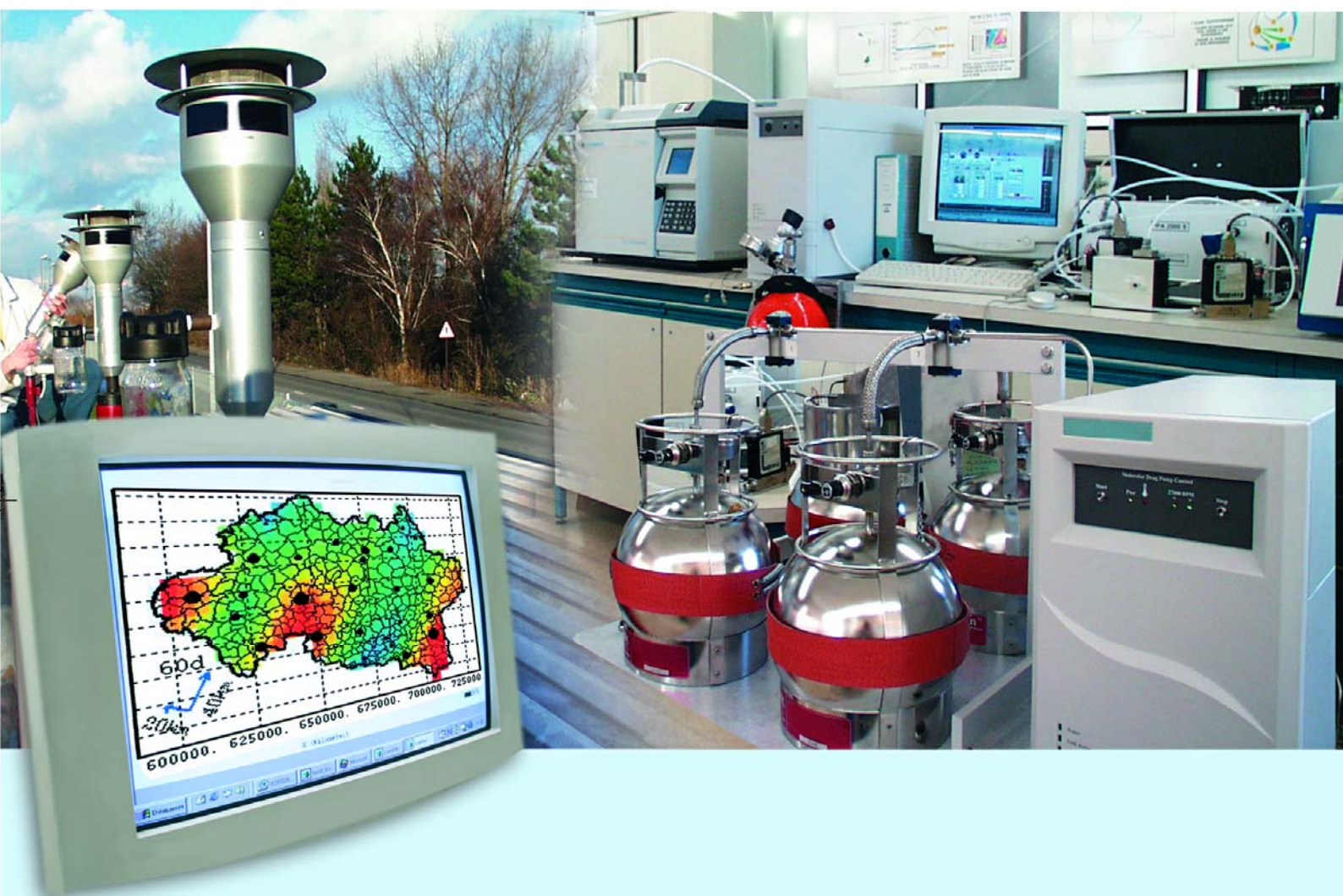




Laboratoire Central de Surveillance de la Qualité de l'Air



Traitements numériques

**Assistance relative à l'exploitation de données de campagnes
et à la réalisation de cartographies (2/2)**

Décembre 2010

Programme 2010

L. MALHERBE, L. LETINOIS





PREAMBULE

Le Laboratoire Central de Surveillance de la Qualité de l'Air

Le Laboratoire Central de Surveillance de la Qualité de l'Air est constitué de laboratoires de l'Ecole des Mines de Douai, de l'INERIS et du LNE. Il mène depuis 1991 des études et des recherches finalisées à la demande du Ministère chargé de l'environnement, et en concertation avec les Associations Agréées de Surveillance de la Qualité de l'Air (AASQA). Ces travaux en matière de pollution atmosphérique ont été financés par la Direction Générale de l'Energie et du Climat (bureau de la qualité de l'air) du Ministère de l'Ecologie, du Développement durable, des Transports et du Logement. Ils sont réalisés avec le souci constant d'améliorer le dispositif de surveillance de la qualité de l'air en France en apportant un appui scientifique et technique au MEDDTL et aux AASQA.

L'objectif principal du LCSQA est de participer à l'amélioration de la qualité des mesures effectuées dans l'air ambiant, depuis le prélèvement des échantillons jusqu'au traitement des données issues des mesures. Cette action est menée dans le cadre des réglementations nationales et européennes mais aussi dans un cadre plus prospectif destiné à fournir aux AASQA de nouveaux outils permettant d'anticiper les évolutions futures.



Assistance relative à l'exploitation de données de campagnes et à la réalisation de cartographies (2/2)

Laboratoire Central de Surveillance
de la Qualité de l'Air

Traitements numériques

Programme financé par la
Direction Générale de l'Energie et du Climat (DGEC)

2010

L. MALHERBE, L. LETINOIS, F. TOGNET, H. BARRE

Ce document comporte 17 pages (hors couverture et annexes)

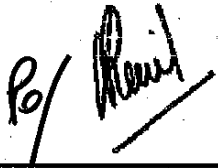
	Rédaction	Vérification	Approbation
NOM	L. MALHERBE	B. BESSAGNET	M. RAMEL
Qualité	Ingénieur de l'Unité Modélisation Atmosphérique et Cartographie Environnementale (MOCA) Direction des Risques Chroniques	Responsable de l'unité Modélisation Atmosphérique et Cartographie Environnementale (MOCA) Direction des Risques Chroniques	Responsable LCSQA/INERIS Direction des Risques Chroniques
Visa			

TABLE DES MATIÈRES

RESUME	6
1. INTRODUCTION	7
2. ECHANTILLONNAGE	8
2.1 Echantillonnage spatial	8
2.1.1 Travail méthodologique	8
2.1.2 Assistance ponctuelle	8
2.2 Echantillonnage temporel	9
3. EXPLOITATION STATISTIQUE DE DONNEES	10
3.1 Organisation d'une formation en statistique	10
3.2 Assistance ponctuelle	11
3.3 Reconstitution de données	11
3.3.1 Retour d'expérience des AASQA	12
3.3.2 Amélioration des outils de calcul	12
4. CARTOGRAPHIE	14
4.1 Réalisation d'un audit	14
4.2 Aide à la réalisation de cartes analysées	14
4.3 Assistance ponctuelle	14
5. CONCLUSION	15
6. REFERENCES	16
7. LISTE DES ANNEXES	17

RESUME

Le LCSQA assure chaque année une assistance technique et méthodologique dans l'élaboration de plans d'échantillonnage, l'exploitation de données et la réalisation de cartographies. La présente note synthétise les actions réalisées en 2010 dans ces différents domaines pour le compte des AASQA.

Ces actions sont de nature diverse : travail de méthode, aide ponctuelle, audit, formation, mais toutes ont pour fin le développement d'une approche cohérente, qui va de la collecte de données à l'analyse statistique ou géostatistique de ces dernières.

Parmi les travaux conduits en 2010, on relèvera plus particulièrement :

- l'approfondissement et la valorisation d'une méthode d'optimisation de l'échantillonnage spatial mise au point en 2009 en collaboration avec le Centre de Géosciences/géostatistique de Mines ParisTech ;
- l'organisation d'une formation de statistique appliquée à la qualité de l'air, avec la participation de Capacités (Université de Nantes) et de Stat-Consult ;
- la réalisation d'un audit sur les outils de modélisation et de cartographie mis en œuvre par AIRFOBEP, en réponse à une sollicitation de cette AASQA ;
- une aide à l'élaboration de cartes analysées avec le logiciel R.

1. INTRODUCTION

En complément de la mesure fixe, la mesure temporaire permet d'enrichir l'information sur la qualité de l'air d'un territoire et de répondre plus précisément aux exigences de la surveillance. Dans ce contexte, les AASQA conduisent régulièrement des campagnes de mesure à l'aide de moyens mobiles équipés d'analyseurs, de tubes à échantillonnage passif ou de préleveurs. Les données recueillies font généralement l'objet d'un traitement statistique et/ou géostatistique afin d'estimer des indicateurs annuels (moyennes, dépassements de seuils), d'évaluer la variabilité temporelle et spatiale des concentrations, ou encore d'établir des cartographies.

D'autre part, grâce au développement des outils de simulation numérique, la modélisation déterministe prend une importance croissante dans l'élaboration de cartographies. La pratique consiste à combiner les résultats de modélisation avec les observations des stations fixes de manière à produire des cartes le plus réalistes et précises possible.

Une application cohérente et efficace de ces différents moyens de surveillance impose de réfléchir sur les stratégies de mesure à mettre en place et la façon d'exploiter au mieux les données disponibles. Le LCSQA contribue depuis plusieurs années à ces réflexions à l'occasion d'études méthodologiques ou au sein de groupes de travail. Chaque année, il propose aux AASQA un appui scientifique et technique sur ces questions. En 2010, l'assistance fournie par le LCSQA a porté sur :

- la conception de plans d'échantillonnage spatial ;
- l'exploitation statistique de données de qualité de l'air ;
- l'élaboration de cartographies.

Les chapitres suivants présentent le travail réalisé sur chacun de ces sujets.

2. ECHANTILLONNAGE

2.1 ECHANTILLONNAGE SPATIAL

2.1.1 TRAVAIL METHODOLOGIQUE

En 2007, le LCSQA a publié un guide de recommandations sur l'échantillonnage spatial, lorsque l'objectif est de produire des cartographies (Wroblewski et al., 2007). Ces préconisations sont fonctions du polluant et du domaine considéré. Elles ont été complétées en 2008, avec la mise au point d'une méthodologie d'échantillonnage, en collaboration avec le Centre de Géosciences/géostatistique de Mines ParisTech (Malherbe et Cárdenas, Fauchaux et Fouquet, 2008). Cette méthode distingue plusieurs stades dans l'élaboration d'un plan d'échantillonnage : conception d'un premier schéma ; ajustement de ce schéma à la suite de la première campagne ; réduction de l'échantillonnage pour un suivi sur le long terme. Les points d'échantillonnage sont disposés de manière à parcourir à la fois l'espace géographique et l'espace des variables auxiliaires utilisées dans la cartographie (Fauchaux et al., 2009).

En 2009 et 2010, la collaboration s'est poursuivie avec le Centre de Géosciences qui a testé et adapté un algorithme d'optimisation de l'échantillonnage (Romary et Fouquet, 2010, cf. annexe 2). A la suite d'échanges techniques entre le Centre de Géosciences et l'INERIS, la méthode a été réajustée en 2010. La principale préoccupation était que la répartition des points de mesure prenne mieux en compte la localisation des zones habitées. L'ensemble du travail a été présenté lors du congrès annuel de la Société française de statistique (Romary et al., 2010). Une publication a été également soumise à la revue *Atmospheric Environment*.

En 2011, le LCSQA offrira son appui à toute AASQA qui souhaiterait appliquer pour ses besoins les méthodologies développées.

2.1.2 ASSISTANCE PONCTUELLE

L'association ATMO Champagne-Ardenne a demandé conseil au LCSQA pour la mise au point de plans d'échantillonnage (tubes à diffusion passive) à l'échelle urbaine ou régionale. Ces campagnes de mesure du NO₂ ont été ou seront conduites par l'AASQA pour ses besoins propres mais elles présentent également un intérêt méthodologique dont le LCSQA, en collaboration avec ATMO Champagne-Ardenne, pourra tirer parti dans ses travaux de cartographie (étude des profils de concentration perpendiculairement aux axes routiers, représentation de la pollution de fond et de proximité, prise en compte des données d'années antérieures dans l'estimation).

