

Note technique

Travaux financés par le ministère chargé de l'environnement

ÉVALUATION DES CARTES ANALYSEES

Test d'une méthode d'évaluation proposée par FAIRMODE

Laure MALHERBE (INERIS), Charline PENNEQUIN (INERIS),
Maxime BEAUCHAMP (INERIS)

SYNTHESE

Le développement de méthodologies permettant de caractériser la performance des modélisations constitue l'une des activités centrales de FAIRMODE, le forum européen sur la modélisation de la qualité de l'air. Cette activité s'est portée récemment sur l'évaluation des cartes analysées, obtenues en combinant des données de modélisation et de mesure par une approche statistique ou géostatistique. L'évaluation de ces cartes s'effectue le plus souvent par validation croisée selon l'algorithme « leave-one-out ». Cette technique, aisée à mettre en œuvre, constitue un moyen pratique d'évaluer la précision relative de différentes cartographies, en particulier dans un cadre opérationnel. Les scores obtenus en chaque station sont toutefois sensibles à la répartition spatiale des sites de mesure. C'est pourquoi des évaluations plus approfondies - par tirages aléatoires - sont parfois réalisées de manière à caractériser plus finement la qualité de l'estimation sur l'ensemble du domaine d'étude.

FAIRMODE a ainsi proposé de tester une méthode d'évaluation fondée sur la réalisation de tirages aléatoires multiples afin de se faire une meilleure idée de ce type de technique et de sa possible valeur ajoutée par rapport à la validation croisée usuelle. Les résultats des tests effectués par le LCSQA sont consignés dans la présente note.

Les principales conclusions de ce travail et les questions associées s'accordent dans l'ensemble avec celles des autres participants à cet exercice. La méthodologie testée et la validation croisée présentent des résultats cohérents. Dans tous les cas étudiés, ceux-ci respectent l'objectif de qualité de l'outil d'évaluation *Delta tool*. A la différence de la validation croisée, la réalisation de nombreux tirages fournit des distributions de valeurs qui, dans la procédure de test proposée par FAIRMODE, ne sont que partiellement exploitées

(seul le cas associé à la plus grande erreur quadratique est retenu). Si l'usage de cette méthode devait être généralisé, il conviendrait d'approfondir les tests et l'exploitation qui en est faite et de s'accorder sur le choix de certains paramètres.

Dans l'attente de nouvelles discussions, la validation croisée, qui est cohérente avec la méthode testée et a l'avantage d'une plus grande rapidité de mise en œuvre, reste une méthode pertinente. Toutefois, en fonction de l'objectif du travail et pour les raisons mentionnées en conclusion, le LCSQA met en avant l'intérêt d'une évaluation par tirages aléatoires.

1. CONTEXTE

FAIRMODE (<http://fairmode.jrc.ec.europa.eu>), forum européen sur la modélisation de la qualité de l'air, a pour mission d'élaborer un cadre scientifique et technique harmonisé pour l'utilisation des outils de modélisation en Europe et de promouvoir l'usage de la modélisation dans la mise en œuvre des Directives. L'assurance qualité constitue l'un de ses principaux domaines d'activité, avec la conception de méthodologies d'évaluation des modélisations et le développement de l'application *Delta tool*, qui permet de réaliser des comparaisons entre modélisations et mesures et d'établir des scores de performance des modélisations.

Les données de modélisation considérées dans ces travaux sont issues de simulations déterministes mais peuvent aussi résulter d'une combinaison statistique ou géostatistique entre des données d'observation et des données simulées (voire d'autres variables auxiliaires). Dans ce cas, des précautions s'imposent quant à la manière de comparer les données de modélisation avec les mesures. Ce problème n'a pas encore été traité spécifiquement par FAIRMODE qui, à l'issue de la réunion technique de 2014, a décidé d'engager une première étude sur ce sujet. A la fin de l'année 2014, FAIRMODE a ainsi sollicité ses membres (laboratoires de référence des états membres, laboratoires de recherche) afin de tester une méthode d'évaluation des cartes analysées. Le LCSQA s'est porté volontaire pour participer à cet exercice et les conclusions qu'il en a tirées font l'objet de la présente note.

2. INTRODUCTION

Les cartes analysées sont obtenues en combinant des résultats de simulation avec des données de mesure selon une approche statistique ou géostatistique. Dans PREV'AIR, la méthode d'analyse employée est un krigeage des observations avec CHIMERE en dérive externe (Malherbe et Ung, 2009¹). Pour évaluer la fiabilité de ces cartes, des données d'observation différentes de celles qui ont servi à leur construction doivent être utilisées. Trois procédures sont couramment appliquées :

- La validation croisée dite *leave-one-out*, qui est la plus fréquente, consiste à retirer une station de mesure des données d'entrée de l'analyse, à estimer la concentration en ce point à l'aide des stations restantes et à répéter ce processus pour l'ensemble des stations. Les concentrations ainsi estimées sont alors comparées aux valeurs mesurées.
- La validation croisée à k plis (*k-fold*) consiste à diviser aléatoirement le jeu de données en k groupes de même taille. Chaque groupe de stations est successivement retiré de l'analyse et les concentrations sont estimées en ces points par krigeage à l'aide des *k-1* groupes restants. Notons que la validation croisée *leave-one-out* est un cas particulier de la validation croisée à k plis (cas où k est égal au nombre total de stations).
- L'évaluation sur un jeu de données indépendant consiste à exclure une partie des stations des données d'entrée de l'analyse et à estimer les concentrations en ces points.

Ces trois méthodes ont leurs limites. Les résultats des deux premières sont sensibles à la distribution spatiale des points de mesure ; pour la troisième, se pose la question du choix des sites à exclure.

La méthode proposée par FAIRMODE cherche à pallier ces défauts par une approche de type Monte-Carlo. Elle repose sur la réalisation d'un grand nombre de tirages aléatoires constitués chacun d'un sous-ensemble de stations d'analyse et du sous-ensemble complémentaire de stations de validation. Elle revient donc à répéter la seconde procédure (évaluation sur un jeu de données indépendant) de multiples fois, ce qui permet de s'affranchir du problème relatif à la sélection des stations de validation.

Les principes de la méthode sont décrits au chapitre 3. Les résultats obtenus, présentés au chapitre 4, sont comparés à ceux de la validation croisée. Les discussions de FAIRMODE consécutives aux tests réalisés sont évoquées en conclusion.

