

Note technique

Travaux financés par le ministère chargé de l'environnement

NOTE DE SYNTHÈSE SUR LES DÉVELOPPEMENTS RÉCENTS EN MATIÈRE DE CARTES ANALYSÉES DES RÉSULTATS DE MODÉLISATION

Cartographie des concentrations de PM_{10} et de $PM_{2.5}$

Maxime Beauchamp, Laure Malherbe, Frédéric Meleux
Laurent Létinois

SYNTHÈSE

La présente note synthétise les travaux récents du LCSQA en matière de modélisation géostatistique. Elle aborde des points de méthodologie et fournit des scores de performance des méthodes développées.

Un modèle de cokrigage pour la cartographie des PM_{10} et des $PM_{2.5}$ a été ainsi mis au point. **Il permet à la fois d'améliorer la précision des cartographies de $PM_{2.5}$ et la cohérence entre les estimations de $PM_{2.5}$ et de PM_{10} .** Un lissage a été également introduit dans le krigage afin d'améliorer la continuité des cartes dans les zones peu contraintes par les stations de mesure. Ces travaux sont destinés à être implantés de manière opérationnelle dans le système PREV'AIR pour améliorer la qualité de la production quotidienne de cartographies de la qualité de l'air pour le jour d'avant. Ils participent donc à la démarche d'assurance qualité du système national de prévision et de cartographie de la qualité de l'air, pour une plus grande précision et une plus grande fiabilité des informations diffusées auprès du Ministère, des AASQA et du grand public.

Ces travaux ont été conduits à l'initiative du LCSQA. Certains d'entre eux ont fait l'objet d'échanges techniques avec le Centre de Géosciences de l'Ecole des Mines de Paris, dans le cadre d'une convention de collaboration scientifique avec l'INERIS.

1. CONTEXTE

Le LCSQA réalise des études en statistique et géostatistique qui contribuent à l'élaboration d'un cadre méthodologique harmonisé pour la cartographie et la prévision de la qualité de l'air, ce qui passe aussi par l'amélioration des performances de PREV'AIR. Ces études s'inscrivent dans les programmes annuels du LCSQA relatifs au traitement des données de qualité de l'air. Si leur objectif est de produire des méthodologies opérationnelles, elles n'en nécessitent pas moins des développements théoriques, comme c'est le cas des traitements géostatistiques exposés ci-après.

2. INTRODUCTION

Cette note synthétise les travaux réalisés en 2015 par le LCSQA sur la construction de cartes analysées à partir de résultats de modélisation régionale et de données d'observation.

Ces travaux s'inscrivent à la fois dans le champ d'activités de la Commission de Suivi *Emissions, Modélisation et Traitement de Données* et du Comité des Utilisateurs de PREV'AIR (hors LCSQA).

Ils ont porté sur :

- l'élaboration d'un nouveau modèle de cokrigage pour la construction d'une carte analysée des concentrations de PM_{10} et de $PM_{2.5}$ (chapitre 3). Il s'agissait à la fois d'améliorer la précision des cartographies de particules et de contraindre les situations aberrantes où l'estimation $PM_{2.5}$ est supérieure à l'estimation PM_{10} à l'issue des traitements mathématiques de krigage
- le lissage de l'estimateur, lorsque celui-ci est produit à partir d'un voisinage glissant, afin de prévenir l'apparition de discontinuités sur les cartes dans les régions peu contraintes par les observations (chapitre 4). Le lissage est obtenu en introduisant l'hypothèse qu'une observation est perturbée par un bruit aléatoire et que ce bruit aléatoire a une variance d'autant plus forte que le site de mesure est plus éloigné du nœud de grille.

Ils ont nécessité d'importants développements, tant du point de vue méthodologique qu'au niveau de la programmation informatique. La qualité des résultats issus de ces développements a été évaluée au moyen des scores usuels de performance.

3. APPORTS DU COKRIGEAGE DES DONNÉES DE PM_{10} ET $PM_{2.5}$ DANS L'ESTIMATION DES CONCENTRATIONS DE PARTICULES

3.1 Contexte

Les mesures de $PM_{2.5}$ (particules de diamètre inférieur à $2.5 \mu m$) ont commencé historiquement plus tardivement que les mesures de PM_{10} (particules de diamètre inférieur à $10 \mu m$), et bénéficient de ce fait d'un réseau moins développé, ce qui peut affecter la qualité des cartographies produites à partir de ces mesures.

Dans PREV'AIR, les cartes analysées d'ozone et de PM₁₀ sont construites selon la méthodologie géostatistique décrite par le LCSQA dans Malherbe et Ung (2009), qui consiste en un krigeage des observations incluant les simulations du modèle CHIMERE en dérive externe. Appliquée telle quelle aux PM_{2.5}, cette approche ne prend pas en compte la relation d'inégalité¹ avec les PM₁₀ et peut donc conduire sur certaines mailles du domaine à estimer des niveaux de concentration en PM_{2.5} supérieurs aux niveaux de concentration en PM₁₀. C'est pourquoi le système PREV'AIR n'utilise pas cette méthodologie pour les PM_{2.5} et se contente actuellement de corriger la carte analysée des PM₁₀ par le ratio PM_{2.5}/PM₁₀ simulé par CHIMERE. Cette méthode n'est toutefois pas toujours satisfaisante compte tenu des incertitudes de CHIMERE dans le calcul de ce ratio, en particulier lors des épisodes de pollution.

3.2 Modèle

Une nouvelle approche, développée en 2015, consiste à utiliser les mesures de PM₁₀ pour améliorer l'estimation des PM_{2.5} dans un modèle géostatistique adapté.

Un modèle de cokrigeage entre les PM₁₀ et les PM_{2.5} a été développé : il intègre CHIMERE en dérive externe via un modèle de dérive additive ; les PM_{2.5} ne seraient plus estimées à partir d'un ratio modélisé par CHIMERE mais comme la différence entre PM₁₀ et PM_{2.5} modélisées par CHIMERE conduisant à l'expression du résidu W_x :

$$Z_x = Y_x + W_x$$

Z_x est la concentration en PM₁₀, Y_x est la concentration en PM_{2.5} et W_x est la différence entre les deux

Les 2 composantes Y_x et W_x sont non-stationnaires, c'est-à-dire que leur moyenne et leur covariance spatiale dépendent du point de l'espace, mais on suppose qu'elles peuvent être expliquées par des variables explicatives :

$$Y_x = m_Y(x) + R_Y(x) = a_0 + \sum_i^{N_Y} a_i f_x^i + R_Y(x)$$

$$W_x = m_W(x) + R_W(x) = b_0 + \sum_j^{N_W} a_j g_x^j + R_W(x)$$

Dans cette étude, $N_Y = N_W = 1$ (seul CHIMERE est utilisé en dérive de Y et W), f^1 et g^1 sont respectivement les fonctions de dérive utilisées dans le calcul des moyennes locales $m_Y(x)$ et $m_W(x)$, i.e. les concentrations en PM_{2.5} et la différence de concentrations entre PM₁₀ et PM_{2.5} simulées par CHIMERE.

Du fait que le modèle CHIMERE respecte évidemment la relation d'ordre naturelle entre PM_{2.5} et PM₁₀, on a : $0 < m_Y(x) \leq m_W(x) \leq m_Z(x)$

L'estimateur de Z_x s'écrit donc :

¹ Par définition les concentrations de PM_{2.5} sont inférieures à celles de PM₁₀

$$Z_x^{CK} = \sum_{\alpha} \lambda_{\alpha} Z_{\alpha} + \sum_{\beta} \nu_{\beta} Y_{\beta} \quad (1)$$

où $\alpha = (1, \dots, N_1)$ est l'indice associé aux N_1 mesures PM_{10} et $\beta = (1, \dots, N_2)$ est l'indice associé aux N_2 mesures $PM_{2.5}$.

Les conditions de non-biais et la minimisation de la variance de l'erreur aboutissent au système linéaire suivant. Le détail des calculs est fourni en Annexe A.

$$\begin{bmatrix} \mathbf{K} & \mathbf{A} \\ \mathbf{A}' & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{W} \\ \mathbf{M} \end{bmatrix} = \begin{bmatrix} \mathbf{K}_x \\ \mathbf{F}_x \end{bmatrix} \quad (2)$$

avec

$$\mathbf{K} = \begin{bmatrix} C^Z & C^{Z,Y} \\ C^{Z,Y} & C^Y \end{bmatrix} \text{ matrice des covariances simples et croisées}$$

$$\mathbf{A} = \begin{bmatrix} 1 & g_{\alpha}^j & 0 & f_{\alpha}^i \\ 0 & 0 & 1 & f_{\beta}^i \end{bmatrix} \text{ matrice des dérivées aux points expérimentaux}$$

$$\mathbf{W} = \begin{bmatrix} \Gamma \\ \mathbf{L} \end{bmatrix} \text{ poids de cokrigage, } \Gamma = (\lambda_{\alpha}) \text{ et } \mathbf{L} = (\nu_{\beta})$$

$$\mathbf{M} = \begin{bmatrix} \mu_j \\ \nu_i \end{bmatrix} \text{ poids de Lagrange}$$

$$\mathbf{K}_x = \begin{bmatrix} C_{\alpha x}^Z \\ C_{\beta x}^{Z,Y} \end{bmatrix} \text{ vecteur des covariances simples et croisées entre les points}$$

expérimentaux et le point à estimer (x)

$$\mathbf{F}_x = \begin{bmatrix} 1 \\ g_x^j \\ 0 \\ f_x^i \end{bmatrix} \text{ vecteur des dérivées en x}$$

