathbf{u}_i\mathbf{u}_i^T$, où $\lambda_i < \cdots < \lambda_P$ sont les valeurs propres de $\mathbf{\Sigma}$ et $\mathbf{u}_i$ ses vecteurs propres.

Algorithme Le processus itératif de l'estimation rang faible est décrit dans le pseudo-code ci-contre (algorithme 5). À la sortie de l'algorithme EM, les solutions (4.37) et (4.38) sont calculées à partir de l'EVD de la matrice de covariance. La condition d'arrêt est fixée sur la convergence de la NMSE (4.35) entre les estimés à l'itération courante et à l'itération précédente vers un seuil de tolérance tol défini à l'avance.

Algorithme 5 Estimation de la covariance avec structure rang faible (forme inspirée de [Sun2016]).

Entrée: -
Sortie: $\hat{\mathbf{R}}$

1: **tant que** NMSE < tol **faire**
2: Estimer $\hat{\mathbf{\Sigma}}^{(m)}$ par l'algorithme 4 ▷ Algorithme EM
3: $\hat{\mathbf{\Sigma}}^{(m)} \overset{\text{EVD}}{=} \sum_{i=1}^{P}\lambda_i\mathbf{u}_i\mathbf{u}_i^T$
4: Calculer $\hat{\sigma}^2$, $\hat{\lambda}_i$ ▷ équations (4.37) et (4.38)
5: $\hat{\mathbf{R}}^{(m)} = \sum_{i=1}^{R}\hat{\lambda}_i\mathbf{u}_i\mathbf{u}_i^T + \hat{\sigma}^2\mathbf{I}$
6: Calculer NMSE
7: $m \leftarrow m+1$
8: **fin tant que**

4.4.3 Simulations numériques

Afin de valider les performances de l'estimateur proposé, celui-ci est calculé sur des données incomplètes simulées de dimension $P = 15, 10$ et $N \in [60, 330]$. Les données ont une distribution gaussienne multivariée de matrice de covariance $\mathbf{\Sigma}$ dont les éléments sont communément définis par la structure de Toeplitz suivante :

$$\Sigma_{ij} = \rho^{|i-j|} \tag{4.39}$$

pour $i, j \in [1, P]$ et $0 \leq \rho \leq 1$. Les données manquantes, dont la quantité varie entre 10% et 50% du total des données, sont générées aléatoirement afin de garantir l'obtention d'une forme générale. La covariance estimée $\hat{\mathbf{\Sigma}}$ est comparée à la vraie la matrice de covariance $\mathbf{\Sigma}$ en calculant la distance riemannienne (géodésique) [Bhatia2009] suivante :

$$\delta^2_{\mathcal{SH}_{++}}(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}) = ||\log(\mathbf{\Sigma}^{-1/2}\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1/2})||_2^2 \tag{4.40}$$

Cette distance correspond à la métrique de Fisher pour les distributions gaussiennes composées sur l'ensemble des matrices symétriques (hermitiennes dans le cas complexe) définies semi-positives $\mathcal{SH}_{++}$. Cette distance possède notamment des propriétés d'invariance aux transformations potentielles que peuvent subir les données [Skovgaard1984]. Les estimateurs suivants sont considérés pour une comparaison des performances :

— L'estimé $\hat{\mathbf{\Sigma}}$ issu de l'algorithme EM (pseudo-code 4)
— La SCM définie par l'équation (4.3) ;

— La SCM estimée sur une partie des données correspondant à la quantité de données observées en pourcentage $\boldsymbol{\Sigma}^{\text{obs}\%}$:

$$\hat{\boldsymbol{\Sigma}}_{\text{SCM}}^{\text{obs}\%} = \sum_{i=1}^{\frac{100}{\text{obs}}N} \mathbf{y}_i \mathbf{y}_i^T \tag{4.41}$$

Par exemple, si les données contiennent 10% de données manquantes, on estimera la SCM sur 90% des données $\hat{\boldsymbol{\Sigma}}^{90\%}$. Même si cet estimé n'est pas strictement équivalent à l'estimé $\hat{\boldsymbol{\Sigma}}$, il fournit néanmoins la possibilité, lorsque les données sont simulées, de comparer deux matrices de covariance estimées à partir de la même quantité de données observées ;

— L'estimé $\hat{\mathbf{R}}$ issu de l'algorithme EM rang faible (pseudo-code 5) ;

— L'estimation de (4.3) en rang faible $\hat{\mathbf{R}}_{\text{SCM}}$;

— L'estimation de (4.41) en rang faible $\hat{\mathbf{R}}_{\text{SCM}}^{\text{obs}\%}$.

Les résultats (figure 4.2) permettent de remarquer que pour une faible quantité de données manquantes (10%), la SCM "incomplète" est légèrement plus performante que l'estimé proposé ici, écart qui s'accentue dans le cas rang faible (rang = 4). Lorsque l'on fait augmenter la quantité de données manquantes (30%), les performances des estimés $\hat{\mathbf{R}}_{\text{SCM}}^{70\%}$ et $\hat{\mathbf{R}}$ sont équivalentes (rang = 3). En définitive, ces simulations numériques fournissent un indicateur de la performance de l'estimé proposé par la mise en évidence de sa cohérence avec les EMV de données gaussiennes multivariées. En section 4.6, cette estimé sera comparé à la matrice de covariance estimée par la méthode EM-EOF (chapitre 2).

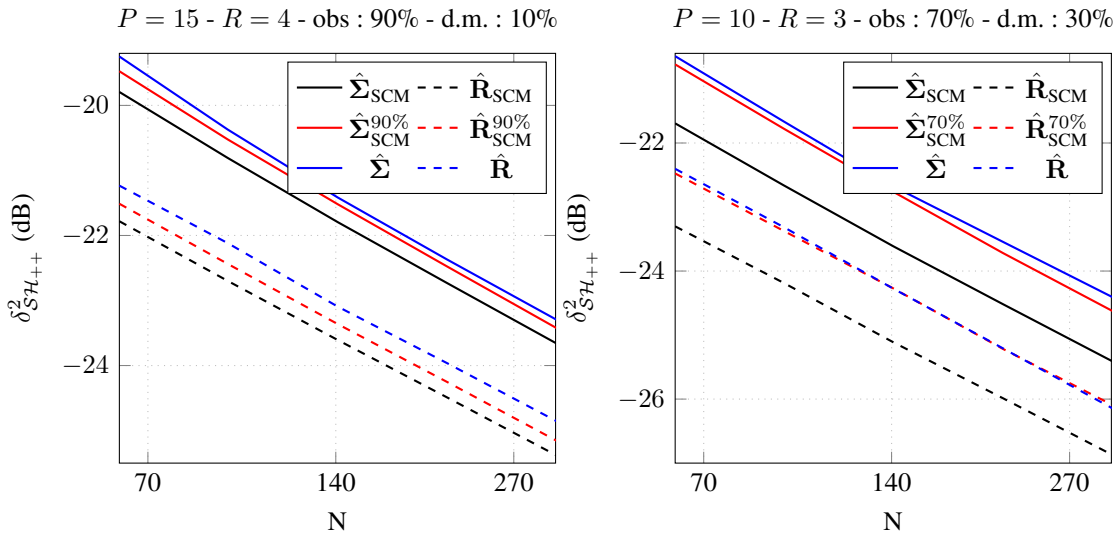


Figure 4.2 – Distance naturelle en fonction du nombre d'observations dans deux configurations (obs : pourcentage de données observées ; d.m. : pourcentage de données manquantes).

4.5 Estimation robuste de la matrice de covariance

Il peut arriver que les données manquantes soit regroupées en un seul bloc (figure 4.1). Considérons par exemple un ensemble de stations mesurant le déplacement terrestre grâce à un signal GPS (*Global Positioning System*) reçu quotidiennement depuis un satellite. Si trois de ces stations nécessitent une maintenance en même temps sur une période donnée, on peut aisément comprendre que les mesures manquantes seront regroupées. Si ces mesures sont insérées dans une

matrice de données $\mathbf{Y}$ de sorte que chaque station soit sur les lignes de $\mathbf{Y}$ et chaque observation soit sur ses colonnes, on peut obtenir une forme en bloc en permutant les lignes. Alors que nous avons étudié des données gaussiennes en section précédente, nous nous situons à présent dans le cas où les données présentent une distribution gaussienne composée, hypothèse plus adaptée à la mesure de déplacement potentiellement soumise à diverses perturbations. Même si l'hypothèse de données manquantes en bloc est probable, elle constitue une étude préliminaire à l'hypothèse de données manquantes de forme générale, qui est dominante dans beaucoup d'applications.

4.5.1 Estimation avec données manquantes en bloc

Nous nous situons à présent dans la configuration où les données contiennent un bloc de données manquantes (Fig. 4.1 (a)). Soit $\mathbf{Y} = \{\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}\} = \{\mathbf{y}_i\}_{1\leq i\leq N} \in \mathbb{R}^P$ une matrice obéissant à une telle configuration. On rappelle que $\mathbf{y}_{i,\text{obs}}$ désigne la partie observée des données et $\mathbf{y}_{i,\text{mis}}$ la partie manquante. Les données manquantes $\mathbf{y}_{i,\text{mis}}$ trouvent position sur les variables aux indices entre $p+1$ et P et aux observations entre $n+1$ et N (figure 4.3), soit :

$$\mathbf{y}_{i,\text{mis}} = (y_{p+1,i}, \ldots, y_{P,i})^T \quad i = n+1, \ldots, N$$

et

$$\mathbf{y}_{i,\text{obs}} = \begin{cases} (y_{1,i}, \ldots, y_{P,i})^T & i = 1, \ldots, n \\ (y_{1,i}, \ldots, y_{p,i})^T & i = n+1, \ldots, N \end{cases}$$

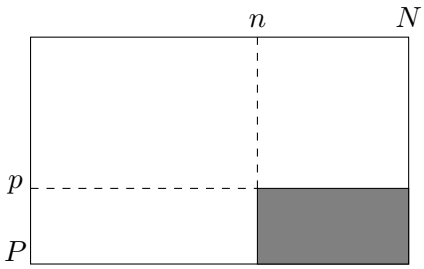


Figure 4.3 – Données manquantes ordonnées en bloc. Gris : données manquantes. Blanc : données observées.

On suppose que tous les $\mathbf{y}_i$ suivent une distribution gaussienne composée :

$$\mathbf{y}_i \sim \mathcal{N}(\mathbf{0}, \tau_i \mathbf{\Sigma}), \quad \tau_i > 0 \tag{4.42}$$

Le but est d'estimer la matrice de covariance de la distribution ci-dessus. La fonction de vraisemblance des données complètes suivant un tel modèle est donnée par :

$$\begin{aligned} \mathcal{L}_{\text{c}} &= \mathcal{L}(\{\mathbf{y}_{i,\text{obs}}\}_{1\leq i\leq N}, \{\mathbf{y}_{i,\text{mis}}\}_{n\leq i\leq N} | \tau_i \mathbf{\Sigma}) \\ &\propto -N \log |\mathbf{\Sigma}| - P \sum_{i=1}^{N} \log \tau_i - \sum_{i=1}^{N} \mathbf{y}_i^T \frac{1}{\tau_i} \mathbf{\Sigma}^{-1} \mathbf{y}_i \end{aligned} \tag{4.43}$$

La séparation du troisième terme en deux parties correspondant aux variables observées et manquantes permet d'obtenir la forme détaillée suivante :

$$\mathcal{L}_{\text{c}} = -N \log |\mathbf{\Sigma}| - P \sum_{i=1}^{N} \log \tau_i - \sum_{i=1}^{n} \mathbf{y}_{i,\text{obs}}^T \frac{1}{\tau_i} \mathbf{\Sigma}^{-1} \mathbf{y}_{i,\text{obs}} - \sum_{i=n+1}^{N} \begin{bmatrix} \mathbf{y}_{i,\text{obs}} \\ \mathbf{y}_{i,\text{mis}} \end{bmatrix}^T \frac{1}{\tau_i} \mathbf{\Sigma}^{-1} \begin{bmatrix} \mathbf{y}_{i,\text{obs}} \\ \mathbf{y}_{i,\text{mis}} \end{bmatrix} \tag{4.44}$$

Les deux premier termes étant déterministes, l'étape E de l'algorithme EM revient à calculer les termes :

$$\mathbb{E}[\mathbf{y}_i^T \frac{1}{\tau_i}\mathbf{\Sigma}^{-1}\mathbf{y}_i] = \begin{cases} \mathbb{E}[\mathbf{y}_{i,\text{obs}}^T \frac{1}{\tau_i}\mathbf{\Sigma}^{-1}\mathbf{y}_{i,\text{obs}}] & 1 \leq i \leq n \\ \mathbb{E}\left[\begin{bmatrix} \mathbf{y}_{i,\text{obs}} \\ \mathbf{y}_{i,\text{mis}} \end{bmatrix}^T \mathbf{\Sigma}_i^{-1} \begin{bmatrix} \mathbf{y}_{i,\text{obs}} \\ \mathbf{y}_{i,\text{mis}} \end{bmatrix} \right] & n+1 \leq i \leq N \end{cases} \tag{4.45}$$

Pour les variables situées entre $n+1 \leq i \leq N$, on utilise la fonction trace $\text{Tr}(.)$ afin de développer l'espérance ci-dessus :

$$\begin{aligned} &\mathbb{E}\left[\text{Tr}\left(\begin{bmatrix} \mathbf{y}_{i,\text{obs}} \\ \mathbf{y}_{i,\text{mis}} \end{bmatrix} [\mathbf{y}_{i,\text{obs}}^T \mathbf{y}_{i,\text{mis}}^T] \mathbf{\Sigma}_i^{-1} \right)\right] \\ &= \frac{1}{\tau_i}\text{Tr}\left(\mathbb{E}\left[\begin{bmatrix} \mathbf{y}_{i,\text{obs}}\mathbf{y}_{i,\text{obs}}^T & \mathbf{y}_{i,\text{obs}}\mathbf{y}_{i,\text{mis}}^T \\ \mathbf{y}_{i,\text{mis}}\mathbf{y}_{i,\text{obs}}^T & \mathbf{y}_{i,\text{mis}}\mathbf{y}_{i,\text{mis}}^T \end{bmatrix} \right] \mathbf{\Sigma}^{-1} \right) \\ &= \frac{1}{\tau_i}\text{Tr}\big(\mathbf{B}_i \mathbf{\Sigma}^{-1}\big) \end{aligned}$$

où $\mathbf{B}_i$ est une matrice de taille $P \times P$ dont l'estimation à l'étape m de l'EM est définie par :

$$\mathbf{B}_i^{(m)} = \begin{bmatrix} \mathbf{y}_{i,\text{obs}}\mathbf{y}_{i,\text{obs}}^T & \mathbf{0} \\ \mathbf{0} & \tau_i^{(m)}(\mathbf{\Sigma}_{22}^{(m)} - \mathbf{\Sigma}_{12}^{(m)}\mathbf{\Sigma}_{11}^{-1(m)}\mathbf{\Sigma}_{12}^{T(m)}) \end{bmatrix} \tag{4.46}$$

La forme finale du log-vraisemblance est donnée par :

$$\mathcal{L}_c = -N\log|\mathbf{\Sigma}| - P\sum_{i=1}^{N}\log\tau_i - \sum_{i=1}^{n}\mathbf{y}_{i,\text{obs}}^T\mathbf{\Sigma}_i^{-1}\mathbf{y}_{i,\text{obs}} - \sum_{i=n+1}^{N}\frac{1}{\tau_i}\text{Tr}\big(\mathbf{B}_i^{(m)}\mathbf{\Sigma}^{-1}\big) \tag{4.47}$$

Théorème 4.5.1. *Soit $\hat{\tau}_i$ et $\hat{\mathbf{\Sigma}}$ les EMV de τ_i et $\mathbf{\Sigma}$ respectivement. Alors :*

$$\hat{\tau}_i = \begin{cases} \frac{\mathbf{y}_i^T\mathbf{\Sigma}^{-1}\mathbf{y}_i}{P} & \textit{pour } i \in \{1,\ldots,n\} \\ \frac{\text{Tr}(\mathbf{B}_i^{(m)}\mathbf{\Sigma}^{-1})}{P} & \textit{pour } i \in \{n+1,\ldots,N\} \end{cases} \tag{4.48}$$

et

$$\hat{\mathbf{\Sigma}} = \frac{P}{N}\left[\sum_{i=1}^{n}\frac{\mathbf{y}_i\mathbf{y}_i^T}{\mathbf{y}_i^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{y}_i} + \sum_{i=n+1}^{N}\frac{\mathbf{B}_i^{T(m)}}{\text{Tr}\big(\mathbf{B}_i^{(m)}\hat{\mathbf{\Sigma}}^{-1}\big)}\right] \tag{4.49}$$

Preuve. Le but est dans un premier temps de différencier $\mathcal{L}_c$ par rapport à la variable τ_i. On sépare pour cela l'équation (4.47) entre les termes variant entre 1 et n puis entre ceux variant entre $n+1$ et N. La résolution de l'équation $\frac{\delta\mathcal{L}_c}{\delta\tau_i} = 0$ est alors triviale et nous permet d'obtenir l'expression de $\hat{\tau}_i$.

Si l'on remplace τ_i par $\hat{\tau}_i$ dans (4.47), on obtient la fonction de log-vraisemblance suivant :

$$\mathcal{L}_c = -N\log|\mathbf{\Sigma}| - P\sum_{i=1}^{n}\log\frac{\mathbf{y}_i^T\mathbf{\Sigma}^{-1}\mathbf{y}_i}{P} - P\sum_{i=n+1}^{N}\log\frac{\text{Tr}(\mathbf{B}_i^{(m)}\mathbf{\Sigma}^{-1})}{P} - NP \tag{4.50}$$

La résolution de l'équation de vraisemblance $\frac{\delta \mathcal{L}_c}{\delta \hat{\mathbf{\Sigma}}} = 0$ mène ensuite à :

$$-N\hat{\mathbf{\Sigma}}^{-1} - P\sum_{i=1}^{n} \frac{\hat{\mathbf{\Sigma}}^{-1}\mathbf{y}_i\mathbf{y}_i^T\hat{\mathbf{\Sigma}}^{-1}}{\mathbf{y}_i^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{y}_i} - P\sum_{i=n+1}^{N} \frac{\hat{\mathbf{\Sigma}}^{-1}\mathbf{B}_i^{T(m)}\hat{\mathbf{\Sigma}}^{-1}}{\mathrm{Tr}\big(\mathbf{B}_i^{(m)}\hat{\mathbf{\Sigma}}^{-1}\big)} = 0$$

Un arrangement des termes, suivi d'une multiplication à droite et à gauche par $\hat{\mathbf{\Sigma}}$ nous conduit au résultat escompté. ■

On peut à présent décrire l'algorithme EM mis en oeuvre pour estimer la matrice de covariance, où m désigne l'itération courante.

Initialisation : À l'itération 0, $\mathbf{\Sigma}^{(0)}$ est initialisée par l'estimateur de Tyler (4.10) obtenu par l'algorithme du point fixe sur les données observées, c'est-à-dire sur le bloc de données observées entre 1 et n. Les textures ne pouvant être initialisées par l'estimateur de Tyler à cause du bloc manquant, nous fixons $\tau_i^{(0)} = (1, \ldots, 1)$.

Étape E : Calcul des espérances de $\mathbf{y}_{i,\text{mis}}|\mathbf{y}_{i,\text{obs}}$, ce qui consiste à calculer les coefficients de la matrice $\mathbf{B}_i^{(m)}$.

$$\begin{aligned}
\mathcal{Q}(\theta|\theta^{(m)}) &\propto E_{\mathbf{y}_{i,\text{mis}}|\mathbf{y}_{i,\text{obs}},\boldsymbol{\theta}^{(m)}}(\mathcal{L}_\text{c}) \\
&= \sum_{i=1}^{N} E_{\mathbf{y}_{i,\text{mis}}|\mathbf{y}_{i,\text{obs}},\boldsymbol{\theta}^{(m)}}(\mathcal{L}_\text{c}) \\
&= \sum_{i=1}^{N} \mathcal{Q}_i(\theta|\theta^{(m)}) \\
&= \sum_{i=1}^{n} (\mathcal{L}_c)_i + \sum_{i=n+1}^{N} \mathcal{Q}_i(\theta|\theta^{(m)})
\end{aligned} \tag{4.51}$$

La première somme correspond au log-vraisemblance négatif des données observées alors que la seconde somme correspond aux données manquantes $\begin{bmatrix} \mathbf{y}_{i,\text{obs}} \\ \mathbf{y}_{i,\text{mis}} \end{bmatrix}, n+1 \leq i \leq N$.

Étape M :

$$\underset{\mathbf{\Sigma},\tau_i}{\arg\min} \; -N\log|\mathbf{\Sigma}| - P\sum_{i=1}^{N}\log\tau_i - \sum_{i=1}^{n}\mathbf{y}_{i,\text{obs}}^T\mathbf{\Sigma}_i^{-1}\mathbf{y}_{i,\text{obs}} - \sum_{i=n+1}^{N}\frac{1}{\tau_i}\mathrm{Tr}\big(\mathbf{B}_i^{(m)}\mathbf{\Sigma}^{-1}\big) \tag{4.52}$$

$$\text{avec } \mathbf{B}_i^{(m)} = \begin{bmatrix} \mathbf{y}_{i,\text{obs}}\mathbf{y}_{i,\text{obs}}^T & 0 \\ 0 & \tau_i^{(m)}(\mathbf{\Sigma}_{22}^{(m)} - \mathbf{\Sigma}_{12}^{(m)}\mathbf{\Sigma}_{11}^{-1(m)}\mathbf{\Sigma}_{12}^{T(m)}) \end{bmatrix}.$$

Algorithme Afin d'estimer tour à tour les paramètres τ_i et $\mathbf{\Sigma}$, on injecte dans l'EM un algorithme du point fixe. On rappelle la forme de la matrice $\mathbf{B}_i$ à l'itération m de l'algorithme EM :

$$\mathbf{B}_i^{(m)} = \begin{pmatrix} \mathbf{y}_i\mathbf{y}_i^T & 0 \\ 0 & \mathbf{Q}_i^{(m)} \end{pmatrix}, \qquad i = n+1, \ldots, N \tag{4.53}$$

avec $\mathbf{Q}_i^{(m)} = \tau_i^{(m)}(\mathbf{\Sigma}_{22}^{(m)} - \mathbf{\Sigma}_{12}^{(m)}\mathbf{\Sigma}_{11}^{-1(m)}\mathbf{\Sigma}_{21}^{(m)})$. L'algorithme est décrit par le pseudo-code 6. Les principales étapes sont les suivantes. Les paramètres $(\tau_i^{(0)}, \mathbf{\Sigma}^{(0)})$ sont d'abord initialisés : $\tau_i^{(0)}$ par 1 et $\mathbf{\Sigma}^{(0)}$ par l'estimé de Tyler (4.10). L'équation du point fixe (4.49) avec $\tau_i^{(0)}$ et $\mathbf{B}_i^{(0)}$ fixé est ensuite imbriqué au sein de l'EM. Une fois que la convergence du point fixe est atteinte, $\tau_i^{(m)}$ est à son tour estimé, puis $\mathbf{Q}_i^{(m)}$ est mis à jour à partir des estimations de $\tau_i^{(m)}$ et $\mathbf{\Sigma}^{(m)}$. Ce processus est ainsi répété jusqu'à ce l'algorithme EM converge. Comme pour les algorithmes précédents, la convergence est évaluée en calculant la NMSE. Celle-ci est calculée pour le point fixe entre les itérations k et $k+1$ (NMSE_{PF}) et pour l'EM entre les itérations m et $m+1$ sur les paramètres $\mathbf{\Sigma}$ et τ_i ($\text{NMSE}_{EM,\mathbf{\Sigma}}$ et NMSE_{EM,τ_i}). Notons qu'à chaque itération k du point fixe, la matrice de covariance est normalisée par son déterminant, soit $\mathbf{\Sigma}^{(m)} = \mathbf{\Sigma}^{(m)}/|\mathbf{\Sigma}^{(m)}|^{1/P}$ [Tatsuoka2000].