¹ 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->

3. METHODE

3.1 Principe

Soit un ensemble de N stations utilisées pour l'analyse d'un champ de concentration. La précision de cette analyse est évaluée par la réalisation d'un grand nombre (noté n) de tirages aléatoires :

- D'après le protocole établi par FAIRMODE2, 20% des stations sont sélectionnées à chaque tirage : ces sites n'interviennent pas dans l'analyse mais servent de points de validation, c'est-à-dire que les concentrations sont estimées en ces points afin de pouvoir être comparées aux observations.
- Il s'agit de tirages avec remise. Le protocole ne fixe pas le nombre n de tirages. Celui-ci peut être choisi par l'utilisateur selon sa propre expérience, de manière empirique ou selon des considérations probabilistes.
- A l'issue des n tirages, chaque station i doit avoir été tirée au moins une fois. k_i valeurs de concentration estimées, avec $k_i \geq 1$, lui sont donc associées. FAIRMODE propose de conserver la valeur qui correspond à l'erreur quadratique maximale.
- Le résultat se compose ainsi de N concentrations estimées qui sont comparées aux N données de mesure.

3.2 Application

La méthodologie précédemment décrite (§3.1) a été appliquée au cas d'étude suivant :

- Echelle spatiale : France
- Période : Année 2012
- Résolution temporelle : horaire
- Polluant : PM_{10}
- Données de modélisation : champs de concentration simulés par CHIMERE (environ 4 km de résolution spatiale)
- Données d'observation : mesures issues des stations rurales, périurbaines et urbaines de fond ($N=213$)
- Méthode d'analyse : krigeage des observations avec CHIMERE en dérive externe

En ce qui concerne les tirages aléatoires :

- Le nombre de tirages à réaliser n'étant pas fixé, des tests ont été réalisés pour trois valeurs de n : $n=200$, $n=300$ et $n=500$.
- Lorsque l'analyse est effectuée pour différentes dates, comme c'est ici le cas (8784 situations horaires), le protocole de FAIRMODE ne précise pas si les n tirages aléatoires sont réalisés une fois pour toutes au début de la période ou s'ils sont

² Document de travail du Working Group Assessment, élaboré d'après une proposition de C. Carnevale (Carnevale C., Finzi G., Pederzoli A., Turrini E., Volta M. Models & measurements for AQD: application to Northern Italy. FAIRMODE technical meeting, 2014).

réitérés pour chaque date. L'une ou l'autre de ces options a été retenue par les participants ; la seconde a été ici adoptée.

- Aucune solution n'a été encore mise en place afin de garantir que pour chaque date, chacune des stations serait sélectionnée au moins une fois comme station de validation. Toutefois, le respect de cette exigence a été systématiquement vérifié à posteriori.

Comme il a été mentionné en 3.1, pour chaque station i , sur l'ensemble des k_i tirages qui la concernent ($1 \leq k_i \leq n$), FAIRMODE propose de ne garder que la valeur estimée correspondant au cas le plus défavorable. Afin d'évaluer l'incidence de ce choix, trois valeurs de concentration ont été ici conservées : il s'agit des valeurs estimées correspondant respectivement au maximum, au 90^e centile (ou à une estimation de ce dernier si $k_i < 10$) et à la médiane des erreurs quadratiques sur ces k_i tirages. Une autre possibilité, qui allait au-delà du protocole, aurait été d'exploiter les k_i valeurs afin de tirer profit de toute l'information disponible. Ce type d'analyse est évoqué dans la discussion finale.

L'application de la méthodologie est illustrée ci-après (Figure 1 et Tableau 1). Les tirages aléatoires et le krigeage ont été réalisés à l'aide du logiciel R, version 3.1.1 (<https://www.r-project.org/>) et des modules RGeostat et gstat³.

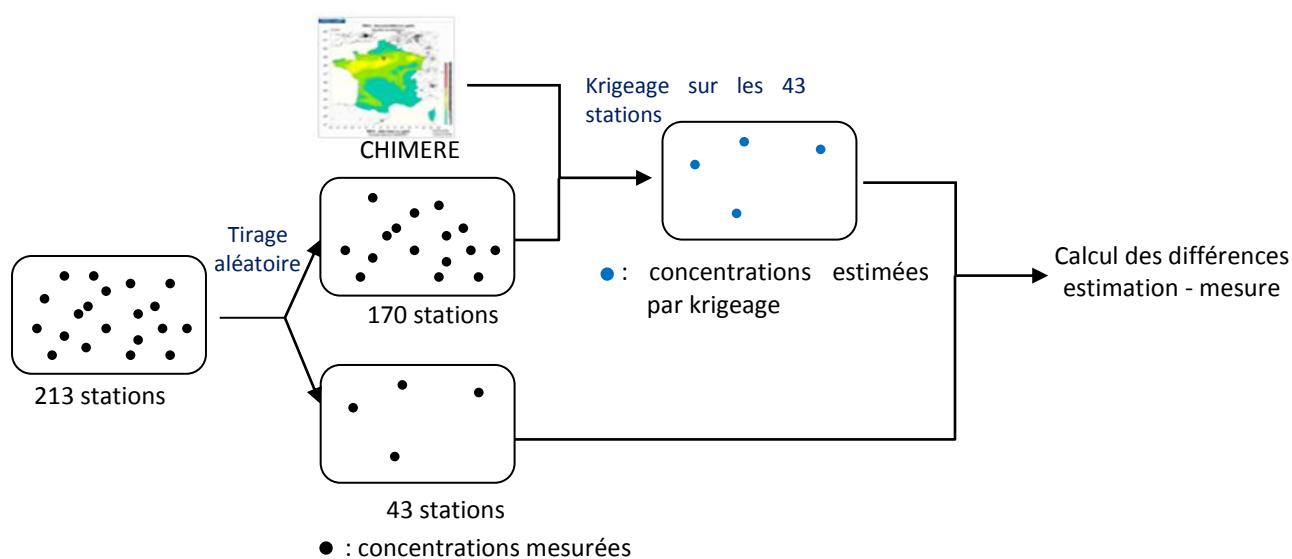


Figure 1 : Application de la méthode à l'échelle de la France. Exemple pour une date et heure donnée et un tirage aléatoire. Ce schéma est mis en œuvre autant de fois qu'il y a de tirages.

³ Renard D., Bez N., Desassis N., Beucher H., Ors F., Laporte F. RGeostats: The Geostatistical R package [Version 2]. MINES-ParisTech / ARMINES.
Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Computers & Geosciences, 30: 683-691.

Tableau 1 : Application de la méthode à l'échelle de la France. Exemple pour une date et heure donnée. Résultats obtenus en chaque station i à l'issue des n tirages.