Les équations de cokrigage restent similaires si l'on cherche à produire une estimation des $PM_{2.5}$. Toutefois, la matrice et le vecteur des dérivées doivent être modifiés en conséquence :

$$\mathbf{A}' = \begin{bmatrix} 1 & 0 \\ 0 & g_{\alpha}^j \\ 0 & 1 \\ f_{\beta}^i & f_{\alpha}^i \end{bmatrix}, \mathbf{F}_x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ f_x^i \end{bmatrix}$$

Ce modèle spécifique a nécessité la réécriture complète d'un programme de krigage original afin de résoudre le système linéaire. En effet, les librairies existantes ne permettent pas de créer une fonction de dérive intégrant une relation d'ordre entre ses différentes composantes. Les librairies usuelles de géostatistique (geoR, gstat, RGeostats, etc.) peuvent toutefois être utilisées pour modéliser les variogrammes des résidus. Une fois les variogrammes expérimentaux ajustés, il est possible de récupérer les paramètres de ces modèles et de les utiliser dans un programme indépendant qui résout le système matriciel (2).

Ce modèle a été testé sur les moyennes journalières de $PM_{2.5}$ portant sur la période janvier-mars 2015. Ce modèle permet notamment de réduire à 0.5% le pourcentage de mailles où l'estimation $PM_{2.5}$ est supérieure à l'estimation PM_{10} , alors que pour un cokrigage classique sans dérive additive ce pourcentage est de 5%. De plus, pour ces quelques mailles qui ne respectent pas la contrainte naturelle imposée par la définition des $PM_{2.5}$ et des PM_{10} , la valeur de l'estimation $PM_{2.5}$ se situe dans la marge d'incertitude attribuée à l'estimation PM_{10} et pragmatiquement définie à partir de l'écart-type de krigeage (voir par ex. Beauchamp et al. 2010). L'estimateur est donc considéré comme cohérent statistiquement. Toutefois, pour des raisons d'affichage, ce type de situation est géré par une troncature de l'estimation $PM_{2.5}$ qui est ramenée à la valeur physiquement majorante donnée par l'estimation PM_{10} .

3.3 Résultats

Les Figure , 2 et 3 montrent les cartes $PM_{2.5}$ analysées pour les dates du 10, 15 et 20 mars 2015, lors de l'épisode de pollution printanier de 2015. La Figure utilise une simple dérive externe (les données PM_{10} ne sont pas utilisées dans l'estimation). La Figure utilise la version actuellement opérationnelle dans PREV'AIR pour fournir des analyses $PM_{2.5}$ (carte de PM_{10} analysée, multipliée par le ratio $PM_{2.5}/PM_{10}$ de CHIMERE afin de forcer la carte à ne pas avoir de valeurs supérieures à la carte analysée de PM_{10}). Enfin, la Figure présente les cartographies obtenues à partir du cokrigage $PM_{2.5}$ - PM_{10} avec modèle de dérive additive.

Il ressort que les analyses $PM_{2.5}$ actuellement produites dans PREV'AIR sous-estiment fortement les concentrations réelles en situation d'épisode : les Tableau 1 et 2 présentent les scores pour les mois de janvier à mars 2015 pour chacun des trois estimateurs testés. Si la correction de l'analyse PM_{10} par le ratio $PM_{2.5}/PM_{10}$ de CHIMERE a des scores corrects en janvier et en février, ce n'est plus du tout le cas en mars, où les épisodes de pollution sont survenus. De plus, cet estimateur introduit un biais systématique important qui se révèle non satisfaisant. Son nombre de fausses alertes est toutefois le meilleur en janvier, ce qui peut s'expliquer par le très fort biais négatif de cet estimateur sur ce mois.

Tableau 1 : Tableaux de scores (en validation croisée) journaliers moyens pour les analyses $PM_{2.5}$ sur les mois de janvier à mars 2015 pour chacun des trois estimateurs testés. Scores calculés sur les stations rurales, périurbaines et urbaines de fond.

		Corrélation	RMSE	NMB
$\sum_{\alpha} \lambda_{\alpha} Z_{\alpha}$ Dérive externe	01-15	0.88	4.95	0.36
	02-15	0.91	4.56	0.30
	03-15	0.96	4.91	0.34
$\sum \lambda_{\alpha} Z_{\alpha} + \sum \nu_R Z_R^{10}$		0.9	3.12	0.34

Cokrigage en dérive externe additive	02-15	0.96	3.12	0.07
	03-15	0.98	3.4	0.04
$\left(\sum_{\alpha} \lambda_{\alpha} Z_{\alpha}^{10}\right) \times Z_M Z_M^{10}$ Correction de l'analyse PM ₁₀ par le ratio PM _{2.5} /PM ₁₀ de CHIMERE	01-15	0.94	4	-13.1
	02-15	0.93	3.92	-4.62
	03-15	0.94	6.73	-13.7

RMSE : racine carrée de l'erreur quadratique moyenne

NMB : biais moyen normalisé

Tableau 2 : Tableaux des scores (en validation croisée) relatifs aux dépassements de seuil pour les analyses PM_{2.5} sur les mois de janvier à mars 2015 pour chacun des trois estimateurs testés. Scores calculés sur les stations rurales, périurbaines et urbaines de fond. Pour le calcul des scores relatifs à la détection des dépassements, puisqu'il n'existe pas de valeur limite journalière PM_{2.5}, le seuil de 25 µg.m⁻³ a été utilisé

		Bonnes détections	Fausse alertes	Dépassements manqués	Non dépassements bien prévus
$\sum_{\alpha} \lambda_{\alpha} Z_{\alpha}$ Dérive externe	01-15	270	81	149	2413
	02-15	425	82	127	2117
	03-15	829	101	136	1943
$\sum_{\alpha} \lambda_{\alpha} Z_{\alpha} + \sum_{\beta} \nu_{\beta} Z_{\beta}^{10}$ Cokrigage en dérive externe additive	01-15	335	58	84	2436
	02-15	468	61	84	2138
	03-15	880	65	85	1979
$\left(\sum_{\alpha} \lambda_{\alpha} Z_{\alpha}^{10}\right) \times Z_M Z_M^{10}$ Correction de l'analyse PM ₁₀ par le ratio PM _{2.5} /PM ₁₀ de CHIMERE	01-15	268	29	151	2465
	02-15	409	73	143	2126
	03-15	753	71	212	1973

Bien que moins marquées, les différences entre la simple dérive externe et le cokrigage avec dérive additive n'en sont pas moins significatives. Il apparaît donc clairement que le nouvel estimateur basé sur un cokrigage en dérive externe additive est le plus robuste et donne de meilleurs résultats..