2.2 ECHANTILLONNAGE TEMPOREL

Le GT *Plans d'échantillonnage et reconstitution de données* (ADEME/AASQA/LCSQA) s'est achevé au début de l'année 2009 par des sessions de formation, la publication d'un guide technique (<http://www.lcsqa.org/echantillonnage-reconstitution-donnees/documentation>) et la mise à disposition d'outils de calcul. Pour en faciliter la mise en œuvre, une interface a été développée sur le site Internet du LCSQA.

L'outil spécifique de la planification de l'échantillonnage et la documentation associée sont accessibles à l'adresse suivante :

<http://www.lcsqa.org/page/echantillonnage>

Le LCSQA n'a pas été interrogé en 2010 sur ce sujet.

3. EXPLOITATION STATISTIQUE DE DONNEES

3.1 ORGANISATION D'UNE FORMATION EN STATISTIQUE

Au-delà du simple constat, qui consiste à caractériser la qualité de l'air de la zone d'étude en comparant les valeurs mesurées aux seuils réglementaires, les AASQA cherchent de plus en plus à interpréter leurs données par une analyse approfondie. Toutes ne disposent pas des compétences nécessaires à ces études. Par l'intermédiaire du CPT, elles ont ainsi exprimé le besoin d'une formation en statistique appliquée à la qualité de l'air. Afin de mieux cerner leurs attentes, un questionnaire a été mis en ligne à leur attention sur le site Internet du LCSQA.

27 réponses issues de 22 AASQA ont été recueillies et examinées. Les besoins relèvent à la fois de la statistique descriptive élémentaire et de la statistique multivariée plus avancée. D'un point de vue pratique, les AASQA ont fait part de leur intérêt pour le logiciel libre R, avec le souhait de se former à son utilisation. Pour un bilan plus détaillé des réponses obtenues, on se reportera au rapport intermédiaire (Malherbe, 2010).

A partir de ce bilan, le LCSQA, en collaboration avec Frédéric Lavancier (université de Nantes) et la société Stat-Consult, a établi un programme de formation. Celle-ci s'adresse à toute personne chargée d'exploiter des données de qualité de l'air. Elle a pour objet de montrer aux participants de quels outils statistiques ils peuvent faire usage pour répondre simplement aux questions habituellement soulevées :

- comment peut-on avoir une vue d'ensemble des données ?
- comment déceler l'effet de tel ou tel facteur ?
- comment étudier les relations entre plusieurs variables ?
- comment détecter un cycle, une tendance ?

Il s'agit aussi d'initier les participants à l'application de ces outils.

Chaque point théorique est illustré sur des cas réels issus de données fournies par les AASQA. La formation accorde une large place à la pratique avec différents exercices sur le logiciel libre R (<http://cran.r-project.org>). L'apprentissage de ce logiciel et du langage qui lui est propre est facilitée par l'usage de l'interface R Commander.

La formation s'étend sur trois jours ; elle s'organise selon le programme suivant.

1ère partie

Introduction à R et R Commander

Analyses graphiques:

- univariées
- bivariées
- multivariées (ACP, classification)

2e partie

Lien significatifs, modélisation:

- Significativité
- Tests statistiques, comparaison de moyennes
- Généralisation : analyse de variance (ANOVA)

Régression linéaire

3e partie

Aspects temporels :

- Séries temporelles et dépendance temporelle
- Analyse des cycles et de la tendance
- Précautions à prendre en présence de dépendance temporelle

4e partie

Selon les besoins: retour sur certains points, questions complémentaires, application des méthodes aux jeux de données apportés par chacun

Deux sessions de formation ont été organisées par le LCSQA au mois de novembre (8-10 et 24-26 novembre 2010). Elles ont été animées par Frédéric Lavancier, avec le concours du LCSQA. Y ont assisté des représentants des associations AIRLOR, AIRPARIF, ATMO Poitou-Charentes, ATMO Champagne-Ardenne, l'AIR de l'Ain et des Pays de Savoie, LIMAIR, ATMO Auvergne et du LCSQA.

Le document de formation sera téléchargeable sur le site du LCSQA à l'issue des sessions de 2011 (12-14 janvier et 26-28 janvier).

3.2 ASSISTANCE PONCTUELLE

Le LCSQA a été sollicité par ATMO Champagne-Ardenne pour une aide sur le traitement statistique de données d'échantillonnage (prélèvements par canister). Cette aide portait sur la réalisation de boîtes de dispersion avec le logiciel R.

3.3 RECONSTITUTION DE DONNEES

Les travaux du GT Plans d'échantillonnage et reconstitution de données (2006-2009) ont été évoqués dans le paragraphe relatif à l'échantillonnage temporel (§ 2.2).

Comme pour la planification de l'échantillonnage, un outil de calcul fondé sur ces travaux et consacré à la reconstitution de données a été mis en ligne au début de l'année 2009. Il est accessible avec sa documentation à l'adresse suivante :

<http://www.lcsqa.org/page/reconstitution>

Rappelons que trois techniques de reconstitution sont proposées : la méthode dite des « plans de sondage », fondée sur la théorie des sondages, la méthode « ISO », issue de la norme ISO 9359, et la régression linéaire.

3.3.1 RETOUR D'EXPERIENCE DES AASQA

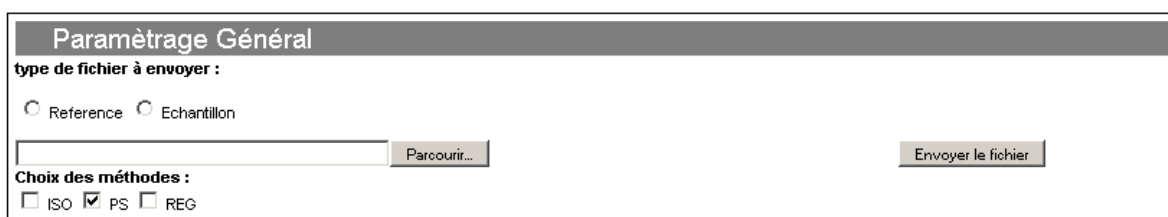
ATMO Champagne-Ardenne s'est servi du module de régression pour reconstituer une concentration moyenne annuelle de NO₂ sur un site de trafic à Reims. Des mesures par moyen mobile ont été réalisées en cet endroit. Une station permanente de fond urbain a été choisie comme site de référence. Afin de valider partiellement la reconstitution, une partie des données a été éliminée de la série d'échantillonnage. La moyenne reconstituée sur toute la période d'échantillonnage à partir de la série réduite est très proche de la moyenne calculée sur la totalité de l'échantillon (écart inférieur à 1 µg/m³).

L'expérience d'ATMO Champagne-Ardenne concorde avec les résultats de tests effectués par le GT sur des données de stations toulousaines. Elle confirme qu'en s'aidant d'une station urbaine de fond implantée dans la même ville, il est possible d'estimer de manière satisfaisante une concentration moyenne annuelle sur un site urbain de trafic. Pour concevoir un plan d'échantillonnage adapté à une telle reconstitution, on se réfèrera aux travaux du GT (cf. références § 2.2).

3.3.2 AMELIORATION DES OUTILS DE CALCUL

En 2009, l'INERIS a recueilli les remarques des utilisateurs relatives aux options de calcul et aux fonctionnalités de l'interface (Malherbe et Debry, 2009). Ces remarques ont été prises en compte ; elles ont conduit aux développements suivants :

- Dans la fenêtre générale de paramétrage, ne sont conservés que les paramètres communs aux trois méthodes (Figure 1). Les paramètres spécifiques de chacune ont été déplacés dans les fenêtres correspondantes (Figure 2 à Figure 4). L'utilisateur voit ainsi plus clairement quelles données d'entrée sont requises par chaque méthode.
- En ce qui concerne la régression, l'utilisateur peut désormais désactiver la procédure de sélection automatique des variables (stepwise) (Figure 4).



Paramétrage Général

type de fichier à envoyer :

☐ Reference ☐ Echantillon

Parcourir...

Choix des méthodes :

☐ ISO ☒ PS ☐ REG

Envoyer le fichier

Figure 1 – Reconstitution d'une moyenne annuelle. Fenêtre générale.

Méthode ISO

type de fichier à envoyer :

☐ _isoH ☐ _isoJ ☐ _iso Met1 ☐ _iso Met2

Parcourir... Envoyer le fichier

Taille des grappes en jours : Redressement : ☒

Intérêt : Polluant : Météo 1 : Météo 2 : Auxiliaire : Strate :

Figure 2 – Reconstitution d'une moyenne annuelle. Paramétrage propre à la méthode ISO.

Méthode PS

Taille des grappes en jours : Redressement : ☒

Intérêt : Auxiliaire :

Strate : Début : Début : Début :

Figure 3 – Reconstitution d'une moyenne annuelle. Paramétrage propre à la méthode Plans de sondage.

Méthode REG

Intérêt :

Autoriser la sélection automatique de variables : ☒ Oui ☐ Non

Sélectionner une ou plusieurs variables auxiliaires (touche Ctrl) :

- ☒ PS_ZAY
- ☒ PS_FER
- ☒ N2_ZAY
- ☒ NO_ZAY
- ☒ O3_ZAY
- ☒ fvNior
- ☒ dvNior
- ☒ poNior
- ☒ huNior
- ☒ tpNior

Lorsque la sélection automatique des variables est autorisée, les variables explicatives dont la contribution n'est pas significative sont éliminées de la régression de façon automatique. Lorsqu'elle ne l'est pas, toutes les variables auxiliaires présélectionnées par l'utilisateur sont conservées.

Figure 4 – Reconstitution d'une moyenne annuelle. Paramétrage propre à la régression.

Ces modifications sont en cours de test. La nouvelle version de l'interface sera mise en service sur le site du LCSQA dans le courant du mois de janvier.

4. CARTOGRAPHIE

4.1 REALISATION D'UN AUDIT

Au début de l'année 2010, faisant suite à des échanges téléphoniques avec l'INERIS, AIRFOBEP a sollicité le LCSQA pour un audit de ses outils de modélisation. L'objet de cette démarche, qui s'inscrivait dans le système qualité de l'association, était de bénéficier de conseils et de recommandations sur les outils utilisés par AIRFOBEP pour produire des cartographies à destination du public.

L'audit a porté sur les aspects méthodologiques de la cartographie. Il s'est attaché aux échelles, variables et méthodes suivantes :

- Echelles : régionale, urbaine, locale (sites industriels)
- Variables : concentrations de polluants (O_3 , NO_2 , SO_2 , PM_{10}), variables météorologiques utilisées dans la modélisation
- Méthodes :
 - o géostatistique (correction des résultats de CHIMERE ; exploitation de campagnes de mesure) ;
 - o modélisation de la dispersion en zone urbaine ou industrielle ;
 - o reconstruction de champs météorologiques.

L'audit s'est déroulé sur une journée (26/03/2010) dans les locaux d'AIRFOBEP. Il a mobilisé deux auditeurs. Il a été le lieu de discussions fructueuses à l'issue desquelles les points forts, les points sensibles et des pistes de progrès ont été identifiés.

4.2 AIDE A LA REALISATION DE CARTES ANALYSEES

L'INERIS a répondu aux demandes de plusieurs AASQA concernant la réalisation de cartes analysées avec R.

Deux séances de travail d'une à deux journées ont été respectivement organisées :

- à ATMO PACA, au mois de janvier
- à ATMO Rhône-Alpes, au mois de septembre, en présence d'AIRPARIF.

4.3 ASSISTANCE PONCTUELLE

ATMO Champagne-Ardenne a réalisé une campagne de mesure sur la ville de Troyes. Cette AASQA a demandé conseil à l'INERIS sur le choix et le prétraitement des variables auxiliaires. Cette assistance se prolongera en 2010 et contribuera à l'élaboration de la cartographie.

5. CONCLUSION

Les demandes reçues en 2010 par le LCSQA concernent aussi bien l'échantillonnage que le traitement de données et l'élaboration de cartographies. En plus des réponses apportées à ces demandes, le LCSQA a approfondi une étude méthodologique sur l'échantillonnage spatial, en collaboration avec le Centre de Géosciences (équipe de géostatistique) de Mines ParisTech. Il a également mis en place une formation de statistique appliquée à la qualité de l'air, avec la participation de l'université de Nantes et de Stat-Consult. Deux sessions de formation ont été proposées aux AASQA en novembre 2011.