Algorithme 6 Imbrication de l'EM et du PF pour l'estimation de $\hat{\mathbf{\Sigma}}, \hat{\tau}_i$.

Entrée: $\mathbf{Y} = \{\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}\} \sim \mathcal{N}(\mathbf{0}, \tau_i\mathbf{\Sigma})$
Sortie: $\hat{\mathbf{\Sigma}}, \hat{\tau}_i$

Initialisation : $\mathbf{\Sigma}^{(0)} = \hat{\mathbf{\Sigma}}_{\text{Tyler}}, \tau_i^{(0)} = (1, \ldots, 1), \mathbf{B}_i^{(0)}$
tant que $\text{NMSE}_{EM,\mathbf{\Sigma}} > tol$ **faire** ▷ boucle EM, m varie
  tant que $\text{NMSE}_{PF} > tol$ **faire** ▷ boucle PF, k varie
    $\mathbf{\Sigma}_{(k+1)}^{(m)} = f(\mathbf{\Sigma}_{(k)}^{(m)}, \mathbf{B}_i^{(m)})$ ▷ équation (4.49)
    $\mathbf{\Sigma}_{(k+1)}^{(m)} = \mathbf{\Sigma}_{(k+1)}^{(m)}/|\mathbf{\Sigma}_{(k+1)}^{(m)}|^{1/P}$ ▷ normalisation par le déterminant
    Calculer NMSE_{PF}
    $k \leftarrow k+1$
  fin tant que
  $\mathbf{\Sigma}^{(m+1)} \leftarrow \mathbf{\Sigma}^{(m)}$
  $\tau_i^{(m+1)} = f(\mathbf{\Sigma}^{(m+1)}) \quad i = 1, \ldots, n$ ▷ équation (4.48)
  $\tau_i^{(m+1)} = f(\mathbf{\Sigma}^{(m+1)}), \mathbf{B}_i^{(m)}) \quad i = n+1, \ldots, N$ ▷ équation (4.48)
  $\mathbf{Q}_i^{(m+1)} = f(\mathbf{\Sigma}^{(m+1)}, \tau_i^{(m+1)}) \quad i = n+1, \ldots, N$ ▷ équation (4.53)
  Calculer $\text{NMSE}_{EM,\mathbf{\Sigma}}$
  Calculer NMSE_{EM,τ_i}
  $m \leftarrow m+1$
fin tant que

4.5.2 Estimation en rang faible

Nous reprenons la procédure décrite en section 4.4.2 pour l'appliquer au cas de la distribution gaussienne composée. Le problème consiste alors à reprendre la minimisation (4.52) sujette à la contrainte sur la structure de la matrice de covariance $\mathbf{\Sigma} = \mathbf{\Sigma}_R + \sigma^2\mathbf{I}$. Tout comme précédemment, la solution globale à ce problème est donnée par (4.37) et (4.38). La procédure d'exécution de l'algorithme consiste alors à injecter le calcul de cette solution dans l'algorithme 6 comme décrit dans le pseudo-code ci-après (algorithme 7).

4.5.3 Simulations numériques

L'algorithme EM est évalué sur des données incomplètes de dimension $P = 10$ et $N \in [44, 251]$, dont la taille du bloc de données manquantes varie. Les données sont tirées à partir d'une distribution gaussienne composée de matrice de covariance $\mathbf{\Sigma}$ dont les éléments sont définis par $\Sigma_{ij} = \rho^{|i-j|}$ pour $i, j \in [1, P]$ et $0 \leq \rho \leq 1$, et de textures $\{\tau_i\}_{i\in[1,N]}$ obéissant à une distribution Gamma $\Gamma(\alpha, 1/\alpha)$, où α et $1/\alpha$ sont les paramètres dits de forme et d'échelle (strictement positifs). Notons que pour $\alpha = 1$, on retrouve une distribution exponentielle, alors que pour α grand, la

Algorithme 7 Estimation rang faible de $\hat{\mathbf{\Sigma}}, \hat{\tau}_i$.

Entrée: $\mathbf{Y} = \{\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}\} \sim \mathcal{N}(\mathbf{0}, \tau_i(\mathbf{\Sigma}_R + \sigma^2\mathbf{I}))$
Sortie: $\hat{\mathbf{\Sigma}}, \hat{\tau}_i$

Initialisation : $\mathbf{\Sigma}^{(0)} = \hat{\mathbf{\Sigma}}_{\text{Tyler}}, \tau_i^{(0)} = 1, \mathbf{B}_i^{(0)}$
tant que $\text{NMSE}_{EM,\mathbf{\Sigma}} > tol$ **faire** ▷ boucle EM, m varie
  tant que $\text{NMSE}_{PF} > tol$ **faire** ▷ boucle PF, k varie
    $\mathbf{\Sigma}_{(k+1)}^{(m)} = f(\mathbf{\Sigma}_{(k)}^{(m)}, \mathbf{B}_i^{(m)})$
    $\mathbf{\Sigma}_{(k+1)}^{(m)} \overset{\text{EVD}}{=} \sum_{i=1}^{P} \lambda_i \mathbf{u}_i \mathbf{u}_i^T$
    Calculer $\hat{\sigma}^2$, $\hat{\lambda}_i$ ▷ équations (4.37) et (4.38)
    $\mathbf{\Sigma}_{(k+1)}^{(m)} = \sum_{i=1}^{R} \hat{\lambda}_i \mathbf{u}_i \mathbf{u}_i^T + \hat{\sigma}^2\mathbf{I}$
    Calculer NMSE_{PF}
    $k \leftarrow k + 1$
  fin tant que
  Calculer $\tau_i^{(m+1)}$, $\mathbf{Q}_i^{(m+1)}$, $\text{NMSE}_{EM,\mathbf{\Sigma}}$, NMSE_{EM,τ_i} ▷ voir Algorithme 6
fin tant que

distribution Gamma converge vers une loi gaussienne. Plusieurs configurations sont étudiées selon la valeur de α ainsi que la taille du bloc de données manquantes $(P \times N)_{\text{manquant}} = (P-p) \times (N-n)$.

Convergence de l'EM

Afin de valider expérimentalement la convergence de l'algorithme EM proposé, les quantitées $\text{NMSE}_{EM,\mathbf{\Sigma}}$, NMSE_{EM,τ_i} sont sauvegardées à chaque itération. Les paramètres de simulation sont ici fixés à $P = 10$, $N = 100$ et $\alpha = 1$. La variation des erreurs (figure 4.4) montre que l'EM converge plus ou moins rapidement selon les différentes configurations sur la taille du bloc de données manquantes. De manière générale, la convergence est plus lente lorsque la taille du bloc de données manquantes augmente, tant sur les observations (N) que sur les variables (P). La convergence est tout de même atteinte pour une large partie d'observations manquantes (80 sur 100) et de variables manquantes (7 sur 10).

Estimation des paramètres

On considère les estimateurs suivants pour comparaison :

— Les estimations $\hat{\tau}$ (4.48) et $\hat{\mathbf{\Sigma}}$ (4.49) ;

— Les estimateurs de Tyler "complets" (4.10) ;

— Les estimateurs de Tyler "incomplets" définis par :

$$\hat{\tau}_i = \frac{\mathbf{y}_i^T \mathbf{\Sigma}_n^{-1} \mathbf{y}_i}{P}, \qquad \hat{\mathbf{\Sigma}}_n = \frac{P}{n} \sum_{i=1}^{n} \frac{\mathbf{y}_i \mathbf{y}_i^T}{\mathbf{y}_i^T \hat{\mathbf{\Sigma}}_n^{-1} \mathbf{y}_i} \tag{4.54}$$

Les performances sont évaluées à l'aide de la distance riemannienne sur la matrice de covariance (4.40) et sur la texture. Cette dernière est définie par [Bouchard2020] :

$$\delta^2_{\mathcal{R}++}(\tau, \hat{\tau}) = ||\log(\tau^{-1} \odot \hat{\tau})||_2^2 \tag{4.55}$$

où $\odot$ est le produit de Hadamard (produit terme à terme). Les résultats de validation sont présentés en figure 4.5 pour l'estimation de la matrice de covariance. Quelle que soit la taille du bloc de données manquantes, l'erreur sur l'estimation de l'algorithme EM se situe entre les erreurs des estimateurs de Tyler "complet" et "incomplet". Les résultats montrent également qu'aucun

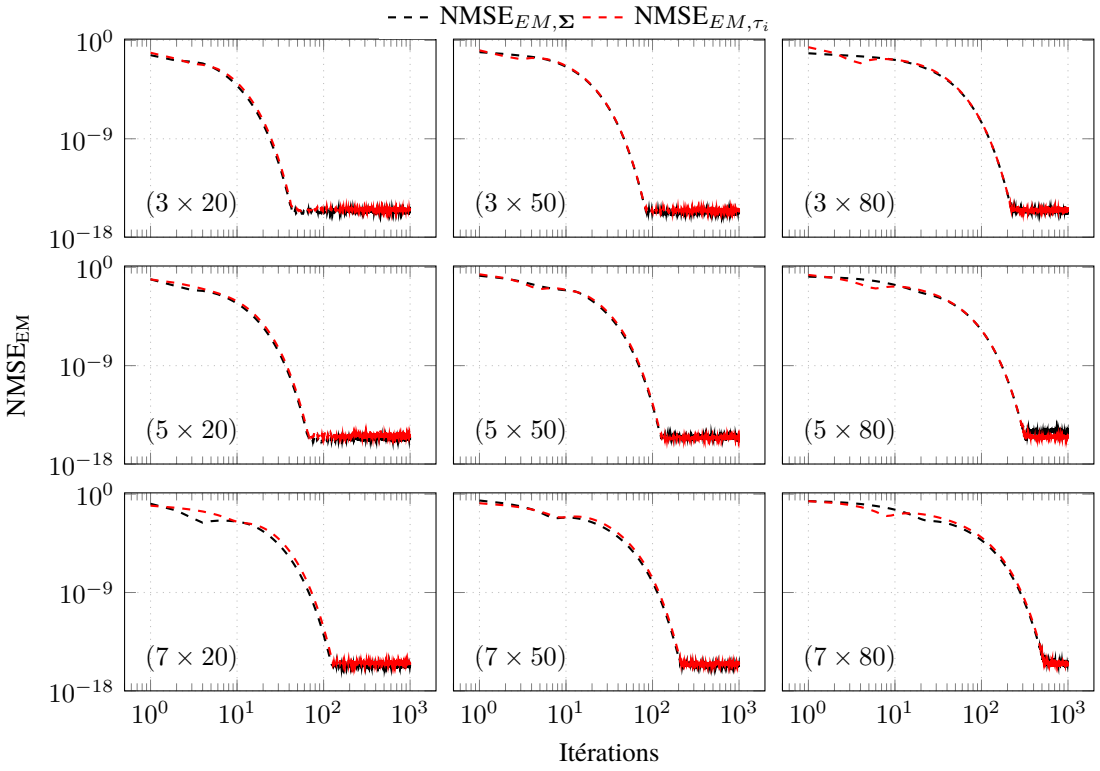
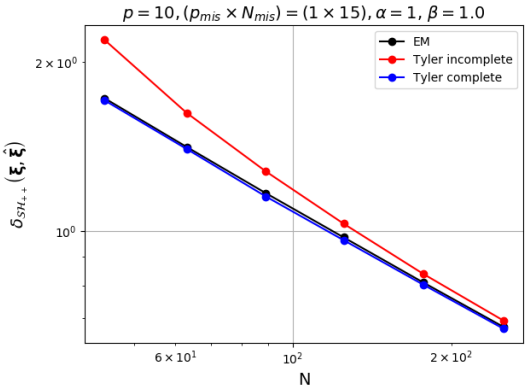
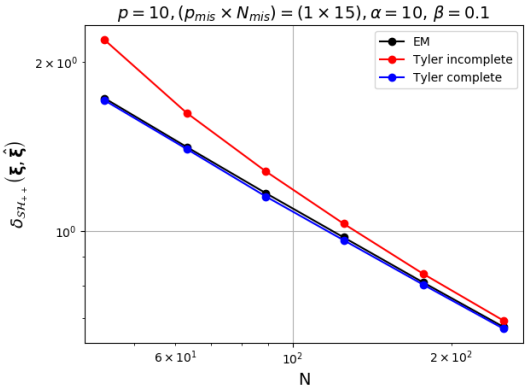


Figure 4.4 – Convergence de l'EM pour différentes tailles de bloc de données manquantes $(P \times N)_{\text{manquant}}$ sur un jeu de données de taille (10×100).

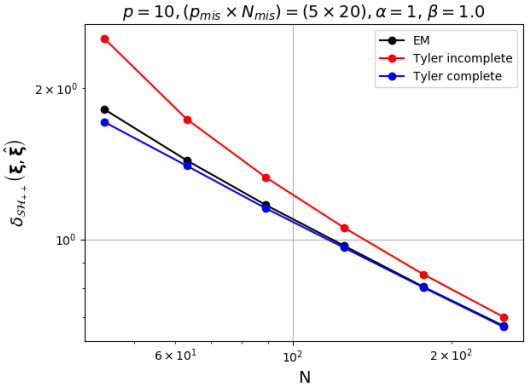
changement visible n'a lieu lorsque α varie, ce qui témoigne de la robustesse de l'estimation. Concernant l'estimation des textures, les résultats sont présentés en figure 4.6. L'augmentation de la distance en fonction du nombre d'observations N est simplement due au fait qu'augmenter N provoque également une augmentation du nombre de textures à estimer sans que la dimension P ne change. Les résultats sur l'estimation de la matrice de covariance avec structure rang faible (figure 4.7) pour une configuration en bloc de données manquantes fournissent également une validation du comportement asymptotique de l'estimation EM et des estimateurs de Tyler.



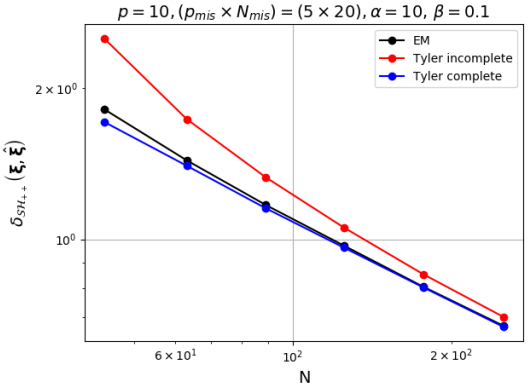
(a) $\alpha = 1$; $(P \times N)_{\text{manquant}} = (1 \times 15)$



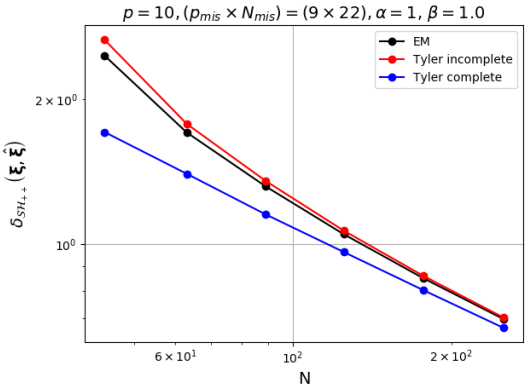
(b) $\alpha = 10$; $(P \times N)_{\text{manquant}} = (1 \times 15)$



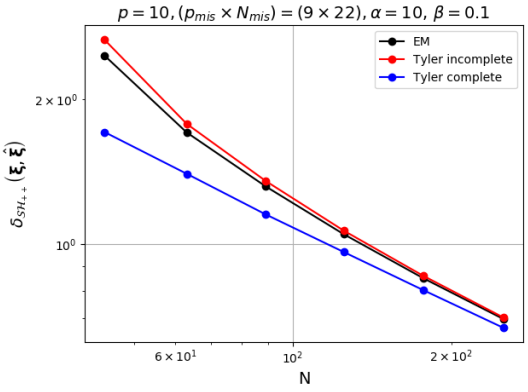
(c) $\alpha = 1$; $(P \times N)_{\text{manquant}} = (5 \times 20)$



(d) $\alpha = 10$; $(P \times N)_{\text{manquant}} = (5 \times 20)$

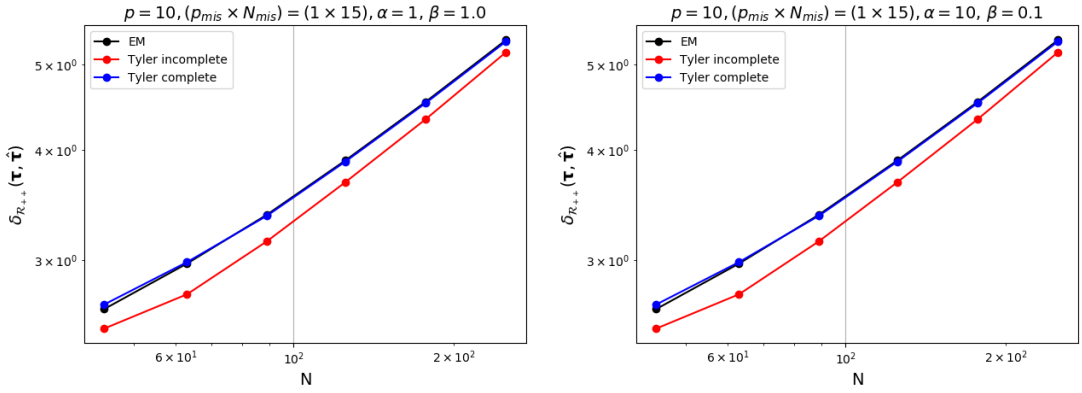


(e) $\alpha = 1$; $(P \times N)_{\text{manquant}} = (9 \times 22)$



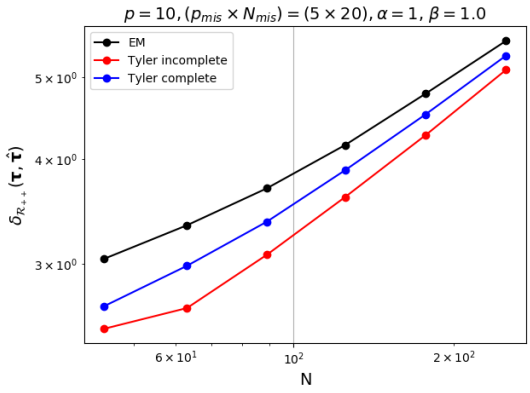
(f) $\alpha = 10$; $(P \times N)_{\text{manquant}} = (9 \times 22)$

Figure 4.5 – Distance naturelle sur la matrice de covariance en fonction du nombre d'observations (N) selon différents α et différentes tailles de bloc de données manquantes.

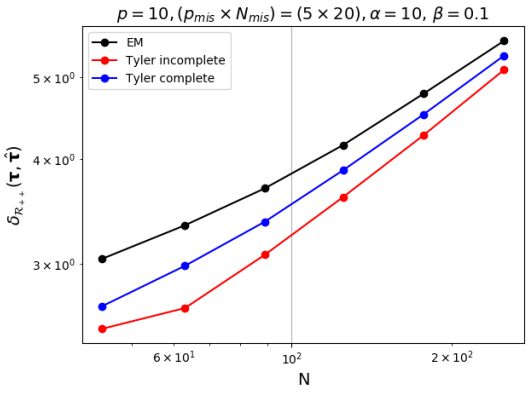


(a) $\alpha = 1$; $(P \times N)_{\text{manquant}} = (1 \times 15)$

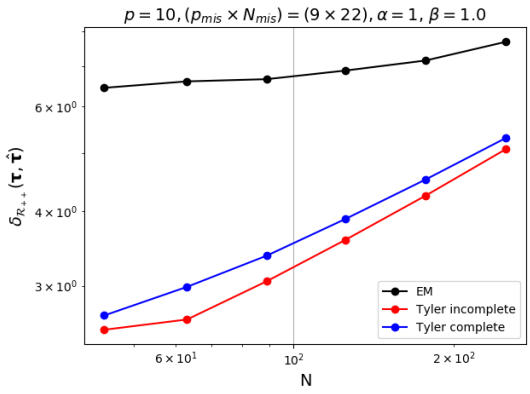
(b) $\alpha = 10$; $(P \times N)_{\text{manquant}} = (1 \times 15)$



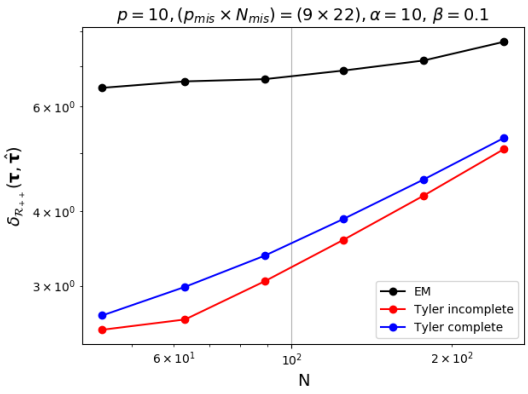
(c) $\alpha = 1$; $(P \times N)_{\text{manquant}} = (5 \times 20)$



(d) $\alpha = 10$; $(P \times N)_{\text{manquant}} = (5 \times 20)$

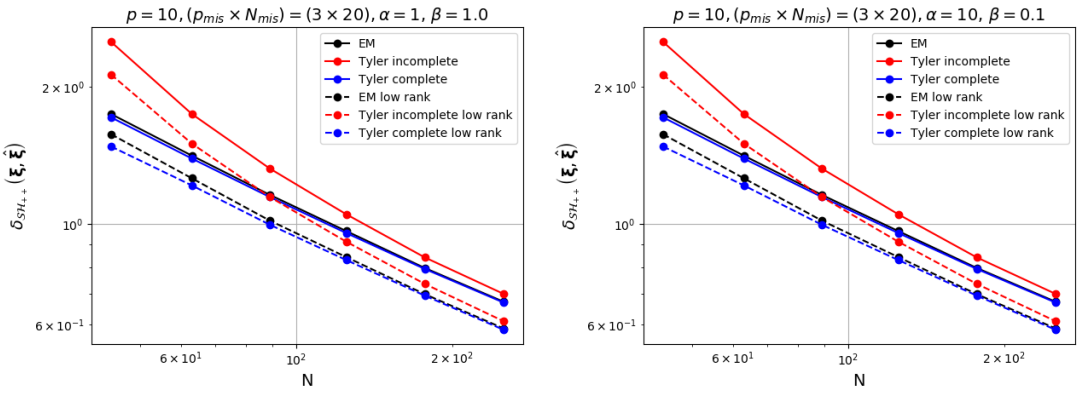


(e) $\alpha = 1$; $(P \times N)_{\text{manquant}} = (9 \times 22)$



(f) $\alpha = 10$; $(P \times N)_{\text{manquant}} = (9 \times 22)$

Figure 4.6 – Distance naturelle sur les paramètres de texture en fonction du nombre d'observations (N) selon différents α et différentes tailles de bloc de données manquantes.



(a) $\alpha = 1$ **(b)** $\alpha = 10$

Figure 4.7 – Distance naturelle sur la matrice de covariance sans et avec structure rang faible, en fonction du nombre d'observations (N) selon différents α et un bloc de données manquantes de taille $(P \times N)_{\text{manquant}} = (3 \times 20)$.

4.6 Comparatif avec la méthode EM-EOF dans le cas Gaussien

La méthode EM-EOF (chapitre 2) est une méthode itérative permettant d'interpoler des données incomplètes comportant de multiples caractéristiques (complexité du comportement de déplacement, caractéristiques du bruit, type de données manquantes). Nous avons vu que cette méthode repose sur la décomposition de la covariance temporelle $\mathbf{\Sigma}_{\text{EM-EOF}}$ des données en fonctions empiriques orthogonales. Une partie seulement de ces fonctions (nombre optimal de modes) sont ensuite sélectionnées pour reconstruire les données manquantes. Les données finales possèdent donc une structure rang faible car celles-ci sont reconstruites à partir des modes les plus représentatifs du comportement des données. Comme la qualité de la reconstruction dépend largement de l'estimation de la covariance temporelle, on s'intéresse donc ici à la comparaison, en terme d'erreurs, de l'estimé $\hat{\mathbf{\Sigma}}_{\text{EM-EOF}}$ avec la covariance de structure rang faible $\hat{\mathbf{R}}$ estimée en section 4.4.2. Nous supposons dans cette étude comparative que les données suivent une distribution gaussienne centrée.

