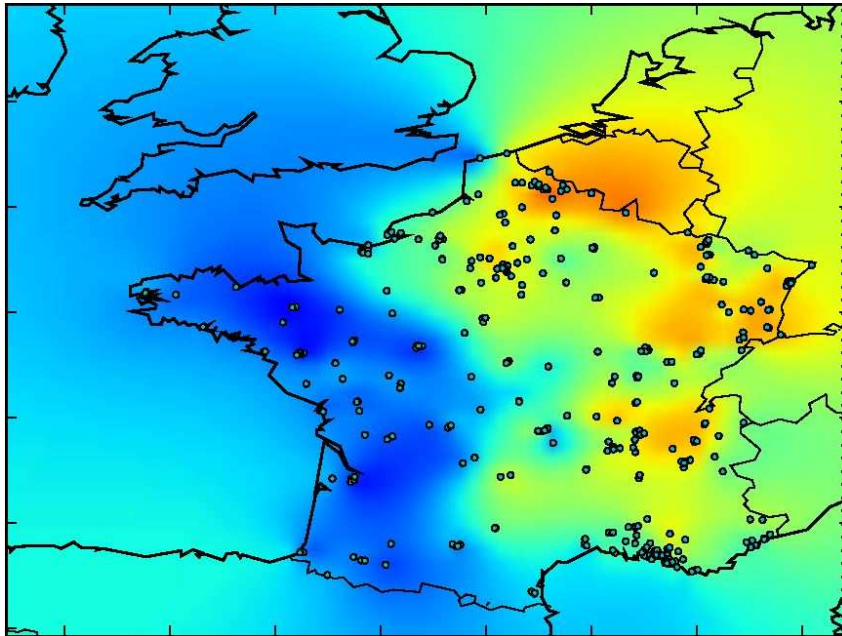


Projet d'ingénieur en laboratoire

Conception optimale de réseaux d'observation pour la qualité de l'air

Matthieu Chevallier
Ingénieur-élève des Ponts et Chaussées

20 juin 2008



Numéro rapport CEREa : 10

Table des matières

1	Introduction	4
2	Principe du <i>network design</i>	5
2.1	Les étapes	5
2.2	Plan du travail	6
3	L'interpolation statistique (<i>krigeage</i>)	7
3.1	La technique du krigeage	7
3.2	L'analyse statistiques des observations	9
4	Validation	14
4.1	Résultat du krigeage ordinaire	14
4.2	Validation	16
4.3	Sensibilité et <i>network design</i>	23
5	Conclusion	24
A	Indicateurs statistiques utilisés	25
B	Krigeage simple/krigeage ordinaire	25
C	Le modèle Polyphemus	27
D	Figures	29
D.1	Krigeage simple	29
D.2	Krigeage ordinaire	33

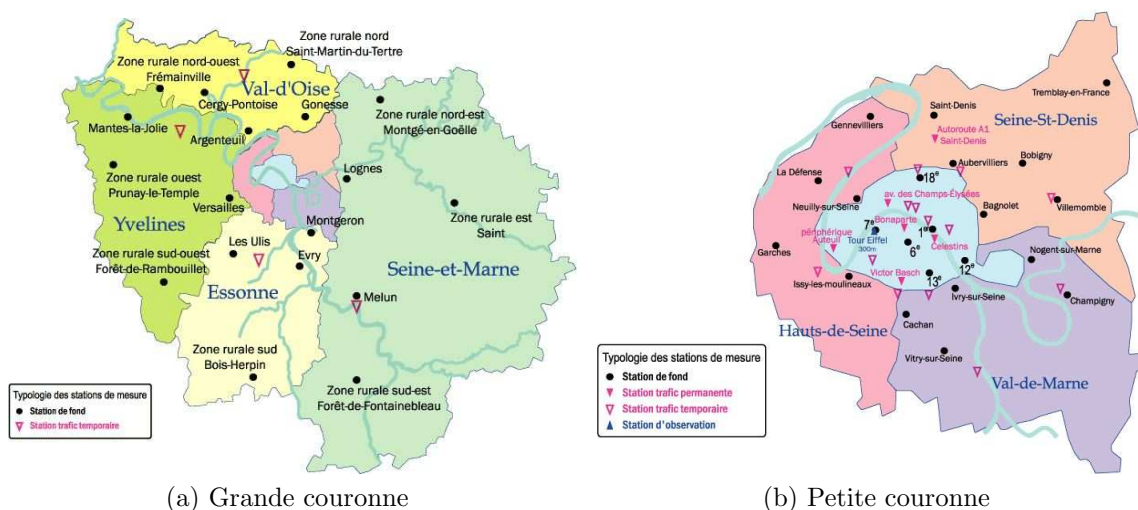


FIG. 1: Réseau de surveillance de la qualité de l'air Airparif

1 Introduction

L'intérêt d'un réseau d'observation de la qualité de l'air sur une région n'est plus à démontrer. L'Ile-de-France, et son réseau de mesure (figure 1) géré par l'agence Airparif en est un bon exemple. Les enjeux d'une bonne connaissance des concentrations des principaux polluants présents dans l'air sont reconnus et désormais incontournables :

- le contrôle du respect des normes en concentrations décrétees ;
- l'activation des mesures d'urgence rendue nécessaire par le dépassement de seuils critiques ;
- la surveillance à long terme de l'évolution des concentrations en polluants, dans le cadre notamment d'une réduction des émissions ;
- la constitution d'une base de données, pour la recherche, ou pour la prévision à court terme grâce à l'assimilation de données.

Les polluants récurrents couramment analysés sont l'ozone (O_3), les oxydes d'azote (NO , NO_2), le dioxyde de soufre (SO_2), le monoxyde de carbone (CO), les particules fines et les composés organiques volatils. La mesure de ces concentrations s'effectue principalement aux moyens de stations terrestres de mesure.

Le réseau, bien que nécessaire, a souvent été conçu sans plan d'ensemble. Le coût financier élevé des installations et du traitement des données pose la question d'une *rationalisation* de ces réseaux d'observation. L'objet du *Network design* est la conception optimale de réseaux d'observation d'un phénomène (radioactivité, pollution) à l'aide d'un critère quantifié.

La quantification de l'optimalité du réseau est liée à la nature du phénomène, ainsi que de l'opération à réaliser qui peut être :

- la réduction d'un réseau existant, tout en maintenant l'efficacité, dans un souci de réduction de coût ;
- l'accroissement d'un réseau existant, pour répondre à un besoin croissant en information ;
- la construction *ab nihilo* d'un réseau.

Il existe à ce jour peu d'applications à la conception de réseaux de mesures en environnement. On citera toutefois quelques travaux cherchant à optimiser le réseau existant de mesures de la pollution et de la pluviométrie de l'Environment Protection Agency des États-Unis. Il existe toutefois un corpus théorique élaboré.

Le CEREAS (Centre d'Enseignement et de Recherche en Environnement Atmosphérique) est un laboratoire commun ENPC/EDF R&D. Il travaille sur la modélisation de la pollution atmosphérique aux échelles locale (rue, ville, site industriel), régionale (pollution photochimique) et continentale (pollution transfrontière). Il développe et opère son propre système de modélisation de la qualité de l'air, *Polyphemus*. Le groupe assimilation de données/modélisation inverse, auquel se rattache l'activité *network design*, est partie intégrante du projet CLIME (Chaînes Logicielles Intégrées pour la Modélisation en Environnement) de l'INRIA.

Le projet *network design* est lui-même un projet au sein du Réseau de Recherche sur le Développement Durable (R2DS), au sein de la thématique *modélisations d'interactions*. Créé en 2006 à l'initiative du Conseil Régional d'Ile-de-France et géré par le CNRS, ce réseau soutient des projets de recherche sur les problèmes de l'environnement et mobi-

lise autour du développement soutenable une communauté scientifique interdisciplinaire (sciences de l'univers, du vivant, de l'ingénieur, sciences humaines et sociales). L'objectif du projet *network design* est de proposer des méthodes et de les appliquer pour le cas de la modélisation de la qualité de l'air à l'échelle régionale/nationale. Les partenaires du CEREAs sont l'INERIS (Institut National de l'Environnement Industriel et des Risques), l'ADEME (Agence De l'Environnement et de la Maîtrise d'Énergie) et Airparif.

2 Principe du *network design*

2.1 Les étapes

Le *Network design* est un volet récent de l'assimilation de données et de la modélisation. Les premiers travaux sur le sujet concernaient déjà un réseau de mesure de l'ozone dans le Midwest des États-Unis [5], et envisageait la réduction d'un réseau en zone urbaine essentiellement, et l'augmentation d'un réseau en zone rurale pour une meilleure couverture. Les auteurs proposaient alors une approche par les techniques de géostatistiques, utilisant le *krigeage* à l'aide de variogrammes et covariogrammes stationnaires. Le CEREAs, en partenariat avec l'IRSN a mené une étude pour la mise en place d'un réseau de surveillance des concentrations en radionucléides dans l'atmosphère [1], un tel réseau permettant la diffusion rapide d'une alerte en cas d'accident, ainsi que la prise en compte de l'ampleur de l'accident éventuel par une mesure correcte de l'étendue du nuage radioactif, et la prévision de son évolution. Les corrélations étaient alors calculées à partir d'une base de données d'événements singuliers, ajustés par un modèle exponentiel :

$$\Gamma_t(d) = A_t \exp\left(-\frac{d}{L_t}\right).$$

Le caractère non stationnaire des phénomènes étudiés apparaissaient alors dans les paramètres de fitting, en particulier la longueur de corrélations L_t , supposée croissante à partir de l'heure de l'accident. Toutefois, la dispersion d'un tel phénomène est très singulière, à la différence de l'ozone. Les fonctions statistiques sont calculées à partir d'une base de donnée d'accidents, créant ainsi une base statistique artificielle à l'aide du modèle. La construction du réseau fait appel à la technique du recuit simulé, avec une fonction de coût du type « norme de Hölder ».