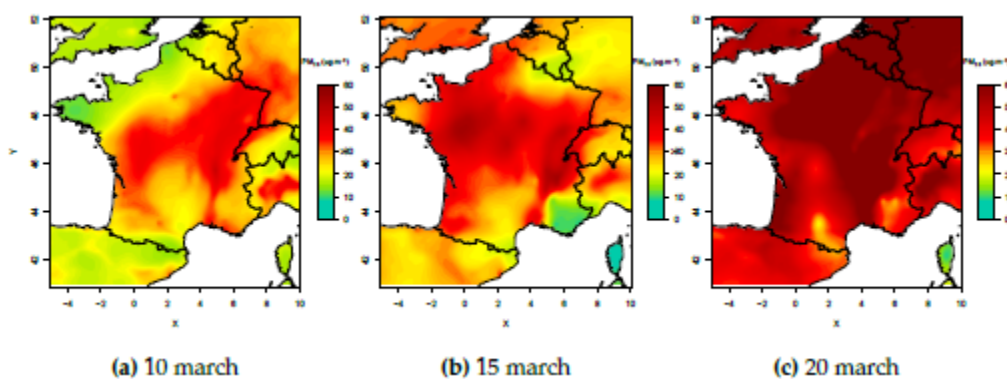


Figure 1 : Cartes PM_{2.5} analysées avec CHIMERE en dérive externe

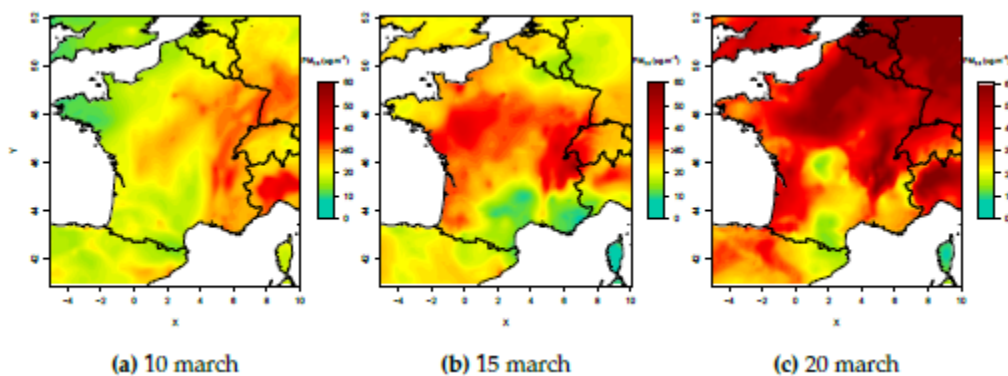


Figure 2 : cartes PM₁₀ analysées corrigées par le ratio PM_{2.5}/PM₁₀ de CHIMERE

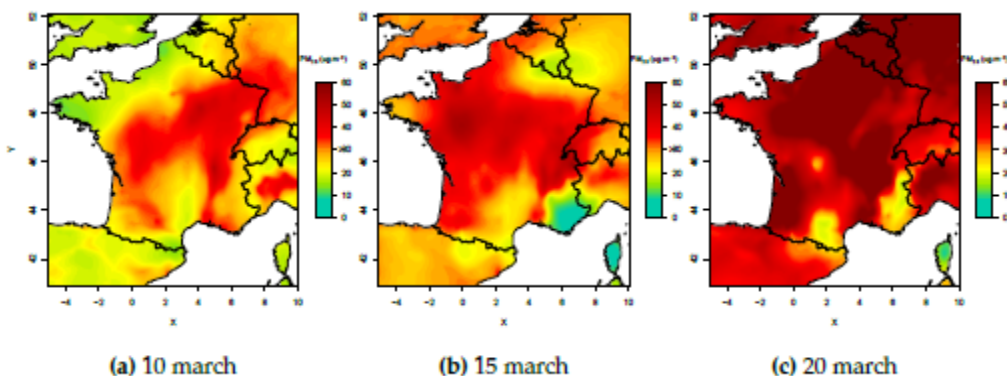


Figure 3 : Cartes PM_{2.5} analysées par un cokrigage avec les PM₁₀ qui introduit CHIMERE dans un modèle de dérive additive

La Figure montre les distributions de scores (corrélation, RMSE, NMB et taux de bonne détection de dépassement) obtenues sur la période janvier-mars 2015.

Globalement la meilleure méthode identifiée est le cokrigage PM_{2.5}-PM₁₀ avec modèle de dérive additive (courbe bleue). Elle apporte un gain très significatif dans la performance de l'analyse PM_{2.5}.

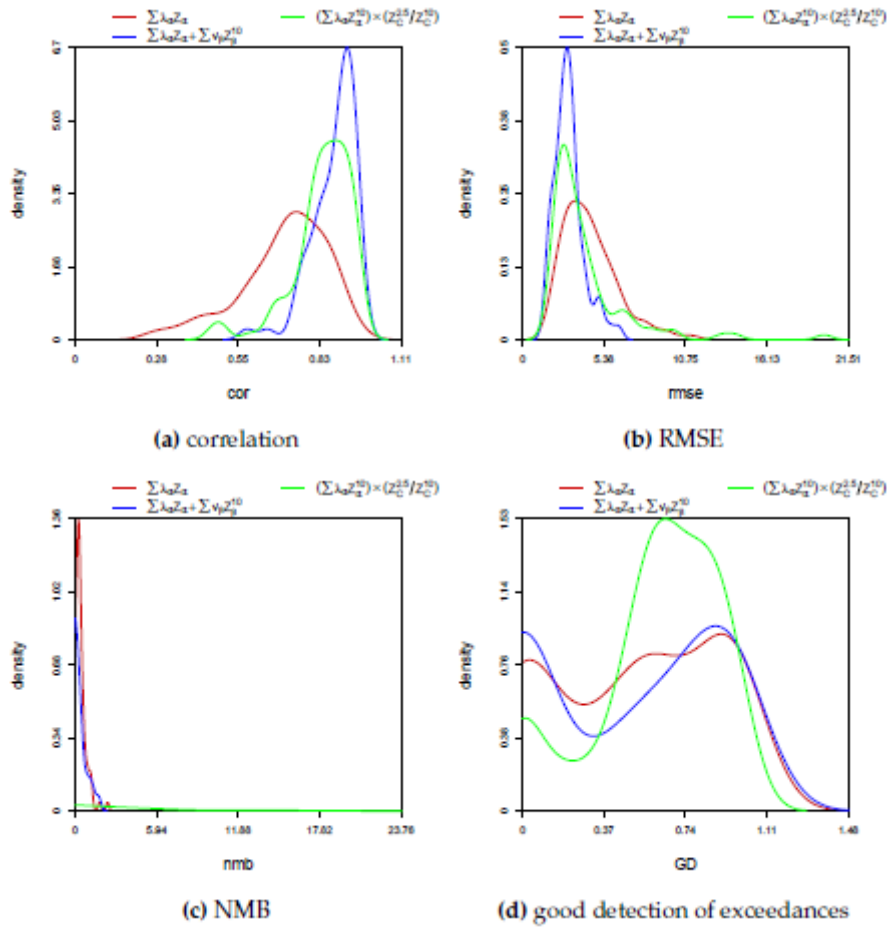


Figure 4 : Distributions (ajustées) des scores (corrélation, RMSE, NMB, taux de bonne détection) avec les 3 familles d'analyse testées (rouge : krigeage en dérive externe des $PM_{2.5}$; bleu : cokrigeage en dérive externe additive des $PM_{2.5}$; vert : correction de l'analyse PM_{10} par le ratio $PM_{2.5}/PM_{10}$ de CHIMERE)

La Figure montre les stations $PM_{2.5}$ sur lesquelles le cokrigeage avec modèle de dérive additive améliore ou dégrade les scores (en corrélation ici mais les résultats sont les mêmes pour la RMSE) : la quasi-intégralité du réseau de mesures français voit ses scores améliorés à l'exception de quelques stations isolées. Pour ces dernières, il s'agit d'un artefact lié à toute méthode de validation croisée dans les régions avec une forte densité de mesure où certaines observations représentatives d'une situation locale ne peuvent être bien estimées à partir des stations voisines.

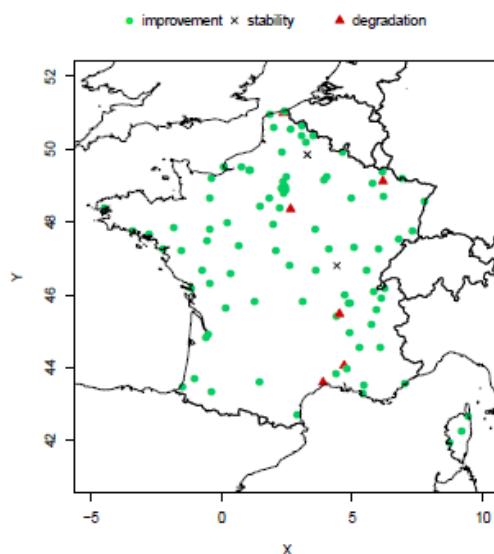


Figure 5 : Amélioration/dégradation des scores par station (vert : amélioration ; rouge : dégradation ; croix : pas de changement)

Les travaux menés ont donc conduit à développer une méthode de krigeage qui améliore très sensiblement les cartographies de $PM_{2.5}$ surtout là où le nombre de stations de mesure de $PM_{2.5}$ est faible par rapport au nombre de stations de PM_{10} .

Même si l'amélioration apportée par le cokrigeage est moins prononcée pour les PM_{10} que pour les $PM_{2.5}$, elle est tout de même significative. Les Figure 6 et 7 montrent les cartes PM_{10} analysées pour les dates du 10, 15 et 20 mars 2015, lors de l'épisode de pollution printanier de 2015. La Figure 6 utilise une simple dérive externe (les données $PM_{2.5}$ ne sont pas utilisées dans l'estimation). La Figure 7 présente les cartographies obtenues à partir du cokrigeage PM_{10} - $PM_{2.5}$ avec modèle de dérive additive.