4.6.1 Estimation de la matrice de covariance sur données synthétiques

Les données complètes sont synthétisées à partir d'une distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_R + \sigma^2\mathbf{I})$, avec $P = 12$, $R = 3$ et $N \in [60, 330]$. Les données manquantes, dont la quantité varie entre 10% et 50% pour les besoins de la simulation, sont générées aléatoirement de sorte qu'elles admettent une forme générale (figure 4.1). La NMSE est calculée sur $\mathbf{\Sigma}_R$:

$$\delta = \frac{||\mathbf{\Sigma}_R - \hat{\mathbf{\Sigma}}||_F}{||\mathbf{\Sigma}_R||_F} \tag{4.56}$$

où $\hat{\mathbf{\Sigma}}$ correspond aux estimés comparés suivants :

— L'estimation rang faible de (4.3) $\hat{\mathbf{R}}_{\text{SCM}}$;

— L'estimation rang faible de la SCM "incomplète" (4.41) $\hat{\mathbf{R}}_{\text{SCM}}^{\text{obs\%}}$;

— L'estimation rang faible issue de l'algorithme EM (section 4.4.2) $\hat{\mathbf{R}}$;

— La matrice de covariance empirique estimée à partir des données reconstruites à l'issue l'étape 1 de la méthode EM-EOF :

$$\hat{\mathbf{\Sigma}}_{\text{EM-EOF}}^{\text{step1}} = \frac{1}{N}\hat{\mathbf{X}}_R\hat{\mathbf{X}}_R^T \tag{4.57}$$

— La matrice de covariance empirique estimée à partir des données reconstruites à l'issue l'étape 2 de la méthode EM-EOF :

$$\hat{\mathbf{\Sigma}}_{\text{EM-EOF}}^{\text{step 2}} = \frac{1}{N}\hat{\mathbf{X}}_r\hat{\mathbf{X}}_r^T \tag{4.58}$$

Notons que le nombre de modes estimé par la méthode EM-EOF est de 3, ce qui correspond au rang des données. Les résultats sont présentés en figure 4.8. On observe que les performances d'estimation des covariances de l'EM et de la méthode EM-EOF sont similaires à la SCM "incomplète" lorsque les données contiennent une faible quantité de données manquantes (10%). Lorsque la quantité de données manquantes augmente (30%), l'étape 2 de la méthode EM-EOF permet d'obtenir une meilleure estimation par rapport à l'étape 1, ce qui pourrait être dû à la mise à jour itérative des données manquantes lors de l'étape 2. L'écart est plus largement visible lorsque la quantité de données manquantes atteint 50% des données complètes, où $\hat{\mathbf{\Sigma}}_{\text{EM-EOF}}^{\text{step2}}$ est plus proche de la vraie covariance que $\hat{\mathbf{R}}$. On observe également que la variation des erreurs de la méthode EM-EOF a tendance à s'aplanir lorsque N augmente, ce qui pourrait s'expliquer par des modes de convergences différents entre l'EM et la méthode EM-EOF. Le premier converge vers les solutions de l'EMV par le calcul d'une erreur sur les paramètres estimés à chaque itération de l'EM, alors que la seconde converge vers une prédiction des données manquantes en se basant sur une erreur calculée directement sur les données. Ceci est lié à une remarque soulignée en introduction, d'ordre plus générale, à savoir que les deux méthodes poursuivent deux objectifs différents : l'estimation de la covariance (EM) et la prédiction de données manquantes (EM-EOF). Il est toutefois intéressant de constater que ces méthodes fournissent des résultats tout à fait comparables en terme d'estimation de la covariance. Concernant la prédiction des données manquantes, la prochaine section a pour objet d'en effectuer la comparaison.

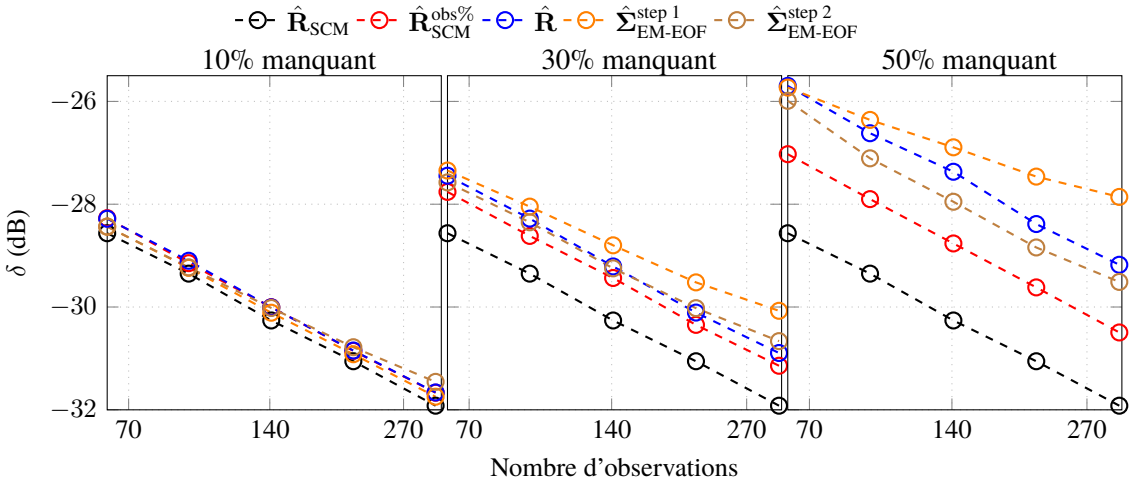


Figure 4.8 – Erreur (moyenne sur 200 simulations) en fonction du nombre d'observations N pour trois configurations (10%, 30%, 50%) de données manquantes.

4.6.2 Reconstruction de données manquantes sur données réelles

Lors du chapitre 1, puis en introduction de ce chapitre, nous avons évoqué la distinction entre problème d'estimation paramétrique et problème d'interpolation. En effet, si l'EM décrit en section 4.3 permet d'estimer la matrice de covariance à partir de données incomplètes, cet algorithme ne permet pas directement de prédire les données manquantes, autrement dit d'interpoler. Le but est ici tenter une stratégie permettant de combiner les deux méthodes "sur le terrain" de la méthode EM-EOF, c'est-à-dire en terme d'interpolation des données.

La procédure adoptée consiste à injecter l'estimation de la covariance issue de l'EM dans la reconstruction utilisée par la méthode EM-EOF. La matrice de covariance estimée est ensuite décomposée par une EVD, ce qui permet de récupérer ses vecteurs propres (ou EOFs). La reconstruction consiste ensuite en une projection des EOFs sur les composantes principales $\mathbf{A}$ (voir équations (2.8) et (2.9)) :

$$\hat{\mathbf{Y}} = \mathbf{A}\mathbf{U}^T \tag{4.59}$$

où la matrice $\mathbf{U}$ contient en colonne les vecteurs propres issus de l'EVD de la matrice de covariance des données observées. S'agissant d'un travail préliminaire, la stratégie adoptée n'est pas optimale car elle ne permet pas de comparer *directement* la méthode EM-EOF et l'algorithme EM en terme d'interpolation, mais plutôt de combiner les deux approches.

Description des données

Les données consistent en des mesures de déplacement de surface issues du réseau de stations GNSS (Global Navigation Satellite System) de l'observatoire volcanologique du Piton de la Fournaise (OVPF) situé sur l'île de la Réunion (figure 4.9). Le réseau est constitué de 22 stations dont l'altitude varie entre 67 et 2590 mètres au-dessus du niveau de la mer. Chaque station mesure les déplacements de surface dans trois directions correpondant aux axes est-ouest, nord-sud et vertical dans la référence terrestre [Smittarello2019b, Smittarello2019a]. Pour les besoins de l'étude, seul le déplacement vertical est traîté. Les mesures considérées s'étendent entre janvier 2014 et mars 2017, à raison d'une mesure par jour. L'ensemble des mesures pour chaque station sont présentées en annexe B.2 (figure B.1). Chaque station GNSS (P) contient 1086 observations (N), dont le taux de données manquantes varie entre 5.1% et 54.9%. Par son caractère irrégulier, le motif des données manquantes s'apparente largement à une forme générale (figure 4.10). Les données manquantes sont essentiellement causées par l'inactivité temporaire des capteurs sur une période temporelle. Notons également que les données, de par la précision des mesures GNSS, contiennent peu de bruit.

Application de la méthode EM-EOF et détermination du rang

Lorsque la méthode EM-EOF est appliquée sur des séries temporelles d'images, la dimension P (nombre de pixels par image) surpasse la dimension N (nombre d'observations temporelles). Dans ce cas, la méthode se base naturellement sur la décomposition de la covariance temporelle, ce qui a notamment l'avantage de minimiser le coût temporel du traitement (voir section 2.2.2). Dans le cas présent, N est bien plus grand que P. Plutôt que de calculer la covariance empirique temporelle, nous calculons donc la covariance empirique spatiale. Si $\mathbf{Y} = (\mathbf{y}_i, \dots, \mathbf{y}_P)$ désigne la matrice rectangulaire de taille $(N \times P)$, où $\{\mathbf{y}_i\}_{i\in[1,P]}$ désigne une station GNSS, la matrice de covariance empirique spatiale est définie par :

$$\mathbf{\Sigma} = \frac{1}{N}(\mathbf{Y} - \mathbf{1}_N\bar{\mathbf{y}})^T(\mathbf{Y} - \mathbf{1}_N\bar{\mathbf{y}}) \tag{4.60}$$

où $\bar{\mathbf{y}} = (\mu_1, \dots, \mu_P)$ est le vecteur des moyennes empiriques temporelles de chaque station et $\mathbf{1}_P = (1, \dots, 1)^T$ est un vecteur de taille P composé de 1. À la différence du traitement d'origine, qui consistait à retirer la moyenne spatiale aux données (équation (2.2)), on retire ici la moyenne temporelle de chaque station $\mathbf{y}_i$.

La méthode EM-EOF est appliquée sur les données en suivant les changements décrits ci-dessus. Le nombre optimal de modes estimé est de 3, ce qui correspond au minimum de l'erreur de validation croisée que l'on a définie par l'équation (2.12). Une approche visuelle du spectre (figure 4.11) permet d'observer que les trois premiers modes correspondent aux trois valeurs

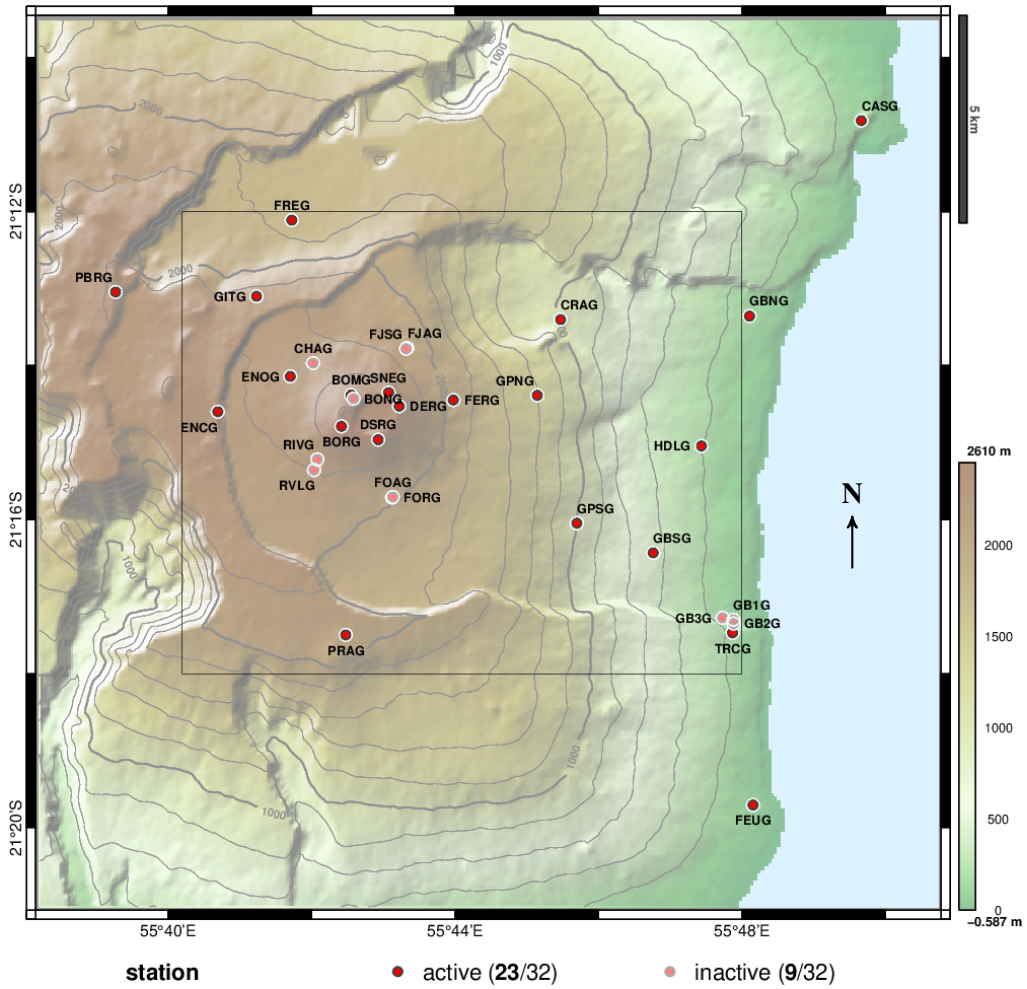


Figure 4.9 – Réseau GNSS permanent de l'OVPF et état de fonctionnement des stations en décembre 2019 (communication personnelle avec Virginie Pinel de l'ISTerre). ©WEBOBS / IPGP. Modèle numérique de terrain : SRTM/NASA.

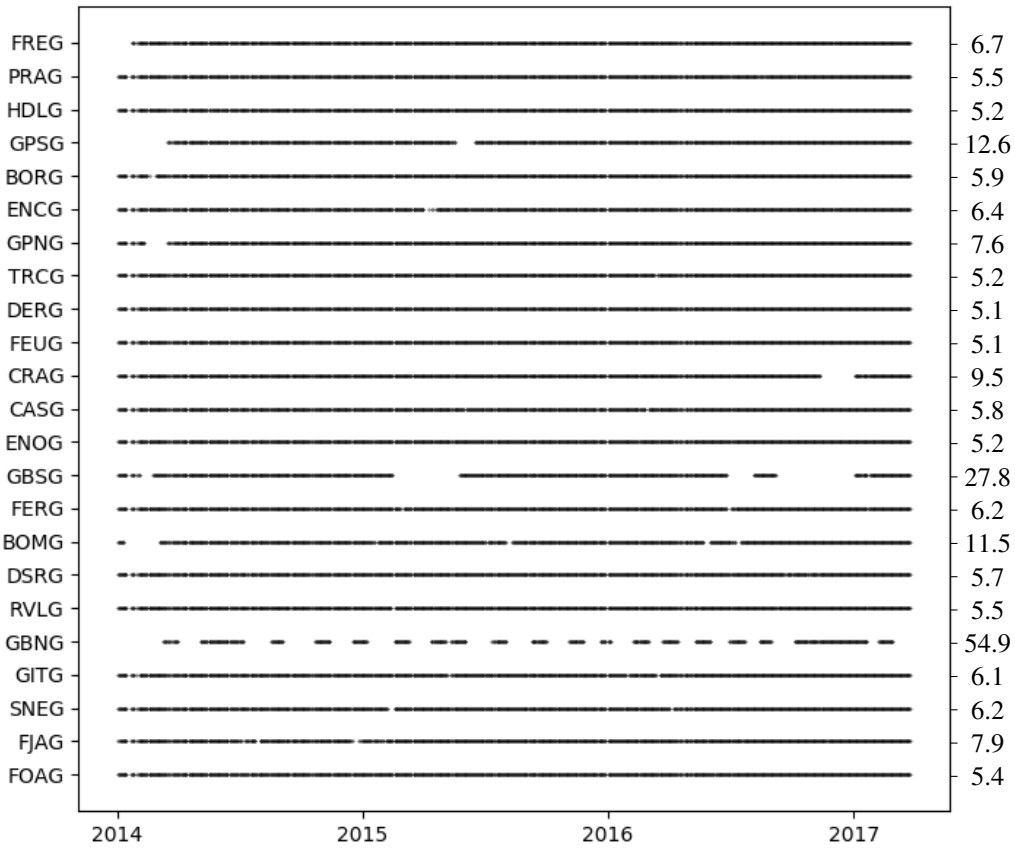


Figure 4.10 – Observations GNSS au cours du temps et pourcentage de données manquantes par station. Noir : observé ; blanc : manquant.

propres dominantes du spectre et rassemblent 75% de la variance totale [5]. Le rang choisi pour l'application de l'algorithme EM est donc fixé à $R = 3$.

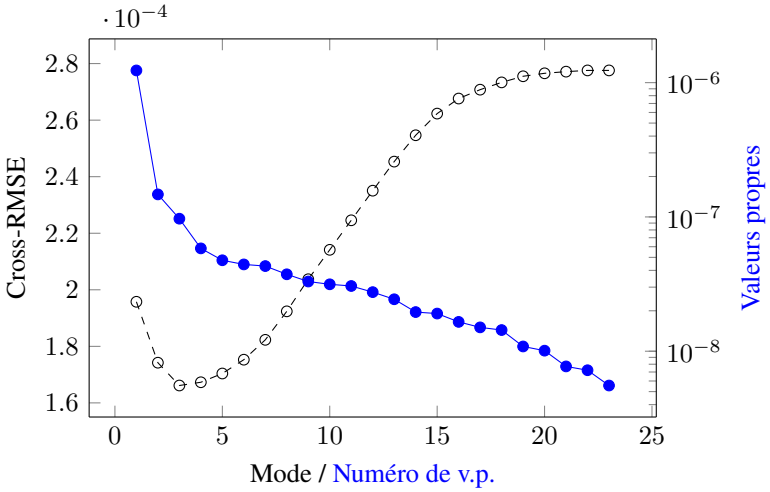


Figure 4.11 – Cross-RMSE (m) en fonction du nombre de mode à l'issue de l'étape 1 de la méthode EM-EOF (moyenne sur 100 simulations) et spectre de valeurs propres (v.p.). Le minimum moyen de l'erreur se situe à l'indice 3.

Résultats préliminaires

L'algorithme EM et la méthode EM-EOF sont à présent appliqués sur les données GNSS. La reconstruction est menée sur les stations GNSS situées en marge du cratère et peu affectées par les événements éruptifs dans la période étudiée (2014-2017). Du fait de leur position, ces stations peuvent contribuer à comprendre les déformations en profondeur en agissant comme contrainte d'un modèle de déformation. Une sélection de résultats est présentée en figure 4.12, comprenant les stations contenant le plus de données manquantes (GBNG, GBSG, GPSG, respectivement 54.9, 27.8 et 12.6% de données manquantes). Dans l'ensemble, les séries temporelles reconstruites suivent les tendances de déplacement, y compris dans les zones de données manquantes. Les séries temporelles reconstruites contenant de plus importantes quantités de données manquantes sont également cohérentes avec la tendance de déplacement. Les zooms 1 et 2 sur les stations GBNG et GBSG permettent de remarquer la proximité entre les deux reconstructions et par rapport à la série temporelle initiale.

Analyse des erreurs de reconstruction

La RMSE sur les données observées et la cross-RMSE sur des données retirées artificiellement parmi les données observées (40 points par station, ce qui représente au total 3% des observations) sont calculées. La première erreur permet de mesurer les potentiels changements induits sur les données observées du fait de la sélection d'un rang inférieur à la dimension P (effet potentiel de filtrage). La seconde erreur permet de mesurer l'erreur d'interpolation des données manquantes [Beckers2003]. Les erreurs par station sont présentées en figure 4.13, ainsi que la différence entre les erreurs $\Delta_{\text{RMSE}} = \text{RMSE}_{\text{EM}} - \text{RMSE}_{\text{EM-EOF}}$ (figure 4.14). Globalement, l'erreur est stable par station GNSS, ce qui laisse présumer que la quantité de données manquantes n'a pas d'influence sur la qualité de la reconstruction.

5. Cette valeur est simplement obtenue par le calcul de la mesure $100 \times \sum_{i=1}^{3} \lambda_i / \sum_{i=1}^{P} \lambda_i$.

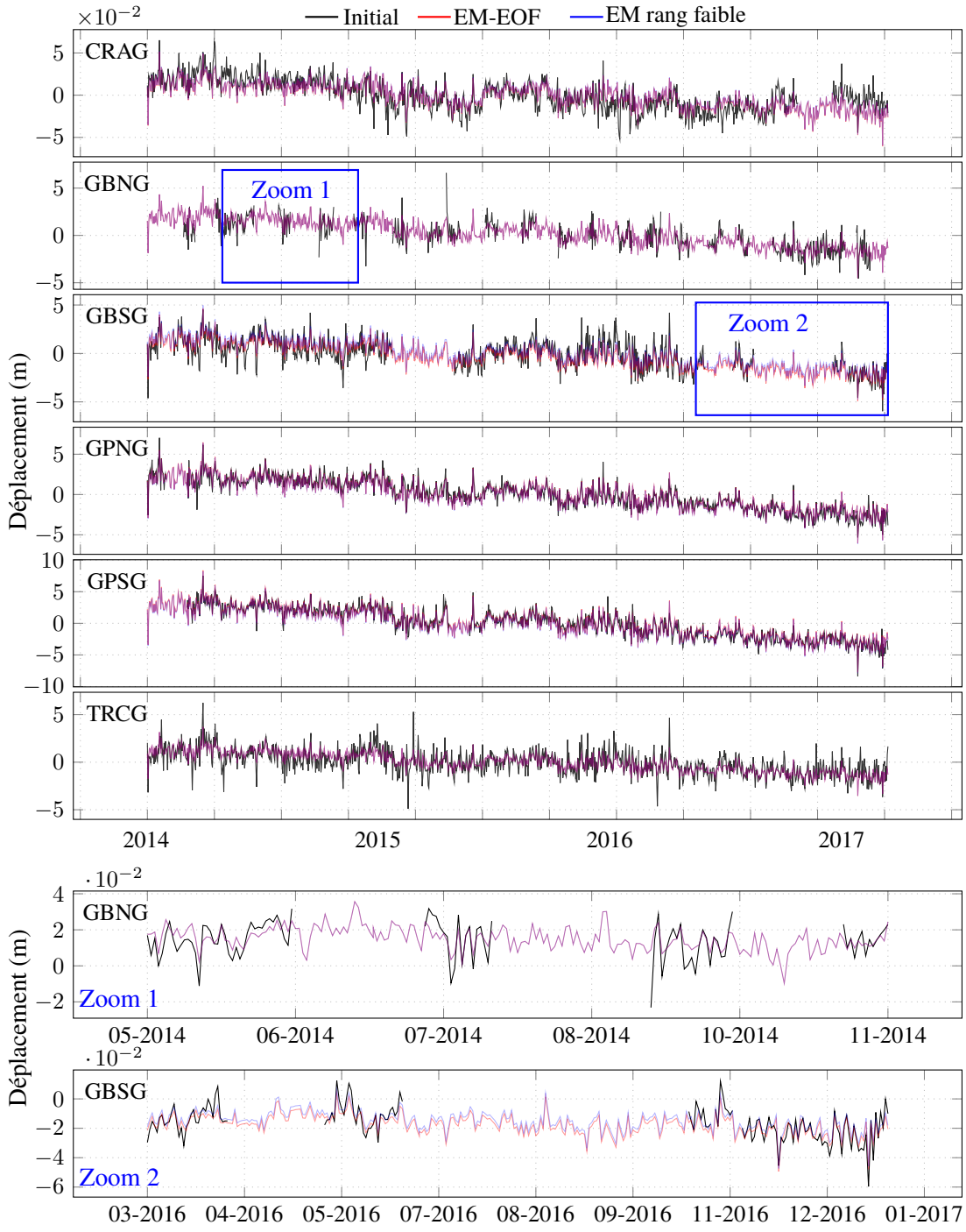


Figure 4.12 – Exemples de séries temporelles de mesure de déplacement (m) reconstruites sur huit stations GNSS et zoom sur les stations GBNG (54.9% de données manquantes) et GBSG (27.8% de données manquantes).