Ces différentes actions ont été l'occasion de nombreux échanges techniques entre les AASQA et le LCSQA. Ces discussions et les besoins qu'elles ont fait ressortir nous conduisent à poursuivre en 2011 une assistance dans les différents domaines évoqués dans ce rapport.

Pour permettre au plus grand nombre de participer à la formation de statistique, deux nouvelles sessions sont offertes au mois de janvier 2011.

En décembre 2009, une journée de formation théorique sur la cartographie à partir de données de modélisation et de mesure a été organisée. Si le besoin se confirme, une à deux journées de formation pratique sur la réalisation de cartes analysées seront proposées en fin d'année aux utilisateurs avancés de R. Enfin, des audits sur les outils de modélisation et de cartographie pourront être conduits à la demande.

6. REFERENCES

ADEME/AASQA/LCSQA, GT *Plans d'échantillonnage et reconstitution de données*, 2009. Guide d'élaboration de plans d'échantillonnage temporel et de reconstitution de données. Editions ADEME, Réf. : 6535, Collection : Données et références, Septembre 2009 - 172p.

<http://www2.ademe.fr/servlet/getDoc?cid=96&m=3&id=63725&p1=30&ref=12441>

Faucheux C., Fouquet (de) C., Cárdenas G., Malherbe L., 2009. How to build an initial sampling scheme: recommendations for measurement surveys of air quality. TIES 2009, congrès annuel de la société TIES (The International Environmetrics Society), Bologne, Italie, 5-9 juillet 2009.

Fouquet (de) C., Faucheux C., 2008. Recommandations pour les schémas d'échantillonnage des campagnes de mesure de la qualité de l'air. Annexe du rapport *Echantillonnage et reconstitution de données dans l'espace et le temps*. www.lcsqa.org

Malherbe L., 2010. Assistance relative à l'exploitation de données de campagnes et à la réalisation de cartographies. Rapport 1/2. Organisation d'une formation en statistique. Rapport LCSQA, ref. DRC-10-111609-05224A, www.lcsqa.org.

Malherbe L., 2010. Bilan sur l'audit sur la modélisation réalisé par AIRFOBEP. Note, DRC-111609-0847A.

Malherbe L. et Debry E., 2009. Assistance à l'exploitation de données de campagnes et à la réalisation de cartographies. Rapport LCSQA, DRC-10-103349-00553A, www.lcsqa.org.

Malherbe L., Cárdenas G., 2008. Echantillonnage et reconstitution de données dans l'espace et le temps. Rapport LCSQA, www.lcsqa.org.

Romary T., Fouquet (de) C., 2010. Recommandations sur le schéma d'échantillonnage pour les campagnes de mesure de la qualité de l'air : une approche par optimisation. Rapport ARMINES/LCSQA.

Romary T., Fouquet (de) C., Malherbe L., 2010. Schéma d'échantillonnage pour les campagnes de mesures de la qualité de l'air : une approche par optimisation. 42^e Journées de statistique, Marseille, 24-28 mai 2010.

Romary T., Fouquet (de) C., Malherbe L., 2010. Sampling design for air quality measurement survey: an optimization approach. Publication soumise à Atmospheric Environment.

Wroblewski A., Riffault V., Perdrix E., Malherbe L., 2007. Adaptation des plans d'échantillonnage aux objectifs des campagnes. Echantillonnage spatial – guide de recommandations. Rapport LCSQA, www.lcsqa.org.

7. LISTE DES ANNEXES

Repère	Désignation	Nombre de pages
Annexe 1	Fiche descriptive de l'étude	2
Annexe 2	Etude d'un algorithme d'optimisation de l'échantillonnage spatial. Romary et Fouquet, 2010. Rapport final.	24

Annexe 1

Fiche descriptive de l'étude

THEME 6 : Modélisation – traitements numériques

ETUDE N° 6/2 : ASSISTANCE A L'ELABORATION DE PLANS D'ECHANTILLONNAGE ET A L'EXPLOITATION DE DONNEES DE CAMPAGNES

Responsable de l'étude : INERIS

Objectif

Les travaux proposés regroupent différentes activités ayant pour objet d'assister les AASQA dans **l'exploitation de données de campagnes et l'élaboration de cartographies**. Il s'agit de fournir aux AASQA des recommandations méthodologiques qui leur permettent de valoriser au mieux les données qu'elles produisent et de répondre aux exigences de la surveillance réglementaire.

Contexte et travaux antérieurs

Depuis plusieurs années, le LCSQA est impliqué dans des études relatives à la conception de stratégies d'échantillonnage spatial et temporel et à l'analyse statistique et géostatistique de données de campagnes, notamment pour établir des cartographies. De janvier 2006 à janvier 2009, il a participé activement aux travaux du GT *Plans d'échantillonnage et reconstitution de données* (guide de recommandations, sessions de formation, développement et mise en ligne d'un logiciel).

En 2009, le LCSQA a été sollicité par plusieurs AASQA pour les aider dans l'utilisation du logiciel de planification et de reconstitution. En ce qui concerne l'échantillonnage spatial, il a poursuivi ses travaux méthodologiques ; une méthode d'optimisation de l'échantillonnage en fonction des variables auxiliaires a été en particulier évaluée et pourra être diffusée auprès des AASQA si les résultats se révèlent concluants.

Par ailleurs, à partir des besoins recensés auprès des AASQA et d'aides ponctuelles apportées au cas par cas, le LCSQA a défini la structure d'une formation sur l'analyse statistique de données de campagnes.

Travaux proposés pour 2010

Dans la continuité des travaux antérieurs, le LCSQA propose d'assurer une activité d'assistance et de formation sur les questions liées à la stratégie d'échantillonnage et à l'exploitation statistique de données de campagne.

1) Assistance

A toute AASQA qui lui en fait la demande, le LCSQA apportera son assistance méthodologique dans le domaine de l'échantillonnage temporel ou spatial. Une aide à l'utilisation du logiciel développé les années précédentes par le LCSQA sera également assurée.

S'il est nécessaire, le logiciel sera mis à jour en fonction des remarques des utilisateurs.

2) Organisation d'une formation sur l'analyse statistique de données

Une formation en statistique ciblée sur les besoins des AASQA (traitement de données de campagne, recherche des déterminants des niveaux de concentrations, corrélations, analyse préalable à l'implantation de nouveaux points de mesure...) sera organisée conformément aux orientations définies en 2009. Selon le nombre d'intéressés, une à deux sessions de formation alternant points théoriques et cas pratiques seront proposées au cours du premier semestre. Les exemples traités dans cette formation seront tirés exclusivement d'études sur la qualité de l'air.

Renseignements synthétiques

Titre de l'étude	Assistance à l'exploitation de données de campagnes et à la réalisation de cartographies		
Personne responsable de l'étude	INERIS : Laure Malherbe		
Travaux	annuels		
Durée des travaux pluriannuels	1 an		
Collaboration AASQA	OUI		
Heures d'ingénieur	EMD :	INERIS : 350	LNE : -
Heures de technicien	EMD :	INERIS : 200	LNE : -
Document de sortie attendu	Mise à jour de la notice d'utilisation du logiciel Supports de formation		
Lien avec le tableau de suivi CPT	Demande d'une formation en statistique		
Lien avec un groupe de travail			
Matériel acquis pour l'étude			

Annexe 2

Recommandations sur le schéma d'échantillonnage pour les campagnes de mesure de la qualité de l'air : une approche par optimisation

Etude réalisée par le Centre de Géosciences/géostatistique de Mines ParisTech
pour le compte du LCSQA

Rapport final (8 avril 2010)

Auteurs : Thomas ROMARY et Chantal de FOUQUET



Recommandations sur le schéma d'échantillonnage pour les
campagnes de mesures de la qualité de l'air :
une approche par optimisation

Thomas Romary et Chantal de Fouquet

8 avril 2010

1 Introduction

Suite au rapport [6], l'Ineris a souhaité développer l'approche d'échantillonnage optimal décrite dans l'article de Brus et Heuvelink [2], dans l'idée de comparer les résultats à ceux offerts par la méthode proposée précédemment. La méthode proposée par Brus et Heuvelink vise à générer automatiquement un plan d'expérience optimal minimisant la variance de prédiction du krigeage universel. Elle repose d'abord sur une modélisation a priori de la concentration. En particulier elle nécessite de connaître à la fois les variables auxiliaires et le variogramme des résidus, avant d'obtenir les données. L'optimisation est ensuite réalisée par recuit-simulé.

Ce rapport est divisé en trois parties principales. Dans la première, nous décrirons en détail la méthodologie en question. Dans la seconde, nous rendrons compte de l'analyse des données de concentration en benzène issues de campagnes réalisées sur les sites de Reims et Lille, effectuée dans l'optique de définir une modélisation a priori. Enfin, nous appliquerons la méthodologie sur les sites de Lille, Reims et Bordeaux et discuterons de ses résultats. En particulier, sur Bordeaux, les résultats seront comparés à ceux de l'approche décrite dans [6].

2 Méthodologie

2.1 Modélisation

La modélisation choisie dans [2] est une modélisation en dérive externe :

$$Z(x) = Y'(x)\beta + S(x), \quad (1)$$

où Z représente la variable d'intérêt, $x \in \mathcal{X}$ est la coordonnée spatiale, Y est un vecteur de variables auxiliaires, β un vecteur de paramètres, $S(x)$ un champ aléatoire, appelé résidu. Etant donné un n -échantillon $\mathbf{Z} = (Z(x_1), \dots, Z(x_n))$ et la matrice $\mathbf{Y}$ des variables auxiliaires associées, on peut prédire Z en un point x_0 par krigeage universel :

$$\widehat{Z}(x_0) = (c_0 + \mathbf{Y}(\mathbf{Y}'\mathbf{C}^{-1}\mathbf{Y})^{-1}(\mathbf{Y}_0 - \mathbf{Y}'\mathbf{C}^{-1}c_0))' \mathbf{C}^{-1}\mathbf{Z}, \quad (2)$$

où $\mathbf{C}$ est la matrice de covariance associée à S , c_0 est le vecteur des covariances des résidus au point de prédiction et les observations. A cet estimateur est associée une variance de prédiction donnée par la formule suivante :

$$\mathbb{V}(Z(x_0) - \widehat{Z}(x_0)) = c(0) - c_0'\mathbf{C}^{-1}c_0 + (\mathbf{Y}_0 - \mathbf{Y}'\mathbf{C}^{-1}c_0)'(\mathbf{Y}'\mathbf{C}^{-1}\mathbf{Y})^{-1}(\mathbf{Y}_0 - \mathbf{Y}'\mathbf{C}^{-1}c_0), \quad (3)$$

où $c(0)$ est la variance ponctuelle de Z . On peut remarquer que (3) prend en compte à la fois la variance de l'erreur de prédiction du résidu (deux premiers termes) et la variance d'estimation de la dérive (terme de droite). Elle est en outre indépendante des données. Elle dépend du plan d'échantillonnage η à travers c_0 , $\mathbf{C}$ et $\mathbf{Y}$. Minimiser cette grandeur intégrée sur le domaine d'étude, noté $\mathcal{X}$ permettrait d'obtenir automatiquement un plan équilibré par sa couverture du domaine spatial et du nuage des auxiliaires. Il s'agit ainsi de minimiser la quantité suivante :

$$O(\eta) = \int_{\mathcal{X}} \mathbb{V}(Z(x) - \widehat{Z}(x)) dx \quad (4)$$

Pour simplifier les choses, cette intégrale sera évaluée sur un maillage régulier de $\mathcal{X}$. Il est intéressant de noter que ce critère est proportionnel à la variance ponctuelle du résidu, on peut s'en rendre compte à la lecture de (3). Ainsi, le plan optimal selon (4) est indépendant d'un facteur multiplicatif de la fonction de covariance.

2.2 Optimisation par Recuit-simulé

Minimiser (4) est un problème d'optimisation non trivial. La dimension de ce problème est très grande et il est impossible d'implémenter des algorithmes déterministes de type Gauss-Newton. Brus et Heuvelink proposent ainsi le recuit-simulé.