Indices des k_i tirages ($1 \leq k_i \leq n$) pour lesquels la station i est sélectionnée comme station de validation	Valeur estimée en cette station	Valeur mesurée	Ecart quadratique	Valeurs d'estimation conservées pour l'évaluation
Tirage 2	Z_{i2}^*	Z_i	$(Z_{i2}^* - Z_i)^2$	Valeurs de Z_{i2}^* correspondant au maximum, au 90 ^e centile et à la médiane des k_i écarts quadratiques ci-contre
Tirage 3	Z_{i3}^*	Z_i	$(Z_{i3}^* - Z_i)^2$	
Tirage 13	Z_{i13}^*	Z_i	$(Z_{i13}^* - Z_i)^2$	
Tirage 55	Z_{i55}^*	Z_i	$(Z_{i55}^* - Z_i)^2$	
...	...	Z_i	...	
Tirage 432	Z_{i432}^*	Z_i	$(Z_{i432}^* - Z_i)^2$	

Les résultats de cette nouvelle approche ont été comparés à ceux de la validation croisée *leave-one-out* ainsi que de la validation croisée à k plis (*k-fold*). Pour cette dernière, qui est automatisée dans le module *gstat* de R, la valeur par défaut de $k=5$ (division du jeu de données en cinq groupes de même taille, soit 20% de la taille initiale) a été conservée.

Pour chacune des 213 stations considérées, et pour chaque valeur de n testée ($n=200$, $n=300$ et $n=500$), un tableau de sept séries temporelles est ainsi constitué : la série des données de mesure, à titre indicatif celle des données de CHIMERE, et les séries de données estimées par les différentes procédures d'évaluation (Tableau 2).

Tableau 2 : Application de différentes méthodes d'évaluation des cartes analysées. Exemple de résultats obtenus pour une station donnée.

Date (AAAAMMMJJHH)	Obs [$\mu\text{g.m}^{-3}$]	CTM [$\mu\text{g.m}^{-3}$]	CV_LOO [$\mu\text{g.m}^{-3}$]	CV_nfold [$\mu\text{g.m}^{-3}$]	MC_P50 [$\mu\text{g.m}^{-3}$]	MC_P90 [$\mu\text{g.m}^{-3}$]	MC_max [$\mu\text{g.m}^{-3}$]
2012010101	15	7.6	20.0	24.0	20.0	27.1	33.1
2012010102	12	7.9	16.0	23.2	18.8	20.8	22.5
(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)

Obs : valeur mesurée

CTM : valeur simulée par CHIMERE (interpolée à la station)

CV_LOO : valeur estimée par validation croisée de type *leave-one-out*

CV_nfold : valeur estimée par validation croisée de type *k-fold*, avec $k=5$

MC_P50, MC_P90, MC_max : valeurs estimées dans l'approche de type Monte-Carlo et associées respectivement à la médiane, au 90^e centile et au maximum des erreurs quadratiques calculées pour cette station

4. EXPLOITATION DES RESULTATS

A partir des séries temporelles obtenues (Tableau 2), des indicateurs annuels de comparaison modélisation-mesure ont été calculés station par station pour les différentes méthodes d'évaluation considérées. Le Delta tool a été utilisé en complément.

Remarque : les stations considérées dans ce travail présentent une couverture temporelle supérieure à 85%.

4.1 Indicateurs statistiques

Les quatre indicateurs de performance considérés (Tableau 3) sont issus du document de FAIRMODE *Guidance Document on Modelling Quality Objectives and Benchmarking* (Viaene et al., 2016).

Tableau 3 : Indicateurs de performance utilisés dans cette étude (*Modelling performance indicators*)

Racine carrée de l'erreur quadratique moyenne	$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (O_i - M_i)^2}$
Corrélation	$R = \frac{\sum_{i=1}^N (M_i - \bar{M})(O_i - \bar{O})}{\sqrt{\sum_{i=1}^N (M_i - \bar{M})^2} \sqrt{\sum_{i=1}^N (O_i - \bar{O})^2}}$
Biais moyen normalisé	$NMB = \frac{BIAS}{\bar{O}} \text{ where } BIAS = \bar{M} - \bar{O}$
Différence normalisée entre écarts-types	$NMSD = \frac{(\sigma_M - \sigma_O)}{\sigma_O}$

Les graphiques ci-après présentent, sous la forme de boîtes à moustaches, la distribution de ces scores sur l'ensemble des stations.