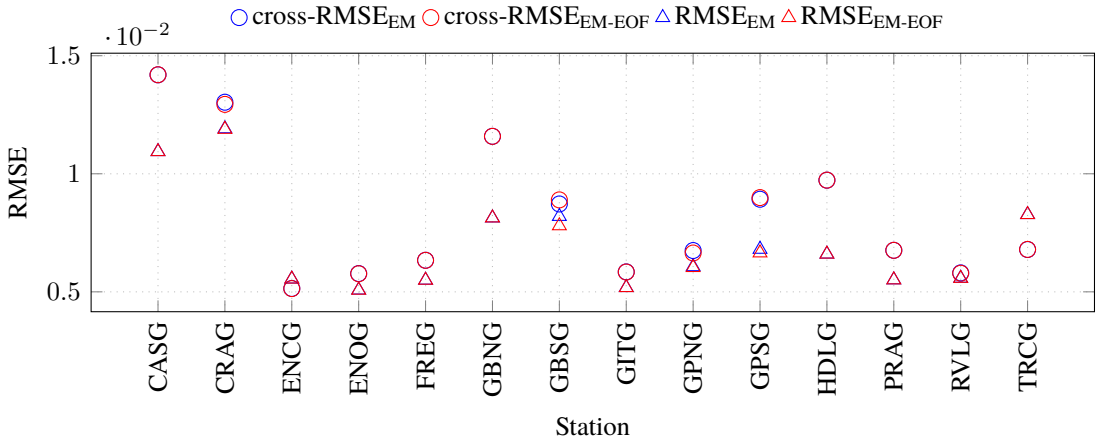


Figure 4.13 – Erreurs (cross-RMSE et RMSE) par station GNSS de l'algorithme EM et de la méthode EM-EOF.

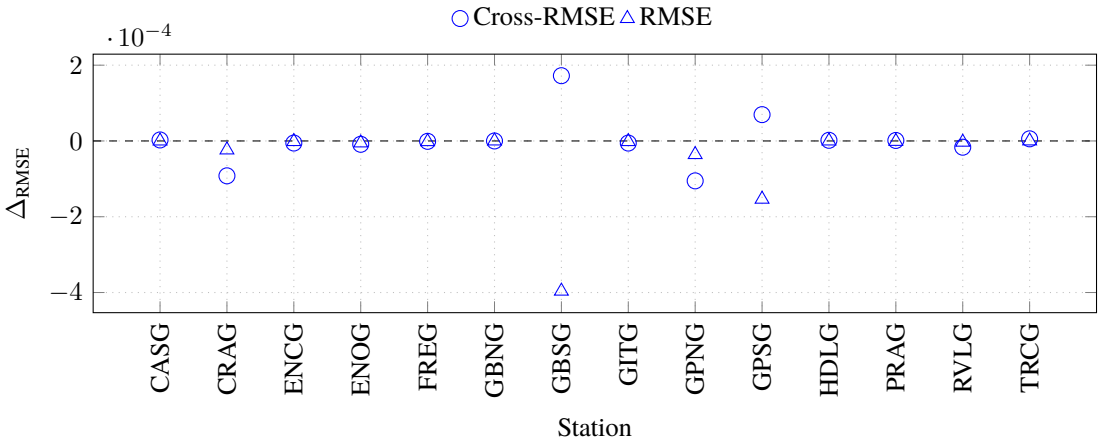


Figure 4.14 – Différence des erreurs de reconstruction entre l'algorithme EM et la méthode EM-EOF.

4.6.3 Discussion

Les simulations et l'application sur des données réelles permettent de soulever quelques points de discussion. Concernant l'estimation de la covariance, les deux méthodes fournissent des résultats similaires pour des quantités de données manquantes moyennes (<30%), alors que la méthode EM-EOF fournit une erreur plus faible lorsque la quantité de données manquantes est plus importante (50%). Il faut cependant rappeler que la méthode EM-EOF n'est pas conçue pour l'estimation de la covariance car elle ne converge pas vers les solutions de l'EMV, ce qui explique la différence de comportement asymptotique lorsque N augmente. Concernant la reconstruction des données manquantes, les deux méthodes fournissent un résultat très similaire en terme d'erreur d'interpolation quelle que soit la quantité de données manquantes. Toutefois, on rappelle que la procédure choisie pour l'application de l'EM ne permet pas d'effectuer une comparaison en tant que telle avec la méthode EM-EOF, mais repose plus simplement sur une combinaison.

Perspective de ce travail

Dans l'immédiat, un travail supplémentaire devra être fourni afin de trouver un moyen fiable de comparaison de l'algorithme EM et de la méthode EM-EOF pour l'interpolation de données manquantes. Ensuite, une suite logique de cette étude consiste à étendre les simulations sur données synthétiques et l'application aux données réelles au cas des distributions gaussiennes composées. Un aperçu de la distribution empirique sur l'ensemble des données de déplacement GNSS (figure 4.15) peut suffire à montrer l'intérêt de cette perspective. En effet, le modèle gaussien ne représente pas fidèlement l'histogramme des données centrées en zéro, qui possède des valeurs hors de la distribution gaussienne à droite et à gauche. Notons que cet histogramme est tracé sur l'ensemble des stations GNSS (vecteurs $\mathbf{y}_i$). Il n'est donc pas représentatif de la distribution des vecteurs $\mathbf{y}_i$ (dont l'indépendance est supposée) mais plutôt des paramètres de textures. Les histogrammes par stations (figure 4.16) montrent des comportements très différents. La distribution gaussienne centrée ne permet pas d'englober la distribution des données avec précision, ce qui pourrait avantager l'hypothèse gaussienne composée dont les paramètres de textures permettent d'intégrer une certaine robustesse aux données atypiques.

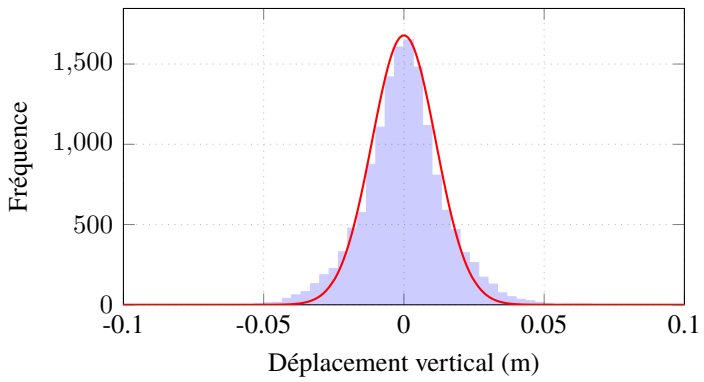


Figure 4.15 – Histogramme de l'ensemble des données de déplacement GNSS et modèle de distribution gaussienne centrée (rouge).

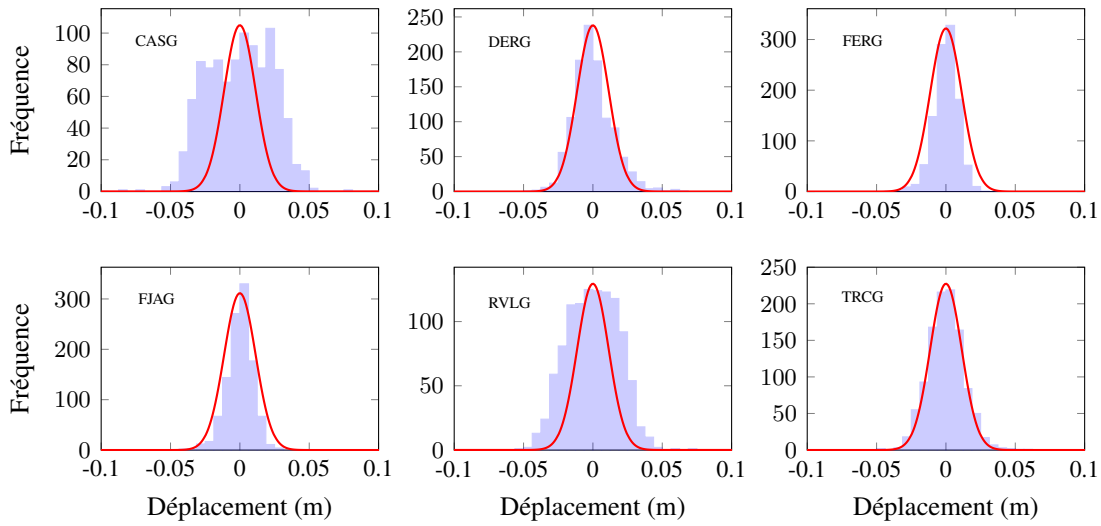


Figure 4.16 – Histogrammes sur six stations GNSS et modèle de distribution gaussienne centrée à variance fixe (rouge).

4.7 Synthèse

Ce chapitre nous a permis d'aborder l'analyse de données manquantes sous l'angle d'un problème d'estimation paramétrique, nécessitant ainsi des hypothèses statistiques sur le modèle régissant les données. Après avoir passé en revue les distributions de probabilité utiles à cette étude (gaussienne, gaussienne composée, modèle par facteur), plusieurs cas d'études ont été examinés selon la forme des données manquantes et la distribution considérée. Afin d'estimer la matrice de covariance, nous avons fait usage de l'algorithme EM, méthode itérative garantissant la convergence vers l'estimation du maximum de vraisemblance (EMV).

La première étude s'est concentrée sur des données manquantes dont la forme est générale et dont la distribution est gaussienne centrée, alors que la seconde a traité d'un cas de données manquantes en bloc avec une distribution gaussienne composé centrée. Dans les deux cas, les performances des estimateurs ont été validées à travers des simulations numériques sur des données incomplètes, y compris dans le cas où la matrice de covariance admet une structure rang faible.

Dans le cas gaussien, une étude comparative de l'algorithme EM avec la méthode EM-EOF a été réalisée. Tout d'abord, la comparaison a été menée sur l'estimation de la covariance sur données synthétiques. Celle-ci a permis de mettre en évidence la similarité des performances d'estimation entre les deux méthodes, bien que les comportements asymptotiques pour N grand soient différents. L'intérêt de l'étape 2 de la méthode EM-EOF, étape itérative de mise à jour des données manquantes, a été mis en évidence, en montrant une amélioration de l'estimation de la covariance par rapport à l'étape 1. D'autre part, les deux méthodes ont été combinées pour l'interpolation de données manquantes au sein de données réelles de mesure de déplacement vertical par GNSS sur le Piton de la Fournaise. Cette comparaison préliminaire a montré que la méthode EM-EOF produit une interpolation équivalente à l'algorithme EM.

Les données réelles observées pouvant présenter des valeurs atypiques, une extension au cas gaussien composé constitue une suite logique de ce travail. Pour cela, il sera nécessaire d'adapter l'opérateur Sweep au cas gaussien composé. À cette fin, on pourra travailler à une généralisation de l'algorithme de [Liu1999], qui a fourni des résultats sur des données bivariées gaussiennes contenant des données manquantes de forme générale, à des données multivariées de distribution gaussienne composée. Cette perspective est également motivée par une inspection de la distribution empirique des données qui ne s'ajuste pas fidèlement au modèle gaussien classique (ce qui pourrait également être vérifié par un test statistique de gaussianité). Ce futur travail soulève l'intérêt que

porte une meilleure caractérisation de la covariance des données de déplacement. Comme déjà mentionné lors du chapitre 1, la covariance des données est en effet utilisée comme entrée de nombreux modèles d'inversion [Tarantola2005] (voir section 1.5.3 et l'équation 1.16), notamment en déplacement de surface en zone volcanique [Smittarello2019b].

Conclusions et perspectives

> L'ineptie consiste à vouloir conclure. Oui, la bêtise consiste à vouloir conclure. Quel est l'esprit un peu fort qui ait conclu, à commencer par Homère ? Contentons-nous du tableau, c'est ainsi, bon.
>
> – Gustave Flaubert, *Correspondance*

> Sur ce, salut les filles, et meilleure route...
>
> – Virginie Despentes, *King Kong Théorie*

Cette thèse est dédiée au développement et à la mise en oeuvre de méthodes de reconstruction de données manquantes au sein de séries temporelles de champs de déplacement estimés par télédétection. Ces méthodes peuvent être classées en deux approches : une approche prédictive et une approche paramétrique. Cette partie résume les contributions de cette thèse, réparties en deux types d'apports :

1. Les *apports méthodologiques*, c'est-à-dire relatifs au développement théorique de méthodes et d'algorithmes de reconstruction des données ;
2. Les *apports applicatifs* ayant trait à l'application des méthodes développées à des cas d'études réels, originaux et diversifiés, dont l'impact ne se limite pas forcément à la télédétection.

Ces contributions ouvrent des perspectives dont une synthèse est proposée en dernier lieu.

Apports méthodologiques

Deux méthodes de reconstruction des données ont été développées lors de cette thèse. La méthode EM-EOF, implémentée sur la base du travail de [Beckers2003], repose sur la corrélation temporelle du champ de déplacement. La méthode EM-EOF étendue, reprenant les travaux de [Golyandina2010] et [von Buttlar2014], est basée sur la corrélation spatio-temporelle du champ de déplacement. Dans ce cas, une formulation empirique des limites inférieures et supérieures du décalage spatial utilisé pour augmenter les données en espace a été proposée pour la première fois en utilisant des outils simples issus de l'analyse de la longueur de corrélation [Ghil2002] des champs de déplacement. Afin d'étudier la performance des méthodes développées, une batterie de tests ont été menés sur différents types de champs de déplacement à complexité variable, contenant plusieurs formes de données manquantes et des bruits de natures différentes modélisant le bruit corrélé présent dans les données réelles de mesure de déplacement. Une telle diversité de cas de simulations n'a, à notre connaissance, pas été proposée jusqu'ici pour l'étude de données manquantes.

Les deux méthodes proposées constituent une alternative crédible et avantageuse à l'omission de données manquantes et aux approches purement spatiales souvent utilisées en mesure de

déplacement : l'étude comparative avec des méthodes d'interpolation spatiales a montré la supériorité des méthodes proposées surtout lorsque le rapport signal-sur-bruit est bas et que le taux de valeurs manquantes est important. La mise en place de la méthode EM-EOF étendue a aussi permis d'améliorer les performances de reconstruction par rapport à la méthode EM-EOF, plus particulièrement dans les cas où la série temporelle est courte (peu d'observations temporelles) et la corrélation spatiale du champ est importante, assurant une certaine complémentarité des deux méthodes et une continuité dans les recherches menées. De plus, le formalisme de l'algorithme EM a été adopté dans le développement des algorithmes, ce qui a déjà été proposé en ACP mais pas en analyse en EOF. Les méthodes proposées sont itératives, nécessitent peu d'information *a priori* et sont seulement dépendantes des données, ce qui les rend potentiellement transportables à d'autres types de données.

Deux critères robustes, l'un basé sur l'erreur de validation croisée (Λ) et l'autre sur l'incertitude des valeurs propres ainsi que l'autocorrélation spatio-temporelle du champ de déplacement ($\mathcal{C}_k$), ont été proposés puis implémentés. Le premier permet de gérer le problème de sur-estimation du nombre de modes lorsque les données sont perturbées par du bruit corrélé, ce qui est un problème établi dans la littérature de l'analyse en EOF [Kondrashov2006]. Le second critère a été construit en formulant une extension de la taille effective d'échantillon (ESS) temporelle originellement proposée par [North1982] pour estimer l'incertitude des valeurs propres du système à une ESS spatio-temporelle à travers des outils provenant de la géostatistique. Ce critère permet d'ajuster la sélection du nombre afin de garder les groupes de valeurs propres significatives dans la reconstruction.

Dans un second temps de ce travail de thèse, une contribution a été apportée en confrontant les approches prédictives précédemment citées à une approche paramétrique. Deux distributions ont été considérées pour cela : la distribution gaussienne et la distribution gaussienne composée. Dans le cas gaussien, l'algorithme EM pour des données manquantes de forme générale a été développé à partir du travail de [DiCesare2006], puis a été adapté au cas d'une covariance à structure rang faible. Dans le cas des données suivant une distribution gaussienne composée, les estimations de la matrice de covariance et des textures au sens du maximum de vraisemblance ont été dérivées pour le cas de données manquantes en bloc, ce qui constitue un résultat théorique. Un second algorithme EM a été développé avec sa version rang faible. L'ensemble de ces algorithmes ont été validés et cette approche a pu être comparée à la méthode EM-EOF pour l'estimation de la covariance. Les résultats obtenus sont similaires, et ce alors que les méthodes ne poursuivent pas la même "philosophie" de traitement des données manquantes, initiant ainsi une comparaison singulière.

Apports applicatifs

Le second volet des contributions de cette thèse est d'ordre applicatif. Les méthodes EM-EOF et EM-EOF étendue, basées sur l'analyse en EOF et en EEOF, ont été appliquées sur des séries temporelles incomplètes de champs de déplacement, ce qui constitue un objet de recherche neuf. Plusieurs données ont été considérées pour cela : deux séries temporelles de champs de déplacement InSAR sur les glaciers de Miage et du Gorner, une série temporelle de champs de déplacement calculés par corrélation d'images SAR sur le glacier d'Argentière et une série temporelle de champs de déplacement issus de la corrélation d'images optique sur le glacier Fox.

L'application sur des champs de déplacement InSAR (interférogrammes déroulés) a montré que la reconstruction des données manquantes est en accord avec les déplacements observés. Lorsqu'il existe une estimation des vitesses de surface dans la littérature, comme c'est le cas du glacier du Gorner [Prébet2019], les résultats obtenus sont en accord avec les résultats existants. De plus, il a été montré la possibilité de corriger les champs de déplacement bruités et/ou contenant des valeurs atypiques résultant d'erreur de traitement (déroulement de phase). Concernant les déplacements issus de la corrélation d'images SAR, l'originalité de notre étude consiste à avoir appliqué une méthode utilisant l'information temporelle alors que la seule étude dédiée utilise l'information

spatiale [Zhang2019] sur un seul champ de déplacement. Dans les deux cas (InSAR et corrélation), les données manquantes corrélées en espace et en temps ont été reconstruites en conservant les motifs des champs de déplacement.

Concernant les champs de déplacement issus de l'imagerie optique, outre l'application nouvelle de ces méthodes sur ce type de données, il a été démontré la possibilité de reconstruire des zones du champ ne possédant aucune observation temporelle, comme les zones de saturation (surface neigeuse) et les zones de déplacement rapide dans la partie basse du glacier Fox (où il est difficile d'obtenir des mesures de vitesse). Une vérification attentive des résultats a montré leur cohérence avec d'autres résultats de la littérature sur le glacier Fox utilisant des images Sentinel-2 [Kääb2016] et des images Venμs à plus haute résolution [Millan2019].

L'ensemble de ces résultats est très prometteur et montre que les méthodes développées peuvent être intégrées au sein d'une chaîne de traitement (par exemple en pré-traitement) pour produire des séries temporelles continues, spatialement et temporellement résolues. Les méthodes peuvent être aussi utilisées à des fins d'augmentation du volume des données, notamment lorsqu'une image est manquante dans la série ou qu'une zone est constamment incomplète, éléments qui ont été illustrés. L'application des méthodes implementées durant cette thèse peut directement profiter au glaciologue modélisateur. En effet, certains modèles d'estimation de l'épaisseur de glace [Fürst2017, Rabatel2018] exigent d'intégrer des vitesses de surface continues afin de contraindre des paramètres peu connus lié à la topographie basale, comme la contrainte de frottement basal, ou des paramètres liés à l'hydrologie sous-glaciaire où les mesures sont rares.

Enfin, les méthodes développées au chapitre 4 ont permis d'entrevoir une autre approche pour l'estimation de la matrice de covariance des données d'observation GNSS, laquelle est utilisée en modélisation de la déformation terrestre en zone volcanique [Smittarello2019a], et plus largement dans de nombreux problèmes d'inversion des données [Bos2004, Cavalié2013]. Une application de l'algorithme EM pour l'interpolation des données manquantes a également été proposée : cette étude se situant encore au stade de l'ébauche, celle-ci s'inscrit dans les perspectives détaillées ci-après.

Perspectives

Cette thèse ouvre des perspectives méthodologiques et applicatives.

D'un point de vue méthodologique, nous avons appliqué jusqu'à maintenant une augmentation de données spatiale à l'aide d'une fenêtre carrée (méthode EM-EOF étendue). Due à la forme polygonale des cibles de déplacement reconstruites, cette fenêtre peut créer des effets de bords lors de la reconstruction, ce qui est dû au moyennage spatial des données. Pour cela, les auteurs de [Golyandina2015] ont proposé une reconstruction par fenêtre adaptative (Figure 4.17) en masquant préalablement les données situées hors de la cible de déplacement. Dans le cas de la mesure de déplacement, une connaissance exacte des limites du champ de déplacement n'existe pas toujours. Pour ce qui est des glaciers, la plupart des contours des zones glaciaires sont repertoriés dans la base du RGI [Consortium2017], ce qui permettrait d'appliquer une telle technique dans de futures études en utilisant les masques du RGI.

De plus, un effort supplémentaire devra être fourni pour réduire le temps de calcul de la méthode EM-EOF étendue, qui, à l'heure actuelle, possède une complexité algorithmique de l'ordre de $\mathcal{O}(\hat{r}MN + c)$. La mise en ligne des codes EM-EOF et EM-EOF étendue ainsi que des données de simulation prêtes à être testées constituent une préoccupation première de la fin de cette thèse afin de garantir une recherche reproductible, transparente, transportable et facilement améliorable par la communauté scientifique.

Enfin, l'extension de l'algorithme EM au cas gaussien composé, ainsi que l'adaptation au rang faible, assure une suite logique du travail exploratoire présenté au chapitre 4. Pour cela, il s'agira d'adapter l'opérateur Sweep au cas gaussien composé, ce qui reste, selon notre connaissance, inédit pour le moment. De plus, une extension de l'algorithme de [Liu1999], qui propose un algorithme

EM à convergence optimisée (*monotone* EM) pour des données bivariées gaussiennes, sera étudiée pour des données multivariées à distribution gaussiennes composées contenant des données manquantes de forme générale. Ce chantier sera rendu possible en prolongeant la collaboration amorcée avec des chercheurs en traitement du signal de l'Université Paris-Nanterre, ce qui est en bonne voie.

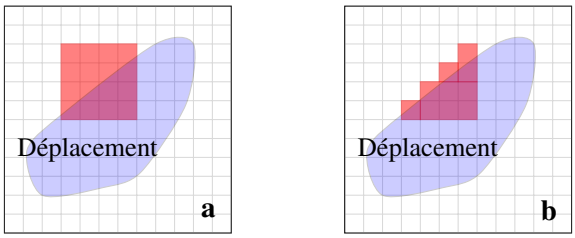


Figure 4.17 – Schéma simplifié d'une fenêtre carrée (a) versus fenêtre adaptative (b) pour l'augmentation spatiale des données (voir figure 3.2).

Du point de vue des applications, nous avons appliqué la méthode EM-EOF sur des interférogrammes déroulés. Ceci a été fait principalement pour une raison : minimiser les biais d'interpolation. On pourrait très bien décider d'appliquer la méthode sur des interférogrammes enroulés, avec les risques que cela peut comporter : en effet, un léger biais d'interpolation peut conduire à un déroulement erroné avec des valeurs décalées. Malgré cette réserve, cette stratégie n'a pas été appliquée et pourrait, à condition de fournir une interpolation suffisamment précise, éviter les erreurs de déroulement de phase auxquelles nous avons été confronté.

L'application sur des mesures GNSS ayant été amorcée dans le cas gaussien, il s'agira par la suite d'appliquer l'EM dans le cas gaussien composé afin d'estimer la matrice de covariance et les textures, puis d'utiliser la covariance pour interpoler les données manquantes. Une première étude a permis de remarquer que la distribution des données n'est pas toujours adaptée au cas gaussien : les éventuels avantages apportés par l'hypothèse gaussienne composée feront donc l'objet de futures discussions.