Le principal enjeu du *network design* est le choix d'un critère de quantification de l'optimalité du réseau, principalement de sa capacité à interpoler. La formalisation de tels critères d'optimalité est ardue, et résulte de la quantification des objectifs du réseau. Dans le cas de l'ozone, la surveillance des concentrations près des zones denses en population et près des zones de forte émission est naturellement prioritaire. Les critères se focalisent sur les statistiques mises en oeuvre pour l'interprétation des données fournies par le réseau. Un réseau, réparti discrètement sur un territoire à décrire entièrement, doit remplir un certain nombre de conditions de nature statistiques : les données recueillies doivent permettre

- l'estimation de valeurs moyennes à une précision déterminée ;
- l'interpolation de la carte des concentrations, à une précision suffisante (prévision spatiale) ;

- l’observation directe, ou le diagnostic des maxima locaux de concentrations ;
- la continuité du réseau en cas de panne d’une station.

L’estimation de l’optimalité du réseau se complique si l’on prend en compte le fait que les émissions diffèrent entre jour et nuit, ou entre différentes heures du jour.

En terme de modélisation, le processus de *network design* se déroule en trois étapes. Tout d’abord, les statistiques du réseau sont évaluées, reflétant les phénomènes étudiés. Ensuite, la technique d’interpolation doit être choisie de manière à reconstruire au mieux la carte des concentrations à partir des mesures, et doit tenir compte de la complexité et de la variabilité de la répartition. Enfin, la capacité à interpoler est évaluée à l’aide d’une fonction de coût à définir. Le réseau peut alors être « construit » (c’est-à-dire étendu, réduit ou créé) de manière itérative, à l’aide d’un algorithme de recuit simulé.

La technique d’interpolation choisie ici est l’interpolation statistique, plus connue sous le nom de *krigeage*. Elle consiste à chercher le meilleur estimateur d’un champ observé, en minimisant la variance de l’erreur modèle. Cette méthode repose sur la technique *BLUE* (Best Linear Unbiased Estimator) : l’estimateur est combinaison linéaire des stations retenues. L’estimateur obtenu est fonction des corrélations entre les différentes stations du réseau, à travers les matrices de covariance.

2.2 Plan du travail

Nous nous sommes concentrés sur le problème de l’amélioration du krigeage du champ d’ozone par les observations du réseau BDQA (Base de Données pour la Qualité de l’Air). Nous avons, dans un premier temps, défini et implémenté une méthode de krigeage, le krigeage ordinaire, que nous avons testé contre le krigeage simple en utilisant des statistiques *a priori*. Nous nous sommes ensuite concentrés sur la modélisation fine des statistiques, et cela en deux temps. Dans un premier temps, nous avons implémenté le calcul des variogramme, corrélogramme et covariogramme, cette dernière matrice étant la matrice statistique utilisée pour le krigeage ordinaire. Dans un second temps, il s’est agi de choisir des modèles simples de covariogrammes, les meilleurs possibles. Nous avons cherché ici à évaluer l’influence de la variabilité temporelle de l’ozone, en calculant des covariogrammes pour chaque heure, puis par tranches de deux, trois, quatre ou six heures. Enfin, la validation a permis d’évaluer la capacité d’un sous-ensemble du réseau à interpoler les observations des stations exclues de ce sous-réseau.

Le travail a été effectué sous la direction de Marc BOCQUET, avec Lin WU, chercheurs au CEREAS. Le modèle utilisé pour l’assimilation des observations est le modèle eulérien de chimie-transport Polair3D, composante de la plateforme Polyphemus, développée par le CEREAS [4] et présenté en annexe C. Les drivers ont été modifiés par Lin WU pour pouvoir être utilisés pour le *network design*. Le jeu de données statistiques tests concerne les observations des 678 stations du réseau BDQA qui couvre le territoire français sur le mois de juillet 2001. Les langages utilisés sont le C++ pour la partie modèle et krigeage, l’essentiel de Polyphemus étant codé en C++, et le Python pour la partie *fitting* des fonctions de covariance, ainsi que pour l’exploitation des résultats.

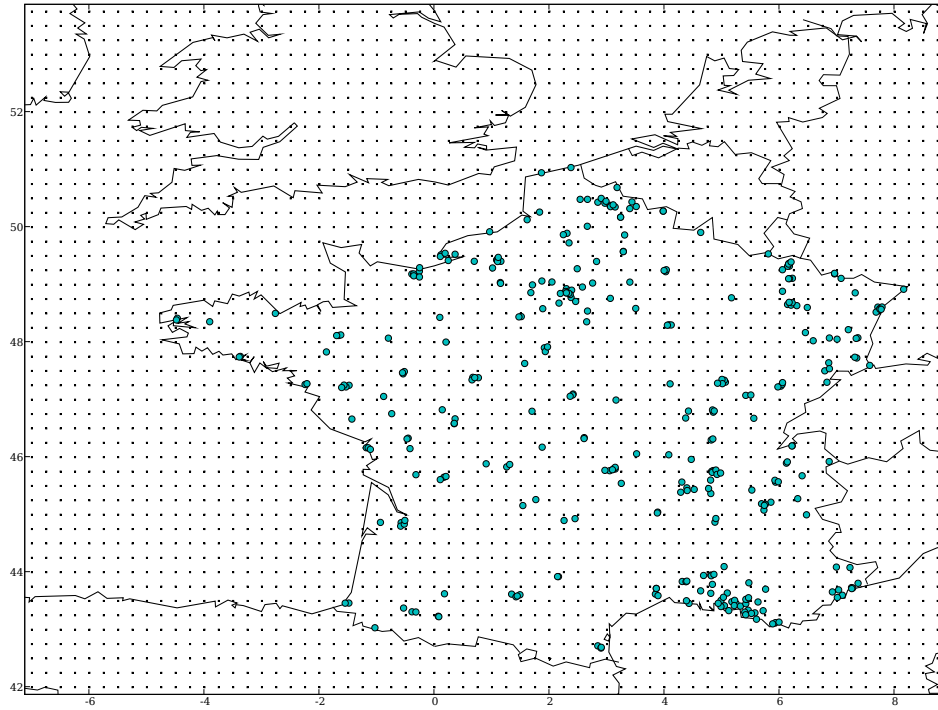


FIG. 2: Domaine étudié et réseau BDQA

3 L'interpolation statistique (*krigeage*)

3.1 La technique du krigeage

Principes généraux Le but du krigeage est d'obtenir une estimation acceptable d'une grandeur à partir de données issues de l'observation. Il s'agit d'un problème typique en géologie et en exploitation minière : D.G. Krige, ingénieur des mines en Afrique du Sud, est à l'origine de la géostatistique et de la méthode d'interpolation qui porte son nom. Cette méthode est basée sur un ensemble de données statistiques concernant les observations, ces données devant être déterminées *a priori*. On part du constat aisé qui voudrait que les observations les plus proches du point de l'estimation se voient donner un poids plus important pour cette estimation. On utilise typiquement un estimateur linéaire, solution d'un problème d'optimisation de moindres carrés.

Le calcul précis des estimateurs est donné en annexe B. Rappelons ici qu'il s'agit d'obtenir un estimateur $\hat{Y}$ des concentrations Y sur $n - p$ stations d'un réseau de n stations, cet estimateur étant combinaison linéaire des observations Z obtenues sur les p stations restantes du réseau potentiel. On cherche alors une matrice $\mathbb{L}$ qui minimise la variance de l'erreur $\text{Tr} [\mathbb{E}((Y - \mathbb{L}Z - \kappa)(Y - \mathbb{L}Z - \kappa)^T)]$.

Krigeage simple Si on connaît parfaitement le champ moyen $\mathbb{E}(Y)$, la minimisation de la variance donne :

$$\hat{Y} = \mathbb{E}(\Delta Y \Delta Z^T) \mathbb{E}(\Delta Z \Delta Z^T)^{-1} (Z - \mathbb{E}(Z)) + \mathbb{E}(Y).$$

Krigeage ordinaire Dans le cas où on suppose la moyenne inconnue, mais uniforme sur le domaine, on résout un problème d'optimisation sous contrainte : la combinaison linéaire est convexe, c'est-à-dire

$$\hat{Y}(s_i) = \sum_{j=1}^p \mathbb{L}_{i,j} Z(s_j), \quad i = p+1, \dots, n$$

avec

$$\sum_{j=1}^p \mathbb{L}_{i,j} = 1.$$

La solution s'écrit :

$$\begin{aligned} \Lambda &= \frac{\mathbb{E}(\Delta Y \Delta Z^T) \mathbb{E}(\Delta Z \Delta Z^T)^{-1} - \mathbb{I}}{\Gamma^T \mathbb{E}(\Delta Z \Delta Z^T)^{-1} \Gamma} \Gamma \\ \hat{Y} &= [\mathbb{E}(\Delta Y \Delta Z^T) - \Lambda \Gamma^T] \mathbb{E}(\Delta Z \Delta Z^T)^{-1} Z. \end{aligned}$$

La solution est, dans tous les cas, fonction des matrices de covariance :

- des mesures effectuées sur les sites instrumentés,

$$\mathbb{E}(\Delta Z \Delta Z^T) = \mathbb{E}((Z - \mathbb{E}(Z))(Z - \mathbb{E}(Z))^T) = (\text{Cov}(Z(s_i), Z(s_j)))_{i,j}$$

- des mesures issues des sites instrumentés avec les mesures issues des sites non-instrumentés,

$$\mathbb{E}(\Delta Y \Delta Z^T) = \mathbb{E}((Y - \mathbb{E}(Y))(Z - \mathbb{E}(Z))^T) = (\text{Cov}(Y(s_i), Z(s_j)))_{i,j}.$$

Pour commencer, regardons la comparaison entre le modèle (simulation Polyphemus) et le krigeage simple effectué à l'aide d'un modèle de covariance *a priori* :

$$\text{Cov}(Z(s_i), Z(s_j)) = 400 \exp\left(-\frac{\|s_i - s_j\|}{100 \text{ km}}\right),$$

en supposant la moyenne à 0. Quelques résultats de l'animation sont données en figure D.1. Dans la suite nous nous concentrons sur le krigeage ordinaire.