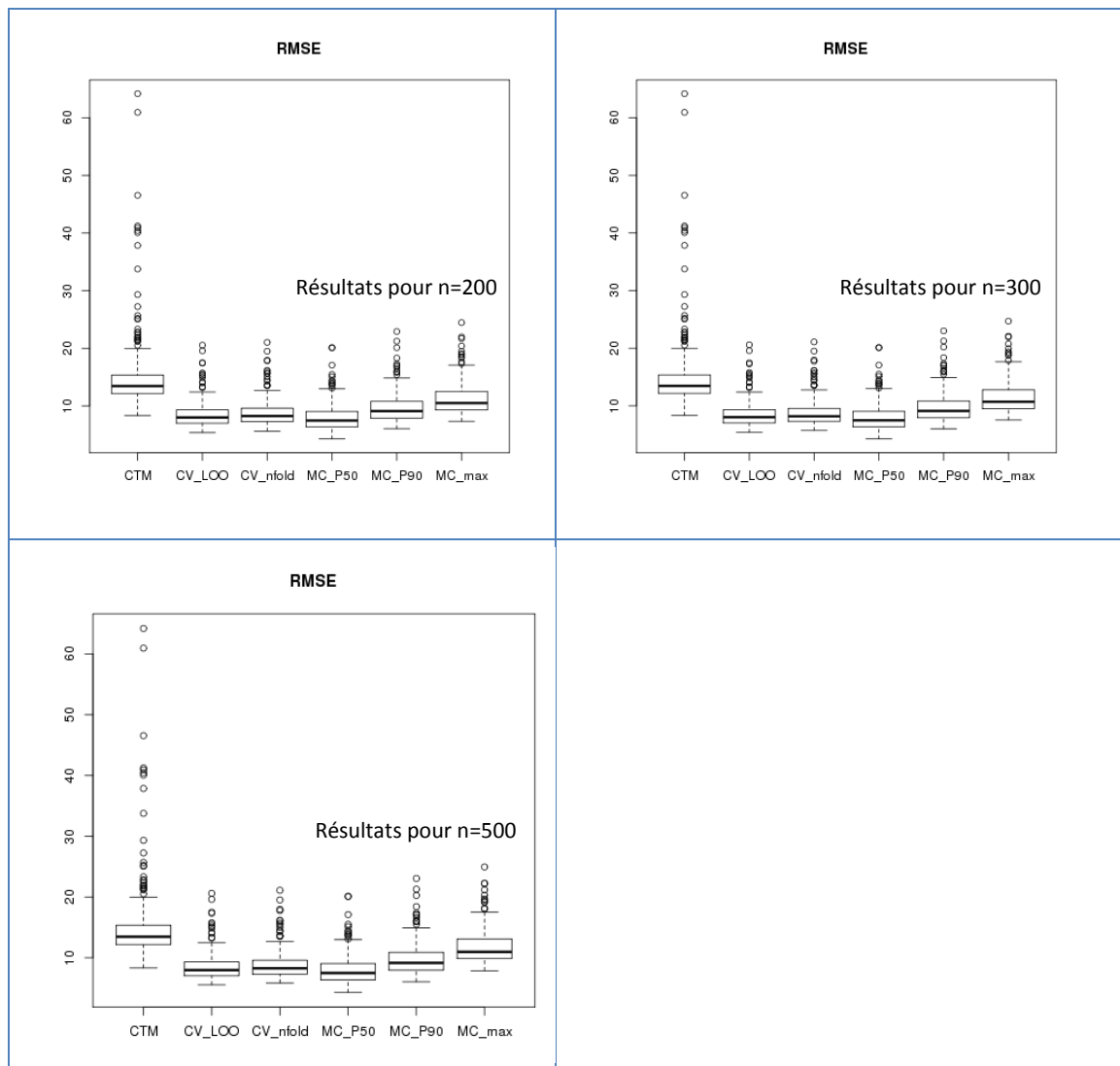


Figure 2 : Distribution de la RMSE [$\mu\text{g.m}^{-3}$] pour les différentes estimations considérées et les trois valeurs de n (nombre de tirages) testées. Chaque boîte à moustaches représente, de bas en haut, le minimum, le premier quartile, la médiane, le troisième quartile et le maximum* de la RMSE (*à l'exception des valeurs potentiellement singulières signalées par les cercles)

Quel que soit l'indicateur, aucune différence notable n'apparaît selon la valeur de n testée. A partir de la figure suivante, la présentation des résultats se limite au cas $n=500$.

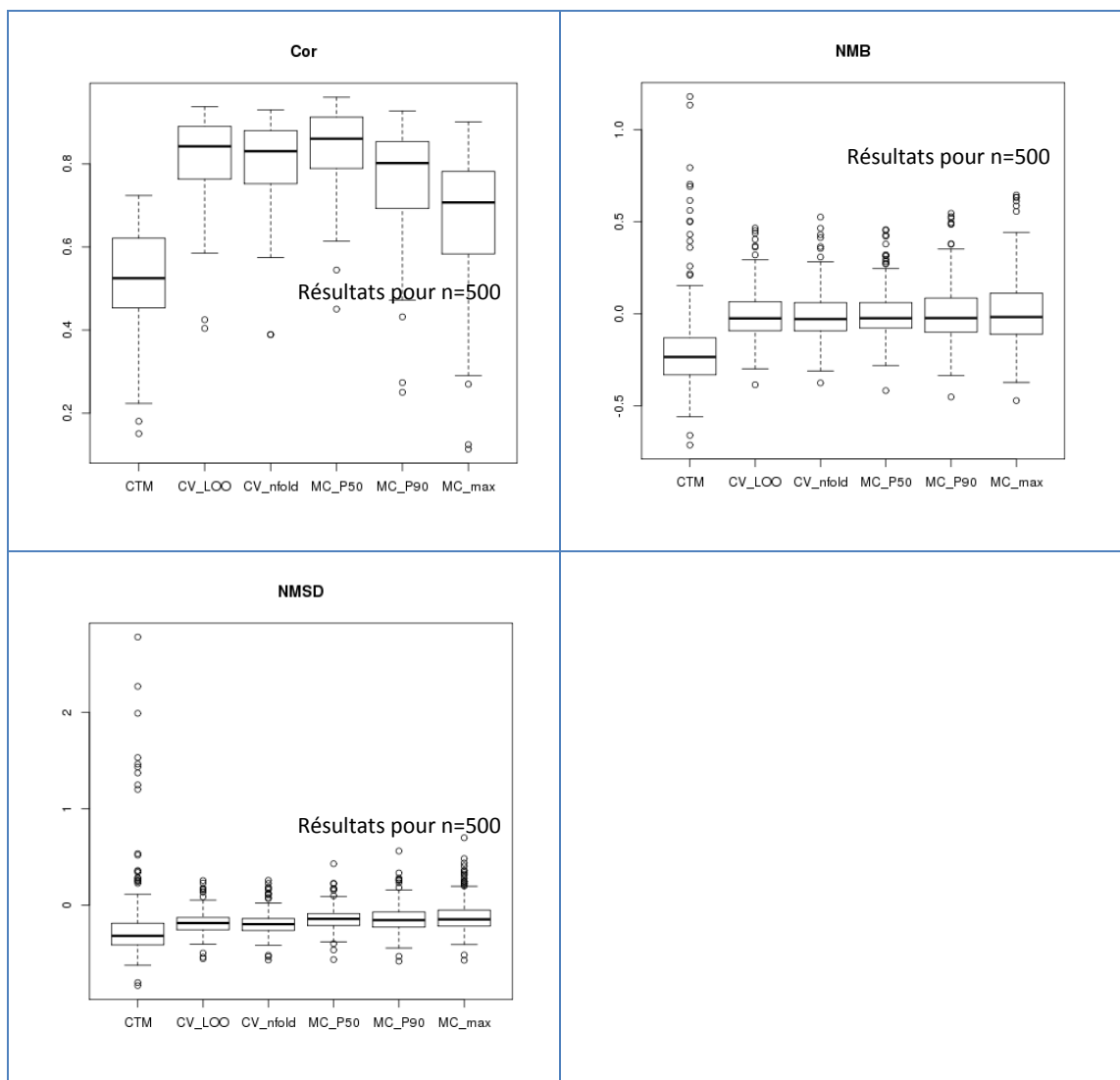


Figure 3 : Distribution de la corrélation temporelle (R), du biais moyen normalisé (NMB) et de la différence normalisée entre écarts-types (NMSD) pour les différentes estimations considérées.

La méthode d'évaluation qui produit les meilleurs scores est l'approche de type Monte-Carlo, lorsque la concentration estimée attribuée à chaque station est celle qui correspond à l'erreur quadratique médiane (MC_P50). La qualité des scores se dégrade logiquement lorsque le 90^e centile (MC_P90) et le maximum (MC_max) de l'erreur sont pris en compte.

Comme on pouvait s'y attendre, la validation croisée à k plis est légèrement plus pénalisante que la validation croisée *leave-one-out* puisque le nombre de stations utilisé pour chaque estimation est plus faible (80%*N pour la validation croisée à 5 plis, N-1 pour la validation croisée *leave-one-out*). Les distributions de scores issues de ces deux variantes sont cependant proches et cohérentes avec celles de la méthode par tirages aléatoires, se situant approximativement entre les performances des estimations associées à la médiane et au 90^e centile de l'erreur quadratique.

4.2 Diagramme de Taylor

Le diagramme de Taylor (Figure 4) est un outil graphique permettant de représenter de manière synthétique la proximité entre une ou plusieurs séries de données modélisées et une série de données de mesure.