La liste d'applications est encore longue. De nombreuses données de mesure de déplacement sont concernées par l'incomplétude de données, comme les cartes de subsidence en milieu urbain qui nécessitent une connaissance très localisée des déplacements. Nous avons commencé pendant cette thèse à collaborer avec des chercheurs du laboratoire ISTerre travaillant sur la caractérisation de séismes lents au Mexique. Des résultats préliminaires ont été obtenus sur des interférogrammes incomplets sur la zone montagneuse du sud-ouest de Mexico (voir [Maubant2020]), où les déplacements issus d'un séisme lent sont particulièrement difficiles à extraire des données InSAR.

Bibliographie

[Alvera-Azcárate2005] Aïda Alvera-Azcárate, Alexander Barth, Michel Rixen et Jean-Marie Beckers. *Reconstruction of incomplete oceanographic data sets using empirical orthogonal functions : application to the Adriatic Sea surface temperature*. Ocean Modelling, vol. 9, no. 4, pages 325–346, 2005.

[Alvera-Azcarate2007] A. Alvera-Azcarate, A. Barth, J-M Beckers et 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.

[Anderson1965] T.W. Anderson. An introduction to multivariate statistical analysis. Wiley, New York, 1965.

[Anuta1970] Paul E Anuta. *Spatial registration of multispectral and multitemporal digital imagery using fast Fourier transform techniques*. IEEE transactions on Geoscience Electronics, vol. 8, no. 4, pages 353–368, 1970.

[Aslan2018] Gokhan Aslan, Ziyadin Cakır, Semih Ergintav, Cécile Lasserre et François Renard. *Analysis of Secular Ground Motions in Istanbul from a Long-Term InSAR Time-Series (1992–2017)*. Remote Sensing, vol. 10, no. 3, 2018.

[Aslan2020] Gokhan Aslan, Michael Foumelis, Daniel Raucoules, Marcello De Michele, Severine Bernardie et Ziyadin Cakir. *Landslide Mapping and Monitoring Using Persistent Scatterer Interferometry (PSI) Technique in the French Alps*. Remote Sensing, vol. 12, no. 8, page 1305, 2020.

[Beckers2003] J. M. Beckers et M. Rixen. *EOF calculations and data filling from incomplete oceanographics datasets.* J. Atmos. Oceanic Technol., vol. 20(12), pages 1836–1856, 2003.

[Beckers2006] J.-M. Beckers, A. Barth et A. Alvera-Azcárate. *DINEOF reconstruction of clouded images including error maps - application to the Sea-Surface Temperature around Corsican Island*. Ocean Sci., vol. 2, no. 2, pages 183–199, 2006.

[Berthier2005] E. Berthier, H. Vadon, D. Baratoux, Y. Arnaud, C. Vincent, K.L. Feigl, F. Rémy et B. Legrésy. *Surface motion of mountain glaciers derived from satellite optical imagery*. Remote Sensing of Environment, vol. 95, no. 1, pages 14 – 28, 2005.

[Bhatia2009] Rajendra Bhatia. Positive definite matrices, volume 24. Princeton university press, 2009.

[Björnsson1997] H. Björnsson et S. A. Venegas. *A Manual for EOF and SVD Analyses of Climatic Data*. Department of Atmospheric and Oceanic Sciences and Center for Climate and Global Change Research, McGill University, 02 1997.

[Bos2004] A. G. Bos, S. Usai et W. Spakman. *A joint analysis of GPS motions and InSAR to infer the coseismic surface deformation of the Izmit, Turkey earthquake*. Geophysical Journal International, vol. 158, no. 3, pages 849–863, 09 2004.

[Bouchard2020] Florent Bouchard, Ammar Mian, Jialun Zhou, Salem Said, Guillaume Ginolhac et Yannick Berthoumieu. *Riemannian geometry for Compound Gaussian distributions : application to recursive change detection*, 2020.

[Brankart1995] Jean-Michel Brankart et Pierre Brasseur. *Optimal Analysis of In Situ Data in the Western Mediterranean Using Statistics and Cross-Validation*. J. Atmos. Oceanic Technol., vol. 13, pages 477–491, 1995.

[Broomhead1986] D.S. Broomhead et Gregory P. King. *Extracting qualitative dynamics from experimental data*. Physica D : Nonlinear Phenomena, vol. 20, no. 2, pages 217 – 236, 1986.

[Bürgmann2000] Roland Bürgmann, Paul A Rosen et Eric J Fielding. *Synthetic aperture radar interferometry to measure Earth's surface topography and its deformation*. Annual review of earth and planetary sciences, vol. 28, no. 1, pages 169–209, 2000.

[Cavalié2013] O Cavalié, E Pathier, Mathilde Radiguet, M Vergnolle, N Cotte, A Walpersdorf, V Kostoglodov et Fabrice Cotton. *Slow slip event in the Mexican subduction zone : Evidence of shallower slip in the Guerrero seismic gap for the 2006 event revealed by the joint inversion of InSAR and GPS data*. Earth and Planetary Science Letters, vol. 367, pages 52–60, 2013.

[Cazenave2004] Anny Cazenave et Robert S Nerem. *Present-day sea level change : Observations and causes*. Reviews of Geophysics, vol. 42, no. 3, 2004.

[Chang2018] Wen-Yen Chang, Meng-Che Wu, Yang-Lang Chang, Sheng-Yung Shih et Bormin Huang. *GPU acceleration of Adaptive Local Kriging Applied to retrieving slant-range surface motion maps*. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 11, no. 11, pages 4317–4325, 2018.

[Chen2004] Jin Chen, Per Jönsson, Masayuki Tamura, Zhihui Gu, Bunkei Matsushita et Lars Eklundh. *A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky–Golay filter*. Remote sensing of Environment, vol. 91, no. 3-4, pages 332–344, 2004.

[Chen2009] Tao Chen, Elaine Martin et Gary Montague. *Robust probabilistic PCA with missing data and contribution analysis for outlier detection*. Computational Statistics & Data Analysis, vol. 53, no. 10, pages 3706–3716, 2009.

[Chen2011] Jin Chen, Xiaolin Zhu, James E Vogelmann, Feng Gao et Suming Jin. *A simple and effective method for filling gaps in Landsat ETM+ SLC-off images*. Remote sensing of environment, vol. 115, no. 4, pages 1053–1064, 2011.

[Chen2017] Yu Chen, Dominique Remy, Jean-Luc Froger, Aline Peltier, Nicolas Villeneuve, José Darrozes, Hugo Perfettini et Sylvain Bonvalot. *Long-term ground displacement observations using InSAR and GNSS at Piton de la Fournaise volcano between 2009 and 2014*. Remote Sensing of Environment, vol. 194, pages 230 – 247, 2017.

[Cheng2013] Qing Cheng, Huanfeng Shen, Liangpei Zhang et Pingxiang Li. *Inpainting for remotely sensed images with a multichannel nonlocal total variation model*. IEEE Transactions on Geoscience and Remote Sensing, vol. 52, no. 1, pages 175–187, 2013.

[Cheng2019] Qing Cheng, Qiangqiang Yuan, Michael Kwok-Po Ng, Huanfeng Shen et Liangpei Zhang. *Missing Data Reconstruction for Remote Sensing Images With Weighted Low-Rank Tensor Model*. IEEE Access, vol. 7, pages 142339–142352, 2019.

[Choudhury2015] NH. Choudhury, A. Rahman et S. Ferdousi. *Kriging Infill of Missing Data and Temporal Analysis of Rainfall in North Central Region of Bangladesh*. J. Climatol. Weather Forecasting, vol. 3, no. 141, 2015.

[Comiso2002] Josefino C Comiso. *A rapidly declining perennial sea ice cover in the Arctic*. Geophysical Research Letters, vol. 29, no. 20, pages 17–1, 2002.

[Consortium2017] RGI Consortium. *Randolph Glacier Inventory – A Dataset of Global Glacier Outlines : Version 6.0*. Rapport technique, Global Land Ice Measurements from Space, Colorado, USA, 2017.

[Corominas2014] J Corominas, Cees van Westen, P Frattini, L Cascini, J-P Malet, S Fotopoulou, F Catani, M Van Den Eeckhaut, O Mavrouli, F Agliardi*et al. Recommendations for the quantitative analysis of landslide risk*. Bulletin of engineering geology and the environment, vol. 73, no. 2, pages 209–263, 2014.

[Cressie2008] Noel Cressie et Gardar Johannesson. *Fixed Rank Kriging for Very Large Spatial Data Sets*. Journal of the Royal Statistical Society. Series B (Statistical Methodology), vol. 70, no. 1, pages 209–226, 2008.

[Curry1944] Haskell B. Curry. *The method of steepest descent for non-linear minimization problems*. Quart. Appl. Math., vol. 2, pages 258–261, 1944.

[De Oliveira2014] Julio Cesar De Oliveira, José Carlos Neves Epiphanio et Camilo Daleles Rennò. *Window regression : A spatial-temporal analysis to estimate pixels classified as low-quality in MODIS NDVI time series*. Remote Sensing, vol. 6, no. 4, pages 3123–3142, 2014.

[Dehecq2015] Amaury Dehecq. *Analysis of himalayan and alpine glaciers dynamic with the use of 40 years of Earth's observation data.* Thèse de doctorat, Université Grenoble Alpes, November 2015.

[Dempster1977] A. P. Dempster, N. M. Laird et D. B. Rubin. *Maximum Likelihood from Incomplete Data via the EM Algorithm*. J. Royal Statistical Society. Series B (Methodological), vol. 39, no. 1, pages 1–38, 1977.

[DiCesare2006] G. DiCesare. *Imputation, estimation and missing data in finance*. Thèse de doctorat (non publiée), University of Waterloo, 2006.

[Doin2009] M-P Doin, Cécile Lasserre, Gilles Peltzer, Olivier Cavalié et Cécile Doubre. *Corrections of stratified tropospheric delays in SAR interferometry : Validation with global atmospheric models*. Journal of Applied Geophysics, vol. 69, no. 1, pages 35–50, 2009.

[Dong2006] D. Dong, P. Fang, Y. Bock, F. Webb, L. Prawirodirdjo, S. Kedar et P. Jamason. *Spatiotemporal filtering using principal component analysis and Karhunen-Loeve expansion approaches for regional GPS network analysis*. Journal of Geophysical Research : Solid Earth, vol. 111, no. B3, 2006.

[Drašković2018] G. Drašković et F. Pascal. *New Insights Into the Statistical Properties of M-Estimators*. IEEE Transactions on Signal Processing, vol. 66, no. 16, pages 4253–4263, 2018.

[Duquenne2005] Françoise Duquenne, Serge Botton, François Peyret, David Bétaille et Pascal Willis. *GPS : localisation et navigation par satellites*. HERMES Science Publication, page 22, 2005.

[Erten2009] Esra Erten, Andreas Reigber, Olaf Hellwich et Pau Prats. *Glacier velocity monitoring by maximum likelihood texture tracking*. IEEE Transactions on Geoscience and Remote Sensing, vol. 47, no. 2, pages 394–405, 2009.

[Fallourd2011] R. Fallourd, O. Harant, E. Trouvé et P. Bolon. *Monitoring temperate glacier displacement by multi-temporal TerraSAR-X images and continuous GPS measurements*. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 4, no. 2, pages 372–386, 2011.

[Fallourd2012] Renaud Fallourd. *Monitoring alpine glaciers by combination of heterogeneous informations : High Resolution SAR image and ground measurements*. Thèse de doctorat, Université de Grenoble, April 2012.

[Fattahi2014] Heresh Fattahi et Falk Amelung. *InSAR uncertainty due to orbital errors*. Geophysical Journal International, vol. 199, no. 1, pages 549–560, 2014.

[Faugeras1993] Olivier Faugeras, Thierry Viéville, Eric Theron, Jean Vuillemin, Bernard Hotz, Zhengyou Zhang, Laurent Moll, Patrice Bertin, Herve Mathieu et Pascal Fua. *Real-time correlation-based stereo : algorithm, implementations and applications*. Rapport technique RR-2013, INRIA, 1993.

[Ferretti2007] Alessandro Ferretti, Andrea Monti-Guarnieri, Claudio Prati, Fabio Rocca et D Massonet. *InSAR principles-guidelines for SAR interferometry processing and interpretation, TM-19*. The Netherlands : ESA Publications, 2007.

[Fisher1922] R. A. Fisher et Edward John Russell. *On the mathematical foundations of theoretical statistics*. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, vol. 222, no. 594-604, pages 309–368, 1922.

[Frahm2010] Gabriel Frahm et Uwe Jaekel. *A generalization of Tyler's M-estimators to the case of incomplete data*. Computational Statistics & Data Analysis, vol. 54, no. 2, pages 374–393, 2010.

[Fukuoka1951] A. Fukuoka. *A study of 10-day forecast (A Synthetic Report)*. Geophys. Mag., vol. 22, no. 3, pages 177–218, 1951.

[Fukushima2005] Y. Fukushima, V. Cayol et P. Durand. *Finding realistic dike models from interferometric synthetic aperture radar data : The February 2000 eruption at Piton de la Fournaise*. Journal of Geophysical Research : Solid Earth, vol. 110, no. B3, 2005.

[Fürst2017] J. J. Fürst, F. Gillet-Chaulet, T. J. Benham, J. A. Dowdeswell, M. Grabiec, F. Navarro, R. Pettersson, G. Moholdt, C. Nuth, B. Sass, K. Aas, X. Fettweis, C. Lang, T. Seehaus et M. Braun. *Application of a two-step approach for mapping ice thickness to various glacier types on Svalbard*. The Cryosphere, vol. 11, no. 5, pages 2003–2032, 2017.

[Gao2008] Feng Gao, Jeffrey T Morisette, Robert E Wolfe, Greg Ederer, Jeff Pedelty, Edward Masuoka, Ranga Myneni, Bin Tan et Joanne Nightingale. *An algorithm to produce temporally and spatially continuous MODIS-LAI time series*. IEEE Geoscience and Remote Sensing Letters, vol. 5, no. 1, pages 60–64, 2008.

[Gao2010] Gui Gao. *Statistical Modeling of SAR Images : A Survey*. Sensors, vol. 10, no. 1, pages 775–795, 2010.

[Garcia2010] Damien Garcia. *Robust smoothing of gridded data in one and higher dimensions with missing values*. Computational Statistics & Data Analysis, vol. 54, no. 4, pages 1167 – 1178, 2010.

[Gentile2012] M. Gentile, F. Courbin et G. Meylan. *Interpolating point spread function anisotropy*. Astron. Astrophys., vol. 549, page A1, Dec 2012.

[Gerber2018] Florian Gerber, Rogier de Jong, Michael E. Schaepman, Gabriela Schaepman-Strub et Reinhard Furrer. *Predicting Missing Values in Spatio-Temporal Remote Sensing Data*. IEEE Trans. Geosci. Remote Sens., vol. 56, no. 5, pages 2841–2853, 2018.

[Ghil2002] M. Ghil, M.R. Allen, M. D Dettinger, K. Ide, D. Kondrashov, M.E. Mann, A.W. Robertson, A. Saunders, Y. Tian, F. Varadi et P. Yiou. *Advanced spectral methods for climatic time series.* Review of Geophysics, vol. 40, 1, pages 1–41, 2002.

[Giannakis2012] Dimitrios Giannakis et Andrew J Majda. *Nonlinear Laplacian spectral analysis for time series with intermittency and low-frequency variability*. Proceedings of the National Academy of Sciences, vol. 109, no. 7, pages 2222–2227, 2012.

[Gini2000] F. Gini, M. V. Greco, M. Diani et L. Verrazzani. *Performance analysis of two adaptive radar detectors against non-Gaussian real sea clutter data*. IEEE Transactions on Aerospace and Electronic Systems, vol. 36, no. 4, pages 1429–1439, 2000.

[Goldstein1988] Richard M Goldstein, Howard A Zebker et Charles L Werner. *Satellite radar interferometry : Two-dimensional phase unwrapping*. Radio science, vol. 23, no. 4, pages 713–720, 1988.

[Golub1996] Gene H. Golub et Charles F. Van Loan. Matrix computations. The Johns Hopkins University Press, third edition, 1996.

[Golyandina2010] N.E. Golyandina et K.D. Usevich. *2D-Extension of Singular Spectrum Analysis : Algorithm and Elements of Theory*. Matrix Methods : Theory, Algorithms and Applications, pages 449–473, 2010.

[Golyandina2015] Nina Golyandina, Anton Korobeynikov, Alex Shlemov et Konstantin Usevich. *Multivariate and 2D Extensions of Singular Spectrum Analysis with the Rssa Package*. Journal of Statistical Software, Articles, vol. 67, no. 2, pages 1–78, 2015.

[Gomis2001] Damià Gomis, Simón Ruiz et Mike A Pedder. *Diagnostic analysis of the 3D ageostrophic circulation from a multivariate spatial interpolation of CTD and ADCP data*. Deep Sea Research Part I : Oceanographic Research Papers, vol. 48, no. 1, pages 269–295, 2001.

[Goodnight1979] J. H. Goodnight. *A tutorial on the SWEEP operator*. American Statistics, vol. 33, pages 149–158, 1979.

[Goovaerts1997] P. Goovaerts. Geostatistics for natural resources evaluation. Oxford University Press, 1997.

[Graham1974] Leroy C Graham. *Synthetic interferometer radar for topographic mapping*. Proceedings of the IEEE, vol. 62, no. 6, pages 763–768, 1974.

[Graham2012] John W Graham, Patricio E Cumsille et Allison E Shevock. *Methods for handling missing data*. Handbook of Psychology, Second Edition, vol. 2, 2012.

[Greco2007] M. S. Greco et F. Gini. *Statistical Analysis of High-Resolution SAR Ground Clutter Data*. IEEE Transactions on Geoscience and Remote Sensing, vol. 45, no. 3, pages 566–575, 2007.

[Groth2015] Andreas Groth et Michael Ghil. *Monte Carlo Singular Spectrum Analysis (SSA) Revisited : Detecting Oscillator Clusters in Multivariate Datasets*. J. Climate, vol. 28, no. 19, pages 7873–7893, 2015.

[Gualandi2016] Adriano Gualandi, Enrico Serpelloni et Maria Elina Belardinelli. *Blind source separation problem in GPS time series*. Journal of Geodesy, vol. 90, no. 4, pages 323–341, 2016.

[Gudmundsson2002] Sverrir Gudmundsson, Freysteinn Sigmundsson et Jens Carstensen. *Three-dimensional surface motion maps estimated from combined interferometric synthetic aperture radar and GPS data*. J. Geophys. Res., vol. 107, 10 2002.

[Gupta2011] Maya R. Gupta et Yihua Chen. *Theory and Use of the EM Algorithm*. Foundations and Trends in Signal Processing, vol. 4, no. 3, pages 223–296, 2011.

[Hannachi2001] A. Hannachi et A. O'Neill. *Atmospheric multiple equilibria and non-Gaussian behaviour in model simulations*. Quarterly Journal of the Royal Meteorological Society, vol. 127, no. 573, pages 939–958, 2001.

[Hannachi2007] A. Hannachi, I.T. Jollife et D.B. Stephenson. *Empirical orthogonal functions and related techniques in atmospheric science : A review.* Int. J. Climatol., vol. 27, pages 1119–1152, 2007.

[Hanssen2001] Ramon F Hanssen. Radar interferometry : data interpretation and error analysis, volume 2. Springer Science & Business Media, 2001.

[Heimann2017] Sebastian Heimann, Marius Kriegerowski, Marius Isken, Simone Cesca, Simon Daout, Francesco Grigoli, Carina Juretzek, Tobias Megies, Nima Nooshiri, Andreas Steinberg, Henriette Sudhaus, Hannes Vasyura-Bathke, Timothy Willey et Torsten Dahm. *Pyrocko - An open-source seismology toolbox and library. V. 0.3*. GFZ Data Services, 2017.

[Hergert2010] Tobias Hergert et Oliver Heidbach. *Slip-rate variability and distributed deformation in the Marmara Sea fault system*. Nature Geoscience, vol. 3, no. 2, pages 132–135, 2010.

[Herman2011] Frédéric Herman, Brian Anderson et Sébastien Leprince. *Mountain glacier velocity variation during a retreat/advance cycle quantified using sub-pixel analysis of ASTER images*. J. Glaciol., vol. 57, no. 202, page 197–207, 2011.

[Hippert-Ferrer2020] Alexandre Hippert-Ferrer, Yajing Yan et Philippe Bolon. *EM-EOF : gap-filling in incomplete SAR displacement time series*. IEEE Trans. Geosci. Remote Sens., vol. in review, 2020.

[Hocke2009] K. Hocke et N. Kämpfer. *Gap filling and noise reduction of unevenly sampled data by means of the Lomb-Scargle periodogram*. Atmos. Chem. Phys., vol. 9, no. 12, pages 4197–4206, 2009.

[Hotelling1933] H. Hotelling. *Analysis of a complex of statistical variables into principal components*. Journal of Educational Psychology, vol. 24, pages 417–520, 1933.

[Hotelling1935] H. Hotelling. *The most predictable criterion*. Journal of Educational Psychology, vol. 26, pages 139–142, 1935.

[Howell2007] David C Howell. *The treatment of missing data*. The Sage handbook of social science methodology, pages 208–224, 2007.

[Jakeman1980] E Jakeman. *On the statistics of K-distributed noise*. Journal of Physics A : Mathematical and General, vol. 13, no. 6, pages 2251–2251, jun 1980.

[Jakob2012] M Jakob. *The fallacy of frequency. Statistical techniques for debris flow frequency magnitude analysis*. In Proceedings of the International Landslide Conference, Banff, Canada, pages 2–8, 2012.

[Jolivet2011] R. Jolivet, R. Grandin, C. Lasserre, M.-P. Doin et G. Peltzer. *Systematic InSAR tropospheric phase delay corrections from global meteorological reanalysis data*. Geophys. Res. Lett., vol. 38, no. 17, 2011.

[Jollife2002] IT Jollife. Principal component analysis. Springer, New York, 2nd edition edition, 2002.

[Jones2001] Donald R. Jones. *A Taxonomy of Global Optimization Methods Based on Response Surfaces*. J. Global Optim., vol. 21, no. 4, pages 345–383, Dec 2001.

[Jönsson2004] Per Jönsson et Lars Eklundh. *TIMESAT—a program for analyzing time-series of satellite sensor data*. Computers & geosciences, vol. 30, no. 8, pages 833–845, 2004.

[Joughin1998] Ian R Joughin, Ronald Kwok et Mark A Fahnestock. *Interferometric estimation of three-dimensional ice-flow using ascending and descending passes*. IEEE Transactions on Geoscience and Remote Sensing, vol. 36, no. 1, pages 25–37, 1998.

[Julien2010] Yves Julien et José A Sobrino. *Comparison of cloud-reconstruction methods for time series of composite NDVI data*. Remote Sensing of Environment, vol. 114, no. 3, pages 618–625, 2010.

[Karhunen1946] Kari Karhunen. *Zur spektraltheorie stochastischer prozesse*. Ann. Acad. Sci. Fennicae, AI, vol. 34, 1946.

[Khazraei2019] SM Khazraei et AR Amiri-Simkooei. *On the application of Monte Carlo singular spectrum analysis to GPS position time series*. Journal of Geodesy, vol. 93, no. 9, pages 1401–1418, 2019.

[Kim2000] Kwang-Y Kim et Qigang Wu. *Optimal detection using cyclostationary EOFs*. Journal of climate, vol. 13, no. 5, pages 938–950, 2000.

[Kimoto1991] M. Kimoto, M. Ghil et KC. Mo. *Spatial structure of the extratropical 40-day oscillation*. In Proceedings of the 8th Conference on Atmospheric and Oceanic waves and Stability, pages 115–116, Boston, MA, 1991. American Meteorological Society.

[Kondrashov2006] D. Kondrashov et M. Ghil. *Spatio-temporal filling of missing points in geophysical data sets*. Nonlinear Processes Geophys., vol. 13, pages 151–159, 2006.

[Korobeynikov2010] Anton Korobeynikov. *Computation- and Space-Efficient Implementation of SSA*. Statistics and Its Interface, vol. 3, no. 3, pages 357–368, 2010.