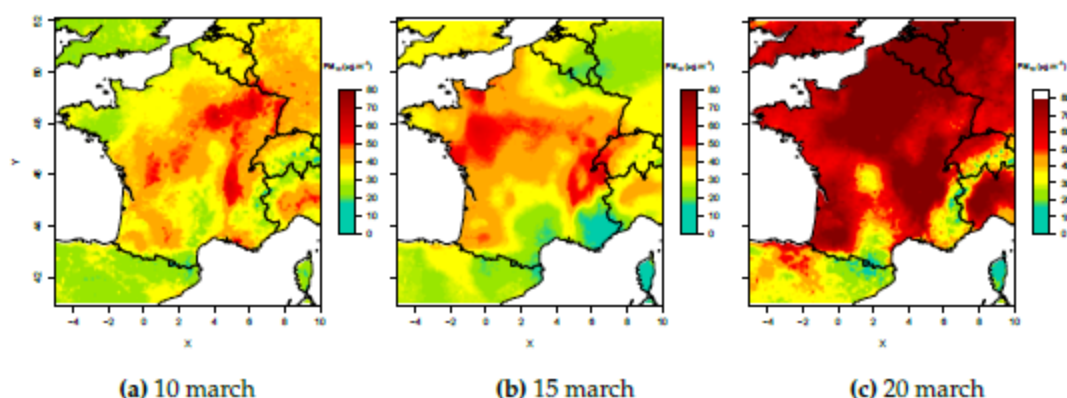


Figure 6 : Cartes PM_{10} analysées avec CHIMERE en dérive externe

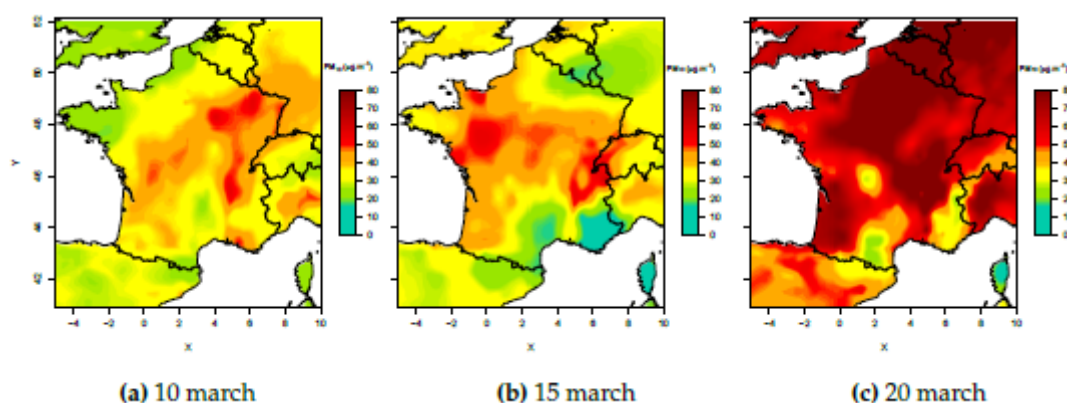


Figure 7 : Cartes PM₁₀ analysées par un cokrigage avec les PM_{2.5} qui introduit CHIMERE dans un modèle de dérive additive

Les Tableaux 3 et 4 présentent les scores pour les mois de janvier à mars 2015 obtenus avec les deux estimateurs testés : krigage des PM₁₀ avec CHIMERE en dérive externe et cokrigage PM₁₀-PM_{2.5} avec modèle de dérive additive.

Tableau 3 : Tableaux de scores (en validation croisée) journaliers moyens pour les analyses PM₁₀ sur les mois de janvier à mars 2015 pour les deux estimateurs testés. Scores calculés sur les stations rurales, périurbaines et urbaines de fond.

		Corrélation	RMSE	NMB
$\sum_{\alpha} \lambda_{\alpha} Z_{\alpha}$ Dérive externe	01-15	0.86	5.78	0.73
	02-15	0.9	5.1	0.64
	03-15	0.96	5	0.71
$\sum_{\alpha} \lambda_{\alpha} Z_{\alpha} + \sum_{\beta} \nu_{\beta} Z_{\beta}^{2.5}$ Cokrigage en dérive externe additive	01-15	0.88	5.56	0.08
	02-15	0.92	4.9	0.32
	03-15	0.96	4.9	0.42

Tableau 4 : Tableaux des scores (en validation croisée) relatifs aux dépassements de seuil pour les analyses PM₁₀ sur les mois de janvier à mars 2015 pour les deux estimateurs testés. Scores calculés sur les stations rurales, périurbaines et urbaines de fond. La valeur limite journalière PM₁₀ fixée à 50 µg.m⁻³ a été utilisée

		Bonnes détections	Fausse alertes	Dépassements manqués	Non dépassements bien prévus
$\sum_{\alpha} \lambda_{\alpha} Z_{\alpha}$ Dérive externe	01-15	90	29	60	6756
	02-15	102	46	84	6079
	03-15	652	97	134	6218

		Bonnes détections	Fausse alertes	Dépassements manqués	Non dépassements bien prévus
Cokrigage en dérive externe additive	$\sum_{\alpha} \lambda_{\alpha} Z_{\alpha} + \sum_{\beta} \nu_{\beta} Z_{\beta}^{2.5}$ 01-15	84	35	66	6750
	02-15	107	40	79	6085
	03-15	655	88	131	6227

4. LISSAGE DU (CO)KRIGEAGE EN VOISINAGE GLISSANT

4.1 Contexte

Pour chaque point de la grille de krigeage, le voisinage représente l'ensemble des données de mesure qui participent à l'estimation. Un voisinage est dit unique s'il inclut la totalité des données, il est dit glissant s'il se déplace avec le point d'estimation. Dans ce cas, il se caractérise par une distance maximale et/ou un nombre maximal de données autour du point. Un voisinage glissant de 80 points de mesure est actuellement utilisé dans PREV'AIR. Il a été défini à l'issue d'analyses de sensibilité. La définition du voisinage de krigeage demeure cependant un problème délicat, en particulier en présence de dérive et lorsque le réseau de mesure présente une densité très hétérogène dans l'espace. Un voisinage glissant trop réduit conduit à des discontinuités dans la cartographie ; un voisinage unique ou un voisinage glissant trop grand fait perdre le bénéfice d'un ajustement local de la dérive.

4.2 Modèle

Le travail présenté en section 3 a été l'occasion de mettre en place le lissage du krigeage lorsque celui-ci se fait en voisinage glissant. Les travaux de Rivoirard et Romary (2011) ont été adaptés au cokrigage et testés sur la période janvier-mars 2015.

En reprenant le modèle introduit en section 3.2, on ajoute au lagrangien deux nouvelles contraintes :

$$L = \text{Var}(Z_x - Z_x^{\text{CK}}) + 2\mu_0 \left(\sum_{\alpha} \lambda_{\alpha} - 1 \right) + 2\nu_0 \sum_{\beta} \nu_{\beta} + 2 \sum_j \mu_j \left(\sum_{\alpha} \lambda_{\alpha} g_{\alpha}^j - g_x^j \right) \\ + 2 \sum_i \nu_i \left(\sum_{\beta} \nu_{\beta} f_{\beta}^i - f_x^i \right) + \sum_{\alpha} \lambda_{\alpha}^2 n_{\alpha} + \sum_{\beta} \nu_{\beta}^2 n_{\beta}$$

où les poids λ_{α} et ν_{β} sont pénalisés respectivement par un bruit n_{α} et n_{β} .

En annulant les dérivées partielles de L par rapport à λ_{α} , ν_{β} , μ_1 , et μ_2 , on obtient le nouveau système matriciel :

$$\begin{bmatrix} \mathbf{D}_{11} & \mathbf{C}_{12} & 1 & \mathbf{g}_\alpha^j & 0 & \mathbf{f}_\alpha^i \\ \mathbf{C}_{21} & \mathbf{D}_{22} & 0 & 0 & 1 & \mathbf{f}_\beta^i \\ 1 & 0 & 0 & 0 & 0 & 0 \\ \mathbf{g}_\alpha^j & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ \mathbf{f}_\alpha^i & \mathbf{f}_\beta^i & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_\alpha \\ \nu_\beta \\ \mu_0 \\ \mu_j \\ \nu_0 \\ \nu_i \end{bmatrix} = \begin{bmatrix} \mathbf{C}_{\alpha x}^Z \\ \mathbf{C}_{\beta x}^{Z,Y} \\ 1 \\ g_x^j \\ 0 \\ 0 \end{bmatrix}$$

dans lequel $\mathbf{D}_{11}$ et $\mathbf{D}_{22}$ sont les matrices de covariances $\mathbf{C}_{11}$ et $\mathbf{C}_{22}$ pour lesquelles les éléments diagonaux sont augmentés d'un bruit additionnel, respectivement n_α et n_β . Plus précisément, on considère qu'une observation est une variable aléatoire dont la valeur est perturbée par un bruit aléatoire. Ce bruit aléatoire a une variance d'autant plus forte que l'observation est éloignée du nœud de grille où l'on veut estimer une valeur de concentration. La forme du bruit est la même que celle utilisée dans Rivoirard et Romary (2011) : elle augmente de 0 à $+\infty$ entre $r = 2/3R$ et R , où R est la valeur du voisinage en distance choisi. De telles fonctions sont par exemple :

$$n_\alpha = C_Z(0) \left(\frac{h_\alpha - r}{R - h_\alpha} \right)^2$$

$$\text{et } n_\beta = C_Y(0) \left(\frac{h_\beta - r}{R - h_\beta} \right)^2$$

avec h_α et h_β la distance entre les points expérimentaux et x .

En qualité de l'air, le problème de la discontinuité survient principalement pour les grands domaines d'étude (par ex. Europe) dans lesquels les réseaux de mesures ne sont pas distribués de manière homogène. Ainsi, on travaille plutôt avec un voisinage en nombre car un voisinage en distance a tendance à inclure trop de stations là où le réseau est dense, et trop peu dans les zones peu informées.

Le modèle présenté pour lisser l'estimateur est pourtant construit à partir d'un voisinage en distance. Pour régler ce problème, puisque les estimations sont menées sur des grilles, on calcule pour chaque maille la distance minimale nécessaire pour inclure N stations dans le voisinage (N est le nombre de stations défini pour le voisinage en nombre). Cette distance varie dans l'espace puisque le réseau n'est pas homogène, mais elle varie de manière continue puisque la grille est régulière. Il est ainsi possible de conserver l'approche avec voisinage en distance dans laquelle le rayon R utilisé n'est plus fixe mais varie continûment dans l'espace : $R(x)$ est la distance entre le centre de maille (x) et le N ème plus proche voisin de x .