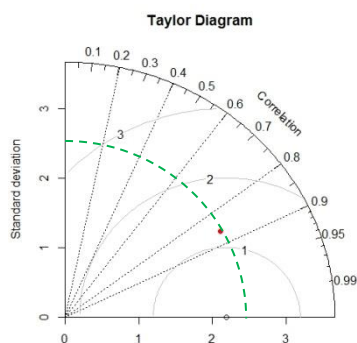


Figure 4 : Exemple de diagramme de Taylor

Le cas idéal correspond au point situé sur l'axe des abscisses : RMSE=0, corrélation =1, écart-type des données modélisées égal à l'écart-type des données de mesure, soit 2,2 dans cet exemple. La série des données modélisées, synthétisée par le point rouge, présente une RMSE comprise entre 1 et 2 (arcs en gris), une corrélation d'environ 0,86 avec la mesure et un écart-type de 2,5 (arc en pointillé vert).

Un diagramme de Taylor a été calculé pour chaque station individuellement puis sur l'ensemble des stations. Des exemples sont présentés ci-après (Figure 5). Ces figures s'accordent avec les boîtes à moustaches. Les résultats des différentes méthodes testées s'ordonnent de manière cohérente et sont homogènes entre les stations.

En général, les séries temporelles estimées dans l'évaluation par tirages aléatoires multiples (cf. Tableau 2) ont, en comparaison de la validation croisée, un écart-type, donc une variabilité légèrement plus grande. S'agissant de la RMSE et de la corrélation, le résultat dépend de la valeur d'estimation retenue à l'issue des tirages. Les performances des estimations sélectionnées d'après la médiane et le 90^e centile des erreurs quadratiques encadrent celles de la validation croisée. L'application du critère de l'erreur maximale diminue assez sensiblement les scores.

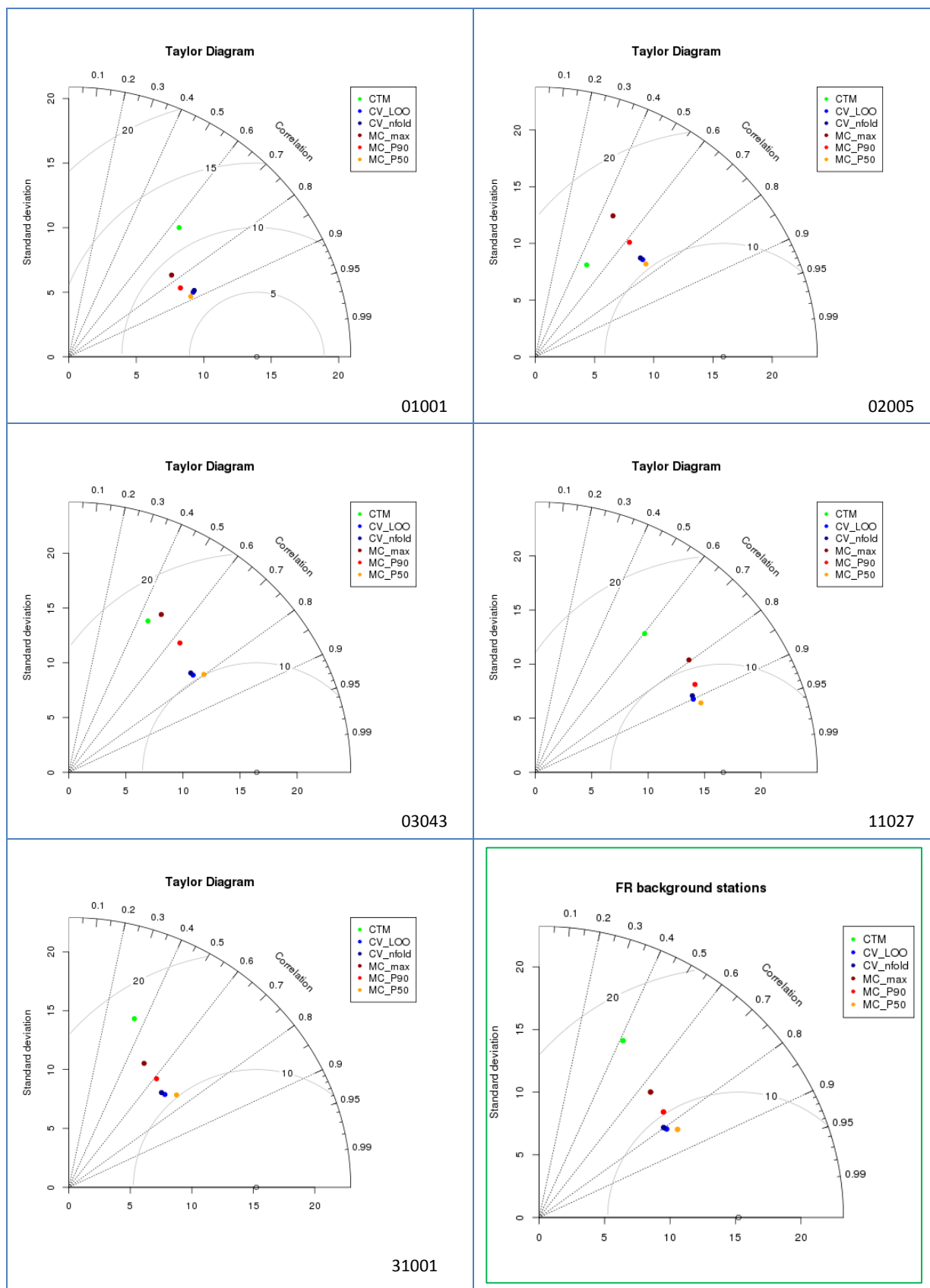
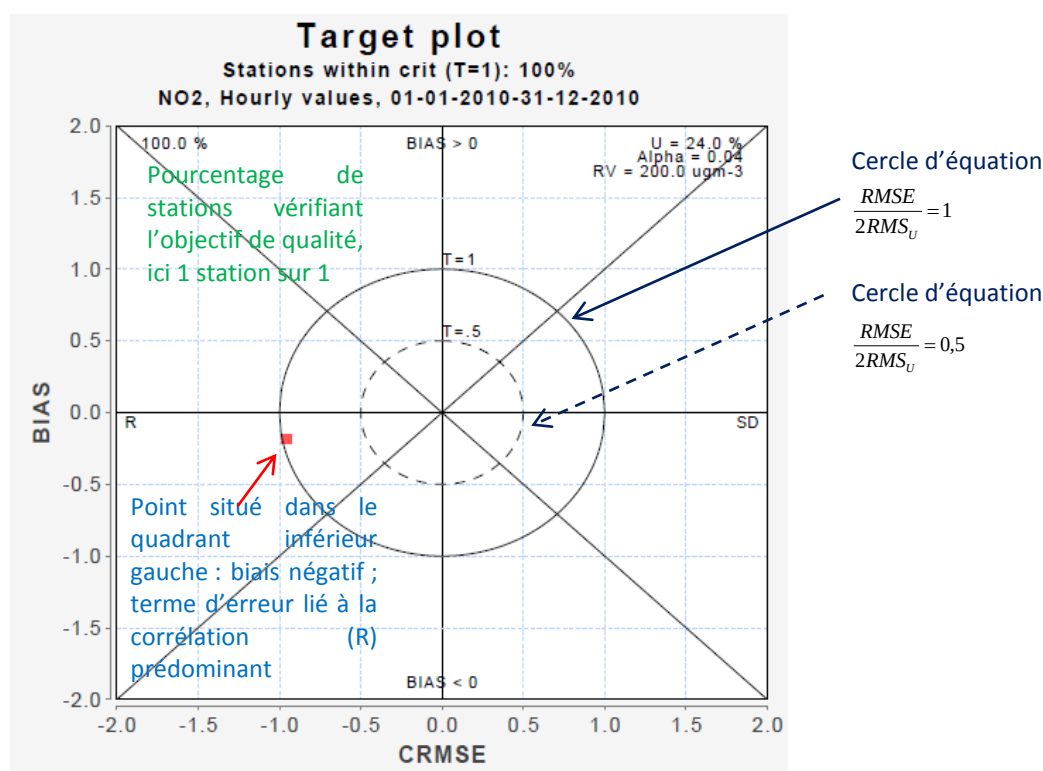