Les corrélations, données en figure (3.1), sont plutôt bonnes. On comprend toutefois la limite du krigeage simple aux longues distance, où le champ est homogénéisé à 0, notre moyenne.

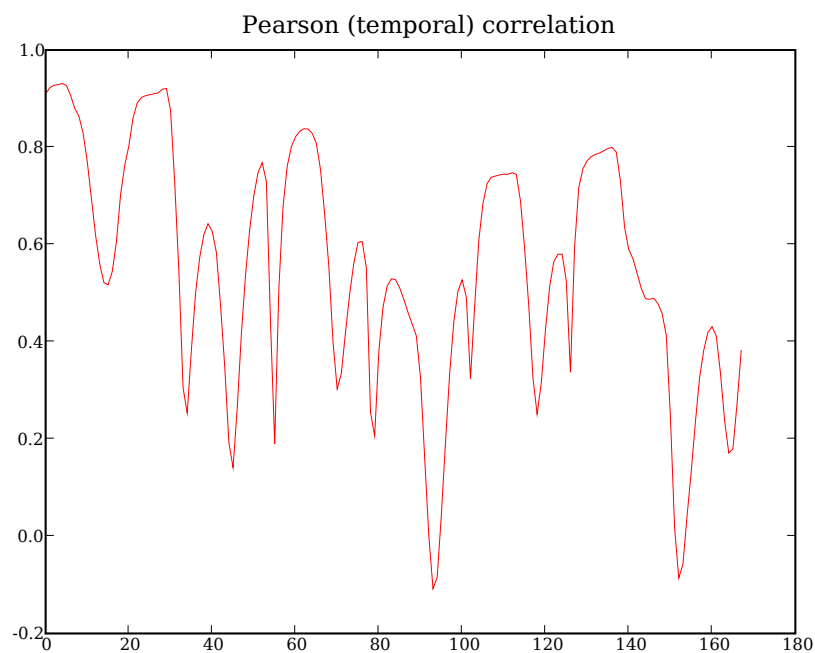


FIG. 3: Corrélation de Pearson : krigeage simple (abscisse : temps, ordonnée : corrélation)

3.2 L'analyse statistique des observations

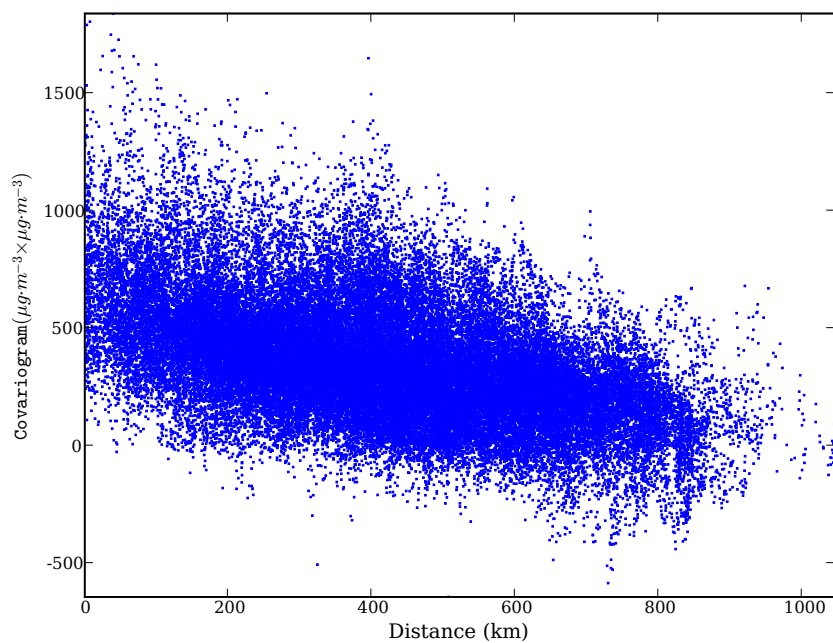


FIG. 4: Nuage de covariances, à partir des observations sur un mois.

Les covariances entre deux stations quelconques du réseau sont représentées en figure 4. Nous nous attachons à représenter les matrices de covariance à l'aide d'un modèle « continu ». Nous supposons, classiquement, que les covariances, représentées par une fonction de covariance ou *covariogramme* C , sont isotropes, c'est-à-dire que C ne dépend de s_i et s_j que par la distance $\|s_i - s_j\|$:

$$\text{Cov}(Z(s_i), Z(s_j)) = C(\|s_i - s_j\|).$$

Le choix de la fonction C , c'est-à-dire d'un covariogramme admissible, est assez étendu. Toutefois il doit obéir à une propriété de positivité, liée au fait que la matrice $\mathbb{E}(\Delta Z \Delta Z^T)$ est définie et positive. Un certain nombre de modèles classiques existe toutefois, cités dans [3]. Le plus couramment utilisé est un modèle exponentiel :

$$C(d) = \sigma^2 \exp\left(-\frac{d}{L}\right),$$

où σ^2 est la variance *a priori*, et L la portée, ou longueur de corrélation. Cette dernière grandeur est importante : la longueur de corrélation donne une estimation du rayon dans lequel un point quelconque de l'espace est porteur d'information pour toute estimation. Le modèle exponentiel implique que les mesures en deux points situés à grande distance l'un de l'autre seront parfaitement décorrélées, ce qui n'est pas tout à fait le cas. Le cycle journalier de l'ozone implique des corrélations évidentes à grande distance, typiquement entre la région parisienne et la côte méditerranéenne, régions fortement urbanisées. Il nous faut tenir compte de ce « bruit de fond ».

Nous avons choisi un modèle *exponentiel translaté* :

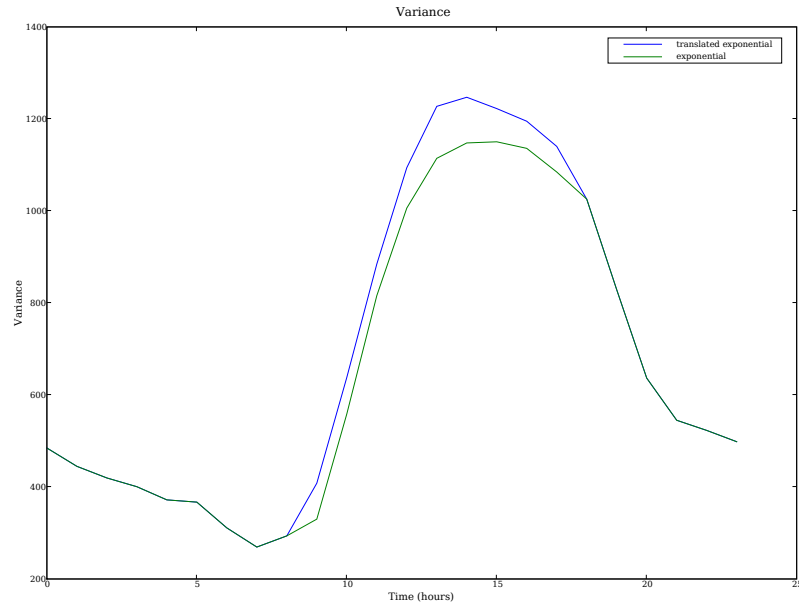
$$C(d) = a \exp\left(-\frac{d}{L}\right) + c,$$

qui est un covariogramme admissible comme somme de covariogrammes admissibles si c est positif. Les grandeurs statistiques caractéristiques d'un tel modèle sont la variance « a priori » $\sigma^2 = a + c$ et la longueur de corrélation L .