4.3 Résultats

Les Figures 9 et 10 montrent les résultats d'un cokrigeage PM_{10} - $\text{PM}_{2.5}$ en dérive externe additive sur les dates du 15 et 20 mars 2015, sans et avec lissage d'un voisinage glissant fixé ici à 20 stations pour marquer les effets de voisinage.

Les Figures 11 et 12 montrent les mêmes cartographies avec un voisinage de 80 stations, comme celui utilisé dans PREV'AIR. Les différences entre les cartographies sont moins marquées mais elles peuvent toutefois rester très importantes sur certaines journées en fonction de l'ajustement automatique qui est fait du variogramme.

A titre indicatif, la Figure 8 fournit les observations disponibles en Europe sur ces mêmes dates ainsi que les sorties brutes du modèle CHIMERE.

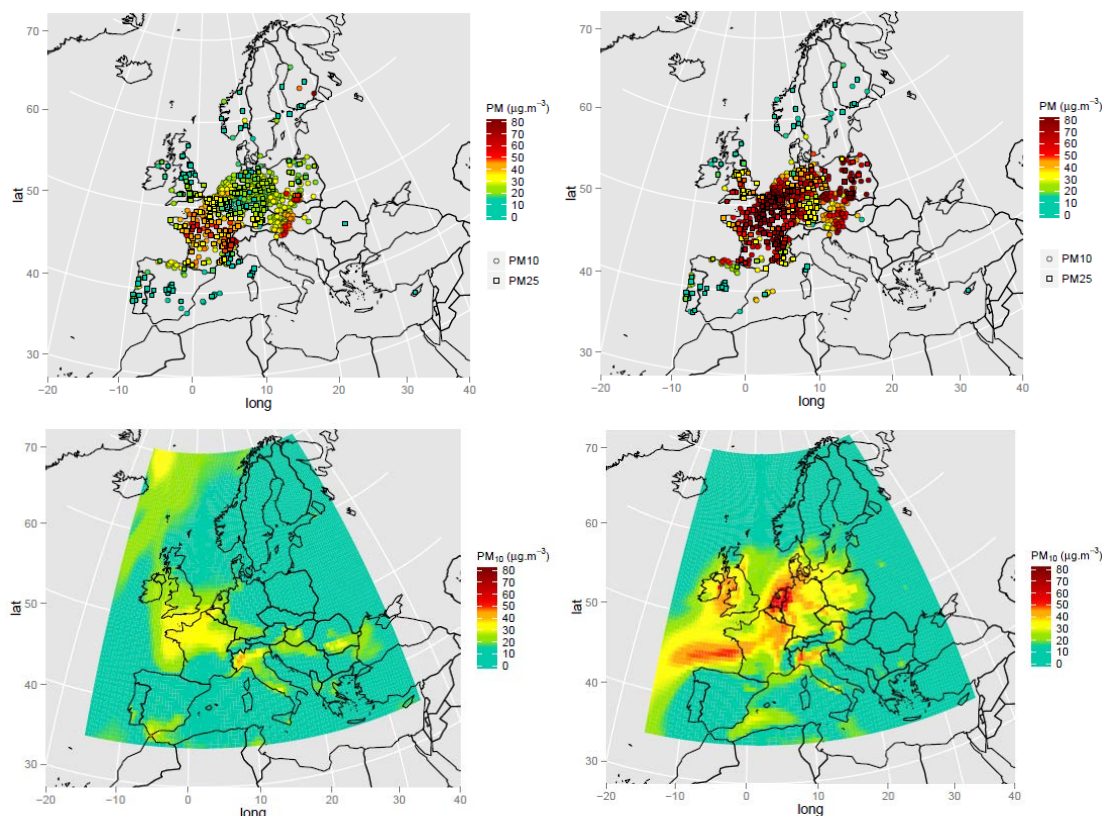


Figure 8 : Observations PM_{10} et $PM_{2.5}$ (15 et 20 mars 2015) et sorties brutes du modèle CHIMERE sur ces 2 journées

Le panache au Nord-Ouest de la cartographie est un artefact lié au faible nombre de stations considérées dans le voisinage (20). Il disparaît lorsqu'on passe à 80 (voir Figures 11 et 12). Les très fortes concentrations estimées sur le golfe de Gascogne s'expliquent à la fois par l'épisode de poussières sahariennes prévus par CHIMERE (voir Figure 8) et la régression localement ajustée par le krigeage sur les 20 (Figures 9 et 10) ou 80 (Figures 11 et 12) stations les plus proches de cette zone. On peut se poser la question de la pertinence de cette régression lorsque les stations les plus proches sont déjà relativement éloignées et qu'aucune validation n'est possible. A défaut, l'ajout de points d'observations fictifs basés sur les valeurs CHIMERE sur les domaines océaniques ou non informés (partie Ouest de la cartographie) est une piste d'amélioration. Une autre option, qui a été choisie dans PREV'AIR, est de construire une carte composite à partir de la carte analysée et de la carte brute CHIMERE. Pour un centre de maille localisé en x , la valeur estimée sera :

- 1- La valeur du (cokrigeage) si le k -ème voisin le plus proche est distant de moins de 500km

- 2- La valeur de CHIMERE si le k-ème voisin le plus proche est distant de plus de 1000km
- 3- Une combinaison linéaire de la valeur du (co)krigeage et de la sortie brute CHIMERE si le k-ème voisin le plus proche est distant de plus de 500km et de moins de 1000km.

Cette carte composite permet d'éviter les problèmes de régression ajustées localement à partir de voisins trop éloignés et donc peu représentatifs de la vraie situation locale, pouvant conduire à des estimations aberrantes. Après quelques tests de sensibilité, k a été fixé à 10. Enfin, des méthodes similaires de cartes composites utilisant l'écart-type de krigage comme critère de composition pourraient également être utilisées.

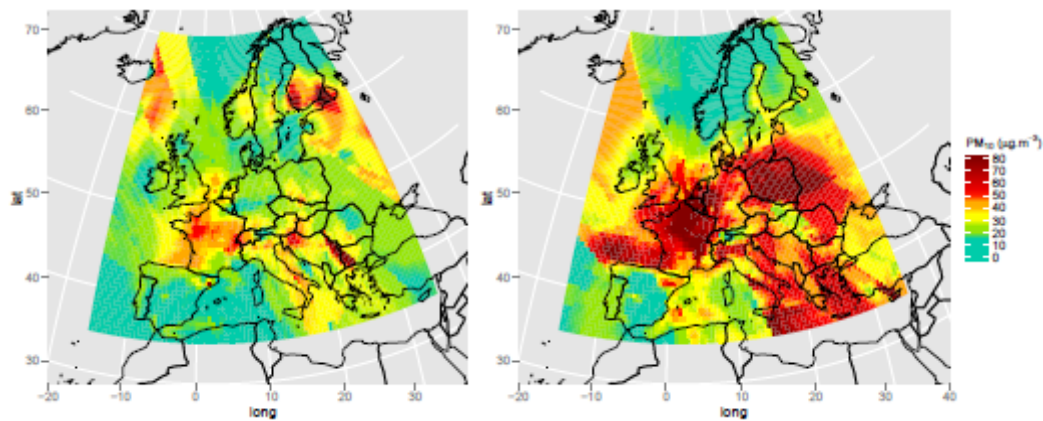


Figure 9 : Analyse PM_{10} (15 et 20 mars 2015) en voisinage glissant (20 stations) sans lissage

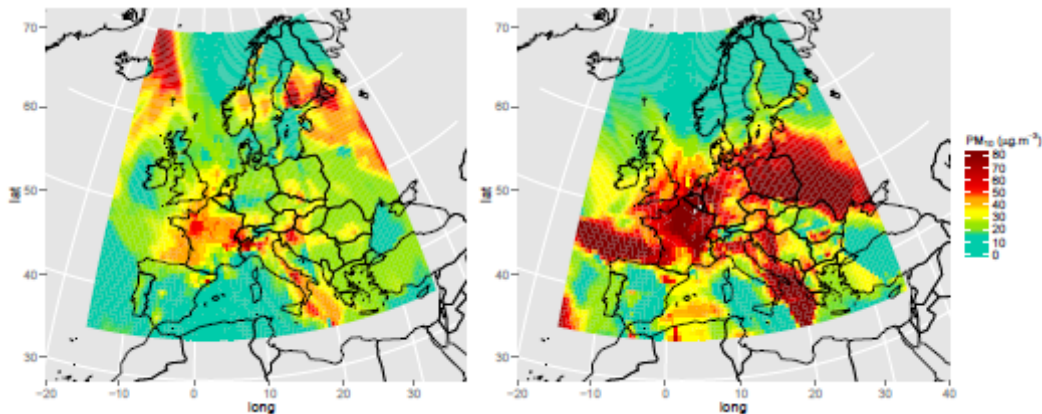


Figure 10 : Analyse PM_{10} (15 et 20 mars 2015) en voisinage glissant (20 stations) avec lissage