Figure 5 : Exploitation des résultats à l'aide du diagramme de Taylor. Exemple de quelques stations et diagramme calculé sur l'ensemble des stations.

4.3 Delta tool

Les données d'estimation dans le cas le plus restrictif (tirages aléatoires avec critère de l'erreur maximale) et en validation croisée *leave-one-out* ont été évaluées avec le Delta tool, outil d'évaluation des modélisations développé par le Centre Commun de Recherche de la Commission Européenne. La version 5.2 du Delta tool accessible via le site du LCSQA (<http://www.lcsqa.org/delta-tool-outil-evaluation-modelisations>) a été utilisée.

Le principal résultat de cet outil est le diagramme dénommé *Target plot*, qui synthétise différents critères de qualité (Figure 6). Des indicateurs de performance complémentaires sont également fournis. Pour une explication plus détaillée des graphes et des statistiques, on se reportera à la présentation réalisée en Commission de Suivi *Emissions, Modélisation et Traitement de Données* (Tognet, novembre 2015)⁴.



Objectif de qualité : $T \leq 1$ avec $T = \frac{RMSE}{2RMS_U}$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (O_i - M_i)^2} \quad . \quad O_i : \text{valeur observée} ; M_i : \text{valeur modélisée}$$

$$RMS_U = \sqrt{\frac{1}{N} \sum_{i=1}^N ((U(O_i)))^2} \quad . \quad U(O_i), \text{incertitude de mesure}$$

Figure 6 : Exemple de *target plot*

⁴ http://www.lcsqa.org/system/files/cs_emtd_25nov2015_evaluation_des_modelisations.pdf

Quelle que soit la méthode d'évaluation, toutes les stations respectent l'objectif de qualité (Figure 7) : l'indicateur de qualité T vérifie l'inégalité $T \leq 1$. Comme précédemment, la validation croisée conduit aux meilleurs scores : la majorité des stations vérifient $T \leq 0,5$, ce qui traduit une plus grande proximité entre estimation et mesure.

Les différents scores temporels (biais, corrélation, écart-type) et spatiaux (corrélation, écart-type) se situent dans leurs zones d'acceptabilité respectives qui ont été fixées par FAIRMODE (Figure 8).

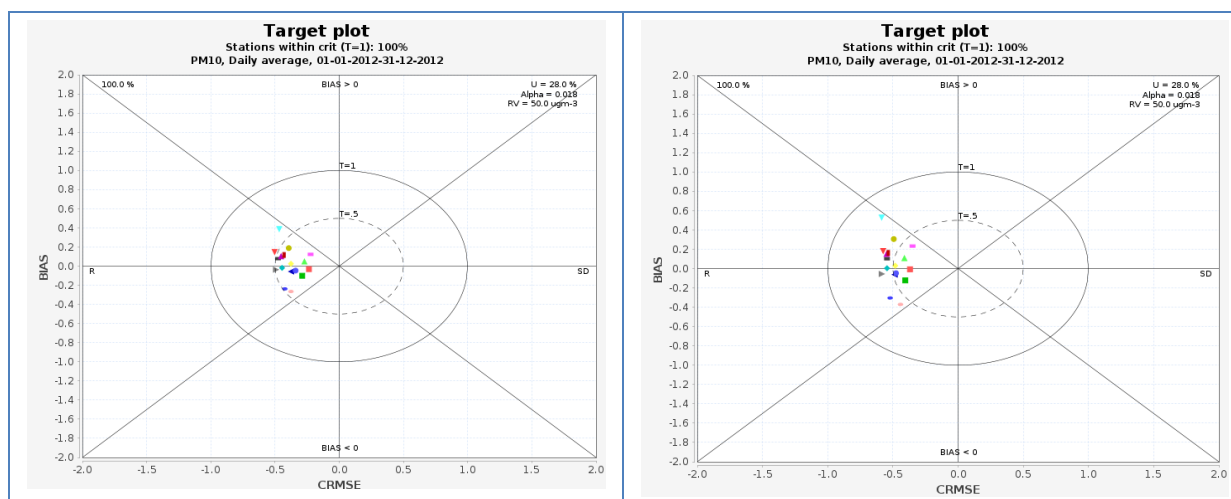


Figure 7 : Diagramme obtenu par le Delta tool. Exemple pour les stations du Sud-Ouest de la France (un point de couleur par station). A gauche : validation croisée *leave-one-out*; à droite : évaluation par tirages aléatoires.

Note : Pour des raisons de lisibilité, il n'a pas été possible de représenter toutes les stations françaises sur le même graphe. Les résultats obtenus pour les autres stations sont similaires. Tous les points se situent à l'intérieur du cercle $T=1$; elles satisfont donc à l'objectif de qualité.