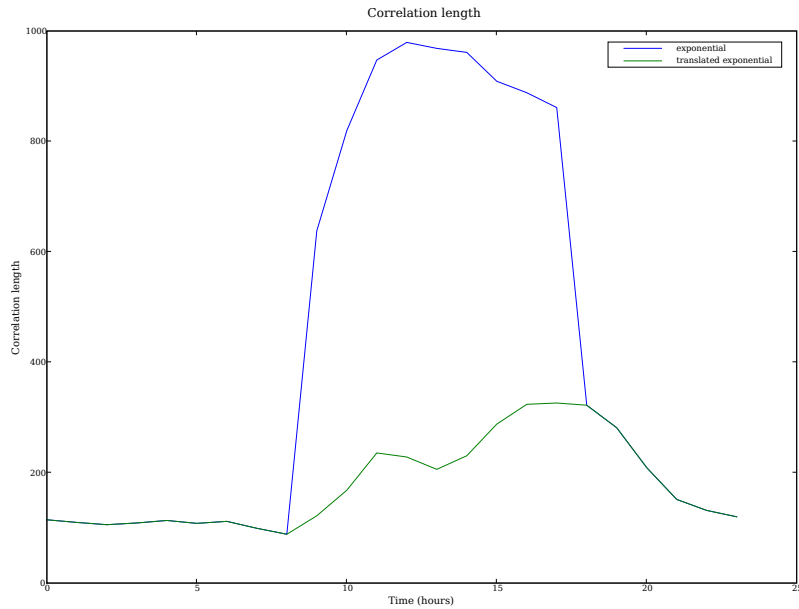
Le *fitting* du covariogramme a été effectué avec l'aide d'un algorithme de moindres carrés (méthode `leastsq`), à partir des distances $(d_i)_{i=0, \dots, N-1}$ entre les couples de stations (au plus $n(n-1)/2$ valeurs) :

$$\min_{a, L, c} \sum_{i=0}^{N-1} \left(C(d_i) - a \exp\left(-\frac{d_i}{L}\right) - c \right)^2.$$

Nous avons choisi d'ajuster les paramètres de façon à tenir compte des variations temporelles de la concentration mesurée en chaque station. Les paramètres sont donc calculés chaque heure, en moyennant les observations de chaque jour pendant le mois de juillet 2001. La figure (5) présente les variations de L et σ^2 alors constatées.



(a)



(b)

FIG. 5: Évolution des paramètres fittés (variance a priori et longueur de corrélation) : comparaison modèle exponentiel/modèle exponentiel translaté

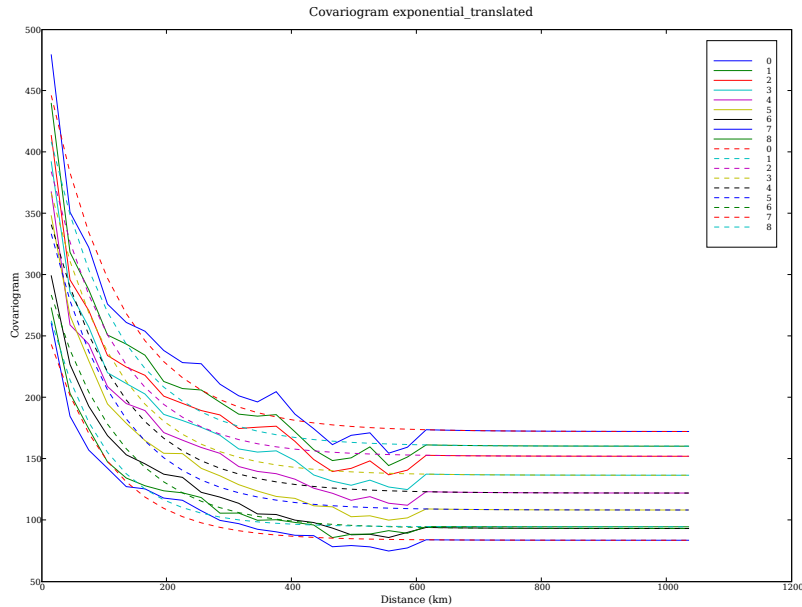
L'existence de ces variations nous conforte dans le fait que ces paramètres de covariance varient au cours du temps. De plus, on observe un maximum de longueur de corrélation à 15 heures, correspondant au pic d'ozone journalier. Cette longueur égale à 400 kilomètres montre qu'à cette heure, on a un grand nombre de phénomènes commun sur une grande

partie du territoire : insolation maximale et zénithale, déplacements journaliers... Toutefois la longueur de corrélation varie assez peu la nuit, avec une valeur assez faible. La variance σ^2 varie de manière remarquable, maximale aussi à 15 heures, puis décroissante durant la fin d'après-midi, la nuit, et croît à partir de 7 heures.

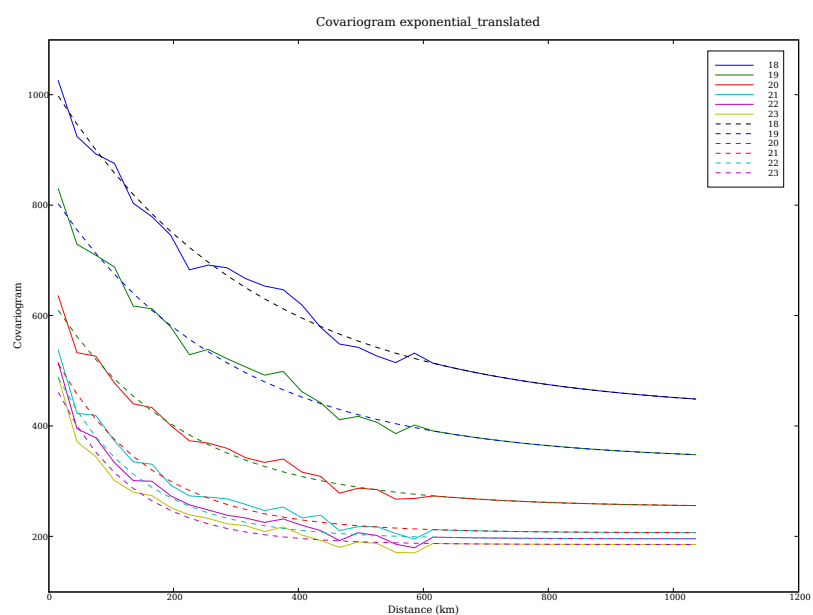
Nous avons choisi de représenter, sur la figure (5), la variation des paramètres L et σ^2 dans le cas d'une fonction de covariance exponentielle entre 9 et 17 heures. Ce modèle est plus classique, mais souffre d'une inadéquation évidente la nuit. La variance *a priori* varie de manière comparable et voisine. Toutefois, ce modèle présente une longueur de corrélation supérieure à 900 kilomètres entre 10 et 15 heures, impliquant que l'ensemble du territoire français, en journée, est en corrélation quasi-parfaite. En corrigeant le modèle exponentiel, on rend compte du caractère plus régional des variations en concentrations d'ozone, en réduisant la longueur de corrélation.

Une fois ce caractère temporel identifié, on peut se demander, pour des questions d'efficacité de calculs évidentes, s'il est réellement pertinent de calculer un covariogramme par heure. En effet, on peut identifier, sur la figure (5), plusieurs régimes correspondants à des plages de temps plus étendues que l'heure. Une cyclicité de 6 heures est claire, mais il serait bon de comparer les performances du krigeage en ajustant des paramètres pour des covariogrammes correspondant à 2, 3, 4, 6 heures. C'est l'objet de notre étude.

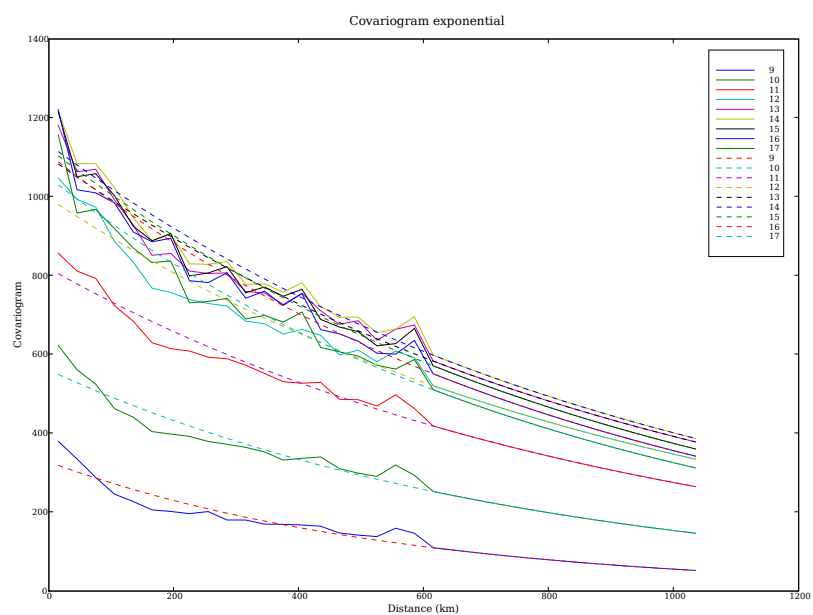
Les fittings définitifs pour des plages de une heure sont donnés en figure 6. Les fonctions ont été coupées en $d = 600$ km pour ne pas prendre en compte les corrélations à trop grande distance, qui existent dans les faits (Paris/Méditerranée), mais qui ne sont pas pertinentes du point de vue de l'interpolation.