L'algorithme de recuit simulé a été introduit par [11], puis généralisé par [9] à des problèmes d'optimisation. Il peut s'appliquer à des problèmes d'optimisation globale et de simulation (cf. [14] et sa bibliographie). Sa dénomination vient d'une analogie avec la métallurgie. En effet, un métal est une forme cristalline au sein de laquelle des atomes peuvent venir se loger ; un métal obtenu par une lente décroissance de la température après sa mise en solution (chauffage) sera plus solide qu'un métal obtenu en la réduisant rapidement. Elle vise en particulier à générer des échantillons issus d'une distribution de Gibbs :

Définition 2.1 *Un champ aléatoire Z , à valeur dans $\mathcal{Z}$, est un champ de Gibbs d'énergie E , s'il est régi par la loi de probabilité de densité (par rapport à la mesure de Lebesgue) :*

$$f(z) = \frac{1}{\mathcal{C}} e^{-E(z)}, \quad (5)$$

$$\mathcal{C} = \int_{\mathcal{Z}} e^{-E(z)} dz, \quad (6)$$

appelée densité de Gibbs.

La constante $\mathcal{C}$ est généralement incalculable en raison de la dimension de $\mathcal{Z}$, c'est pourquoi les méthodes de Monte-Carlo par chaîne de Markov (MCMC) et la variante qu'est le recuit simulé sont particulièrement adaptées à la simulation de telles distributions.

L'idée fondamentale de cet algorithme est qu'un changement d'échelle, appelée *température*, permet des mouvements plus rapides sur la surface de la fonction E à minimiser/explorer que les méthodes d'optimisations déterministes comme la méthode du gradient, ou que les méthodes MCMC classiques. Ainsi, ce changement d'échelle permet d'éviter, en théorie, de rester bloqué dans le bassin d'attraction d'un minimum/mode local. Décrivons maintenant l'algorithme de recuit simulé.

Etant donné une suite de températures croissantes positives (T_n) , une chaîne $Z_1 = z_1, Z_2 = z_2, \dots$ est générée à partir des densités de Gibbs suivantes :

$$\pi_i(z) \propto \exp(-E(z)/T_i).$$

Alors que T_i décroît, les valeurs simulées selon cette densité se concentrent dans un voisinage toujours plus restreint du minimum de E . En effet, T_i diminuant, la densité cible va être de moins en moins "plate" : à haute température, elle aura une forme très dilatée, ce qui permettra à la chaîne générée de parcourir aisément le paysage énergétique, tandis que lorsque T_i tendra vers 0, la densité cible se rapprochera d'une masse de dirac au minimum de E .

Cet algorithme peut être vu comme une version non homogène de l'algorithme de Metropolis Hastings. En effet, dans sa version la plus courante, (T_n) décroît à chaque itération, le noyau de la chaîne varie donc au cours du temps. Les résultats de convergence classiques sur les chaînes de Markov homogènes ne s'appliquent donc pas dans ce cadre. L'algorithme se présente de la manière suivante :

Algorithm 1 Recuit simulé

```
Générer  $z_0$ 
for  $i = 1 \dots N$  do
  Générer  $y$  selon une distribution instrumentale de densité  $q(|y - z_{i-1}|)$ ,
  Générer  $u \sim \mathcal{U}_{[0,1]}$ 
  if  $u < \alpha = \inf(\exp((E(z_{i-1}) - E(y))/T_i), 1)$  then
     $z_i = y$ 
  else
     $z_i = z_{i-1}$ 
  end if
   $T$  prend la valeur  $T_{i+1}$ 
end for
```

On peut noter que lorsque $E(y)$ sera inférieure à $E(z_{t-1})$, la transition sera systématiquement acceptée. Dans le cas contraire, elle sera acceptée avec une probabilité non nulle. C'est ce mécanisme qui assure la validité théorique de l'algorithme. En pratique, sa paramétrisation nécessite néanmoins un certain doigté, un mauvais choix de paramètres pouvant conduire à des résultats décevants. On applique généralement des règles heuristiques pour choisir la paramétrisation de cet algorithme : il est nécessaire que la température de départ soit suffisamment élevée et sa décroissance suffisamment lente. Un résultat de convergence existe, sous conditions sur E , avec l'hypothèse que la température décroisse en $1/\log(n)$. Dans la pratique néanmoins, on emploie généralement une suite à décroissance géométrique.

Dans notre problème, l'énergie à minimiser est la moyenne sur $\mathcal{X}$ de la variance de prédiction du krigeage universel (4) en fonction du plan d'échantillonnage η . Comme dans [2], nous considérons le cas où le nombre n de points à échantillonner est fixé.

Brus et Heuvelink proposent une méthode issue de [9] pour choisir la température initiale. Elle est choisie de sorte que la probabilité moyenne d'acceptation d'une transition empirant le critère soit égale à 0,8. On peut la calculer en lançant l'algorithme avec une température égale à 1 pour un certain nombre d'itérations et en acceptant toutes les transitions. On estime la moyenne des augmentations du critère : $\overline{\Delta V^+}$. La température de départ est alors donnée par

$$T_0 = \frac{\overline{\Delta V^+}}{\log(0.8)}$$

Les auteurs de [2] proposent un schéma particulier de descente de température : la garder constante durant un certain nombre d'itérations (ou d'acceptations) avant de la faire décroître. La distribution instrumentale q , autrement dit le mécanisme permettant de proposer un nouveau plan en fonction du précédent consiste à choisir aléatoirement un point de η puis à le déplacer dans une direction aléatoire, à une distance tirée uniformément entre 0 et une borne h qui se réduira au cours de l'algorithme. Cette méthode a été proposée dans [8]. Nous proposons ici de mélanger cette distribution instrumentale avec une autre consistant à remplacer un point de η , sélectionné aléatoirement, par un point tiré uniformément parmi ses n plus proches voisins dans le nuage des auxiliaires, la valeur de n décroissant au fur et à mesure de l'algorithme. Ceci peut permettre une exploration plus rapide du nuage des auxiliaires et ainsi rendre l'algorithme plus efficace. Il convient donc de définir la probabilité d'effectuer une proposition selon l'une ou l'autre des deux distributions instrumentales.

Il reste enfin à choisir le nombre d'itérations pendant lesquelles la température demeure constante et un critère d'arrêt (nombre d'itérations fixé, température minimale atteinte, critère de non-progression,...).

Avant d'appliquer cette méthodologie, il convient de réaliser une étude des données disponibles

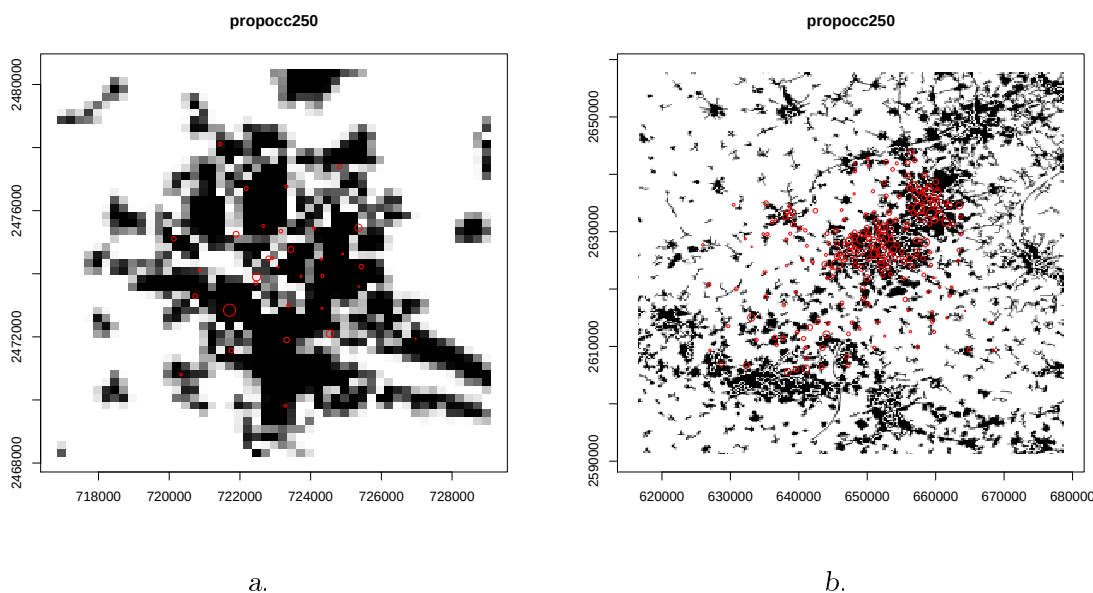


FIG. 1 – Plans d’échantillonnage et proportion d’occupation des sols utilisés pour les campagnes de Reims (a.) et Lille (b.)

afin de déterminer une modélisation a priori pour la variable d’intérêt : choix des variables auxiliaires et du modèle de variogramme.

3 Analyse des données disponibles

Pour cette étude, nous disposons de données de différentes campagnes sur la concentration de benzène pour les agglomérations de Lille et Reims. Précisément, nous disposons à Lille de deux campagnes la première ayant eu lieu à l’été 2003 et la seconde durant l’hiver 2004, toutes deux selon le même plan d’échantillonnage comprenant 234 points de mesure, une quinzaine étant indisponibles à chaque campagne. En ce qui concerne Reims, nous avons à disposition les résultats de 4 études qui ont eu lieu aux mois de février, mai, août et novembre d’une même année, échantillonnés en 33 points. A noter que 3 points correspondent à des mesures de proximité automobile, ce qui entraîne des concentrations sensiblement plus élevées. Les prédictors Y_i choisis sont la densité de population et la proportion de tissu urbain relativement à une maille de 250 m.

Nous présentons en figure 1 a. et b. respectivement les cartographies de densité de tissu urbain de Reims et Lille, où l’on a représenté en rouge les points d’échantillonnages (la taille du cercle est proportionnelle à la moyenne des valeurs relevées).

3.1 Modélisation de la dérive

Nous rappelons le modèle de la concentration en benzène dans l’air ambiant pour ces deux villes :

$$Z(x) = \sum_{i=1}^n \beta_i Y_i(x) + S(x), \quad (7)$$

Nous employons les logarithmes translats ($\log(1+y)$) des 2 prédictors suivants : la proportion de tissu urbain dans un rayon de 250 m (classe agrégée 1 de la classification CORINE Land

Cover) et la densité de population dans un rayon de 250 m. La dérive est estimée par moindres carrés ordinaires. Nous présentons ici les résultats des régressions pour chacune des deux agglomérations et les campagnes de mesures associées.

3.1.1 Reims

En ce qui concerne Reims, nous avons estimé une régression pour chacune des 4 campagnes ainsi que pour la moyenne des 4 en expliquant la concentration de benzène par la proportion de tissu urbain et la densité de population ; nous en présentons les résultats en annexe. Les deux paramètres sont fortement significatifs dans les 5 cas de figure. Le pouvoir explicatif du modèle linéaire estimé est globalement élevé (R^2).

En revanche, les estimations des régressions diffèrent nettement lorsque l'on inclut ou non les 3 points de mesure de proximité automobile : les valeurs estimées des paramètres et les variances des résidus changent significativement. A noter que pour l'estimation des régressions pour le jeu de données sans les trois points de proximité automobile, nous avons également soustrait le point de mesure numéro 6 qui présente alors un fort effet levier.

3.1.2 Lille

Nous avons estimé trois régressions linéaires expliquant la concentration de benzène (hiver, été, moyenne) par la proportion de tissu urbain et la densité de population, nous en présentons les résultats en annexe. Globalement, on peut voir que les deux paramètres sont ici aussi fortement significatifs. D'autre part, on peut noter que le modèle de régression a un fort pouvoir explicatif, avec des R^2 supérieurs à 0,85.

A noter la différence entre les résultats des campagnes d'hiver et d'été : les concentrations en benzène sont plus élevées en été, ce qui se traduit par des coefficients de régression plus élevés ainsi qu'une variance des résidus plus élevée en hiver.

3.1.3 Bilan

La principale conclusion qui ressort de ces régressions est le fort pouvoir explicatif des régresseurs choisis, avec des R^2 systématiquement supérieurs à 0.8. Il est en revanche difficile de tirer des conclusions sur un comportement a priori des β_i . En effet, les valeurs estimées varient énormément selon la ville considérée et, dans une moindre mesure, selon la campagne de mesure. Ceci peut être dû à la différence typologique entre les deux villes : Lille est une des plus grandes agglomérations françaises, avec une forte densité de population (6500 hab/km², pour plus d'1 million d'habitants dans l'agglomération) et un tissu industriel dense tandis que Reims est une agglomération de taille bien moins importante (moins de 300000 habitants dans l'agglomération). Ceci dit, ces paramètres n'entrent pas en jeu dans (4). Notons enfin que nous n'avons pas inclus la moyenne (intercept ou ordonnée à l'origine) dans l'estimation de la régression. Les résultats des régressions où nous avons laissé ce paramètre donnaient des R^2 beaucoup plus faibles. On peut justifier ceci par un argument simple : dans une zone non habitée où la proportion de surface artificielle est nulle, la concentration en benzène devrait logiquement être nulle à une incertitude près.