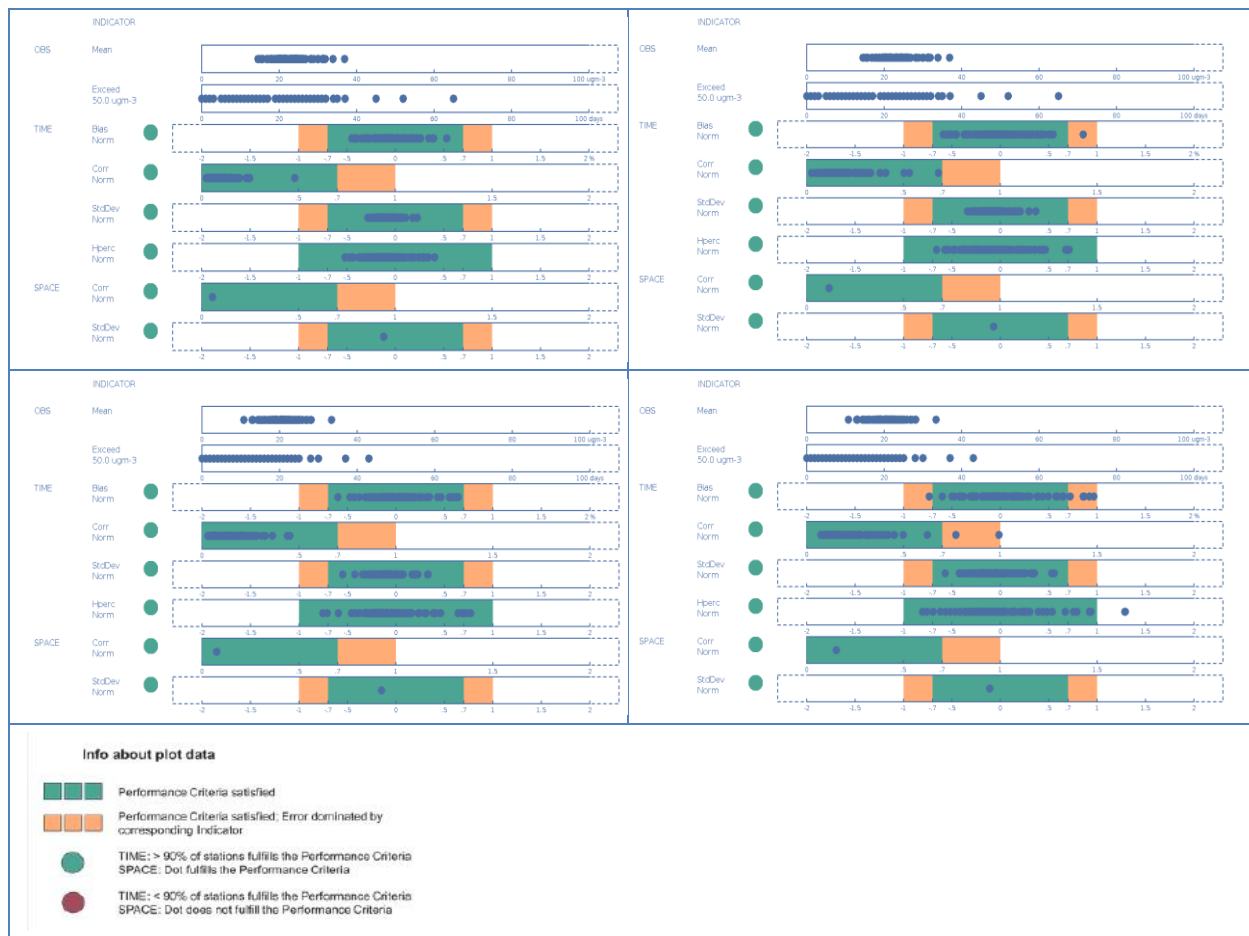


Figure 8 : Figures complémentaires issues du Delta tool. A gauche : validation croisée *leave-one-out* ; à droite : évaluation par tirages aléatoires. Ligne du haut : première moitié des stations, ligne du bas : seconde moitié (il n'a pas été possible de traiter simultanément toutes les stations françaises avec la version en ligne de l'outil, celles-ci ont donc été divisées en deux ensembles).

5. CONCLUSION ET DISCUSSION

5.1 Bilan des tests

Le rôle accru de la modélisation dans la surveillance de la qualité de l'air suppose le développement d'un système d'assurance et de contrôle de la qualité adapté aux méthodes de modélisation et à leur contexte d'utilisation. FAIRMODE, et depuis 2015, le CEN, contribuent à l'élaboration d'un tel système en Europe. Pour faciliter le transfert des méthodologies et recommandations issues de ces travaux, et afin de s'assurer qu'elles répondent aux besoins nationaux, le LCSQA participe aux activités de ces différents groupes européens et en rend compte aux AASQA via la Commission de Suivi *Emissions, Modélisation et Traitement de Données*. En réponse à une sollicitation de FAIRMODE, une méthodologie d'évaluation des cartes analysées fondée sur la réalisation de tirages aléatoires a été ainsi testée et comparée à la validation croisée.

Sa mise en œuvre a nécessité des développements informatiques mais n'a pas posé de problème technique particulier. Les questions qui se sont posées concernent plutôt des points de méthode ou le choix de certains paramètres, ainsi qu'il est évoqué dans la discussion. Le temps de calcul, supérieur à celui de la validation croisée, reste acceptable (4 à 10 h pour l'évaluation d'une année de données horaires selon le nombre de tirages).

Le cas étudié concerne la cartographie des PM₁₀ à l'échelle nationale sur l'année 2012. Des calculs supplémentaires ont été réalisés pour une autre année (2013) et pour d'autres polluants (NO₂, O₃, PM_{2,5}) mais n'ont pu être exploités dans le temps imparti pour cet exercice. Les résultats pourront être traités ultérieurement selon les orientations prises par FAIRMODE.

Deux autres équipes membres de FAIRMODE s'étaient portées volontaires pour effectuer des tests : VITO (Belgique) et l'Université d'Aveiro (Portugal). Les travaux de VITO ont porté sur la cartographie des PM₁₀ en 2009 à l'échelle de la Belgique et ceux de l'Université d'Aveiro sur la cartographie de l'ozone en 2005 à l'échelle du Portugal. Les résultats ont été présentés par chaque équipe lors d'une réunion technique de FAIRMODE (Aveiro, 2015) et ont fait l'objet d'une communication orale conjointe (congrès HARMO, 2016)⁵.

Plusieurs conclusions communes aux trois équipes ressortent de ces travaux :

- La méthodologie est automatisable sans difficulté majeure.
- Les résultats obtenus sont cohérents avec ceux de la validation croisée. Les différences entre les deux méthodes sont plus sensibles dans la présente étude que dans le cas traité par VITO.
- Si elle devait être appliquée par tous de manière homogène, la procédure à suivre nécessiterait quelques précisions. Cette remarque concerne notamment la manière d'effectuer les tirages lorsque l'évaluation porte sur une période et non une date unique.

3 Bino Maiheu, Laure Malherbe, Ana I. Miranda, Alexandra Monteiro, Claudio Carnevale and Stijn Janssen, 2016. Evaluation of a Monte Carlo-based validation technique for data assimilated air quality assessments within FAIRMODE. HARMO 17 conference, Budapest, 9-12 mai 2016.