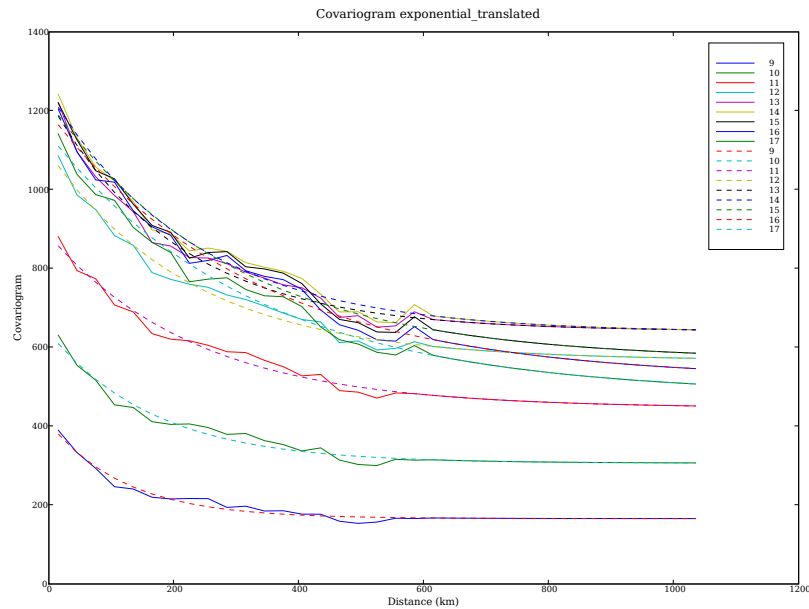
(a) 0h-8h



(b) 18h-23h



(c) 9h-17h : modèle exponentiel



(d) 9h-17h : modèle exponentiel traduit

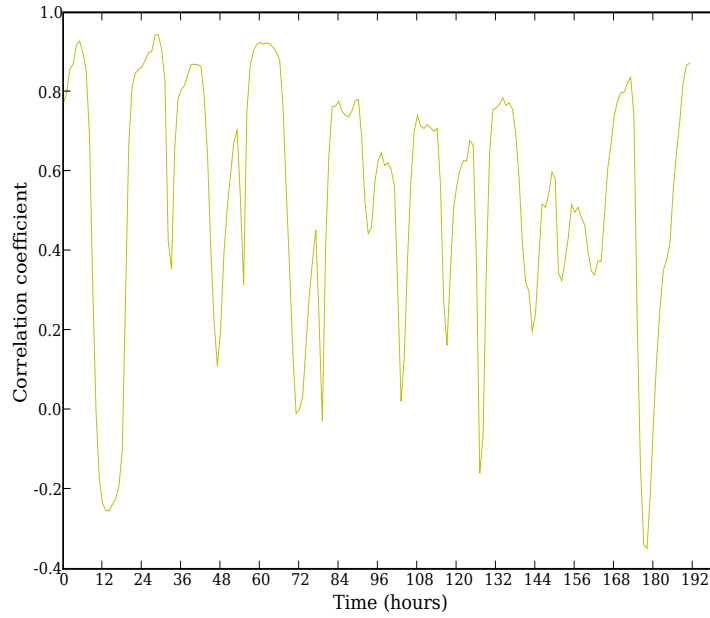
FIG. 6: Covariogrammes et covariogrammes modèles

4 Validation

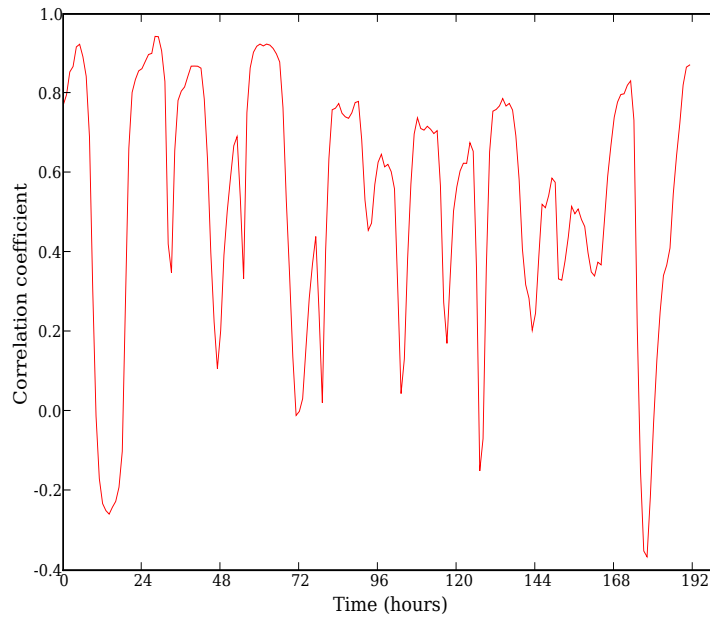
4.1 Résultat du krigeage ordinaire

Quelques figures en annexe D.2 montrent la représentation du champ d'ozone sur notre semaine test à l'aide de la technique de krigeage pour plusieurs plages horaires d'estimation (ou *bins*). Il est certain que l'utilisation du krigeage ordinaire améliore notablement l'interpolation par rapport au krigeage simple. Regardons encore les corrélations avec le champ moyen.

En comparant avec la figure 3.1, on se rend compte que le krigeage ordinaire rend mieux compte des variations du champ d'ozone : la corrélation est plus élevée. Toutefois, le cycle diurne apparaît encore clairement, la nuit étant souvent sujette à de très mauvais scores (inférieurs à 0.2). On est peu étonné de trouver une telle disparité entre le jour et la nuit. On pouvait cependant espérer avoir correctement modélisé la fonction de covariance pour toutes les périodes de la journée.



(a) Bins de 3h



(b) Bins de 6h

FIG. 7: Coefficient de corrélations : krigeage ordinaire par plages de 3 heures et 6 heures

Peu de différences apparaissent entre les deux choix de *bin*. Seule la validation nous permet de discriminer quel *bin* est optimal, et si nous avons réellement amélioré la méthode.

4.2 Validation

La validation a consisté à évaluer le krigeage ordinaire adapté à l'aide des observations. On évalue la capacité d'une partie du réseau à interpoler les valeurs des concentrations aux autres points de mesure de ce même réseau. Les indicateurs statistiques utilisés sont la RMSE (*Root Mean Square Error*) et la corrélation de Pearson (voir annexe A).

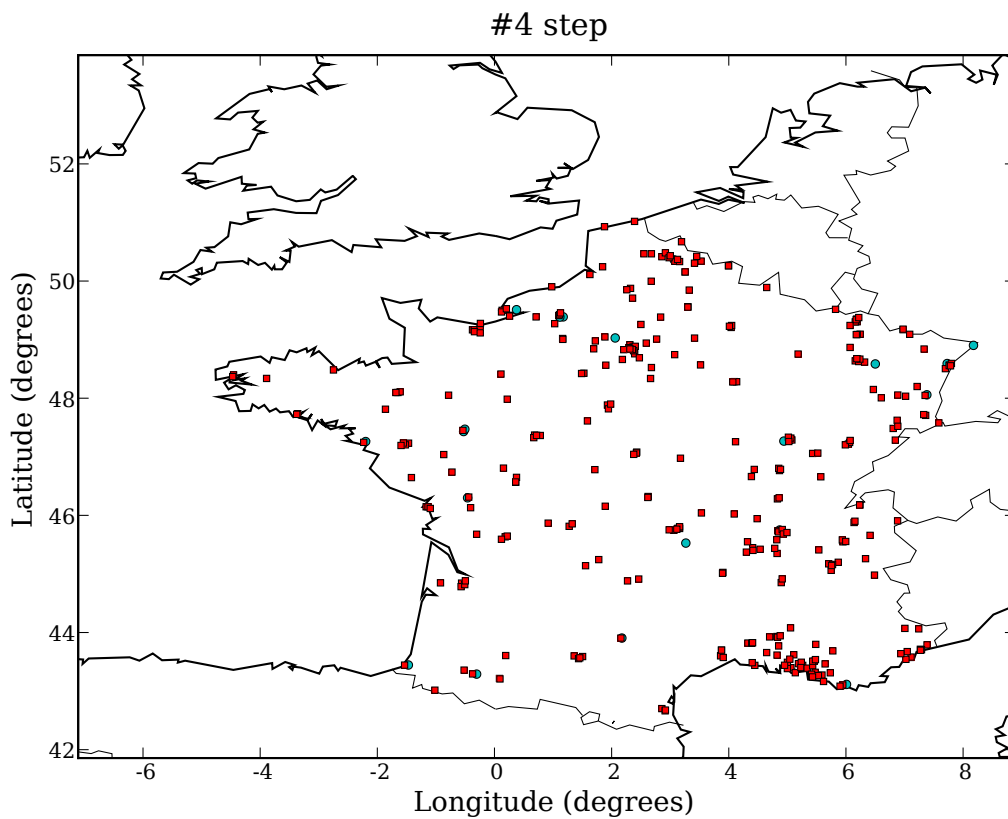
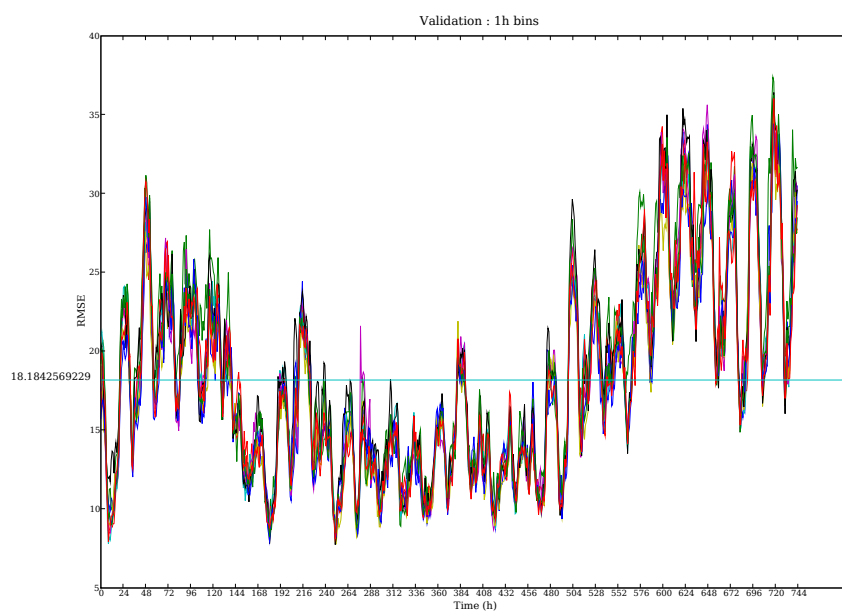
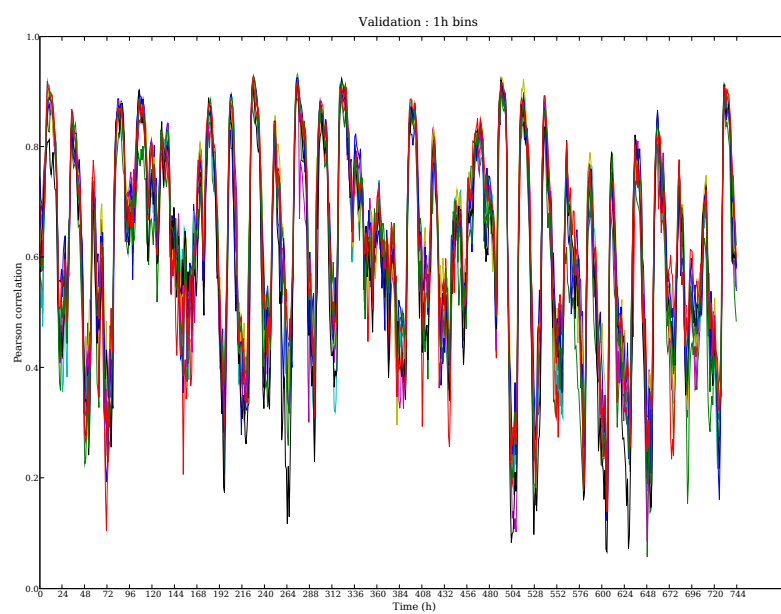