3.2 Modélisation spatiale : analyse des résidus

A partir des résidus issus des régressions, nous allons estimer le variogramme de la composante $S(x)$ de (7). Notons que les résidus $S(x)$ ne sont pas directement observables, les résidus issus de la régression sont les résidus estimés, obtenus en ôtant de la fonction aléatoire

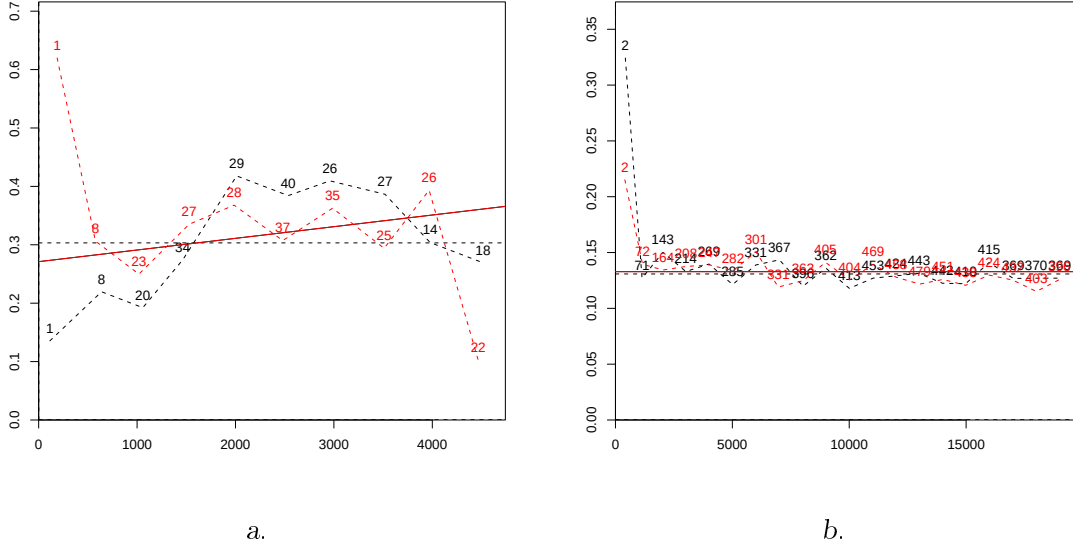


FIG. 2 – Variogramme expérimental et estimé pour les résidus de la concentration moyenne issus de la validation croisée pour Reims (a.) et Lille (b.)

$Z(x)$ l'estimation de la dérive. Ils sont ainsi issus d'une projection sur l'orthogonal de l'espace défini par les régresseurs (Y_i) , voir [10] pour plus de détails. Plusieurs méthodologies peuvent alors être mises en œuvre :

- une estimation des résidus par validation croisée,
- une estimation par dichotomie d'échantillon (test et validation),
- une approche par voisinages glissants ou zones homogènes.

On a choisi ici l'approche par validation croisée. Globalement, l'estimation du variogramme des résidus fait apparaître un effet de pépite associé à un variogramme linéaire de faible pente sur les données de Reims (figure 2, a.). Sur Lille aucune structure n'est visible : le résidu évalué par validation croisée ne montre pas de structure apparente (figure 2, b.). Cette faible structuration apparente est probablement due à la présence d'un effet proportionnel qu'il convient de prendre en compte dans l'évaluation du variogramme.

3.3 Mise en évidence et prise en compte de l'effet proportionnel

Un effet proportionnel est une variation sur le domaine d'étude de la variance ponctuelle de la variable d'intérêt. Il tient son nom du fait que l'on observe régulièrement des variations plus importantes dans les plages de valeurs fortes et semble dépendre de la valeur de la variable régionalisée. Il est généralement complexe à modéliser sur la base des seules données mais peut être parfois expliqué par des variables auxiliaires. Le cas de Lille en est une bonne illustration.

On a représenté en figure 3 la moyenne des concentrations observées en fonction du log de la densité de population. La variance de la concentration moyenne estimée est nettement plus élevée lorsque le logarithme translaté de la densité de population est supérieur à deux. On peut ainsi mettre en évidence deux classes de variabilité différente en fonction des variables auxiliaires : la variance des concentrations appartenant à la seconde classe est trois fois supérieure à celle correspondant à la première.

Lorsque l'on prend en compte cet effet proportionnel dans la régression sur les données de

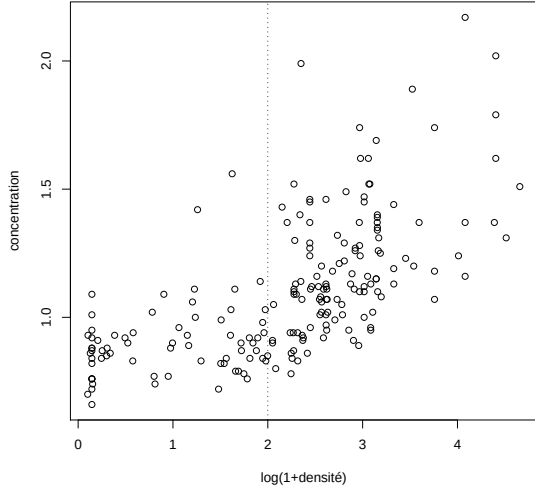


FIG. 3 – Variabilité de la concentration moyenne en fonction de la densité de population.

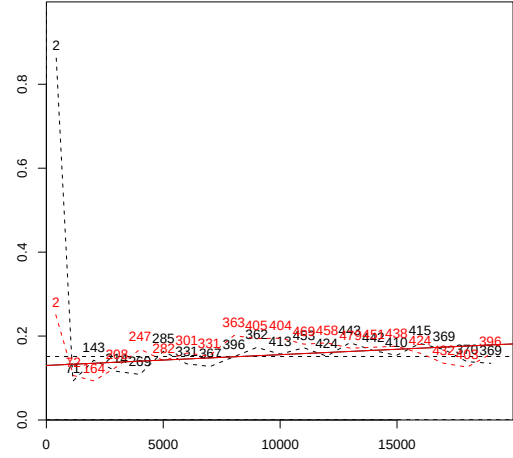


FIG. 4 – Variogramme expérimental et estimé pour les résidus de la concentration moyenne (avec prise en compte de l'effet proportionnel)

Lille en effectuant l'estimation par moindres carrés pondérés, les résidus estimés par validation croisée présentent effectivement une structure pépite + linéaire, semblable à celle observée sur les données de Reims comme le montre la figure 4.

En revanche, les données de Reims, moins nombreuses, ne nous permettent pas d'identifier un éventuel effet proportionnel.

3.4 Conclusion

Au final, cette analyse exploratoire des données a permis de sélectionner les variables auxiliaires, les logarithmes traduits de la densité de population et de la proportion de tissu urbain, qui expliquent de manière satisfaisante ($R^2 > 0.8$) la moyenne observée des concentrations en benzène. D'autre part, elle a permis d'identifier, bien que grossièrement, la structure variographique à employer. Enfin, elle mis en évidence un effet proportionnel pour les données de Lille qu'il a été possible de modéliser grâce à la densité de population. Sur Reims, en revanche, les données n'ont pas permis de trancher quant à la présence éventuelle d'un effet proportionnel. Il conviendra donc de tester différentes modélisations a priori du variogramme pour l'implémentation de la méthode d'échantillonnage optimal.

4 Application

Dans cette section, nous allons décrire les résultats obtenus en appliquant la méthodologie présentée en section 2 aux différents cas d'étude à notre disposition : Reims, Lille et Bordeaux. Dans chacun des cas, nous avons considéré le problème du placement des points d'échantillonnage sur une grille dont le pas de discrétisation est fixé à 250m. Le cas de Bordeaux fera l'objet d'une comparaison avec les résultats donnés par la méthode proposée dans [6].

4.1 Reims

Sur la ville de Reims, nous avons cherché à optimiser le placement de 33 points sur une grille 49×49 selon différentes modélisations a priori. Nous avons initialisé l'algorithme avec une température de $T_0 = 10^{-5}$, puis nous avons fait décroître ce paramètre selon la relation : $T_{t+1} = 0.95T_t$, dès que 100 transitions étaient acceptées. Enfin, nous avons fait varier la taille de l'intervalle où la distance de déplacement du point est simulée au cours de la simulation : celle-ci étant fixée à $20 \times 250m$ au départ, on l'a faite décroître à raison d'une unité toutes les 500 transitions acceptées.

Nous avons principalement fait varier deux paramètres de la modélisation a priori : le rapport pépité sur linéaire, c'est-à-dire le rapport de la valeur de l'effet pépité à la variance maximale d'un champ aléatoire de variogramme linéaire sur le domaine, et l'effet proportionnel. Bien que nous n'ayons pu l'identifier sur les données de Reims, nous avons envisagé cette modélisation à la lumière des analyses effectuées sur les données de Lille. Nous avons ainsi fait varier ces deux paramètres ce qui nous a donné les quatre résultats suivants :

1. rapport pépité/linéaire = 1, sans effet proportionnel, figure 5
2. rapport pépité/linéaire = 1, avec effet proportionnel (effet pépité doublé lorsque le logarithme translaté de la densité de population est supérieur à 2), figure 6
3. rapport pépité/linéaire = 4, sans effet proportionnel, figure 7
4. rapport pépité/linéaire = 4, avec effet proportionnel, figure 8

Chaque figure représente 5 graphiques différents. Dans la partie gauche, on trouve d'abord la valeur du critère lors de l'acceptation d'une transition (a), puis la moyenne empirique du taux d'acceptation de transition (mal), le plan d'échantillonnage minimisant le critère (etabest) et enfin sa représentation (points rouges) dans le nuage des auxiliaires (AUX), en logarithme translaté, avec la densité de population en abscisse et la proportion de tissu urbain en ordonnée. Sur la droite figure le plan optimal superposé à la proportion d'occupation des sols.

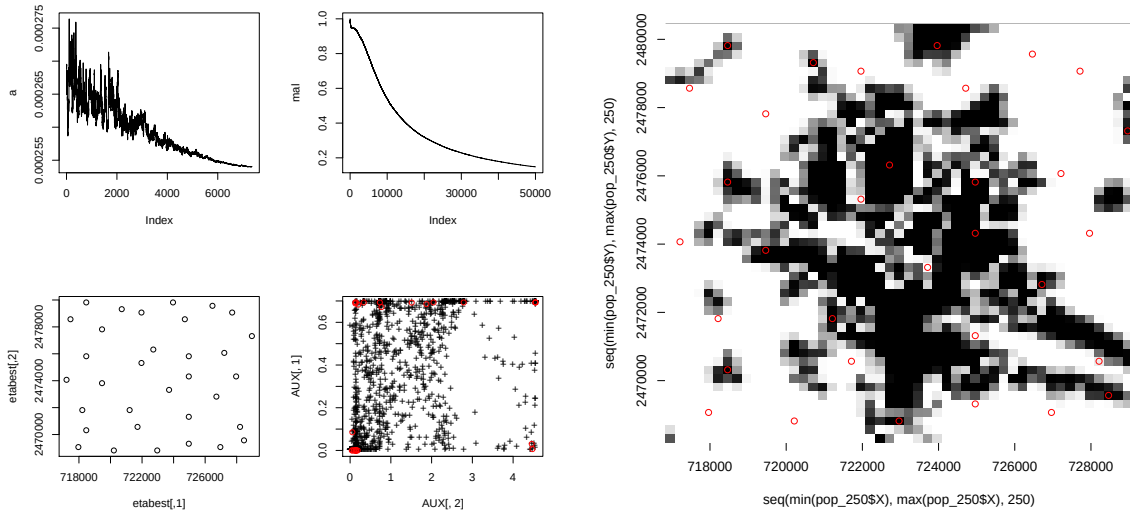


FIG. 5 – Résultats pour un rapport pépité/linéaire = 1, sans effet proportionnel

On peut d'abord voir sur chacune des quatre figures que l'algorithme s'est comporté de manière satisfaisante : les premières transitions sont acceptées avec un taux proche de 1, ce qui