[Kosambi1943] D.D. Kosambi. *Statistics in function spaces*. Journal of the Indian Mathematical Society, vol. 7, pages 73–88, 1943.

[Kositsky2010] A. P. Kositsky et J.-P. Avouac. *Inverting geodetic time series with a principal component analysis-based inversion method.* J. Geophys. Res., vol. 115, no. B03401, 2010.

[Kutzbach1967] John E. Kutzbach. *Empirical Eigenvectors of Sea-Level Pressure, Surface Temperature and Precipitation Complexes over North America*. J. Appl. Meteorol., vol. 6, no. 5, pages 791–802, 1967.

[Kääb2016] Andreas Kääb, Solveig H. Winsvold, Bas Altena, Christopher Nuth, Thomas Nagler et Jan Wuite. *Glacier Remote Sensing Using Sentinel-2. Part I : Radiometric and Geometric Performance, and Application to Ice Velocity*. Remote Sensing, vol. 8, no. 7, 2016.

[Latif2008] Bassam Abdel Latif, Rémi Lecerf, Grégoire Mercier et Laurence Hubert-Moy. *Preprocessing of low-resolution time series contaminated by clouds and shadows*. IEEE Transactions on Geoscience and Remote Sensing, vol. 46, no. 7, pages 2083–2096, 2008.

[Lepot2017] Mathieu Lepot, Jean-Baptiste Aubin et François H.L.R. Clemens. *Interpolation in Time Series : An Introductive Overview of Existing Methods, Their Performance Criteria and Uncertainty Assessment.* Water, vol. 9, no. 796, 2017.

[Lin2013] Chao-Hung Lin, Kang-Hua Lai, Zhi-Bin Chen et Jyun-Yuan Chen. *Patch-based information reconstruction of cloud-contaminated multitemporal images*. IEEE Transactions on Geoscience and Remote Sensing, vol. 52, no. 1, pages 163–174, 2013.

[Lin2014] Chao-Hung Lin, Kean-Hua Lai, Zhi-Bin Chen et Jyun-Yuan Chen. *Patch-Based Information Reconstruction of Cloud-contamined Multitemporal Images*. IEEE Trans. Geosci. Remote Sens., vol. 52, no. 1, pages 163–174, 2014.

[Little1987] R. J. A. Little et D. B. Rubin. *Statistical analysis with missing data*. Hoboken, NJ : Wiley, vol. 65, 1987.

[Little2002] R. J. A. Little et D. B. Rubin. Statistical analysis with missing data. Wiley, New York, 2nd edition, 2002.

[Little2014] Todd D Little, Terrence D Jorgensen, Kyle M Lang et E Whitney G Moore. *On the joys of missing data*. Journal of pediatric psychology, vol. 39, no. 2, pages 151–162, 2014.

[Liu1999] Chuanhai Liu. *Efficient ML estimation of the multivariate normal distribution from incomplete data*. Journal of Multivariate Analysis, vol. 69, pages 206–217, 1999.

[Liu2018a] Ning Liu, Wujiao Dai, Rock Santerre et Cuilin Kuang. *A MATLAB-based Kriged Kalman Filter software for interpolating missing data in GNSS coordinate time series*. GPS Solutions, vol. 22, no. 1, page 25, 2018.

[Liu2018b] X. Liu et M. Wang. *Gap Filling of Missing Data for VIIRS Global Ocean Color Products Using the DINEOF Method*. IEEE Transactions on Geoscience and Remote Sensing, vol. 56, no. 8, pages 4464–4476, 2018.

[Liu2019a] H. Liu, Z. Liu, S. Liu, Y. Liu, J. Bin, F. Shi et H. Dong. *A Nonlinear Regression Application via Machine Learning Techniques for Geomagnetic Data Reconstruction Processing*. IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 1, pages 128–140, 2019.

[Liu2019b] Junyan Liu et Daniel P. Palomar. *Regularized robust estimation of mean and covariance matrix for incomplete data*. Signal Processing, vol. 165, pages 278 – 291, 2019.

[Loève1945] M. Loève. *Fonctions aléatoires de second ordre*. C. R. Acad. Sci., 1945.

[Lorenz1956] E. N. Lorenz. *Empirical orthogonal functions and statistical weather prediction*. Statistical forecasting project, M.I.T., Cambridge, MA, 1956. 48 pp.

[Lu2007] Xiaoliang Lu, Ronggao Liu, Jiyuan Liu et Shunlin Liang. *Removal of noise by wavelet method to generate high quality temporal data of terrestrial MODIS products*. Photogrammetric Engineering & Remote Sensing, vol. 73, no. 10, pages 1129–1139, 2007.

[Mahot2012] Mélanie Mahot. *Robust covariance matrix estimation in signal processing*. Thèse de doctorat, École normale supérieure de Cachan - ENS Cachan, December 2012.

[Maître2013] Henri Maître. Processing of synthetic aperture radar (SAR) images. John Wiley & Sons, 2013.

[Malek2017] Salim Malek, Farid Melgani, Yakoub Bazi et Naif Alajlan. *Reconstructing cloud-contaminated multispectral images with contextualized autoencoder neural networks*. IEEE Transactions on Geoscience and Remote Sensing, vol. 56, no. 4, pages 2270–2282, 2017.

[Manolakis2001] Dimitris Manolakis, David Marden, John Kerekes et Gary Shaw. *On the statistics of hyperspectral imaging data*. In Proc. SPIE, pages 308–316, 2001.

[Mantovani2019] Matteo Mantovani, Giulia Bossi, Gianluca Marcato, Luca Schenato, Giacomo Tedesco, Giacomo Titti et Alessandro Pasuto. *New Perspectives in Landslide Displacement Detection Using Sentinel-1 Datasets*. Remote Sensing, vol. 11, no. 18, 2019.

[Mariethoz2010] Gregoire Mariethoz et Philippe Renard. *Reconstruction of incomplete data sets or images using direct sampling*. Mathematical Geosciences, vol. 42, no. 3, pages 245–268, 2010.

[Maronna1976] Ricardo Antonio Maronna. *Robust M-Estimators of Multivariate Location and Scatter*. Annals of Statistics, vol. 4, no. 1, pages 51–67, 01 1976.

[Maronna2006] R.A. Maronna, R.D. Martin et V.J. Yohai. Robust statistics : Theory and methods. Wiley Series in Probability and Statistics. John Wiley and Sons, first edition, 2006.

[Maubant2020] L. Maubant, E. Pathier, S. Daout, M. Radiguet, M.-P. Doin, E. Kazachkina, V. Kostoglodov, N. Cotte et A. Walpersdorf. *Independent Component Analysis and Parametric Approach for Source Separation in InSAR Time Series at Regional Scale : Application to the 2017–2018 Slow Slip Event in Guerrero (Mexico)*. Journal of Geophysical Research : Solid Earth, vol. 125, no. 3, page e2019JB018187, 2020.

[Melgani2006] F. Melgani. *Contextual Reconstruction of Cloud-Contaminated Multitemporal Multispectral Images*. IEEE Trans. Geosci. Remote Sens., vol. 44, pages 442–455, February 2006.

[Mendez-Rial2011] Roi Mendez-Rial, María Calvino-Cancela et Julio Martin-Herrero. *Anisotropic inpainting of the hypercube*. IEEE Geoscience and Remote Sensing Letters, vol. 9, no. 2, pages 214–218, 2011.

[Mian2019] Ammar Mian. *Contributions to SAR Image Time Series Analysis*. Thèse de doctorat, Université Paris-Saclay, September 2019.

[Millan2019] Romain Millan, Jérémie Mouginot, Antoine Rabatel, Seongsu Jeong, Diego Cusicanqui, Anna Derkacheva et Mondher Chekki. *Mapping Surface Flow Velocity of Glaciers at Regional Scale Using a Multiple Sensors Approach*. Remote Sensing, vol. 11, no. 21, 2019.

[Molnar2008] Frank J. Molnar, Brian Hutton et Dean Fergussen. *Does analysis using "last observation carried forward" introduce bias in dementia research ?* CMAJ : Canadian Medical Association journal, vol. 170, no. 8, page 751–753, 2008.

[Mora2003] Oscar Mora, Jordi J Mallorqui et Antoni Broquetas. *Linear and nonlinear terrain deformation maps from a reduced set of interferometric SAR*

images. IEEE Transactions on Geoscience and Remote Sensing, vol. 41, no. 10, pages 2243–2253, 2003.

[Moreira2013] Alberto Moreira, Pau Prats-Iraola, Marwan Younis, Gerhard Krieger, Irena Hajnsek et Konstantinos P Papathanassiou. *A tutorial on synthetic aperture radar*. IEEE Geoscience and remote sensing magazine, vol. 1, no. 1, pages 6–43, 2013.

[Moreno2014] Álvaro Moreno, Francisco Javier García-Haro, Beatriz Martínez et María Amparo Gilabert. *Noise reduction and gap filling of fapar time series using an adapted local regression filter*. Remote Sensing, vol. 6, no. 9, pages 8238–8260, 2014.

[Nakamura2007] Kazuki Nakamura, Koichiro Doi et Kazuo Shibuya. *Estimation of seasonal changes in the flow of Shirase Glacier using JERS-1/SAR image correlation*. Polar Sci., vol. 1, no. 2-4, pages 73–83, 2007.

[Neteler2010] Markus Neteler. *Estimating daily land surface temperatures in mountainous environments by reconstructed MODIS LST data*. Remote sensing, vol. 2, no. 1, pages 333–351, 2010.

[Ng1997] Andrew Y. Ng. *Preventing "Overfitting" of Cross-Validation Data*. In Proceedings of the Fourteenth International Conference on Machine Learning, ICML '97, pages 245–253, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[North1982] Gerald R. North, Thomas L. Bell, Robert F. Cahalan et Fanthune J. Moeng. *Sampling Errors in the Estimation of Empirical Orthogonal Functions*. Mon. Weather Rev., vol. 110, no. 7, pages 699–706, 1982.

[Obukhov1947] AM Obukhov. *Statistically homogeneous fields on a sphere.* Uspethi Mathematicheskikh Nauk, vol. 2, pages 196–198, 1947.

[Olinsky2003] Alan Olinsky, Shaw Chen et Lisa Harlow. *The comparative efficacy of imputation methods for missing data in structural equation modeling*. European Journal of Operational Research, vol. 151, no. 1, pages 53 – 79, 2003.

[Ollila2012] Esa Ollila, David Tyler, Visa Koivunen et H. Vincent Poor. *Complex Elliptically Symmetric Distributions : Survey, New Results and Applications*. IEEE Transactions on Signal Processing, vol. 60, pages 5597–5625, 11 2012.

[Overland1982] James Overland et R W. Preisendorfer. *A Significance Test for Principal Components Applied to a Cyclone Climatology*. Mon. Weather Rev., vol. 110, page 1, 01 1982.

[Pascal2008] Frédéric Pascal, Yacine Chitour, Jean Philippe Ovarlez, Philippe Forster et Pascal Larzabal. *Covariance Structure Maximum-Likelihood Estimates in Compound Gaussian Noise : Existence and Algorithm Analysis*. IEEE Trans. Signal Process., vol. 56, no. 1, pages 34–48, 2008.

[Pepe2016] Antonio Pepe, Manuela Bonano, Qing Zhao, Tianliang Yang et Hanmei Wang. *The Use of C-/X-Band Time-Gapped SAR Data and Geotechnical Models for the Study of Shanghai's Ocean-Reclaimed Lands through the SBAS-DInSAR Technique*. Remote Sensing, vol. 8, no. 11, 2016.

[Plaut1994] Guy Plaut et Robert Vautard. *Spells of Low-Frequency Oscillations and Weather Regimes in the Northern Hemisphere*. Journal of the Atmospheric Sciences, vol. 51, no. 2, pages 210–236, 1994.

[Poggio2012] Laura Poggio, Alessandro Gimona et Iain Brown. *Spatio-temporal MODIS EVI gap filling under cloud cover : An example in Scotland*. ISPRS journal of photogrammetry and remote sensing, vol. 72, pages 56–72, 2012.

[Preisendorfer1988] Rudolph W. Preisendorfer. Principal Component Analysis in Meteorology and Oceanography. Elsevier, 1988.

[Prébet2017] Rémi Prébet. Extraction du signal de déplacement à partir d'une série temporelle d'interférogrammes sentinel-1 dans des milieux montagneux. Master's thesis, Ecole Normale Supérieure Paris-Saclay, 2017.

[Prébet2019] Rémi Prébet, Yajing Yan, Matthias Jauvin et Emmanuel 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., vol. 57, no. 8, pages 5829–5852, 2019.

[Rabatel2018] Antoine Rabatel, Olivier Sanchez, Christian Vincent et Delphine Six. *Estimation of Glacier Thickness From Surface Mass Balance and Ice Flow Velocities : A Case Study on Argentière Glacier, France*. Frontiers in Earth Science, vol. 6, page 112, 2018.

[Rao1972] C.R. Rao. Linear statistical inference and its applications. Wiley, New York, 1972.

[Ray2008] J Ray, Z Altamimi, X Collilieux et Tonie van Dam. *Anomalous harmonics in the spectra of GPS position estimates*. GPS solutions, vol. 12, no. 1, pages 55–64, 2008.

[Reed1974] I. S. Reed, J. D. Mallett et L. E. Brennan. *Rapid Convergence Rate in Adaptive Arrays*. IEEE Transactions on Aerospace and Electronic Systems, vol. AES-10, no. 6, pages 853–863, 1974.

[Rocca2007] Fabio Rocca. *Modeling interferogram stacks*. IEEE Transactions on Geoscience and Remote Sensing, vol. 45, no. 10, pages 3289–3299, 2007.

[Roerink2000] GJ Roerink, Massimo Menenti et Wout Verhoef. *Reconstructing cloudfree NDVI composites using Fourier analysis of time series*. International Journal of Remote Sensing, vol. 21, no. 9, pages 1911–1917, 2000.

[Rosen2004] Paul A. Rosen, Scott Hensley, Gilles Peltzer et Mark Simons. *Updated repeat orbit interferometry package released*. Eos, Transactions American Geophysical Union, vol. 85, no. 5, pages 47–47, 2004.

[Rubin1976] Donald B. Rubin. *Inference and missing data*. Biometrika, vol. 63, no. 3, pages 581–592, 1976.

[Santos2019] Miriam Seoane Santos, Ricardo Cardoso Pereira, Adriana Fonseca Costa, Jastin Pompeu Soares, João Santos et Pedro Henriques Abreu. *Generating synthetic missing data : A review by missing mechanism*. IEEE Access, vol. 7, pages 11651–11667, 2019.

[Scambos1992] Theodore A Scambos, Melanie J Dutkiewicz, Jeremy C Wilson et Robert A Bindschadler. *Application of image cross-correlation to the measurement of glacier velocity using satellite image data*. Remote sensing of environment, vol. 42, no. 3, pages 177–186, 1992.

[Schneider2001] Tapio Schneider. *Analysis of Incomplete Climate Data : Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values*. J. Climate, vol. 14, pages 853–871, 2001.

[Serafino2006] Francesco Serafino. *SAR image coregistration based on isolated point scatterers*. IEEE Geoscience and Remote Sensing Letters, vol. 3, no. 3, pages 354–358, 2006.

[Serneels2008] Sven Serneels et Tim Verdonck. *Principal component analysis for data containing outliers and missing elements*. Comput. Stat. Data Anal., vol. 52, pages 1712–1727, 2008.

[Shen2015] Huanfeng Shen, Xinghua Li, Qing Chen, Chao Zeng, Gang Yang, Huifang Li et Liangpei Zhang. *Missing Information Reconstruction of Remote Sensing Data : A Technical Review*. IEEE Geosci. Remote Sens. Mag., vol. 3, pages 61–85, 09 2015.

[Shnidman2005] D. A. Shnidman. *Radar detection in clutter*. IEEE Transactions on Aerospace and Electronic Systems, vol. 41, no. 3, pages 1056–1067, 2005.

[Sibson1980] Robin Sibson, editeur. A brief description of natural neighbor interpolation, numéro 24-27, Sheffield, UK, March 1980. Proceedings of the Interpreting Multivariate Data.

[Skovgaard1984] Lene Theil Skovgaard. *A Riemannian Geometry of the Multivariate Normal Model*. Scandinavian Journal of Statistics, vol. 11, no. 4, pages 211–223, 1984.

[Smittarello2019a] D. Smittarello, V. Cayol, V. Pinel, A. Peltier, J-L. Froger et V. Ferrazzini. *Magma Propagation at Piton de la Fournaise From Joint Inversion of InSAR and GNSS*. Journal of Geophysical Research : Solid Earth, vol. 124, no. 2, pages 1361–1387, 2019.

[Smittarello2019b] Delphine Smittarello, Valérie Cayol, Virginie Pinel, Jean-Luc Froger, Aline Peltier et Quentin Dumont. *Combining InSAR and GNSS to Track Magma Transport at Basaltic Volcanoes*. Remote Sensing, vol. 11, no. 19, 2019.

[Solari2018] Lorenzo Solari, Anna Barra, Gerardo Herrera, Silvia Bianchini, Oriol Monserrat, Marta Béjar-Pizarro, Michele Crosetto, Roberto Sarro et Sandro Moretti. *Fast detection of ground motions on vulnerable elements using Sentinel-1 InSAR data*. Geomatics, Natural Hazards and Risk, vol. 9, no. 1, pages 152–174, 2018.

[Srebro2003] Nathan Srebro et Tommi Jaakkola. *Weighted Low-Rank Approximations*. Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), 2003.

[Strozzi2002] T. Strozzi, A. Luckman, T. Murray, U. Wegmuller et C. L. Werner. *Glacier motion estimation using SAR offset-tracking procedures*. IEEE Trans. Geosci. Remote Sens., vol. 40, no. 11, pages 2384–2391, Nov 2002.

[Sun2016] Y. Sun, P. Babu et D. P. Palomar. *Robust Estimation of Structured Covariance Matrix for Heavy-Tailed Elliptical Distributions*. IEEE Transactions on Signal Processing, vol. 64, no. 14, pages 3576–3590, 2016.

[Tarantola1987] A Tarantola. *Inverse Problem Theory Elsevier*. New York, 1987.

[Tarantola2005] Albert Tarantola. Inverse problem theory and methods for model parameter estimation. SIAM, 2005.

[Tatsuoka2000] Kay S. Tatsuoka et David E. Tyler. *On the uniqueness of S-functionals and M-functionals under nonelliptical distributions*. Ann. Statist., vol. 28, no. 4, pages 1219–1243, 08 2000.

[Taylor2013] Marc H. Taylor, Martin Losch, Manfred Wenzel et Jens Schröter. *On the Sensitivity of Field Reconstruction and Prediction Using Empirical Orthogonal Functions Derived from Gappy Data*. Journal of Climate, vol. 26, no. 22, pages 9194–9205, 10 2013.

[Thacker1996] W. C. Thacker. *Metric-based principal components : data uncertainties*. Tellus A, vol. 48, no. 4, pages 584–592, 1996.

[Theiler2005] J. Theiler, B.R. Foy et A.M. Fraser. *Characterizing non-gaussian clutter and detecting weak gaseous plumes in hyperspectral imagery*. In Proc. SPIE, volume 5806, pages 182–193, 2005.

[Thiébaux1984] H. J. Thiébaux et F. W. Zwiers. *The Interpretation and Estimation of Effective Sample Size*. J. Climate Appl. Meteor., vol. 23, no. 5, pages 800–811, 1984.

[Tipping1999] M. E. Tipping et C. M. Bishop. *Probabilistic principal component analysis*. Journal of the Royal Statistical Society. Series B (Statistical Methodology), vol. 61, no. 3, pages 611–622, 1999. Available from http ://www.ncrg.aston.ac.uk/Papers/index.html.

[Trouvé2007] Emmanuel Trouvé, Gabriel Vasile, Michel Gay, Lionel Bombrun, Pierre Grussenmeyer, Tania Landes, Jean-Marie Nicolas, Philippe Bolon, Ivan Petillot, Andreea Julea*et al. Combining airborne photographs and spaceborne SAR data to monitor temperate glaciers : Potentials and limits*. IEEE Transactions on Geoscience and Remote Sensing, vol. 45, no. 4, pages 905–924, 2007.

[Tyler1987] David E. Tyler. *A Distribution-Free M-Estimator of Multivariate Scatter*. The Annals of Statistics, vol. 15, no. 1, pages 234–251, 1987.

[van Buuran2012] S. van Buuran. Flexible imputation of missing data. Chapman and Hall/CRC, New York, 2012.

[Vaughan2013] David G Vaughan, Josefino C Comiso, Ian Allison, Jorge Carrasco, Georg Kaser, Ronald Kwok, Philip Mote, Tavi Murray, Frank Paul, Jiawen Ren*et al. Observations : cryosphere*. Climate change, vol. 2103, pages 317–382, 2013.

[Vautard1992] R. Vautard, P. Youi et M. Ghil. *Singular spectrum analysis : a toolkit for short, noisy chaotic signals*. Physica D, vol. 58, pages 95–126, 1992.

[Verger2013] Aleixandre Verger, Frédéric Baret, Marie Weiss, Sivasathivel Kandasamy et Eric Vermote. *The CACAO Method for Smoothing, Gap Filling and Characterizing Seasonal Anomalies in Satellite Time Series*. IEEE Trans. Geosci. Remote Sens., vol. 51, no. 4, pages 1963–1972, 2013.

[Vernier2011] Flavien Vernier, Renaud Fallourd, Jean Michel Friedt, Yajing Yan, Emmanuel Trouvé, Jean-Marie Nicolas et Luc Moreau. *Fast correlation technique for glacier flow monitoring by digital camera and space-borne SAR images*. EURASIP Journal on Image and Video Processing, vol. 2011, no. 1, page 11, 2011.

[von Buttlar2014] Jannis von Buttlar, Jakob Zscheischler et Miguel D Mahecha. *An extended approach for spatiotemporal gapfilling : Dealing with large and systematic gaps in geoscientific datasets*. Nonlinear Processes in Geophysics, vol. 21, no. 1, pages 203–215, 2014.

[Von Storch2001] Hans Von Storch et Francis W Zwiers. Statistical analysis in climate research. Cambridge university press, 2001.

[Wahba1980] Grace Wahba et James Wendelberger. *Some new mathematical methods for variational objective analysis using splines and cross validation*. Monthly weather review, vol. 108, no. 8, pages 1122–1143, 1980.

[Walczak2001a] B. Walczak et D. L. Massart. *Dealing with missing data : Part I*. Chemom. Intell. Lab. Syst., vol. 58, pages 15–27, 2001.

[Walczak2001b] B. Walczak et D. L. Massart. *Dealing with missing data : Part II*. Chemom. Intell. Lab. Syst., vol. 58, pages 29–42, 2001.

[Wang2015] Di Wang et Andreas Kääb. *Modeling Glacier Elevation Change from DEM Time Series*. Remote Sensing, vol. 7, no. 8, pages 10117–10142, 2015.

[Weare1982] B. C. Weare et J. S. Nasstrom. *Examples of extended empirical orthogonal function analysis.* Mon. Weather Rev., vol. 110, pages 481–485, 1982.

[Weiss2014] Daniel J Weiss, Peter M Atkinson, Samir Bhatt, Bonnie Mappin, Simon I Hay et Peter W Gething. *An effective approach for gap-filling continental scale remotely sensed time-series*. ISPRS Journal of Photogrammetry and Remote Sensing, vol. 98, pages 106–118, 2014.

[Willmott2006] C.J. Willmott et K. Matsuura. *On the use of dimensioned measures of error to evaluate the performance of spatial interpolators*. International Journal of Geographical Information Science, vol. 20, no. 1, pages 89–102, January 2006.

[Wu2013] Meng-Che Wu, Jian Guo Liu et Philippa Jane Mason. *Adaptive local kriging to retrieve slant-range surface motion maps of the Wenchuan earthquake*. Int. J. Remote Sens., vol. 34, no. 21, pages 7589–7606, 2013.