FIG. 8: Exemple de réseau pour la validation : le sous-réseau retenu (20 stations) apparaît en points bleus, le sous-réseau complémentaire (317 stations) en carrés rouges.

On effectue l'expérience sur une période test d'un mois (juillet 2001). Le réseau considéré ($\mathcal{G} \cup \mathcal{U}$) est constitué des stations ayant des observations durant la période test. Le choix du sous-réseau $\mathcal{G}$ est aléatoire, et opéré par échantillonnage. Les corrélations et RMSE sont données ci-après, avec dix réalisations de l'échantillonnage pour chaque bins envisagés.

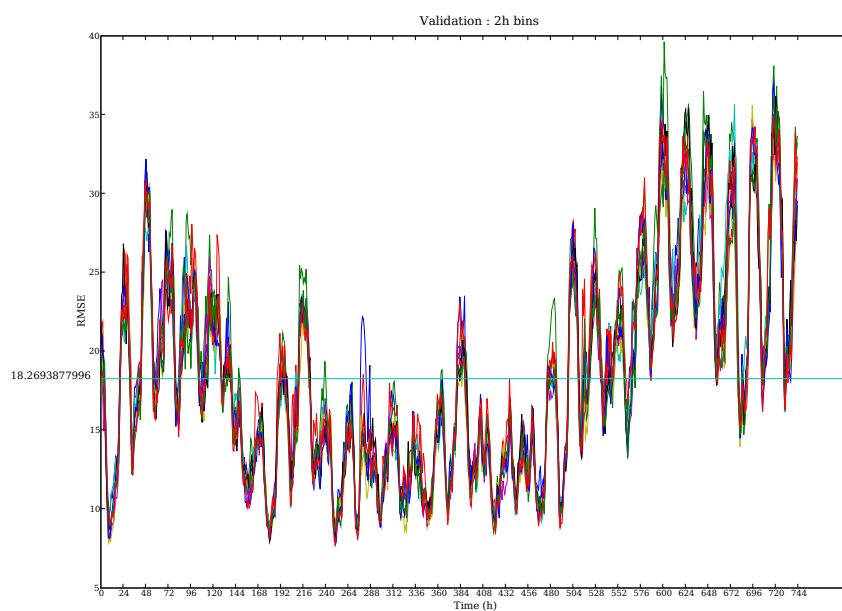


(a) RMSE

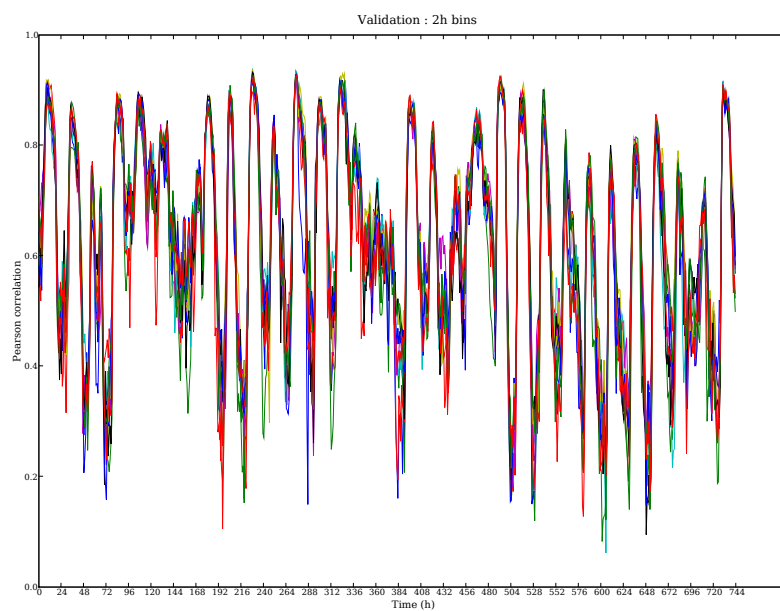


(b) Corrélation

FIG. 9: Validation : *bins* de 1 h

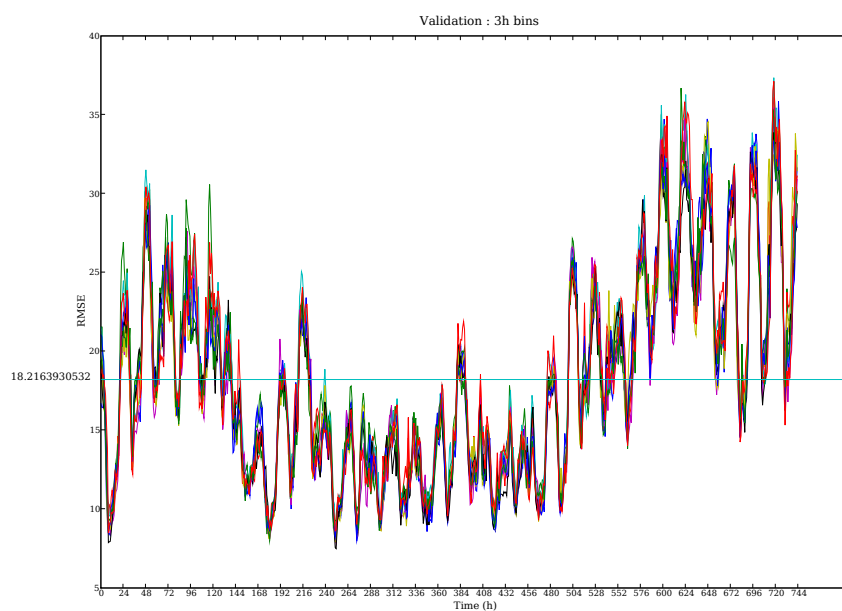


(a) RMSE

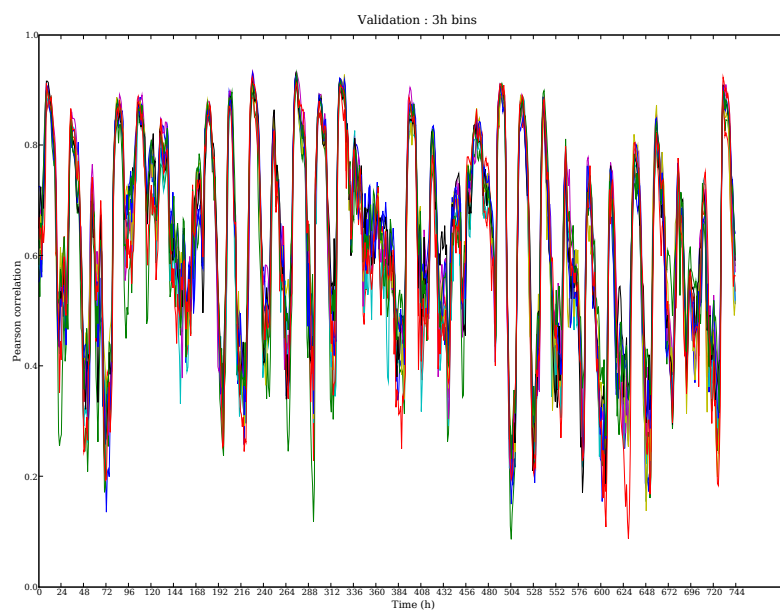


(b) Corrélation