- Des analyses complémentaires sur le nombre de tirages à réaliser (n) et le nombre minimal de fois où une station doit être sélectionnée ($n_{min_station}$) seraient utiles. En particulier la valeur $n_{min_station}=1$ pourrait être insuffisante. Une fois $n_{min_station}$ défini, VITO propose de recourir à la théorie des probabilités (problème « du collecteur de coupons ») afin de déterminer le nombre n de tirages nécessaire.
- Il serait intéressant d'examiner l'influence de la répartition spatiale des stations, selon qu'elle est plus ou moins homogène. Le choix de ne retenir pour chaque station que la valeur la plus pénalisante (associée à la plus grande erreur quadratique) est également discutable et ne permet pas d'exploiter entièrement l'information apportée par les nombreux tirages.

5.2 Discussion

A ce jour, FAIRMODE n'a pas statué sur les suites à donner à ces travaux. Le gain apporté par la méthodologie testée n'a pas encore été précisément identifié. Il faut en particulier évaluer l'intérêt de l'information produite par rapport au temps de calcul supplémentaire nécessaire. Dans l'attente de discussions futures, la validation croisée *leave-one-out*, qui permet une évaluation rapide et satisfaisante des cartographies, demeure une méthode recommandée par FAIRMODE et le LCSQA.

Cette conclusion mérite néanmoins d'être nuancée. La validation croisée *leave-one-out* est une méthode pratique lorsqu'il s'agit de contrôler l'adéquation d'un modèle géostatistique avec les données expérimentales ou de comparer entre eux différents estimateurs possibles. Si l'objectif est d'évaluer la fiabilité d'une cartographie issue d'un seul estimateur (dans notre cas le krigeage des observations avec CHIMERE en dérive externe), alors la réalisation de tirages aléatoires multiples présente plusieurs avantages. En effet, la validation croisée fournit généralement de bons résultats dans les zones où la densité de stations est élevée, ce qui influence favorablement les statistiques de performance. La réalisation de nombreux tirages aléatoires offre une évaluation plus juste de la précision de l'estimation sur l'ensemble du domaine. En outre, comme il est expliqué ci-après, une exploitation de tous les tirages, et non pas seulement de ceux qui correspondent aux erreurs maximales par station (ou à certains quantiles des erreurs), permet de mieux apprécier la variabilité de l'estimation.

En effet, les scores et outils d'évaluation considérés dans cette étude relèvent principalement du domaine de la modélisation déterministe. Ils servent à qualifier en moyenne des résultats de simulation relativement à des données de mesure. Dans cette étude, l'évaluation ne porte pas directement sur les sorties de CHIMERE mais sur les résultats d'une estimation géostatistique par krigeage qui, par définition, réalise un lissage. Ainsi des indicateurs tels que la comparaison des écarts-types entre les séries de concentrations estimées et mesurées ne sont-ils pas entièrement appropriés. La géostatistique introduit d'autres quantités destinées à caractériser le comportement de l'estimateur, en particulier le biais conditionnel $[E(Z_x | Z_x^*) - Z_x^*]$, qui représente l'écart moyen entre la valeur vraie (inconnue) Z_x et la valeur estimée Z_x^* connaissant cette dernière, ainsi que la dispersion conditionnelle $[Z_x - E(Z_x | Z_x^*)]$, qui mesure la précision de l'estimation connaissant la valeur estimée. L'évaluation par tirages aléatoires, qui permet de disposer d'un grand nombre de couples (Z_x, Z_x^*) (Figure 9) est un moyen d'étudier ces deux quantités et notamment d'en estimer la variance (qui doit être faible).

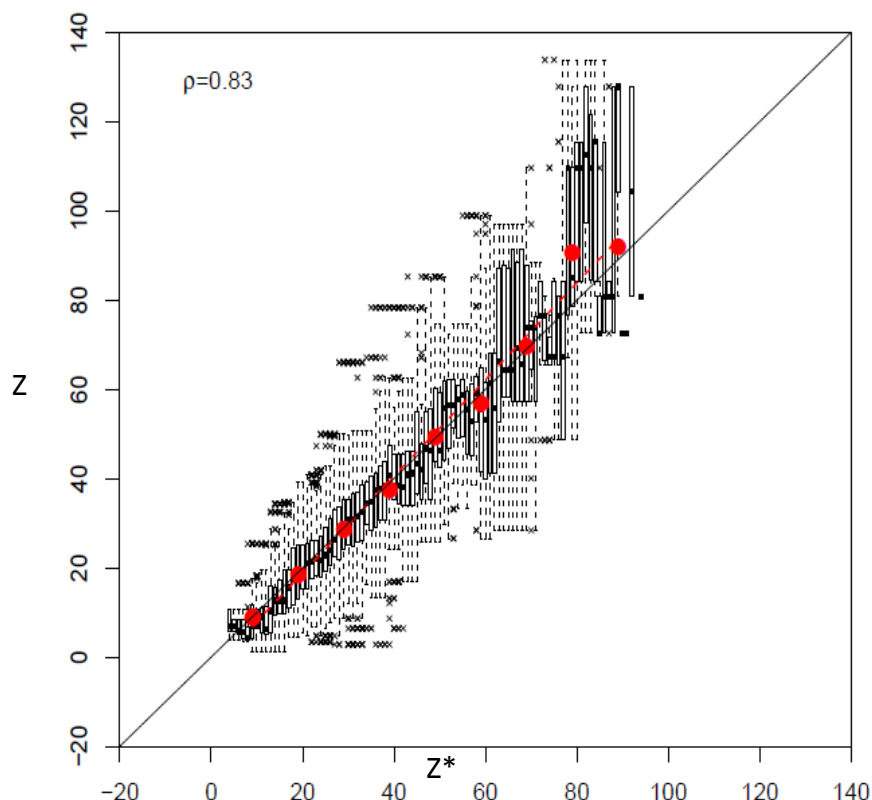


Figure 9 : Evaluation par tirages aléatoires multiples. Exemple issu de travaux internes.

Concentrations moyennes journalières de NO_2 mesurées et estimées par krigeage. La mesure est placée en ordonnée ; elle est supposée représenter la concentration vraie, inconnue hors des stations (l'erreur de mesure n'est pas prise en compte). Les valeurs estimées par krigeage sur l'ensemble des tirages ont été placées en abscisse ; elles sont regroupées par intervalle de $1 \mu\text{g.m}^{-3}$. Chaque rectangle gris est une boîte à moustaches qui, pour une valeur - ici un petit intervalle - de Z^* , montre la dispersion de Z . La courbe rouge représente la moyenne estimée de Z par classe de Z^* (classes allant de 5 à $95 \mu\text{g.m}^{-3}$ par pas de $10 \mu\text{g.m}^{-3}$) : c'est la courbe de régression établie à partir des moyennes expérimentales par classe (points rouges) ; sa distance par rapport à la bissectrice est un indicateur du biais. En pratique, ce type de figure permet de se rendre compte de l'exactitude et de la précision d'une cartographie en fonction du niveau de concentration estimé.