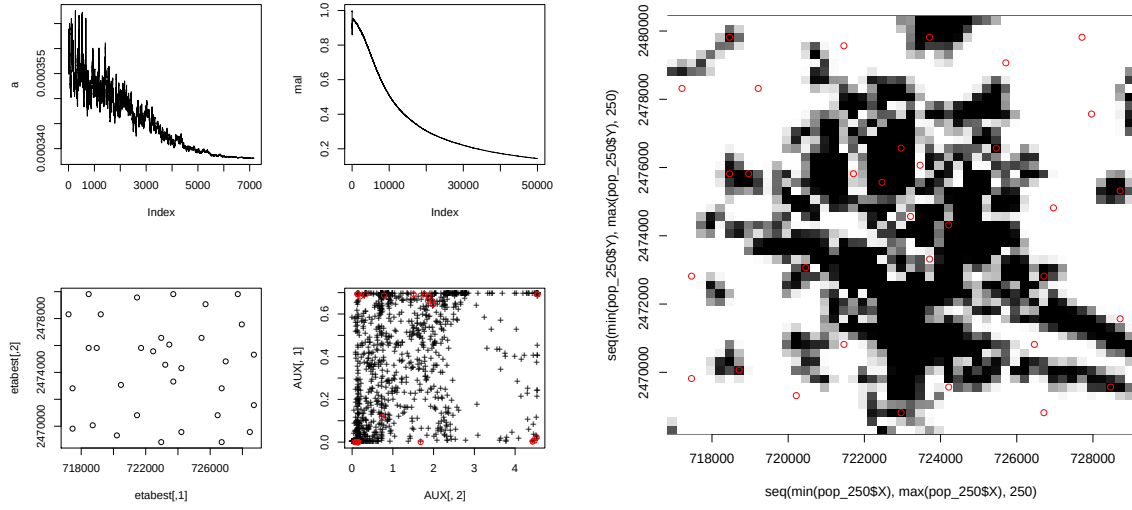


FIG. 6 – Résultats pour un rapport pépite/linéaire = 1, avec effet proportionnel

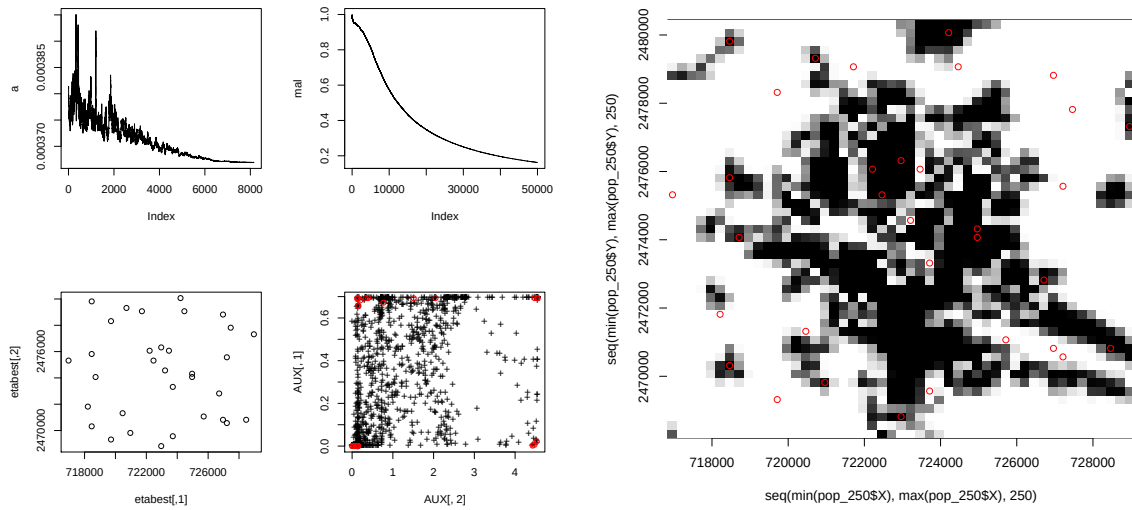


FIG. 7 – Résultats pour un rapport pépite/linéaire = 4, sans effet proportionnel

autorise une bonne exploration de l'espace d'état. Ensuite, ce taux décroît au fur et à mesure des itérations, ce qui va de pair avec la diminution de la température. Parallèlement, la fonction objectif décroît de manière satisfaisante, puis se stabilise, ce qui indique la convergence de l'algorithme. Enfin, des simulations répétées avec le même jeu de paramètres d'entrée et une "graine" différente ont donné des résultats assez similaires, ce qui prouve une bonne stabilité de ces résultats.

L'augmentation du rapport pépite/linéaire, dont l'effet sur le plan optimal est nettement visible sur le domaine physique, engendre mécaniquement une variance plus élevée sur les paramètres de la dérive. Si l'on compare les figures 5 et 7, sans effet proportionnel, on peut

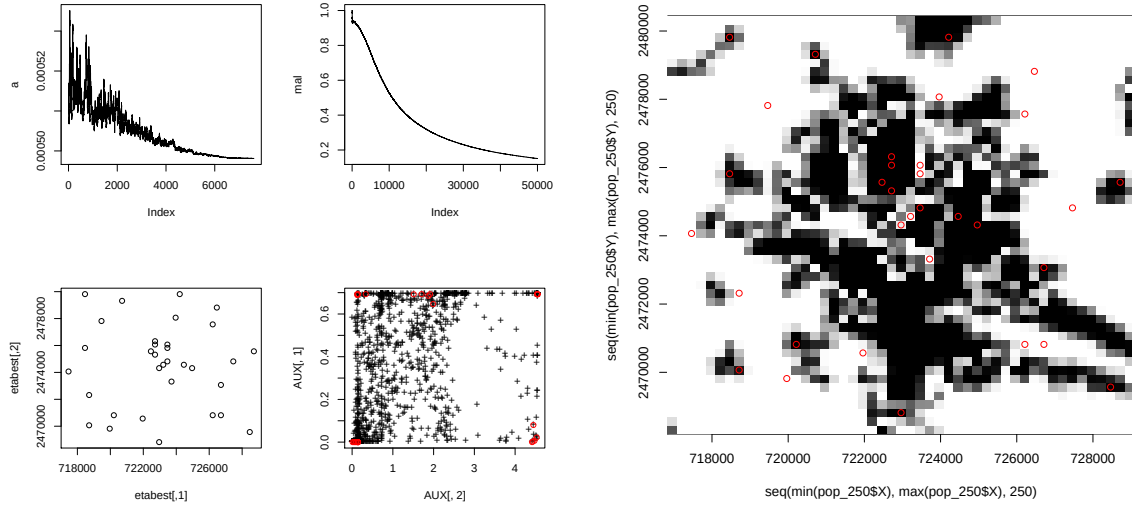


FIG. 8 – Résultats pour un rapport pèpète/linéaire = 4, avec effet proportionnel

voir comment la modélisation de la dérive devient prépondérante : dans le premier cas, on a un plan réparti de manière uniforme sur tout le domaine ; dans le second cas, on a plus de points au centre de l'agglomération, là où l'on est à même de balayer une plus grande plage de valeurs des auxiliaires, tandis que le reste des points se répartit à proximité des bords du champ. Autrement dit, dans le second cas, on va autoriser des distances plus grandes entre les points du plan d'échantillonnage pour mieux ajuster la dérive. Sur la représentation du plan optimal dans le nuage des auxiliaires, on voit ainsi sur la figure 7 que les points sont plus concentrés aux quatre coins du nuage, ceux qui permettent d'ajuster la dérive linéaire de manière optimale.

L'ajout d'un effet proportionnel est nettement visible dans les deux espaces de représentation. Sur le nuage des auxiliaires, on voit ainsi apparaître des points à proximité de la zone de délimitation de cet effet, lorsque le log translaté de la densité de population vaut 2 (figures 6 et 8).

4.2 Lille

Une difficulté supplémentaire apparaît dans le cas de Lille. La taille du domaine est beaucoup plus importante et, par suite, le calcul du critère est beaucoup plus coûteux : la grille comporte 249×266 noeuds. 50000 itérations de l'algorithme prennent ainsi 3 jours contre 3 heures pour Reims. Une simplification du calcul du critère, proposée dans [7], s'est avérée trompeuse : elle consistait à calculer le critère sur une grille plus grossière que la grille d'étude. Mécaniquement, les points d'échantillonnage se plaçaient alors sur la grille grossière, minimisant la variance d'estimation en ces points. Nous avons également envisagé de calculer le critère par voisinage glissant mais les temps de calculs se sont révélés prohibitifs, chaque itération faisant intervenir un nombre d'inversions de matrice (de covariance) égal au nombre de points de l'échantillonnage. Nous avons choisi ici d'optimiser un plan d'échantillonnage de 200 points. En raison des temps de calculs importants, nous n'avons effectué que deux simulations. L'effet proportionnel a été fixé avec un effet pèpète 3 fois plus important dans les zones où le log translaté de la densité de population est supérieur à 2, comme nous l'indiquait les

données (voir paragraphe 3.3, p. 7). Par rapport à Reims, nous avons modifié la température initiale, fixée ici à 10^{-8} , et la distance maximale initiale proposée pour le déplacement du point, fixée à $30 \times 250m$.

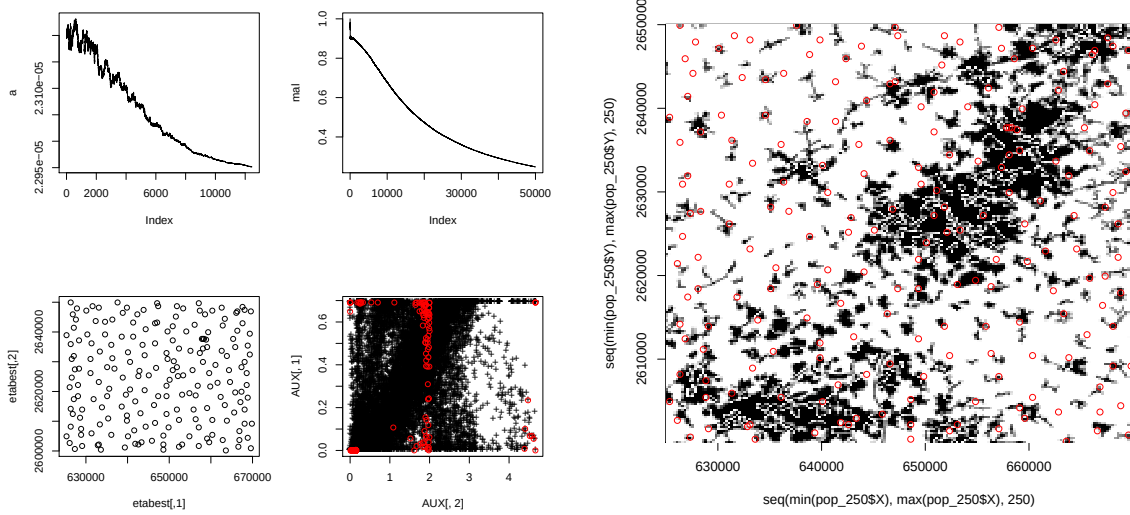


FIG. 9 – Résultats pour un rapport pépité/linéaire = 1, avec effet proportionnel

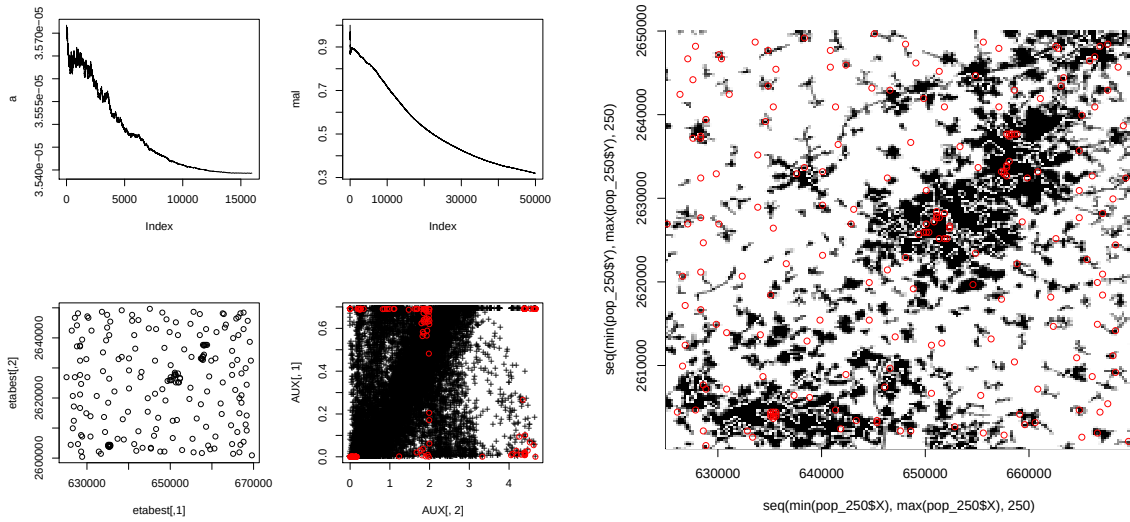


FIG. 10 – Résultats pour un rapport pépité/linéaire = 4, avec effet proportionnel

Ici aussi, l'algorithme s'est bien comporté et semble avoir convergé. La figure 9 fait apparaître une densité importante de points dans les centre des zones urbanisées. La représentation dans le nuage des auxiliaires fait apparaître la répartition caractéristique due à l'effet proportionnel déjà observée sur Reims, les points se positionnant aux extrémités du nuage et de la zone de délimitation de l'effet. D'autre part, la répartition spatiale des points montre une bonne couverture du domaine.