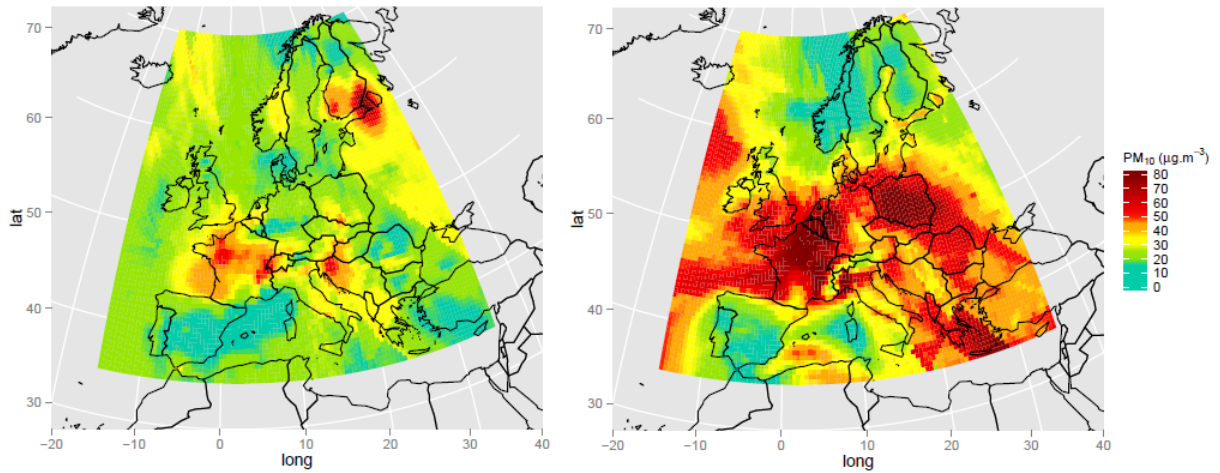


Figure 11 : Analyse PM₁₀ (15 et 20 mars 2015) en voisinage glissant (80 stations) sans lissage

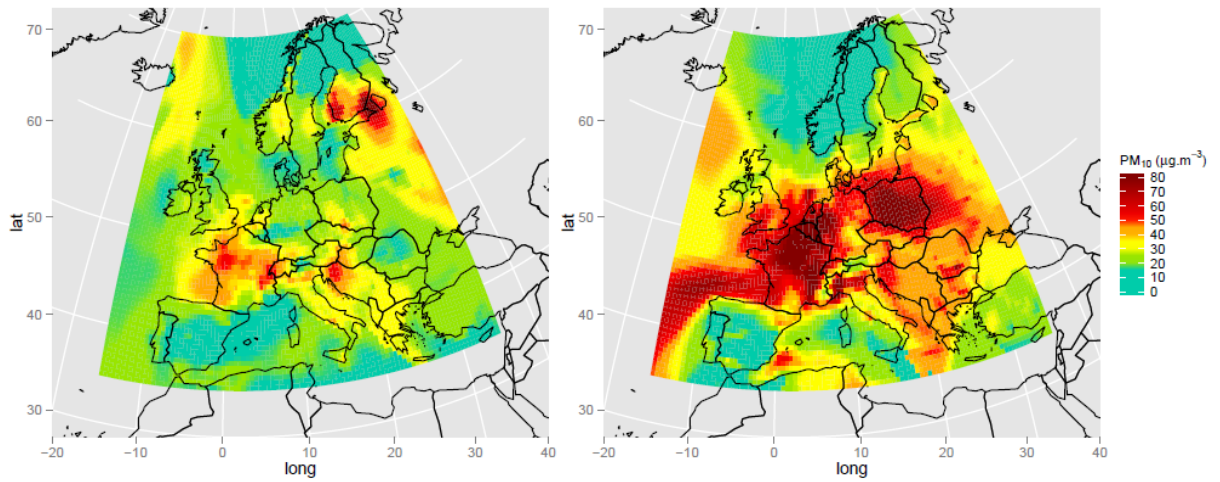


Figure 12 : Analyse PM₁₀ (15 et 20 mars 2015) en voisinage glissant (80 stations) avec lissage

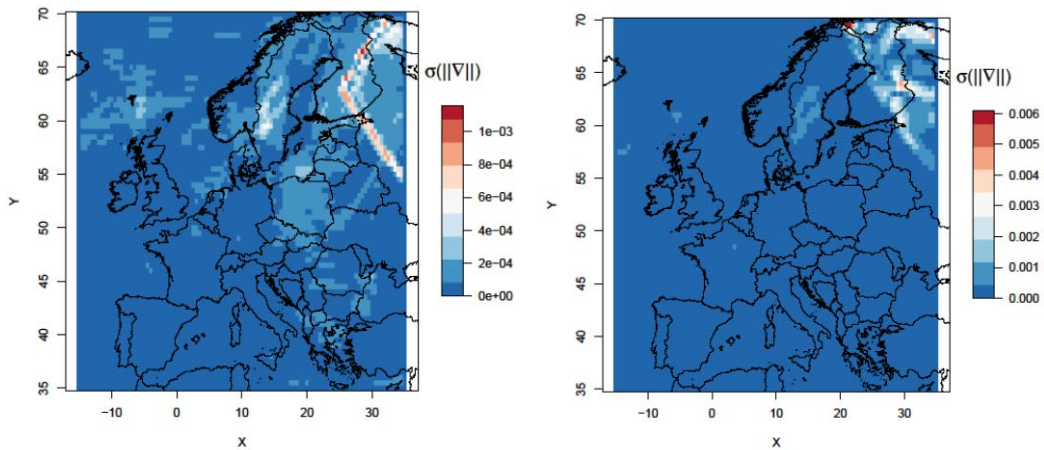
D'un point de vue purement graphique, le réalisme des cartographies est grandement amélioré : les phénomènes de rupture sont très fortement atténués. Cette analyse visuelle est confirmée par l'étude des gradients de concentrations sur les cartographies. En particulier, on s'attend à ce que l'écart-type de la norme des gradients de concentration $\sigma(|\nabla|)$ estimés par krigeage soient moins élevés avec le lissage dans les zones impactées par les effets du voisinage glissant non contraint, en particulier le Nord-Ouest et le Nord-Est de la cartographie Europe.

Le détail du calcul de cet indicateur est décrit ci-dessous :

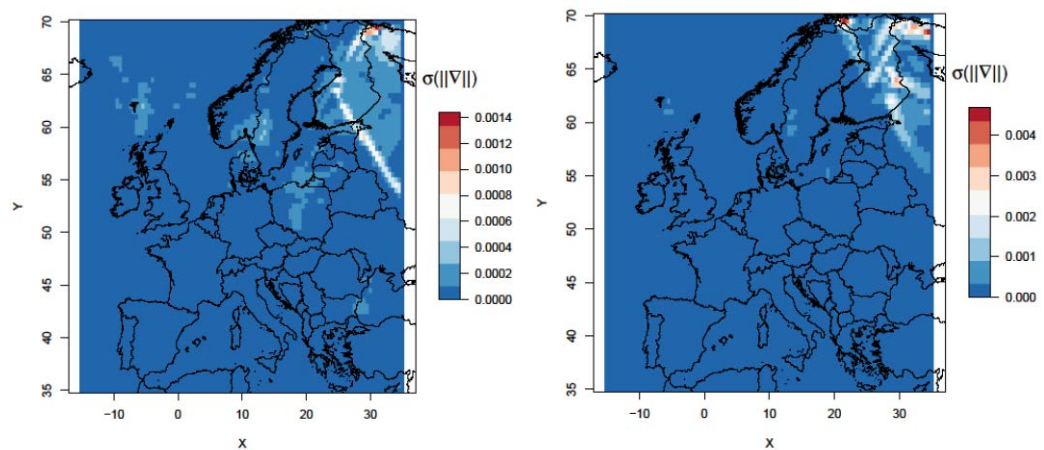
$$\sigma(|\nabla|)(x, y) = \sqrt{\frac{\sum_{i=1}^N \left(|\nabla|(x, y) - \overline{|\nabla|}(x, y) \right)^2}{N}}$$

où $||\nabla|| (x, y) = \sqrt{\left(\frac{\partial Z_i}{\partial x}(x, y)\right)^2 + \left(\frac{\partial Z_i}{\partial y}(x, y)\right)^2}$ et $\overline{||\nabla||} (x, y)$ est la moyenne temporelle de $||\nabla|| (x, y)$; N est le nombre de jours utilisées dans l'étude (90 de janvier à mars) et $Z_i(x, y)$ est la valeur du champ de concentration le jour i au point (x, y) .