FIG. 10: Validation : *bins* de 2 h

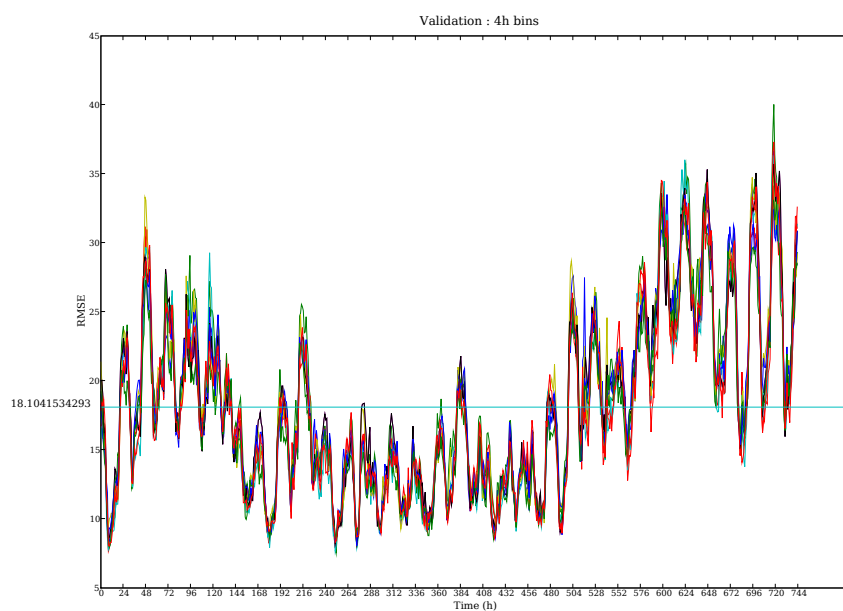


(a) RMSE

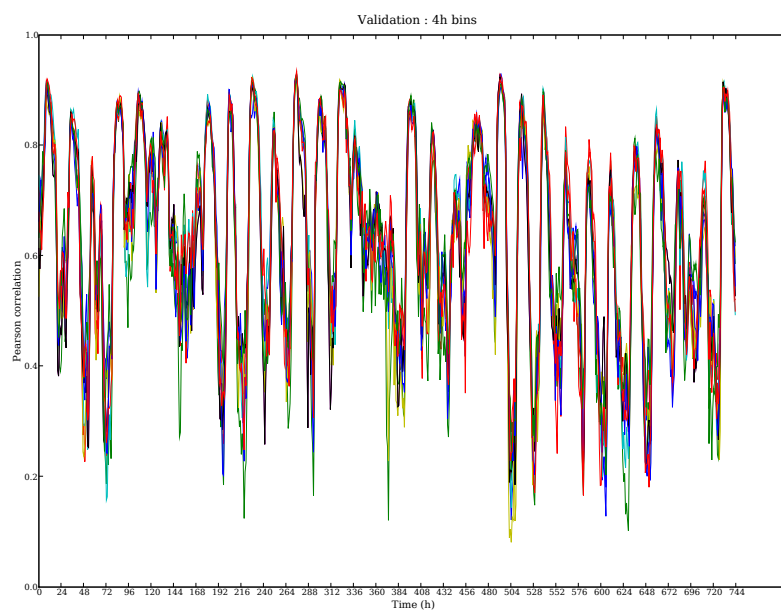


(b) Corrélation

FIG. 11: Validation : *bins* de 3 h

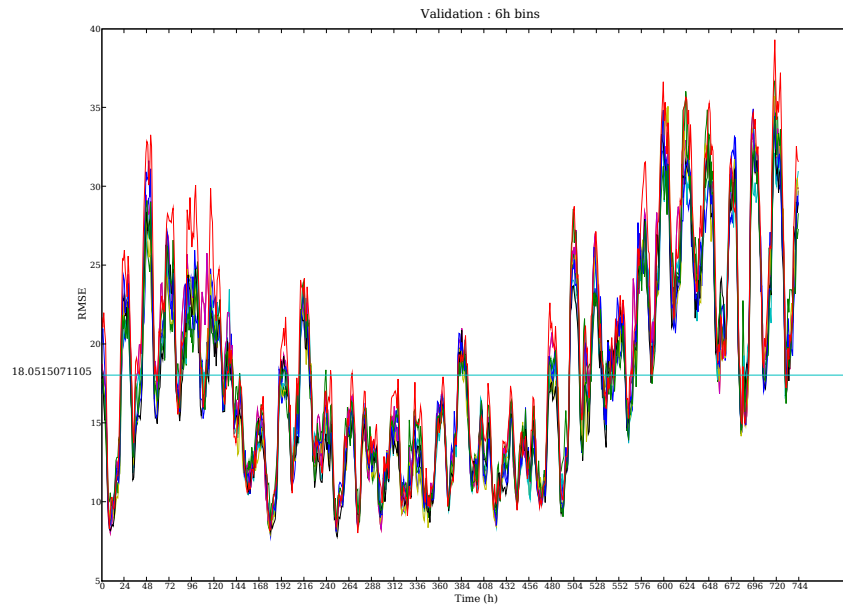


(a) RMSE

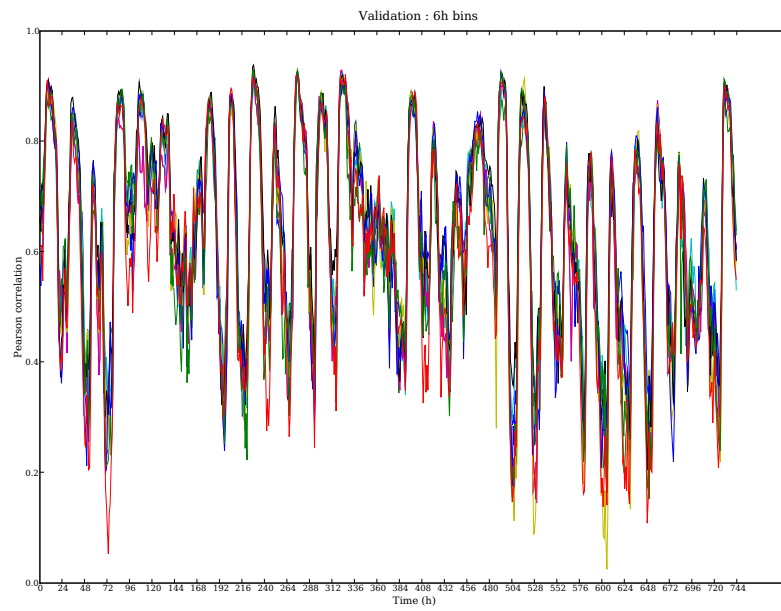


(b) Corrélation

FIG. 12: Validation : *bins* de 4 h

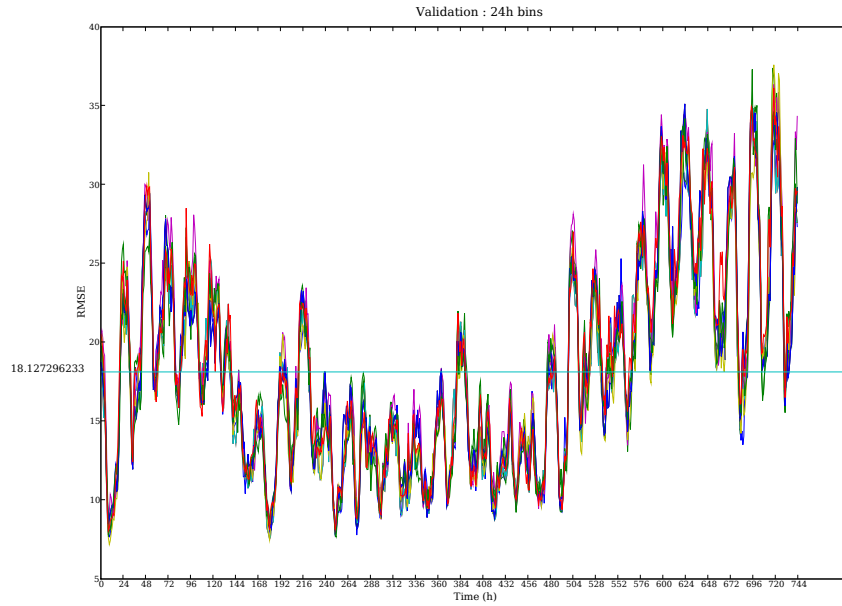


(a) RMSE

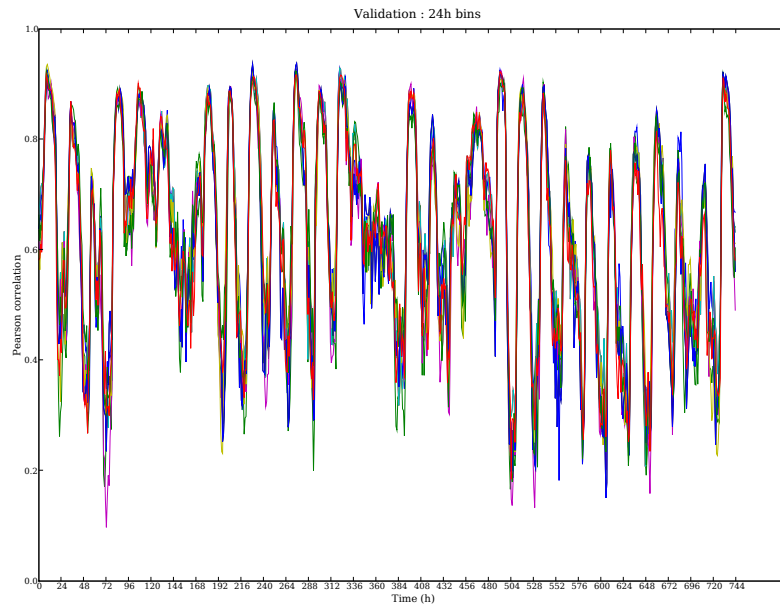


(b) Corrélation

FIG. 13: Validation : *bins* de 6 h



(a) RMSE



(b) Corr lation

FIG. 14: Validation : *bins* de 24 h

On remarque la cyclicit  des indicateurs, les pics de RMSE ayant lieu la nuit, et ceux des corr lations ayant lieu le jour. La corr lation est de mani re g n rale tr s acceptable (pics sup rieurs   0.8), les minima de nuits  tant toutefois une manifestation de l'incertitude des mesures de nuit.