[Wu2018] Wei Wu, Luoqi Ge, Jiancheng Luo, Ruohong Huan et Yingpin Yan. *A Spectral-Temporal Patch-Based Missing Area Reconstruction for Time-Series Images*. Remote Sensing, vol. 10, page 1560, 2018.

[Xu2016] Chang Xu. *Reconstruction of gappy GPS coordinate time series using empirical orthogonal functions*. J. Geophys. Res. Solid Earth, vol. 121, pages 9020–9033, 2016.

[Yan2011] Yajing Yan. *Fusion of displacement measurements from SAR imagery : application to seismo-volcanic modeling*. Thèse de doctorat, Université de Grenoble, 2011.

[Yan2012] Yajing Yan, Marie-Pierre Doin, Pénélope Lopez-Quiroz, Florence Tupin, Bénédicte Fruneau, Virginie Pinel et Emmanuel Trouvé. *Mexico City subsidence measured by InSAR time series : Joint analysis using PS and SBAS approaches*. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 5, no. 4, pages 1312–1326, 2012.

[Yeo2019] Kyongmin Yeo. *Data-driven reconstruction of nonlinear dynamics from sparse observation*. Journal of Computational Physics, vol. 395, pages 671–689, 2019.

[Yu2019] H. Yu, Y. Lan, Z. Yuan, J. Xu et H. Lee. *Phase Unwrapping in InSAR : A Review*. IEEE Geoscience and Remote Sensing Magazine, vol. 7, no. 1, pages 40–58, 2019.

[Zebker1992] Howard A Zebker, John Villasenor*et al. Decorrelation in interferometric radar echoes*. IEEE Transactions on geoscience and remote sensing, vol. 30, no. 5, pages 950–959, 1992.

[Zeng2013a] Chao Zeng, Huanfeng Shen et Liangpei Zhang. *Recovering missing pixels for Landsat ETM+ SLC-off imagery using multi-temporal regression analysis and a regularization method*. Remote Sensing of Environment, vol. 131, pages 182–194, 2013.

[Zeng2013b] Zhaocheng Zeng, Liping Lei, Shanshan Hou, Fei Ru, Xianhua Guan et Bing Zhang. *A Regional Gap-Filling Method Based on Spatiotemporal Variogram Model of CO2 Columns*. IEEE Transactions on Geoscience and Remote Sensing, vol. 52, no. 6, pages 3594–3603, 2013.

[Zhang2018] Q. Zhang, Q. Yuan, C. Zeng, X. Li et Y. Wei. *Missing Data Reconstruction in Remote Sensing Image With a Unified Spatial–Temporal–Spectral Deep Convolutional Neural Network*. IEEE Trans. Geosci. Remote Sens., vol. 56, no. 8, pages 4274–4288, Aug 2018.

[Zhang2019] Kui Zhang, Faming Gong, Zhiyong Li, Shujun Liu et Yuhan Shen. *Recover Glacier Velocity Fields Derived From the SAR Speckle Tracking Technique Using Artificial Neural Network*. IEEE Geoscience and Remote Sensing Letters, vol. 16, no. 8, pages 1250–1253, 2019.

[Zhu2012] Xiaolin Zhu, Desheng Liu et Jin Chen. *A new geostatistical approach for filling gaps in Landsat ETM+ SLC-off images*. Remote sensing of Environment, vol. 124, pages 49–60, 2012.

[Zoubir2018] Abdelhak M Zoubir, Visa Koivunen, Esa Ollila et Michael Muma. Robust statistics for signal processing. Cambridge University Press, 2018.

Table des figures

1.1 Exemples de données manquantes en télédétection. (a) réflectance avec capteur défecteux sur la bande 6 d'Aqua MODIS ; (b) le problème SLC-off sur la plateforme Landsat ETM+ ; (c) concentration en ozone par Aura OMI ; (d) image IKONOS-2 en présence de nuages ; (e) mesure de *Land Surface Temperature* par MODIS en présence de nuages ; (f) NDVI en présence de nuage (MODIS). Image issue de [Shen2015] ©2017 IEEE. 7

1.2 Exemples de formes de données manquantes dans une série temporelle d'images aux dates $t_1, t_2, \ldots, t_N$, où N est le nombre d'images : (a) aléatoire (sans dépendance spatiale ni temporelle) ; (b) corrélée spatialement ; (c) corrélée spatio-temporellement ; (d) corrélée temporellement ; (e) aléatoire et corrélée spatio-temporellement. 8

1.3 Image SAR acquise par le satellite TerraSAR-X en trajectoire ascendante sur le massif du Mont-Blanc [Fallourd2012]. Certains glaciers du massifs sont clairement visibles, comme la Mer de Glace et le glacier d'Argentière côté Chamonix, ou le glacier de Miage côté Courmayeur. Le nord se situe à gauche. 9

1.4 Interférogramme du déplacement cosismique de surface dans la zone d'Oaxaca au Mexique calculé à partir de plusieurs images Sentinel-1 avant et après le séisme du 23 juin 2020 (image ESA). Les franges montrées ici sont dues à la phase du déplacement sismique. 10

1.5 Déplacement DInSAR moyen dans la ligne de visée calculé à partir d'images ENVISAT sur la période 2007-2010 en zone urbaine (Shanghai, Chine) [Pepe2016]. 12

1.6 Interférogrammes déroulés (centimètre dans la ligne de visée) calculés à partir de paires (6 jours) d'images SAR Sentinel-1 A/B sur les glaciers du Gorner (gauche) et de Miage (droite). La représentation est en géométrie radar. 13

1.7 Interférogrammes déroulés non corrigés (centimètres dans la ligne de visée) à trois dates différentes calculés à partir d'image Sentinel-1 A/B sur la région montagneuse de l'ouest et du sud-ouest de Mexico (L. Maubant, ISTerre, communication personnelle). 14

1.8 Estimation du déplacement du glacier d'Argentière par corrélation d'une paire d'images TerraSAR-X entre le 29 septembre et le 10 octobre 2008. a) pics de corrélation ZNCC ; b) magnitude du déplacement ; c) estimation de l'orientation du déplacement. Figure tirée de [Fallourd2011] ©2011 IEEE. 15

1.9 Vitesses de surface (m/an) des glaciers du Mont-Blanc (gauche) et des alpes bernoises (droite) obtenues pour la période 1999-2003 à partir d'images Landsat 7 [Dehecq2015]. 16

1.10 Vitesse de surface (m/an) sur le glacier Fox en Nouvelle-Zélande. Le calcul des vitesses est issu de la chaîne de traitement développée par [Millan2019]. Les contours sont ceux du Randolph Glacier Inventory (RGI) [Consortium2017]. . . . 16

1.11 Aperçu de quelques méthodes récentes classées par approche spatiale, temporelle ou spatio-temporelle, et positionnement de notre étude. 20

1.12 Exemple d'un signal périodique (gauche) et des modes de variabilité qui le composent (droite) : (a) tendance, (b) oscillations et (c) bruit. Ici, la somme des modes de variabilité permet de reconstruire le signal. Spectralement, la tendance correspond à une fréquence basse, tandis que le bruit est (généralement) un événement haute fréquence. 23

1.13 Extrait de la séquence d'images AVHRR (400 pixels) sur la mer Adriatique contenant des nuages utilisées par [Beckers2003]. ©American Meterological Society. Figure utilisée avec permissions. 25

1.14 Interférogramme initial (a), reconstruit (b) et résidus (reconstruit-initial) sur le glacier du Gorner. Figure tirée de [Prébet2019]. ©2019 IEEE. 25

1.15 Illustration du principe de validation croisée sur un champ sans données manquantes. $\hat{X}$ désigne le champ reconstruit et $\hat{\mathbf{x}}_{\text{cv}}$ les points de validation croisée reconstruits. Après reconstruction, l'erreur est calculée entre $\mathbf{x}_{\text{cv}}$ et $\hat{\mathbf{x}}_{\text{cv}}$. 29

2.1 Diagramme simplifié de la méthode EM-EOF. 37

2.2 Exemples de champs de déplacement synthétisés dans cette étude. On notera que le champ g_5 est une composition spatiale de champs d'ordre 1, 3 et 4, avec diverses formes de distances à l'origine (voir tableau 2.1). 43

2.3 Exemples de perturbations : a) bruit blanc gaussien (non simulé dans l'étude) ; b) bruit spatialement corrélé (SCN) ; c) bruit spatio-temporellement corrélé (STCN) ; d) erreur localisée issue du traitement sur champ d'ordre 1 ; e) erreur localisée issue du traitement sur champ d'ordre 1 perturbé par un bruit STCN. 43

2.4 Evolution du degré de corrélation $c(r)$ en fonction de la distance r. Plus l'exposant γ est grand, moins le bruit est corrélé. A l'inverse, plus γ est petit et plus $c(r)$ est invariant à la distance r, ce qui correspond à un bruit plus corrélé. 44

2.5 Types de données manquantes superposées à un champ de déplacement du premier ordre contenant un bruit SCN. A gauche : données manquantes aléatoires ; à droite : données manquantes corrélées. 44

2.6 Reconstruction des champs de déplacement [cm] du premier ordre (a)(b) et second ordre (c)(d). Le SNR varie entre 1.24 et 1.44 et la quantité de données manquantes est de 30%. (a)(c) : données manquantes aléatoires et SCN ; (b)(d) : données manquantes corrélées et STCN. Les résidus sont la différence entre le champ reconstruit et le champ perturbé. 47

2.7 Série temporelle d'un champ du second ordre perturbé par (a) données manquantes corrélées sur 10 dates consécutives et (b) données manquantes aléatoires. Rouge : déplacement vrai ; cercles noirs : déplacement bruité avec données manquantes ; courbe hachée noire : données manquantes ; ligne grise : série temporelle reconstruite. 47

2.8 Reconstruction d'un champ du second ordre perturbé par 30% de données manquantes, un bruit STCN et plusieurs erreurs de déroulement de phase (cercles rouges). 48

2.9 Cartes d'erreur [cm] en fonction du % de données manquantes et du SNR dans le cas d'un champ du premier ordre (a)(b) et du second ordre (c)(d) perturbé par des trous aléatoires (a)(c) et corrélés (b)(d). Tous les déplacement sont également perturbés par un bruit spatio-temporellement corrélé (STCN). 48

2.10 Reconstruction des champs de déplacement [cm] du troisième ordre (a)(b), quatrième ordre (c)(d) et post-sismique (e). Le SNR varie entre 1.24 et 1.61 et la quantité de données manquantes est de 30%. (a)(c)(e) : données manquantes aléatoires et SCN ; (b)(d) : données manquantes corrélées et STCN. Les résidus sont la différence entre le champ reconstruit et le champ perturbé. 49

2.11 Pourcentage de variance expliquée pour trois bruits : bruit blanc gaussien (WGN, couleur orange) et bruit fortement (noir) et faiblement (magenta) corrélé spatialement (respectivement $\gamma = 0.9$ et $\gamma = 0.2$). 50
2.12 Cartes d'erreur [cm] en fonction du % de données manquantes et du SNR dans le cas d'un champ du premier ordre (a)(b) et du second ordre (c)(d) perturbé par des trous aléatoires (a)(c) et corrélés (b)(d). Tous les déplacements sont également perturbés par un bruit STCN. 51
2.13 Cartes d'erreur [cm] en fonction du % de données manquantes et du SNR dans le cas d'un déplacement post-sismique perturbé par des trous aléatoires (a)(c) et corrélés (b)(d). Tous les déplacements sont également perturbés par un bruit STCN. 51
2.14 Série temporelle d'un champ du troisième ordre perturbé par (a) données manquantes corrélées sur 10 date consécutives et (b) données manquantes aléatoires. Rouge : déplacement vrai ; cercles noirs : déplacement bruité avec données manquantes ; courbe hachée noire : données manquantes ; ligne grise : série temporelle reconstruite. 52
2.15 Reconstruction d'un champ de déplacement [cm] multiformes avec 30% de données manquantes et SNR = 1.6. Le champ est composé de plusieurs blocs, de haut en bas : champ linéaire (g_1), deux champs d'ordre 3 (g_3) puis champ d'ordre 4 (g_4) (voir tableau 2.1). Les champs de déplacement sont perturbés par des données manquantes aléatoire et un SCN (a), puis des données manquantes corrélées et un STCN (b). 52
2.16 Reconstruction bloc par bloc du champ de déplacement [cm] multiformes présenté en figure 2.15 (même perturbations). Les chiffres correspondent au nombre optimal de modes par bloc. 53
2.17 Erreur moyenne (cross-RMSE) des méthodes EM-EOF, NNI et krigeage en fonction de la quantité de données manquantes et du SNR (100 simulations). Légende : (a) Données manquantes aléatoires, SNR = 2, SCN ; (b) données manquantes corrélées, SNR = 2, SCN ; (c) SCN, 30% de données manquantes aléatoires ; (d) STCN, 30% de données manquantes aléatoires. 55
2.18 Emplacement géographique des glaciers du Gorner (massif du Mont Rose), Miage et Argentière (massif du Mont-Blanc). 56
2.19 Interférogramme initial (a) et reconstruit (b), et résidus (reconstruit-initial) (c) en géométrie radar sur le glacier du Gorner à trois intervalles de temps (2016/12/23-2016/12/29, 2017/01/10-2017/01/16, 2017/01/16-2017/01/22, format YYYY/MM/JJ). Les séries temporelles aux points P_1, P_2 et P_3 sont présentées en figure 2.20. Les valeurs de déplacement sont en centimètres dans la ligne de visée (line-of-sight, abrégé LOS) du radar. 56
2.20 Série temporelle de mesures de déplacement sur différentes zones P_1, P_2 et P_3 (figure 2.19) situées sur le glacier du Gorner ainsi que leur reconstruction $\hat{P}_1$, $\hat{P}_2$ et $\hat{P}_3$ par la méthode EM-EOF. Les cercles représentent les valeurs existantes alors que les lignes pleines correspondent aux valeurs reconstruites. 57
2.21 Interférogramme initial (a) et reconstruit (b), et résidus (reconstruit-initial) (c) en géométrie radar sur le glacier de Miage à trois intervalles de temps (2016/12/11-2016/12/17, 2016/12/29-2017/01/04 et 2017/02/21-2017/02/27, format YYYY/MM/JJ). Les séries temporelles aux points P_1 et P_2 sont présentées en figure 2.22. Les valeurs de déplacement sont en centimètres dans la ligne de visée (LOS) du radar. . 60
2.22 Séries temporelles de mesures de déplacement sur différentes zones P_1 et P_2 (figure 2.21) situées sur le glacier de Miage, ainsi que leur reconstruction $\hat{P}_1$ et $\hat{P}_2$ par la méthode EM-EOF. Les cercles représentent les valeurs existantes alors que les lignes pleines correspondent aux valeurs reconstruites. 60

2.23 Taille de la fenêtre de corrélation utilisée selon la période d'acquisition des images entre octobre 2016 et décembre 2017 (J : janvier, F : février, etc.). 61
2.24 (a) Cross-RMSE $E(k)$ versus nombre de modes EOF k utilisés pour la reconstruction du jeu de données du glacier d'Argentière. Le minimum de $E(k)$ est atteint pour $k = 58$ (étape 1). Après l'étape 2, le nombre optimal de modes est réduit à 20 modes, ce qui correspond à un minimum local de $E(k)$. (b) Pourcentage de variance expliquée par chaque mode (jusqu'au numéro 32). L'énergie du système est bien répartie sur un grand nombre de modes, rendant ainsi difficile la tâche de sélection du nombre optimal de modes. 62
2.25 Champ de déplacement (corrélation d'amplitude) initial (a), reconstruit (b) et résidus (reconstruit-initial) (c) en géométrie radar sur le glacier d'Argentière à trois intervalles de temps (2017/08/26-2017/09/07, 2017/09/19-2017/10/01 et 2017/10/25-2017/11/06, format YYYY/MM/JJ). Les valeurs de déplacement sont en mètres dans la direction azimuthale. 63
2.26 Interférogramme reconstruit [cm] (2017/01/16-2017/01/22) sur la partie haute du glacier du Gorner par les méthodes NNI (a), krigeage (b) et EM-EOF (c). 64

3.1 Diagramme simplifié de la méthode EM-EOF étendue. 67
3.2 Illustration de l'augmentation spatiale des champs $(\mathbf{X}_t)_{1\leq t\leq N}$ à l'aide d'une fenêtre glissante de taille $M_x \times M_y$. Le champ $\mathbf{X}_1$ est ici augmenté en une matrice $\mathbf{D}_1$ de taille $K_xK_y \times M_xM_y$, laquelle est stockée dans une grande matrice spatio-temporelle $\mathcal{D}$. Chaque $\mathbf{D}_t$ correspondant à $\mathbf{X}_t$ est ensuite ordonné en ligne, ce qui résulte en une matrice de taille $(K \times NM)$. 68
3.3 (a) Spectre de valeurs propres λ_k et (b) mesure de confiance associée $\mathcal{C}_k$. La ligne verte correspond à l'estimation du nombre de modes après l'étape 2. La ligne rouge correspond à l'ajustement du nombre de modes par le calcul de $\mathcal{C}_k$. Les barres verticales sont les intervalles d'incertitude de chaque valeur propre issus de la règle empirique (3.15). 74
3.4 Variation de la cross-RMSE en fonction du nombre de modes pour différent décalages spatiaux sur des données synthétiques incomplètes (N=10, P=144). Le minimum de la cross-RMSE évolue avec la variation du décalage spatial : plus ce dernier est grand, plus le nombre de modes sélectionnés sera grand. 75
3.5 Champs de déplacement considérés dans cette étude. Une description des modèles mathématiques des champs g_0 à g_3 est fournie en tableau 3.1 76
3.6 Valeurs propres d'un bruit SCN pour différentes corrélations. 78
3.7 Spectres de valeurs propres λ_k d'un champ de déplacement synthétique augmenté et mesure $\mathcal{C}_k$ pour différentes quantités de données manquantes (gauche) et différents SNR (droite). Lignes verticales et nombres en couleur : estimation du nombre optimal de modes $\hat{r}$. 79
3.8 (a) Valeurs propres λ_k de la matrice de données augmentées $\mathcal{D}$ (50 premières) du champ g_0 perturbé par des données manquantes aléatoires et un bruit SCN ; (b) mesure de confiance associée $\mathcal{C}_k$ et estimation du nombre de modes à l'issue de l'étape 2 (ligne verte) puis après ajustement (ligne rouge) ; (c) $\mathcal{C}_k$ versus λ_k. Les cercles rouges correspondent au nombre de modes sélectionnés. 79
3.9 Reconstruction [cm] d'un champ d'ordre 1 perturbé par des données manquantes aléatoires et un bruit SCN ((a), (b)) puis des données manquantes corrélées et un bruit STCN ((c), (d)), par les méthodes EM-EOF étendue ((a), (c)) et EM-EOF ((b), (d)). La quantité de données manquantes est fixée à 30% et le SNR à 1.8 ((a)(b)) et 2 ((c)(d)). Le résidu est la différence entre le champ reconstruit et le champ perturbé. 80

3.10 Série temporelle d'un champ d'ordre 3 perturbé par (a) données manquantes aléatoires et bruit SCN ; (b) données manquantes corrélées et bruit STCN. Rouge : déplacement vrai ; cercles noirs : déplacement perturbé observé ; courbe en pointillés noire : déplacement perturbé non observé (données manquantes) ; ligne grise : reconstruction par la méthode EM-EOF ; courbe en pointillés bleue : reconstruction par la méthode EM-EOF étendue. 81

3.11 (a) Valeurs propres λ_k de la matrice de données augmentées $\mathcal{D}$ (100 premières) du champ g_1 perturbé par des données manquantes aléatoires et un bruit SCN ; (b) mesure de confiance associée $\mathcal{C}_k$ et estimation du nombre de modes (ligne rouge) ; (c) $\mathcal{C}_k$ versus λ_k. Les cercles rouges correspondent au nombre de modes sélectionnés. 81

3.12 Reconstruction [cm] d'un champ d'ordre 3 perturbé par des données manquantes de type aléatoire et un bruit SCN, par les méthodes EM-EOF étendue (a) et EM-EOF (b). La quantité de données manquantes est fixée à 30% et le SNR à 2. Le résidu est la différence entre le champ reconstruit et le champ perturbé. 82

3.13 Reconstruction par la méthode EM-EOF étendue d'un champ d'ordre 3 par ajout de modes successifs (jusqu'à 15). 82

3.14 Série temporelle d'un champ d'ordre 3 perturbé par (a) données manquantes aléatoires et bruit SCN ; (b) données manquantes corrélées et bruit STCN. Rouge : déplacement vrai ; cercles noirs : déplacement perturbé observé ; courbe en pointillés noire : déplacement perturbé non observé (données manquantes) ; ligne grise : reconstruction par la méthode EM-EOF ; courbe en pointillés bleue : reconstruction par la méthode EM-EOF étendue. 83

3.15 Reconstruction d'un champ de déplacement synthétique d'ordre 3 [cm] perturbé par des données manquantes corrélées et un bruit STCN, pour les méthodes EM-EOF étendue (a) et EM-EOF (b). La quantité de données manquantes est fixée à 30% et le SNR à 1.8. 83

3.16 (a) Valeurs propres λ_k de la matrice de données augmentées $\mathcal{D}$ (100 premières) du champ g_2 perturbé par des données manquantes aléatoires et un bruit SCN ; (b) mesure de confiance associée $\mathcal{C}_k$ et estimation du nombre de modes à l'issue de l'étape 2 (ligne verte) puis après ajustement (ligne rouge) ; (c) $\mathcal{C}_k$ versus λ_k. Les cercles rouges correspondent au nombre de modes sélectionnés. 84

3.17 Reconstruction par la méthode EM-EOF étendue d'un champ d'ordre n par ajout modes successifs (jusqu'à 22). 84

3.18 Reconstruction d'un champ d'ordre n perturbé par des données manquantes aléatoires et un bruit spatialement corrélé, pour les méthodes EM-EOF étendue (haut) et EM-EOF (bas). La quantité de données manquantes est fixée à 50% et le SNR à 1.8. 85

3.19 Série temporelle d'un champ de déplacement synthétique d'ordre n perturbé par (a) données manquantes aléatoires et bruit SCN ; (b) données manquantes corrélées et bruit STCN. Rouge : déplacement vrai ; cercles noirs : déplacement bruité avec données manquantes ; courbe noire en pointillés : données manquantes ; ligne grise : série temporelle reconstruite par la méthode EM-EOF ; courbe bleue en pointillés : reconstruction par la méthode EM-EOF étendue. 85

3.20 Reconstruction d'un champ de déplacement synthétique [cm] d'ordre n perturbé par des données manquantes corrélées et un bruit STCN, pour les méthodes EM-EOF étendue (a) et EM-EOF (b). La quantité de données manquantes est fixée à 50% et le SNR à 1.5. 86