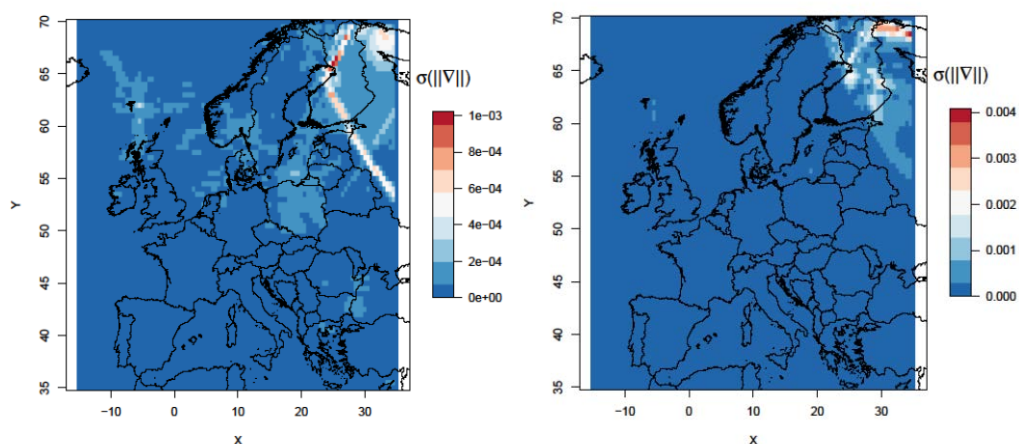
La Figure 12 montre les cartographies de $\sigma(||\nabla||)$ relatives aux PM_{10} pour des voisinages en nombre de 20, 40, 60, 80 et 100 stations, avec et sans lissage. Il apparaît assez clairement que la contrainte de continuité dans le krigeage fait presque disparaître les variations temporelles non réalistes des gradients du nord-ouest de la cartographie. Sur la partie nord-est, ces variations existent toujours mais elles sont spatialement plus continues.



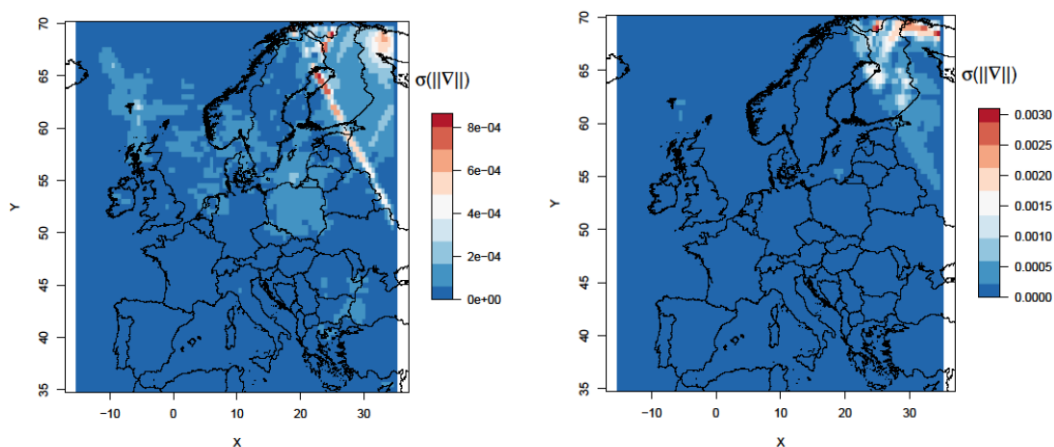
(a) 20 stations sans lissage (à gauche) et avec lissage (à droite)



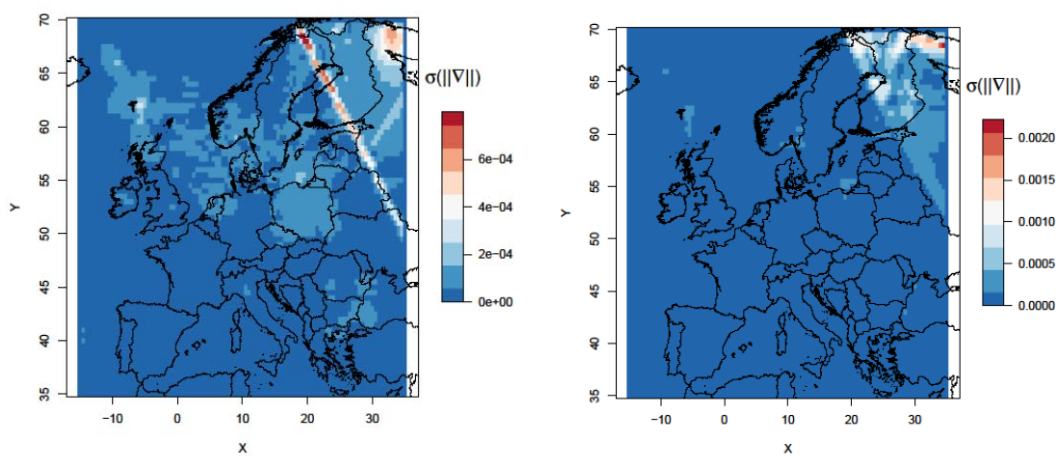
(b) 40 stations sans lissage (à gauche) et avec lissage (à droite)



(c) 60 stations sans lissage (à gauche) et avec lissage (à droite)



(c) 80 stations sans lissage (à gauche) et avec lissage (à droite)



(e) 100 stations sans lissage (à gauche) et avec lissage (à droite)

Figure 13 : cartographies de $\sigma(||\nabla||)$ relatives aux PM_{10}

D'après les scores, il est difficile de savoir laquelle des méthodes est la meilleure. Des validations croisées ont été réalisées avec des voisinages en nombre allant de N=20 à N=100 par pas de 10, avec et sans forçage de la continuité (la lettre s signifie « smoothing » sur l'axe des abscisses de la Figure 14). La validation croisée dans le cas d'un cokrigeage en voisinage unique a également été effectuée. La Figure 10 montre les distributions de scores journaliers (corrélation, RMSE, NMB et taux de bonne détection) sur la période janvier-mars 2015 pour chacun des voisinages, avec et sans lissage. Les distributions semblent équivalentes.

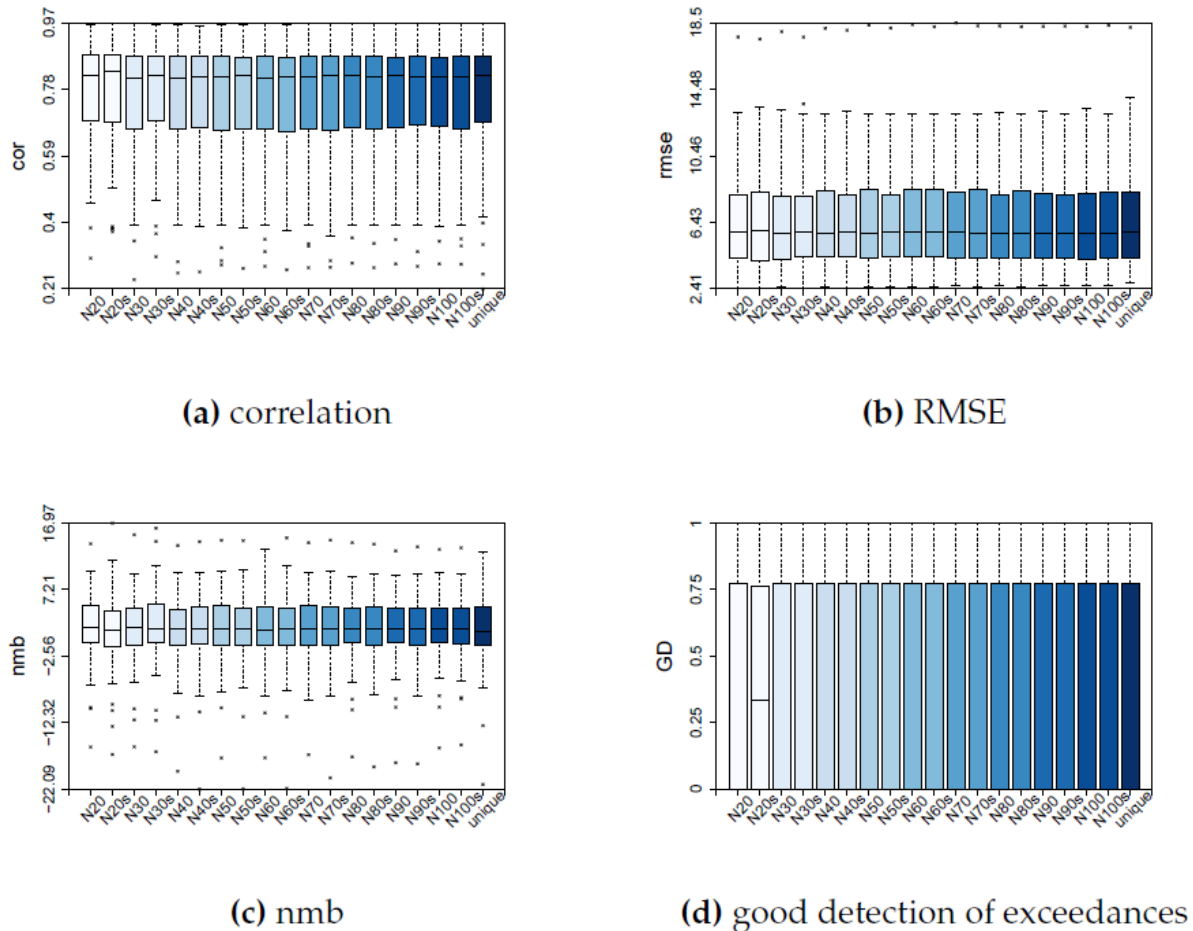


Figure 1 : Impact du voisinage de krigeage et du lissage sur la performance du cokrigeage en validation croisée

* l'échelle de couleur des boxplots va du blanc (N=20) au bleu sombre (voisinage unique incluant toutes les stations)

Il existe certainement des différences de performances entre les voisinages choisis mais elles ne sont probablement significatives que sur les parties océaniques et la partie Est du domaine Europe, là où il n'y a pas de stations de mesures pour valider les modèles.