4.3 Bordeaux : comparaison avec la méthode précédente

Sur cette ville, les données de concentrations en benzène n'ont pas été prises en compte pour construire la modélisation a priori. L'algorithme a été mis en oeuvre sur la base de l'information extraite des jeux de données de Reims et Lille. Le domaine a été discrétisé sur une grille 184×154 de pas 250m, soit un nombre de mailles environ deux fois inférieur à celui de Lille. Le temps de calcul s'en trouve ainsi réduit mais reste conséquent : 2 jours environ. La densité de population n'était pas disponible selon le même mode de calcul que précédemment, nous disposons sur Bordeaux de la densité de population dans un rayon de 100 m. La distribution de cette grandeur est sensiblement différente de la densité de population calculée sur une maille de 250m. Nous prendrons la médiane du log, soit 4.7, comme borne déterminant l'effet proportionnel.

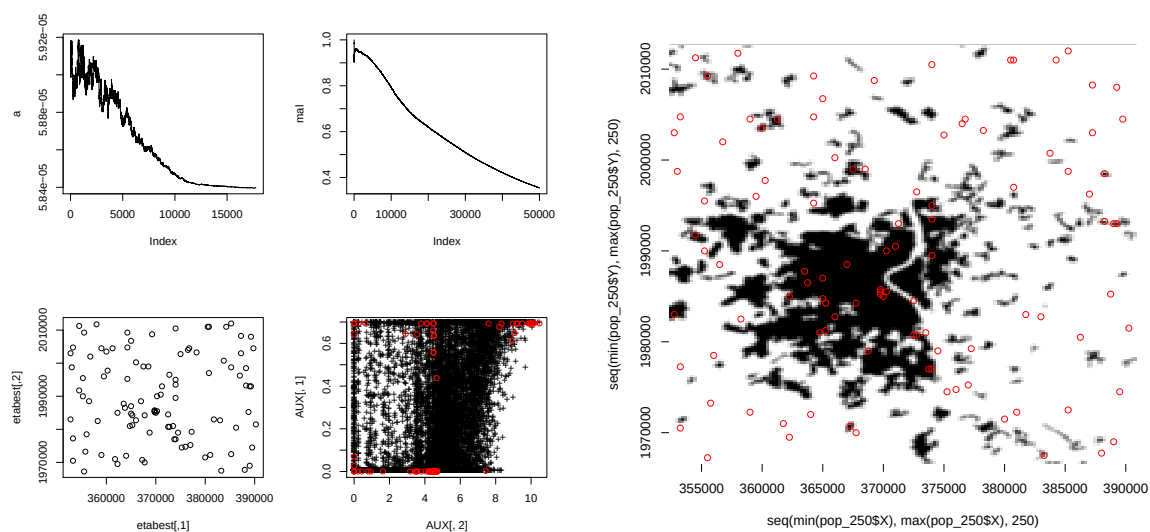


FIG. 11 – Résultats pour un rapport pépite/linéaire = 4, avec effet proportionnel

Les résultats correspondent aux attentes : le plan optimal représenté en figure 11 semble avoir une bonne répartition spatiale et parcourt le nuage des auxiliaires conformément à ce qui a été observé sur Lille et Reims. Au regard du critère calculé, il est bien évidemment meilleur que le plan proposé dans [6], de par la nature de la méthode.

De la même manière, nous avons représenté en figure 12, le plan optimal généré par l'algorithme de recuit-simulé lorsque le rapport pépite sur linéaire est égal à 1.

Ici aussi les résultats ne sont guère surprenants : une part plus forte de la structuration dans la variance du résidu entraîne mécaniquement une répartition spatiale plus uniforme puisque la réduction de la variance offerte par le krigeage à proximité des points de mesure est alors plus importante, induisant une couverture spatiale plus complète.

4.4 Bilan

Globalement, cette méthode présente l'avantage d'être facilement implémentable. Elle semble en outre fournir des résultats satisfaisants et ce de manière automatique, bien que les temps de calculs soient importants, ce dernier point étant lié à l'utilisation d'un langage interprété. On pourrait s'attendre à de bien meilleures performances en codant l'algorithme en C par exemple.

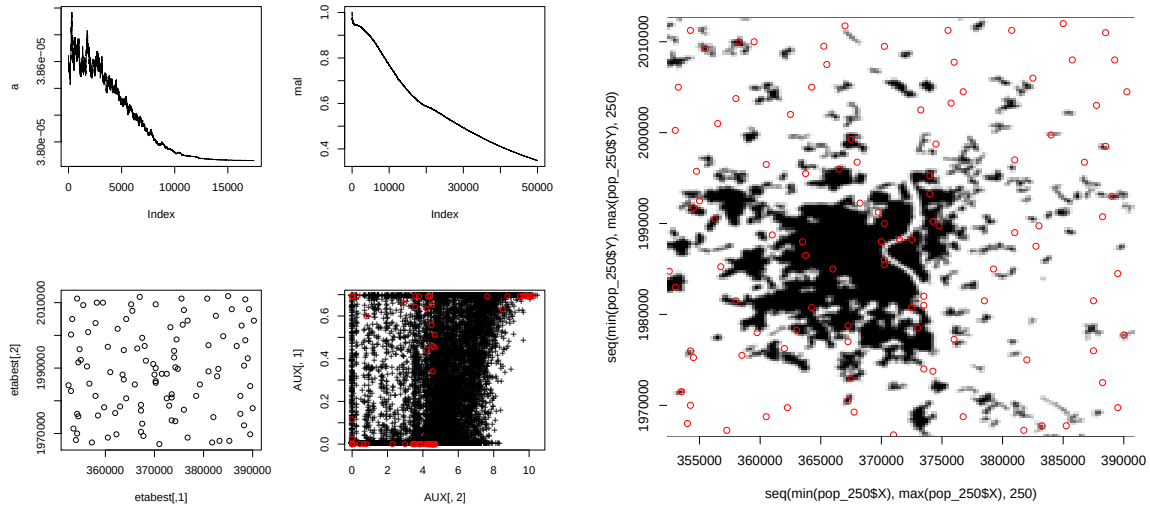


FIG. 12 – Résultats pour un rapport pépite/linéaire = 1, avec effet proportionnel

Notons que la valeur du critère pour le plan optimal était systématiquement nettement inférieure à celle calculée pour les plans existants, cette dernière étant souvent du même ordre que la valeur calculée pour un plan tiré aléatoirement, selon un processus de Poisson homogène. La principale faiblesse de cette approche est d'être associée à une modélisation a priori restrictive : il faut connaître a priori les variables auxiliaires significatives ainsi que leur ordre (linéaire, quadratique,...) ; il faut également fixer le modèle de covariance des résidus. Enfin, le plan obtenu en minimisant ce critère ne permettra pas forcément de remettre en cause la modélisation initiale une fois les mesures effectuées et pourra éventuellement conduire à de fâcheux résultats pour l'estimation. C'est pourquoi une approche bayésienne pourrait se révéler intéressante. Celle-ci pourrait en effet nous permettre d'appréhender une incertitude sur certains paramètres (rapport pépite/linéaire, effet proportionnel, portée,...) et d'employer une gamme plus large de critères.

Ceci dit, le choix d'une approche bayésienne nous inciterait à nous orienter vers d'autres algorithmes numériques que le recuit-simulé : outre sa lenteur, il peut présenter des défauts en présence de nombreux minima locaux : si la température est déjà suffisamment basse, il est possible que la chaîne générée se bloque dans le bassin d'attraction d'un minimum local.

5 Perspectives : approche bayésienne et méthodes numériques alternatives

Un tour d'horizon de la littérature concernant la planification d'expérience permet de se rendre compte que l'approche développée ici entre dans le cadre plus large des méthodes bayésiennes. En effet, plutôt qu'une modélisation a priori figée, il est possible de considérer une distribution a priori sur les paramètres de cette modélisation. On peut par exemple imaginer que le rapport pépite/linéaire est distribué uniformément sur l'intervalle $[1, 4]$. Il est alors possible de chercher le plan optimal selon un certain critère $O(\eta)$, par exemple (4), tout en prenant en compte cette incertitude. Cela revient à optimiser l'espérance de ce critère

relativement au paramètre $\theta \in \Theta$ incertain. Dans le cas de (4), on aura ainsi :

$$\mathbb{E}(O_\theta(\eta)) = \int_{\Theta} \int_{\mathcal{X}} \mathbb{V}(Z(x)|\theta) dx d\theta, \quad (8)$$

où $\mathbb{V}(Z(x)|\theta)$ est la variance de prédiction du krigeage universel conditionnellement à θ . Finalement, l'approche de Brus et Heuvelink revient à faire du bayésien avec des a priori qui sont des dirac en une valeur fixée du paramètre.

L'approche bayésienne pour la planification d'expérience a été formulée dès les années 60 mais il a fallu attendre le milieu des années 90 pour voir apparaître des applications pratiques, notamment en raison du coût numérique de ces méthodes. La très bonne revue de Chaloner et Verdinelli [5] donne un certain nombre d'exemples dans le cadre des modèles linéaires généralisés pour l'estimation de paramètres.

En géostatistique, les premiers articles faisant mention d'une approche bayésienne datent des années 90 et traitent de l'estimation des paramètres du modèle ou de la prédiction de manière indépendante, voir [7] et sa bibliographie. Dans le premier cas, le plan est optimisé afin d'obtenir des paramètres de variance minimale et dans le second cas, les paramètres du modèle étant fixés, on cherche à minimiser la variance de prédiction. Deux articles récents [15] et [7] proposent d'intégrer les incertitudes sur les paramètres et sur la prédiction dans un cadre unique. Dans le premier, l'estimation des paramètres de la covariance est réalisée par maximum de vraisemblance et leur variance estimée par la matrice d'information de Fisher ; on fait donc l'hypothèse que le phénomène étudié suit une distribution gaussienne. Dans le second, une approche bayésienne est adoptée, impliquant le choix d'une loi a priori pour chacun des paramètres. Cette dernière approche semble plus prometteuse car facilement extensible à une grande variété de modèles géostatistique. En outre, elle se passe de l'hypothèse gaussienne. Elle permet enfin d'intégrer facilement l'information disponible préalablement à l'étude.

La planification d'expérience par les méthodes bayésiennes nécessite la définition d'un critère à maximiser en fonction des résultats attendus de l'étude. Dans la littérature, ce critère est appelé fonction d'utilité et l'on cherchera à le maximiser.