On constate que malgré le tirage aléatoire du sous-réseau pour le krigeage, les indicateurs présentent une dispersion assez faible. Cependant, la faible dispersion des valeurs moyennes, et surtout leur répartition assez inattendue montre que certaines plages horaires semblent privilégiées, dans nos calculs.

	1 h	2 h	3 h	4 h	6 h	24 h
RMSE moyenne ($\mu g.m^{-3}$)	18.18	18.27	18.22	18.10	18.05	18.13

Le bon score des *bins* de 6 heures paraît cohérent, puisque les calculs sont effectués sur les plages 0h-6h, 6h-12h, 12h-18h et 18h-0h, plages cohérentes et couvrant de périodes où l'évolution de l'ozone est homogène, et où le jour et la nuit ont un poids égal. Les plages de deux et trois heures sont trop précises : on effectue alors un grand nombre de calculs entachés d'erreur, du fait de l'incertitude des mesures de nuit. L'imperfection du modèle de nuit est visible sur la courbe de corrélation. Un degré de précision trop important dans le choix de *bins* doit être fait avec précaution, et malgré le fitting correct des covariances la nuit, le modèle est insatisfaisant.

4.3 Sensibilité et *network design*

Nous avons effectué un calcul de validation pour des sous-réseaux de 20, 40, 60, 80, 100, 120, 140 et 160 stations, et comparé les RMSE moyennes. Nous avons choisi des bins de trois heures tout d'abord, puis de six heures, et nous avons comparé les résultats avec un modèle exponentiel *a priori*, avec une longueur de corrélation de 100 kilomètres.

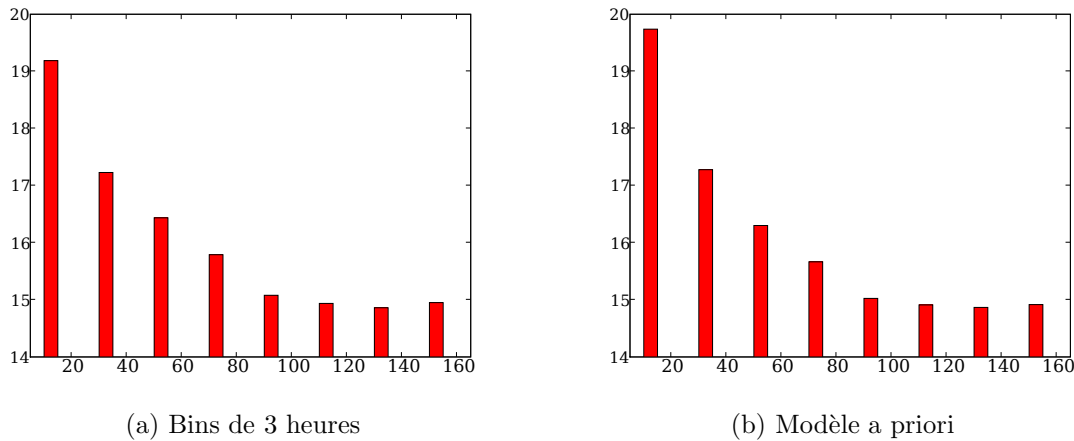


FIG. 15: RMSE moyenne (ordonnées) en fonction du nombre de stations retenues (abscisses)

La RMSE décroît lorsque le nombre de stations augmente. C'est bien conforme à l'intuition, la certitude sur le champ interpolé étant de plus en plus grande à mesure que l'on augmente le nombre de stations avec lesquelles on effectue le krigeage. Toutefois cette décroissance est de moins en moins rapide. En particulier, elle atteint un plateau à partir

de 100 - 120 stations dans les deux cas, le plateau étant atteint plus rapidement dans le cas de bins de 3 heures. En effet, lorsqu'on part de vingt stations, ajouter des stations permet de couvrir une partie plus significative du territoire. Cependant, à partir de 100 stations, on peut espérer que le territoire soit couvert par un réseau dont la maille est en moyenne de l'ordre de grandeur de la longueur de corrélation. Dans les cas que nous avons ajustés, la longueur de corrélation moyenne est environ 230 kilomètres, et ce rayon est atteint plus rapidement que le rayon de 100 kilomètres correspondant au covariogramme *a priori*.

5 Conclusion

Le travail effectué, très ambitieux, nécessite d'être poursuivi. Nous avons réussi à traduire la variabilité temporelle intrinsèque du champ d'ozone à l'aide d'une banque d'outils statistiques, et à inclure cette variabilité dans le calcul d'interpolation. Ce premier pas nous a de plus permis d'esquisser un début de *network design*, en démontrant la décroissance de la RMSE avec le nombre de stations retenues dans notre sous-réseau.

Le calcul rend mieux compte de la variabilité. Toutefois, l'amélioration reste assez peu significative, et on peut apporter deux explications à cela. Tout d'abord, nos calculs incluent la prise en compte des heures de la nuit, au cours desquelles l'incertitude des observations est très grande. C'est ce que nous avons vu sur les courbes de corrélation et de RMSE. Il faut donc adapter un autre modèle statistique pour les émissions de nuit. Les résultats de jour sont cependant bons, les corrélations élevées, et l'étude reste pertinente entre 7h et 17h par exemple.

De plus, les stations sont installées dans des régions différentes (urbaines, périurbaines, trafic ou rurale-régionale), où la variabilité et les caractéristiques des émissions sont très différentes. Nos calculs, dans un but de première approche et de simplicité, ne prennent pas en compte cette différenciation. Il sera alors nécessaire, dans des travaux futurs sur le sujet, de calculer des covariogrammes tenant compte du type des stations, ce qui pose d'importants problèmes d'implémentation. En effet, la fonction de corrélation à calculer ne dépendra plus uniquement de la distance entre deux stations, mais aussi de la nature des stations séparées par cette distance. On devra alors créer une nouvelle famille de covariogrammes $\{C^{u,u}, C^{r,r}, C^{u,r}, \dots\}$ croisés, et adapter l'inversion et le calcul d'interpolation. Cette approche nécessitera en outre une cartographie du territoire permettant de connaître le type de domaine de chaque maille.

Dans tous les cas, ce travail a apporté un éclairage sur les méthodes à mettre en oeuvre pour la conception optimale d'un réseau de surveillance de la qualité de l'air, et sur la nécessité de définir de bons outils statistiques pour évaluer les performances d'un tel réseau à rendre compte de la tendance moyenne d'un champ de concentration d'ozone.

Références

- [1] Abida R., Bocquet M., Vercauteren N., Isnard O., *Design of monitoring network over France in case of radiological accidental release*, submitted, 2007.
- [2] Bocquet M., *Construction optimale de réseaux de mesures : application à la surveillance des polluants aériens*, cours de ParisTech, 2006.
- [3] Cressie N.A.C., *Statistics for spatial data*, Wiley series in probability and mathematical statistics, 1993.
- [4] Mallet V., Quélo D., Sportisse B., Ahmed de Biasi M., Debry É., Korsakissok I., Wu L., Roustan Y., Sartelet K., Tombette M., Foudhil H., *Technical note : the air quality modeling system Polyphemus*, Atmospheric chemistry and physics, 7, pp 5479-5487, 2007.
- [5] Nychka D., Saltzman N., *Design of air-quality monitoring networks*, Case studies in environmental statistics, Ch. 4, edited by Nychka D., Piegorsch W. and Cox L.H., Springer Lecture Notes in Statistics (132), Springer Verlag, NY, pp. 51-76, 1998.

Remerciements

Je tiens à remercier le CEREa de m'avoir accueilli pour ce projet, et de m'avoir donné les moyens techniques d'arriver à des résultats. Je remercie tout particulièrement Marc Bocquet, mon tuteur, de m'avoir permis de prendre part à ce travail, et de m'avoir donné les éléments théoriques pour avancer, ainsi qu'un certain nombre d'astuces d'expérience nécessaire pour l'ingénieur novice des Ponts que je suis. Je remercie aussi Lin Wu, pour les progrès qu'il m'a permis de faire en programmation et modélisation, au cours de nos longues discussions devant l'écran, ainsi que de son aide précieuse pour le débogage. Merci enfin à Pierre Tran et à Vivien Mallet, pour le soutien technique permanent et, comme pour tous au CEREa, leur disponibilité. Merci enfin à Lauriane Saunier, Benoît Vié et Jérôme Favier pour le côté familial d'une entreprise pédagogique aux principes assez obscurs...

*

*

*