A défaut d'améliorer les scores, ce qui n'était de toute façon pas attendu car on contraint plus le système de cokrigeage, la continuité des cartes est améliorée.

4.4 Remarques

Le problème de la continuité des cartographies en voisinage glissant se pose quel que soit le modèle de krigeage utilisé. Pour les autres polluants d'intérêt, et notamment ceux dont la carte est diffusée sur le site PREV'AIR (PM₁₀ et O₃), le krigeage en dérive externe doit également être adapté pour permettre le lissage de l'estimateur. L'approche utilisée pour le cokrigeage est conservée et adaptée à des systèmes de krigeage en dérive externe classique, i.e. un bruit, construit avec une fonction similaire à celle utilisée en section 4.2, est ajouté sur la diagonale de la matrice de krigeage.

5. CONCLUSIONS

Le modèle de cokrigeage en dérive externe additive a démontré sa capacité à mieux reproduire les niveaux de concentrations en particules dans les cartes analysées. Il réduit d'un ordre de grandeur la fraction de points où les estimations PM_{2,5} sont supérieures aux estimations PM₁₀ (de 5 à 0.5%).

La méthode de lissage des estimateurs géostatistiques qui a été développée à la fois pour le modèle de cokrigeage introduit dans cette note, mais également pour les modèles de krigeage en dérive externe utilisés pour les analyses des PM₁₀, du NO₂ et de l'O₃, a aussi montré sa capacité à répondre au problème de discontinuité des cartographies construites en voisinage glissant.

Ces deux nouveautés méthodologiques seront portées dans le système PREV'AIR en 2016.

6. RÉFÉRENCES

Rivoirard J. & Romary Th. (2011). **Continuity for Kriging with Moving Neighborhood**. Mathematical Geosciences, **43**(4), 469–481.

Malherbe L., Ung A. **Travaux relatifs à la plate-forme nationale de modélisation PREV'AIR : réalisation de cartes analysées d'ozone**. Rapport LCSQA 2009 : <http://www.lcsqa.org/rapport/2009/ineris/travaux-relatifs-plate-forme-nationale-modelisation-pre-vair-realisation-cartes->

Beauchamp M., Malherbe L. et Létinois, L. **Application de méthodes géostatistiques pour la détermination de zones de représentativité en concentration et la cartographie des dépassements de seuils**. Rapport LCSQA 2010 disponible sur le site du LCSQA : <http://www.lcsqa.org/rapport/2010/ineris/application-methodes-geostatistiques-determination-zones-representativite-concen>

ANNEXE A

Ecriture du système de krigeage

Les conditions de non-biais sont données par :

$$\begin{aligned}
E(Z_x - Z_x^{CK}) &= E \left[Z_x - \left(\sum_{\alpha} \lambda_{\alpha} Z_{\alpha} + \sum_{\beta} \nu_{\beta} Y_{\beta} \right) \right] \\
&= a_0 + \sum_i a_i f_x^i + b_0 + \sum_j b_j g_x^j - \sum_{\alpha} \lambda_{\alpha} (a_0 + \sum_i a_i f_{\alpha}^i + b_0 + \sum_j b_j g_{\alpha}^j) \\
&\quad - \sum_{\beta} \nu_{\beta} (a_0 + \sum_i a_i f_{\beta}^i) \\
&= a_0 \left(1 - \sum_{\alpha} \lambda_{\alpha} - \sum_{\beta} \nu_{\beta} \right) + \sum_i a_i \left(f_x^i - \sum_{\alpha} \lambda_{\alpha} f_{\alpha}^i - \sum_{\beta} \nu_{\beta} f_{\beta}^i \right) \\
&\quad + b_0 (1 - \sum_{\alpha} \lambda_{\alpha}) + \sum_j b_j (g_x^j - \sum_{\alpha} \lambda_{\alpha} g_{\alpha}^j) = 0
\end{aligned}$$

Les vecteurs f_x^i et g_x^j sont supposées être mutuellement libres (indépendants linéairement). Ainsi, l'estimateur est non biaisé quels que soient a_0 , a_i , b_0 , b_i inconnus si et seulement si les 4 conditions de non-biais suivantes sont respectées:

$$\begin{aligned}
\sum_{\alpha} \lambda_{\alpha} + \sum_{\beta} \nu_{\beta} &= 1 \\
\sum_{\alpha} \lambda_{\alpha} &= 1 \Rightarrow \sum_{\beta} \nu_{\beta} = 0 \\
\sum_{\alpha} \lambda_{\alpha} f_{\alpha}^i + \sum_{\beta} \nu_{\beta} f_{\beta}^i &= f_x^i \\
\sum_{\alpha} \lambda_{\alpha} g_{\alpha}^j &= g_x^j
\end{aligned} \tag{3}$$

La variance d'estimation vaut :

$$\begin{aligned}
\text{Var}(Z_x - Z_x^{CK}) &= \text{Var} \left[Z_x - \left(\sum_{\alpha} \lambda_{\alpha} Z_{\alpha} + \sum_{\beta} \nu_{\beta} Y_{\beta} \right) \right] \\
&= C_Z(0) + \sum_{\alpha, \alpha'} \lambda_{\alpha} \lambda_{\alpha'} C_{\alpha\alpha'}^Z + \sum_{\beta, \beta'} \nu_{\beta} \nu_{\beta'} C_{\beta\beta'}^Y - 2 \sum_{\alpha} \lambda_{\alpha} C_{\alpha x}^Z \\
&\quad - 2 \sum_{\beta} \nu_{\beta} C_{\beta x}^{Z,Y} + 2 \sum_{\alpha\beta} \lambda_{\alpha} \nu_{\beta} C_{\alpha\beta}^{Z,Y}
\end{aligned} \tag{4}$$

Afin de minimiser la variance de l'erreur d'estimation (4), on forme le lagrangien sous contrainte d'égalité en utilisant la méthode des multiplicateurs de Lagrange sur les conditions de non-biais (3) :

$$\begin{aligned}
L = C_Z(0) &+ \sum_{\alpha, \alpha'} \lambda_{\alpha} \lambda_{\alpha'} C_{\alpha \alpha'}^Z + \sum_{\beta, \beta'} v_{\beta} v_{\beta'} C_{\beta \beta'}^Y - 2 \sum_{\alpha} \lambda_{\alpha} C_{\alpha x}^Z - 2 \sum_{\beta} v_{\beta} C_{\beta x}^{Z,Y} \\
&+ 2 \sum_{\alpha \beta} \lambda_{\alpha} v_{\beta} C_{\alpha \beta}^{Z,Y} + 2 \mu_0 \left(\sum_{\alpha} \lambda_{\alpha} - 1 \right) + 2 v_0 \sum_{\beta} v_{\beta} \\
&+ 2 \sum_j \mu_j \left(\sum_{\alpha} \lambda_{\alpha} g_{\alpha}^j - g_x^j \right) + 2 \sum_i v_i \left(\sum_{\alpha} \lambda_{\alpha} f_{\alpha}^i + \sum_{\beta} v_{\beta} f_{\beta}^i - f_x^i \right)
\end{aligned}$$

La dérivation du lagrangien par rapport aux poids inconnus aboutit à :

$$\begin{aligned}
\frac{\partial L}{\partial \lambda_{\alpha}} = 0 &\Rightarrow \sum_{\alpha'} \lambda_{\alpha'} C_{\alpha \alpha'}^Z - C_{\alpha x}^Z + \sum_{\beta} v_{\beta} C_{\alpha \beta}^{Z,Y} + \mu_0 + \sum_i v_i f_{\alpha}^i + \sum_j \mu_j g_{\alpha}^j \\
\frac{\partial L}{\partial v_{\beta}} = 0 &\Rightarrow \sum_{\beta'} v_{\beta'} C_{\beta \beta'}^Y - C_{\beta x}^{Z,Y} + \sum_{\alpha} \lambda_{\alpha} C_{\alpha \beta}^{Z,Y} + v_0 + \sum_i v_i f_{\beta}^i
\end{aligned}$$

Le modèle de cokrigage en dérive externe additive est donc :

$$\begin{bmatrix} C_{\alpha \alpha'}^Z & C_{\alpha \beta}^{Z,Y} & 1 & g_{\alpha}^j & 0 & f_{\alpha}^i \\ C_{\alpha \beta}^{Z,Y} & C_{\beta \beta'}^Y & 0 & 0 & 1 & f_{\beta}^i \\ 1 & 0 & 0 & 0 & 0 & 0 \\ g_{\alpha}^j & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ f_{\alpha}^i & f_{\beta}^i & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_{\alpha'} \\ v_{\beta'} \\ \mu_0 \\ \mu_j \\ v_0 \\ v_i \end{bmatrix} = \begin{bmatrix} C_{\alpha x}^Z \\ C_{\beta x}^{Z,Y} \\ 1 \\ g_x^j \\ 0 \\ f_x^i \end{bmatrix}$$