3.21 (a) Valeurs propres λ_k de la matrice de données augmentées $\mathcal{D}$ (150 premières) du champ g_3 perturbé par des données manquantes corrélées et un bruit STCN ; (b) mesure de confiance associée $\mathcal{C}_k$ et estimation du nombre de modes à l'issue de l'étape 2 (ligne verte) et après ajustement (ligne rouge) ; (c) $\mathcal{C}_k$ versus λ_k. Les cercles rouges correspondent au nombre de modes sélectionnés. 87
3.22 Reconstruction d'un champ de déplacement synthétique [cm] d'ordre n perturbé par des données manquantes aléatoires et un bruit SCN ((a), (b)), puis par des données manquantes corrélées et un bruit STCN ((c), (d)), pour les méthodes EM-EOF étendue ((a), (c)) et EM-EOF ((b), (d)). La quantité de données manquantes est fixée à 30% et le SNR à 1.8. 88
3.23 Série temporelle d'un champ de déplacement synthétique d'ordre n perturbé par (a) données manquantes aléatoires et bruit SCN ; (b) données manquantes corrélées et bruit STCN. Rouge : déplacement vrai ; cercles noirs : déplacement bruité avec données manquantes ; courbe noire en pointillés : données manquantes ; ligne grise : série temporelle reconstruite par la méthode EM-EOF ; courbe bleue en pointillés : reconstruction par la méthode EM-EOF étendue. 89
3.24 RMSE en fonction de la quantité de données manquantes (%). Trait plein : données manquantes aléatoires et bruit SCN ; trait en pointillés : données manquantes corrélées et bruit STCN. SNR = 2. 90
3.25 RMSE en fonction du SNR. Trait plein : bruit STCN et données manquantes corrélées ; trait en pointillés : bruit SCN et données manquantes aléatoires. La quantité de données manquantes est fixée à 30%. 91
3.26 Évolution de la RMSE dans les séries temporelles des champs g_0 (première ligne) à g_3 (dernière ligne). Colonne de gauche : données manquantes aléatoires ; colonne de droite : données manquantes corrélées. La quantité de données manquantes est fixée à 30% et le SNR à 2. 92
3.27 Emplacement géographique du glacier Fox (contours vert et point noir en gradient) dans les Alpes du sud en Nouvelle-Zélande. Les contours sont ceux du Randolph Glacier Inventory (RGI) [Consortium2017]. Figure tirée de [Wang2015]. 94
3.28 Série temporelle de champs de vitesse de surface (mètres/an) obtenus par corrélation d'amplitude d'images optiques Sentinel-2 sur le glacier Fox entre février et mi-septembre 2018. L'emplacement des zones étudiées P1, P2 et P3 est également illustré à la date 2018-02-13. Les contours proviennent du Randolph Glacier Inventory (RGI) [Consortium2017]. 95
3.29 (a) Valeurs propres λ_k (60 premières) du jeu de données augmenté sur le glacier Fox, (b) mesure de confiance associée $\mathcal{C}_k$ et (c) $\mathcal{C}_k$ versus λ_k. Les cercles et ligne rouges représentent les valeurs propres correspondant aux modes sélectionnés. . . 96
3.30 Reconstruction de la série temporelle de champs de vitesse de surface (mètres/an) sur le glacier Fox entre février et mi-septembre 2018. Les contours proviennent du RGI. 97
3.31 Évolution des vitesses de surface entre février et mi-septembre 2018 sur les zones P1 (cercles bleus), P2 (cercles rouges), P3 (cercles noirs) et séries temporelles reconstruites en P2 et P3. Les intervalles d'erreur sont ceux de l'étude de [Millan2019]. 98

4.1 Forme de données manquantes : (a) bloc, (b) monotone et (c) générale. En blanc : valeurs observées ; en gris : valeurs manquantes. 106
4.2 Distance naturelle en fonction du nombre d'observations dans deux configurations (obs : pourcentage de données observées ; d.m. : pourcentage de données manquantes). 113
4.3 Données manquantes ordonnées en bloc. Gris : données manquantes. Blanc : données observées. 114

4.4 Convergence de l'EM pour différentes tailles de bloc de données manquantes $(P \times N)_{\text{manquant}}$ sur un jeu de données de taille (10×100). 119
4.5 Distance naturelle sur la matrice de covariance en fonction du nombre d'observations (N) selon différents α et différentes tailles de bloc de données manquantes. 120
4.6 Distance naturelle sur les paramètres de texture en fonction du nombre d'observations (N) selon différents α et différentes tailles de bloc de données manquantes. 121
4.7 Distance naturelle sur la matrice de covariance sans et avec structure rang faible, en fonction du nombre d'observations (N) selon différents α et un bloc de données manquantes de taille $(P \times N)_{\text{manquant}} = (3 \times 20)$. 122
4.8 Erreur (moyenne sur 200 simulations) en fonction du nombre d'observations N pour trois configurations (10%, 30%, 50%) de données manquantes. 123
4.9 Réseau GNSS permanent de l'OVPF et état de fonctionnement des stations en décembre 2019 (communication personnelle avec Virginie Pinel de l'ISTerre). ©WEBOBS / IPGP. Modèle numérique de terrain : SRTM/NASA. 125
4.10 Observations GNSS au cours du temps et pourcentage de données manquantes par station. Noir : observé ; blanc : manquant. 126
4.11 Cross-RMSE (m) en fonction du nombre de mode à l'issue de l'étape 1 de la méthode EM-EOF (moyenne sur 100 simulations) et spectre de valeurs propres (v.p.). Le minimum moyen de l'erreur se situe à l'indice 3. 127
4.12 Exemples de séries temporelles de mesure de déplacement (m) reconstruites sur huit stations GNSS et zoom sur les stations GBNG (54.9% de données manquantes) et GBSG (27.8% de données manquantes). 128
4.13 Erreurs (cross-RMSE et RMSE) par station GNSS de l'algorithme EM et de la méthode EM-EOF. 129
4.14 Différence des erreurs de reconstruction entre l'algorithme EM et la méthode EM-EOF. 129
4.15 Histogramme de l'ensemble des données de déplacement GNSS et modèle de distribution gaussienne centrée (rouge). 130
4.16 Histogrammes sur six stations GNSS et modèle de distribution gaussienne centrée à variance fixe (rouge). 131
4.17 Schéma simplifié d'une fenêtre carrée (a) versus fenêtre adaptative (b) pour l'augmentation spatiale des données (voir figure 3.2). 136

B.1 Déplacement vertical mesuré par 22 stations GNSS de l'OVPF entre 2014 et 2017. La reconstruction des données (section 4.6.2) est effectuée sur les stations non concernées par les événements éruptifs sur cette période, c'est-à-dire les courbes qui ne présentent pas de motifs en "escalier". XXXII

Liste des tableaux

2 Principales caractéristiques des jeux de données étudiés dans cette thèse. 4

2.1 Récapitulatif des champs déterministes synthétisés. g_1 est linéaire en temps et en espace alors que les chamsp g_2 à g_5 sont non-linéaires et incluent diverses oscillations à fréquences variables. g_6 est un modèle de déformation post-sismique avec un temps de décroissance $\tau_e = 1.5$. Les constantes A et b sont fixées à 0 et 1 respectivement. $r_1 = \sqrt{x^2 + y^2}$, $r_2 = \sqrt{(x-1)^2 + (y-1)^2}$ et $r_3 = \exp(-(x+y)^2) + xy + \tan(x)$ sont les distances à l'origine. Les coordonnées (x, y) varient dans l'intervalle compacte $[-1, 1]^2$ et t est la variable temps. Les valeurs des fréquences sont fixées à : $f_1 = 0.25, f_2 = 0.75, f_3 = 2.5, f_4 = 1.25, f_5 = 5$. 42

2.2 Paramètres des expériences des cas 1, 2 et 3 (voir section 2.3.5 ci-après). Tous les champs de déplacement contiennent 30% de données manquantes. 45

2.3 Temps de calculs moyens de l'algorithme EM-EOF pour une série temporelle de 40 images synthétiques avec 30% de données manquantes. 54

2.4 Moyenne (cm) et écart-type σ (cm) des champs de résidus des points observés et des points de validation croisée (abrégés CV) sur les glaciers du Gorner et de Miage. Le symbol 'x' indique les interférogrammes manquants. Les numéros de date s'étendent entre novembre 2016 et mars 2017. 58

3.1 Modèles des champs déterministes synthétisés. t est la variable temps, les coordonnées (x, y) discrétisent le compact $[-1, 1]^2$, et $r_1 = \sqrt{(x+0.1)^2 + (y+0.3)^2}$ et $r_2 = \exp(-(x+y)^2) + xy + \tan(x)$ sont les distances à l'origine. $w_1, \ldots, w_8 = 2\pi f_1, \ldots, 2\pi f_8$ sont les vitesses angulaires du signal dont les fréquences sont fixées à $\{f_1, \ldots, f_8\} = \{0.25, 0.75, 2.5, 1.25, 5, 7.5, 1.75, 0.5\}$. 76

3.2 Principaux paramètres de simulations des quatre cas d'étude (g_0, g_1, g_2, g_3). . . . 77

3.3 Comparaison des méthodes EM-EOF étendue et EM-EOF selon la grille suivante : $+++$ très adaptée ; $++$ adaptée ; $+$ moyennement adaptée ; $-$ peu adaptée. Abbréviations - $\nearrow$: Grand ; $\searrow$: Petit ; Aléa. : Aléatoire ; Corr. : Corrélé. 93

3.4 Principales caractéristiques du jeu de données de vitesse de surface sur le glacier Fox. 94

Liste des publications

Article dans une revue internationale avec comité de lecture

— Hippert-Ferrer A., Yan Y., Bolon P., Millan R. *Spatio-temporal filling of missing data in remotely sensed displacement measurement time series.* IEEE Geoscience and Remote Sensing Letters, *Accepté en juillet 2020*.

— Hippert-Ferrer A., Yan Y., Bolon P. *EM-EOF : gap-filling in incomplete SAR displacement time series.* IEEE Transactions on Geoscience and Remote Sensing, *Accepté en juillet 2020*.

Communication dans un congrès national ou international avec acte et comité de lecture

— Hippert-Ferrer A., Yan Y., Bolon P. *Gap-filling based on EOF analysis of spatio-temporal covariance of satellite image derived displacement time series.* IGARSS, septembre 2020, symposium virtuel, (oral).

— Hippert-Ferrer A., Yan Y., Bolon P. *Reconstruction de données manquantes par analyse en EOF : application aux séries temporelles de mesures de déplacement InSAR*. GRETSI, août 2019, Lille, France (poster).

— Hippert-Ferrer A., Yan Y., Bolon P. *Gap-filling based on iterative EOF analysis of temporal covariance : application to InSAR displacement time series*. IGARSS, juillet 2019, Yokohama, Japon (oral).

Communication dans un congrès international sans acte

— Hippert-Ferrer A., Yan Y., Bolon P. *Spatio-temporal missing data reconstruction in satellite displacement measurement time series*. EGU General Assembly, mai 2020, Vienne, Autriche (oral).

— Hippert-Ferrer A., Yan Y., Bolon P. *Gap-filling of InSAR displacement time series*. MDIS, octobre 2019, Strasbourg, France (oral).

— Hippert-Ferrer A., Yan Y., Bolon P. *A gap-filing method to reconstruct incomplete SAR displacement measurement time series.* Living Planet Symposium, mai 2019, Milan, Italie (poster).

— Hippert-Ferrer A., Yan Y., Bolon P. *Gap filling based on EOF analysis of temporal covariance of offset tracking displacement measurement time series*. 19th General Assembly of Wegener, septembre 2018, Grenoble, France (oral).

A

Génération d'un bruit corrélé

A.1 Génération d'un bruit spatio-temporel

Le bruit spatio-temporellement corrélé est modélisé comme étant la somme d'un bruit spatialement corrélé et d'un bruit temporellement corrélé. Le premier est obtenu comme décrit en sous-section 2.3.2. Le second est obtenu par une décomposition de Cholesky d'une matrice de covariance positive semi-définie $\mathbf{R}$ vérifiant :

$$\mathbb{E}[\mathcal{Z}\mathcal{Z}^T] = \mathbf{R} \tag{A.1}$$

où $\mathcal{Z} \in \mathbb{R}^{N \times P}$ est le bruit corrélé désiré et $\mathbb{E}[.]$ est l'opérateur espérance. Les éléments de la matrice $\mathbf{R}$ à la position (i, j) sont donnés par :

$$(r_{ij})_{1 \leq i \leq n, 1 \leq j \leq n} = \rho^{|i-j|} \tag{A.2}$$

où le paramètre ρ varie dans l'intervalle $[0, 1]$ et permet de régler la corrélation temporelle : une valeur de 0 correspond à un bruit non corrélé alors qu'une valeur proche de 1 correspond à un bruit totalement corrélé. Comme annoncé ci-dessus, la matrice $\mathbf{R}$ est positive semi-définie, ce qui permet d'y appliquer la décomposition suivante decomposition, appelée décomposition de Cholesky :

$$\mathbf{R} = \mathbf{L}\mathbf{L}^T \tag{A.3}$$

où la matrice $\mathbf{L}$ et sa transposée $\mathbf{L}^T$ sont respectivement des matrices triangulaires inférieure et supérieure. Une matrice aléatoire $\mathbf{Y}$ (dont les colonnes sont indépendantes) de taille $N \times P$ suivant une distribution gaussienne est ensuite générée afin de résoudre l'équation :

$$\mathcal{Z} = \mathbf{L}\mathbf{Y} \tag{A.4}$$

$\mathcal{Z}$ est une matrice $N \times P$ où le degré de corrélation entre chaque ligne est directement paramétrable par ρ. Pour montrer que $\mathcal{Z}$ a bien la covariance temporelle $\mathbf{R}$ désirée (A.1) :

$$\mathbb{E}[\mathcal{Z}\mathcal{Z}^T] = \mathbf{L}\mathbb{E}[\mathbf{Y}\mathbf{Y}^T]\mathbf{L}^T = \mathbf{R} \tag{A.5}$$

B

Opérateur Sweep et données GNSS

B.1 L'opérateur sweep

L'opérateur sweep [1] fournit un moyen simple et pratique de faire les calculs de maximum de vraisemblance pour des données incomplètes.

Définition B.1.1. *Opérateur sweep*

Soit $\mathbf{G}$ *une matrice symétrique de taille* $P \times P$*. On dira que* $\mathbf{H} = SWP[k]\mathbf{G}$ *est la matrice résultant de l'opérateur sweep sur* $\mathbf{G}$ *de sorte que les éléments de* $\mathbf{H}$ *soient définis par :*

$$\begin{aligned} h_{kk} &= -1/g_{kk} \\ h_{jk} &= h_{kj} = g_{jk}/g_{kk} \qquad j \neq k \\ h_{jl} &= g_{jl} - g_{jk}g_{kl}/g_{kk} \qquad j \neq k, l \neq k \end{aligned} \tag{B.1}$$

Supposons que $\mathbf{Y} = \{\mathbf{y}_i\}$ désigne un échantillon de N observations sur P variables (4.13), dont chaque élément est désigné par y_{ij}. Soit $\mathbf{G}$ la matrice de taille $(P+1) \times (P+1)$ suivante :

$$\mathbf{G} = \begin{bmatrix} 1 & \mu_1 & \dots & \mu_P \\ \mu_1 & N^{-1}\sum_i y_{i1}^2 & \dots & N^{-1}\sum_i y_{iP}y_{i1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_P & N^{-1}\sum_i y_{i1}y_{iP} & \dots & N^{-1}\sum_i y_{iP}^2 \end{bmatrix} \tag{B.2}$$

où $\mu_1, \dots, \mu_P$ sont les moyennes empiriques et les autres éléments sont les sommes des produits croisés normalisées par N. L'application de l'opérateur sweep sur la ligne et la colonne à l'indice 0 de $\mathbf{G}$ résulte en une nouvelle matrice définie par :

1. Le lecteur intéressé pourra trouver une description détaillée de l'opérateur sweep, ainsi que de nombreux exemples, dans le livre de [Little2002].

$$\mathbf{H} = \text{SWP}[0]\mathbf{G} = \begin{bmatrix} -1 & \mu_1 & \dots & \mu_P \\ \mu_1 & \sigma_{11} & \dots & \sigma_{P1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_P & \sigma_{1P} & \dots & \sigma_{PP} \end{bmatrix} \tag{B.3}$$

où $(\sigma_{jk})_{j,k=1,\dots,P}$ est la covariance empirique. On remarque donc que l'opération SWP$[0]$ effectue une correction du produit croisé en soustrayant le produit de leur moyenne respective pour obtenir la covariance empirique, soit :

$$\sigma_{jk} = N^{-1} \sum_i y_{ij} y_{ik} - \mu_j \mu_k \tag{B.4}$$

De plus, on note SWP$[k_1, k_2, \dots, k_i]\mathbf{H}$ l'application des opérations successives SWP$[k_1]$, SWP$[k_2], \dots,$ SWP$[k_i]$ sur la matrice $\mathbf{H}$. Cette opération permet d'obtenir les estimés du maximum de vraisemblance correspondant à la régression linéaire multivariée des variables $\{y_{ij}\}_{j \notin \{k_1,\dots,k_i\}}$ sur les variables de prédiction $\{y_{ij}\}_{j \in \{k_1,\dots,k_i\}}$. L'opérateur sweep sert donc à rendre des variables de description (descripteurs) de la distribution normale multivariée en variables de prédiction (prédicteurs).

Si $\mathbf{Y}_{\text{obs}} = \{y_{ij}\}_{j \in \{k_1,\dots,k_i\}}$ et $\mathbf{Y}_{\text{mis}} = \{y_{ij}\}_{j \notin \{k_1,\dots,k_i\}}$, l'opérateur sweep est particulièrement utile puisque les variables de prédiction sont à présent des variables observées. On peut alors déduire les paramètres de la distribution des données manquantes à partir des données observées et des paramètres estimés à l'itération courante.

B.2 Données GNSS

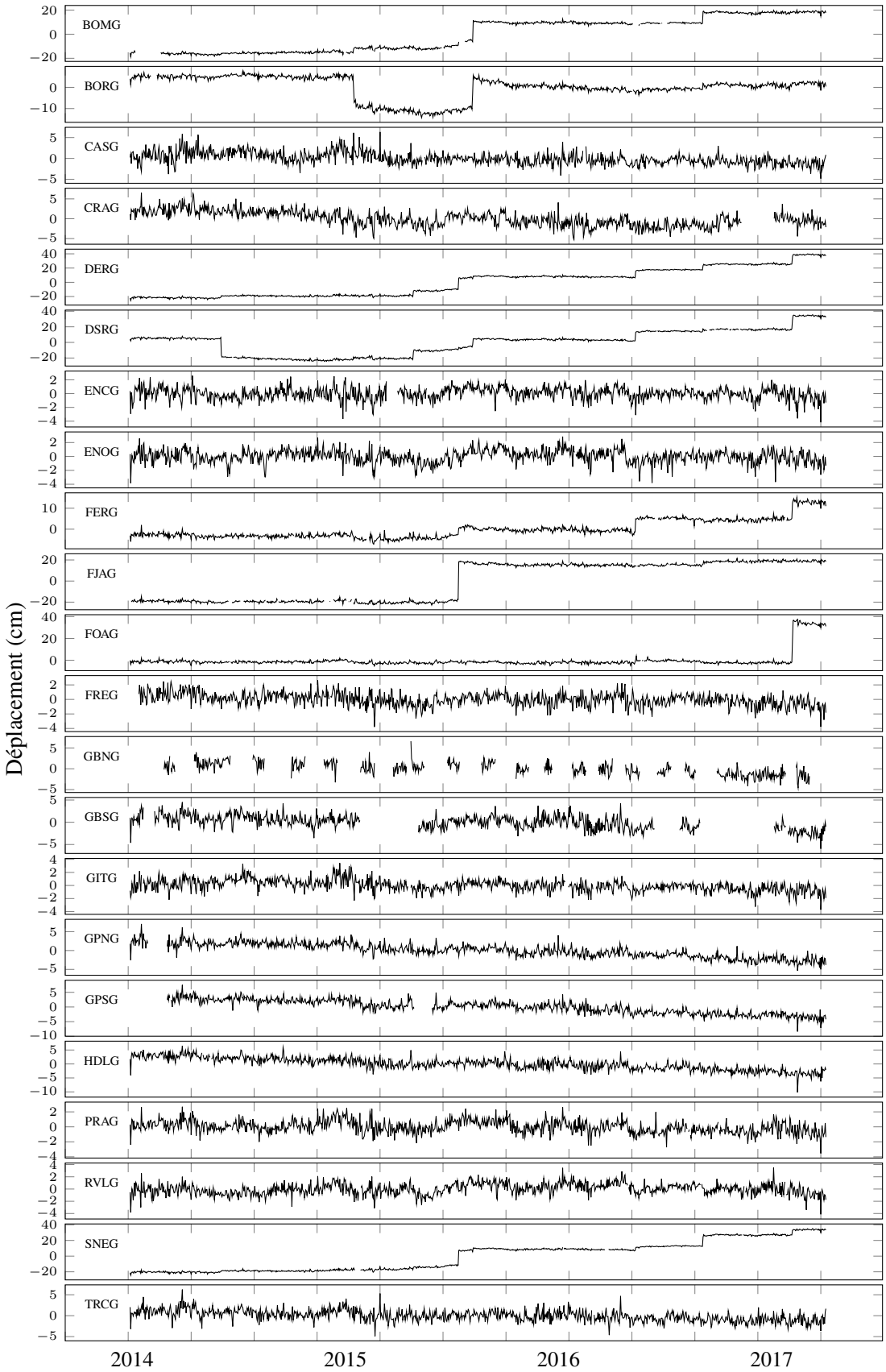


Figure B.1 – Déplacement vertical mesuré par 22 stations GNSS de l'OVPF entre 2014 et 2017. La reconstruction des données (section 4.6.2) est effectuée sur les stations non concernées par les événements éruptifs sur cette période, c'est-à-dire les courbes qui ne présentent pas de motifs en "escalier".

Reconstruction de données manquantes dans des séries temporelles de mesures de déplacement par télédétection.

Résumé

Malgré la masse de données (satellitaires et in-situ) disponibles en mesure de déplacement, l'incomplétude de données reste toujours un problème fréquemment rencontré. Ce phénomène est principalement dû au changement des propriétés de surface de l'objet observé et/ou aux limites techniques des méthodes de calcul de déplacement terrestre (e.g. interférométrie différentielle, corrélation croisée). Rendant les données discontinues en espace et en temps, l'incomplétude de données constitue un écueil vers la compréhension complète des phénomènes physiques sous-jacents liés au déplacement de surface. Malgré ce constat, l'analyse de données manquantes ne bénéficie pas d'une attention sérieuse et dédiée à la mesure de déplacement. Des méthodes de reconstruction adaptées aux données sont ainsi nécessaires pour gérer la présence de données manquantes spatio-temporelles au sein de séries temporelles de mesure de déplacement. Dans cette thèse, nous proposons trois approches pour l'analyse et la reconstruction de données manquantes en mesure de déplacement par télédétection. Les deux premières approches sont basées sur la décomposition de la covariance temporelle et spatio-temporelle du signal de déplacement en fonctions empiriques orthogonales (EOFs). Ces études ont débouchées sur le développement de deux méthodes, appelées EM-EOF et *extended* EM-EOF, nécessitant d'initialiser les valeurs manquantes avant traitement. La troisième étude, plus prospective, est orientée vers l'estimation robuste de la matrice de covariance du signal de déplacement, sans initialisation préalable des valeurs manquantes. Ces trois approches ont en commun de s'appuyer sur un schéma de résolution itératif de type espérance-maximisation (EM) ainsi que sur la sélection d'un nombre réduit de modes décrivant le maximum de variabilité du signal de déplacement. L'ensemble des cas d'études sur données réelles et synthétiques fournissent des résultats prometteurs, renforçant l'intérêt que porte l'étude des données manquantes en mesure de déplacement par télédétection.

Mots-clés : Données manquantes, mesure de déplacement, EOF, covariance, algorithme EM, série temporelle.

Abstract

Despite the large volume of available data (satellite and in-situ) in displacement measurement, data incompleteness is still a commonly encountered issue. This phenomenon is mainly due to surface property changes of the observed object and/or to technical limitations of the displacement extraction methods (e.g. differential interferometry, offset tracking). By generating time and space discontinuity, data incompleteness can hinder a thorough understanding of underlying physical phenomena that induce surface displacement. However, missing data analysis in displacement measurement has not been paid significant and dedicated attention. In this context, advanced reconstruction methods, adapted to the data specificities, are necessary for handling spatio-temporal gaps in displacement measurement time series. In this thesis, we propose three approaches for analysing and imputing missing data in remotely sensed displacement measurement time series. The two first approaches are based on the decomposition of the temporal and spatio-temporal covariance of the displacement signal into Empirical Orthogonal Functions (EOFs). These studies have led to the development of two methods, so-called EM-EOF and extended EM-EOF, both requiring an initialization of the missing values. The third approach intends to explore techniques in robust estimation of the covariance matrix without initialization of the missing values. All approaches rely on an Expectation-Maximization (EM)-type iterative resolution scheme and reckon with the covariance low rank structure, which describe most of the variability of the displacement signal. Both synthetic simulations and real data applications present promising results, bringing to light the interest of the proposed approaches for missing data imputation in remotely sensed displacement measurement time series.

Keywords : Missing data, displacement measurement, EOF, covariance, EM algorithm, time series.