Plus précisément, le problème de l'échantillonnage (ou planification d'expérience) optimal(e) se résout en maximisant l'espérance d'une fonction d'utilité définie selon une modélisation donnée. On optimise cette quantité en fonction du plan $\eta \in H$. L'expérience est définie par un modèle $p_\eta(Z|\theta)$, c'est-à-dire la loi des données observables $Z \in \mathcal{Z}$ conditionnellement au paramètre $\theta \in \Theta$, et dépendant éventuellement du plan $\eta \in H$. Le modèle est complété par une loi a priori sur le paramètre $p(\theta)$. L'utilité est une fonction, par exemple le coût quadratique négatif $u = -\{\theta - \mathbb{E}_\eta(\theta|Z)\}^2$, ou bien une quantité spécifique au problème comme $u =$ nombre total de patients traités, ou encore $1 - O(\eta)$, où $O(\eta)$ est défini dans l'équation (4), p. 2. Puisque le plan d'échantillonnage est choisi avant que l'expérience ne soit menée, il faut maximiser l'espérance de $u(\cdot)$, conditionnellement à (θ, Z) . Formellement, le problème se pose de la manière suivante :

$$\eta^* = \arg \max_{\eta \in H} U(\eta), \text{ où } U(\eta) = \int_{\mathcal{Z}} \int_{\Theta} u(\eta, \theta, z) p_\eta(z|\theta) p(\theta) d\theta dz, \quad (9)$$

$U(\eta)$ est l'espérance de l'utilité pour le plan η . Autrement dit, l'approche bayésienne suggère l'emploi d'une fonction d'utilité qui reflète l'objet de l'expérience, en liant le choix du plan à un problème de décision. On sélectionne ensuite le plan qui maximise l'espérance de l'utilité. Notons également que la fonction d'utilité peut dépendre explicitement du phénomène étudié Z . Il est ainsi nécessaire de le simuler selon le modèle a priori. Ceci rend le calcul de (9) très coûteux si Z est de dimension grande et implique de choisir une distribution a priori pour Z , mais cela permet de construire des fonctions d'utilité très informatives, incluant par exemple

la probabilité d'approcher un seuil.

A moins que le modèle $p_\eta(Z|\theta)$ et l'a priori $p(\theta)$ ne soient choisis de manière à rendre les calculs explicites, la maximisation de U passe par des méthodes numériques. Dans l'approche décrite dans ce rapport, nous avons employé le recuit-simulé. Il existe néanmoins d'autres algorithmes dans la littérature qui ont été proposés pour pallier les faiblesses de ce dernier. On peut citer principalement l'algorithme de Müller [13], qui s'apparente à un recuit-simulé dans le cadre du modèle de probabilité augmenté ([12]), et l'algorithme proposé dans [1] basé sur une approche particulière. Plus récemment, une famille d'algorithmes a été proposée dans [4] (décrit en annexe) et dont une adaptation judicieuse pourrait se révéler très efficace. L'un des avantages serait notamment la possibilité de paralléliser l'algorithme afin de diminuer les temps de calculs. En outre, ces méthodes permettent de définir des algorithmes permettant d'éviter de rester coincer dans un minimum local.

Références

- [1] AMZAL, B., BOIS, F. Y., PARENT, E., AND ROBERT, C. P. Bayesian-optimal design via interacting particle systems. *Journal of the American Statistical Association* 101, 474 (2006), 773–785.
- [2] BRUS, D. J., AND HEUVELINK, G. B. M. Optimization of sample patterns for universal kriging of environmental variables. *Geoderma* 138 (2007), 86–95.
- [3] CAPPÉ, O., GODSILL, S., AND MOULINES, E. An overview of existing methods and recent advances in sequential Monte Carlo. In *Proceedings of IEEE* (2007), vol. 95, pp. 899–924.
- [4] CAPPÉ, O., GUILLIN, A., MARIN, J.-M., AND ROBERT, C. P. Population monte-carlo. *Journal of Computational and Graphical Statistics* 13, 4 (2004), 907–929.
- [5] CHALONER, K., AND VERDINELLI, I. Bayesian experimental design : a review. *Statistical Science* 10, 3 (1995), 273–304.
- [6] DE FOUQUET, C., AND FAUCHEUX, C. Recommandations sur le schéma d'échantillonnage pour les campagnes de mesure de la qualité de l'air. Tech. rep., Mines Paristech, 2009.
- [7] DIGGLE, P. J., AND LOPHAVEN, S. Bayesian geostatistical design. *Scandinavian Journal of Statistics* 33, 1 (2006), 53–64.
- [8] GROENIGEN, J. W., SIDERIUS, W., AND STEIN, A. Constrained optimization of soil sampling for minimisation of the kriging variance. *Geoderma* 87 (1999), 139–259.
- [9] KIRKPATRICK, S., GELATT JR., C. D., AND VECCHI, M. P. Optimization by Simulated Annealing. *Science* 220, 4598 (1983), 671–680.
- [10] MATHERON, G. *La théorie des variables régionalisées et ses applications*. Les cahiers du centre de morphologie mathématiques de Fontainebleau. Ecole des Mines de Paris, 1970. Fascicule 5.
- [11] METROPOLIS, N., ROSENBLUTH, A., ROSENBLUTH, M., AND TELLER, A. T. M. Equations of state calculations by fast computing machines. *Journal of Chemical Physics* 21 (1953), 1087–1091.
- [12] MÜLLER, P. Simulation based optimal design. In *Bayesian statistics* (1999), J. Bernardo, J. Berger, A. David, and A. Smith, Eds., vol. 6, Oxford University Press, pp. 459–474.
- [13] MÜLLER, P., SANSÓ, B., AND DE IORIO, M. Optimal bayesian design by inhomogeneous Markov chain simulation. *Journal of the American Statistical Association* 99, 467 (2004), 788–798.

- [14] ROBERT, C. P., AND CASELLA, G. *Monte-Carlo Statistical Methods*, 2nd edition, ed. Springer, Berlin, 2004.
- [15] ZHU, Z., AND STEIN, M. L. Spatial sampling design for parameter estimation of the covariance function. *Journal of Statistical Planning and Inference* 134, 2 (2005), 583 – 603.

6 Annexe

6.1 Résultats des régressions

6.1.1 Lille

Campagne d'été

Call:

```
lm(formula = dat$z1.été ~ 0 + log(1 + dat$nbhabit250) + log(1 +  
    dat$propocc250))
```

Residuals:

Min	1Q	Median	3Q	Max
-0.63656	-0.10822	0.04915	0.33701	1.15886

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
log(1 + dat\$nbhabit250)	0.31159	0.02309	13.496	< 2e-16 ***
log(1 + dat\$propocc250)	0.36910	0.10817	3.412	0.000769 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.356 on 216 degrees of freedom

Multiple R-squared: 0.877, Adjusted R-squared: 0.8759

F-statistic: 770.1 on 2 and 216 DF, p-value: < 2.2e-16

Campagne d'hiver

Call:

```
lm(formula = dat$z2.hiver ~ 0 + log(1 + dat$nbhabit250) + log(1 +  
    dat$propocc250))
```

Residuals:

Min	1Q	Median	3Q	Max
-0.57111	-0.16570	0.06527	0.47186	1.37925

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
log(1 + dat\$nbhabit250)	0.39797	0.03132	12.708	< 2e-16 ***
log(1 + dat\$propocc250)	0.44226	0.14447	3.061	0.00249 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4669 on 213 degrees of freedom

Multiple R-squared: 0.8634, Adjusted R-squared: 0.8622

F-statistic: 673.4 on 2 and 213 DF, p-value: < 2.2e-16

Moyenne

Call:

```
lm(formula = dat$z3.moyenne ~ 0 + log(1 + dat$nbhabit250) + log(1 +
```

```

dat$propocc250))

Residuals:
      Min       1Q   Median       3Q      Max
-0.56806 -0.11902  0.04752  0.32211  1.03723

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
log(1 + dat$nbhabit250)  0.36339     0.02656   13.681 < 2e-16 ***
log(1 + dat$propocc250)  0.35475     0.12412    2.858  0.00472 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3875 on 197 degrees of freedom
Multiple R-squared:  0.8839,    Adjusted R-squared:  0.8827
F-statistic: 749.9 on 2 and 197 DF,  p-value: < 2.2e-16

```

6.1.2 Reims

Campagne de février

```

Call:
lm(formula = dat$z1.B_4.au.18.02 ~ 0 + log(1 + dat$nbhabit250) +
    log(1 + dat$propocc250))

```

```

Residuals:
      Min       1Q   Median       3Q      Max
-0.7141 -0.2940 -0.1226  0.3555  1.8995

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
log(1 + dat$nbhabit250)  0.3246     0.1085    2.993 0.005844 **
log(1 + dat$propocc250)  2.0164     0.5281    3.818 0.000714 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

Residual standard error: 0.6612 on 27 degrees of freedom
(2372 observations deleted due to missingness)
Multiple R-squared:  0.9261,    Adjusted R-squared:  0.9206
F-statistic: 169.1 on 2 and 27 DF,  p-value: 5.367e-16

```

Campagne de mai

```

Call:
lm(formula = dat$z2.B_5.au.19.05 ~ 0 + log(1 + dat$nbhabit250) +
    log(1 + dat$propocc250))

```

```

Residuals:
      Min       1Q   Median       3Q      Max
-0.886725 -0.055642  0.003316  0.133318  0.555176

```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
log(1 + dat\$nbhabit250)	0.06783	0.04069	1.667	0.107100
log(1 + dat\$propocc250)	0.83417	0.19815	4.210	0.000253 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2481 on 27 degrees of freedom

(2372 observations deleted due to missingness)

Multiple R-squared: 0.9037, Adjusted R-squared: 0.8965

F-statistic: 126.6 on 2 and 27 DF, p-value: 1.911e-14

Campagne d'août

Call:

```
lm(formula = dat$z3.B_4.au.18.08 ~ 0 + log(1 + dat$nbhabit250) +  
    log(1 + dat$propocc250))
```

Residuals:

Min	1Q	Median	3Q	Max
-0.37537	-0.08963	0.03891	0.10891	0.27462

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
log(1 + dat\$nbhabit250)	0.09070	0.02908	3.119	0.00428 **
log(1 + dat\$propocc250)	0.17194	0.14160	1.214	0.23516

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1773 on 27 degrees of freedom

(2372 observations deleted due to missingness)

Multiple R-squared: 0.8361, Adjusted R-squared: 0.8239

F-statistic: 68.86 on 2 and 27 DF, p-value: 2.496e-11

Campagne de novembre

Call:

```
lm(formula = dat$z4.B_12.au.26.11 ~ 0 + log(1 + dat$nbhabit250) +  
    log(1 + dat$propocc250))
```

Residuals:

Min	1Q	Median	3Q	Max
-0.49907	-0.12602	0.05996	0.18251	0.55872

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
log(1 + dat\$nbhabit250)	0.11591	0.03907	2.966	0.006237 **
log(1 + dat\$propocc250)	0.76882	0.19025	4.041	0.000397 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2382 on 27 degrees of freedom
 (2372 observations deleted due to missingness)
 Multiple R-squared: 0.9299, Adjusted R-squared: 0.9247
 F-statistic: 179 on 2 and 27 DF, p-value: 2.627e-16

Moyenne

Call:

```
lm(formula = dat$z5.Moyenne_.B ~ 0 + log(1 + dat$nbhabit250) +
    log(1 + dat$propocc250))
```

Residuals:

Min	1Q	Median	3Q	Max
-0.32000	-0.14547	-0.02795	0.24688	0.80561

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
log(1 + dat\$nbhabit250)	0.1446	0.0494	2.926	0.006875 **
log(1 + dat\$propocc250)	1.0711	0.2405	4.453	0.000132 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3012 on 27 degrees of freedom
 (2372 observations deleted due to missingness)
 Multiple R-squared: 0.9364, Adjusted R-squared: 0.9317
 F-statistic: 198.7 on 2 and 27 DF, p-value: < 2.2e-16

6.2 Population Monte-Carlo

L'idée de départ est celle de l'échantillonnage d'importance dont le principe est le suivant : si l'on cherche à estimer $\mathbb{E}_\pi[f(x)] = \mathcal{J}$ à partir d'un échantillon $(x_1, \dots, x_N)$ simulé selon une loi g (si π , notamment sa constante de normalisation, est inaccessible), l'estimateur non-biaisé par échantillonnage d'importance sera :

$$\widehat{\mathcal{J}} = \sum_{i=1}^N \omega^{(i)} f(x^{(i)}),$$

pour toute f telle que $\int f(x)^2 d\pi(x) < \infty$, où $\omega^{(i)} = \tilde{\omega}^{(i)} / \sum_{i=1}^N \tilde{\omega}^{(i)}$, $\tilde{\omega}^{(i)} = \pi(x^{(i)})/g(x^{(i)})$.

Partant, le principe des méthodes dites de filtrage particulaire visent à simuler selon une suite de densités $\{\pi_t\}_{t=1, \dots, T}$ en appliquant séquentiellement une méthode d'échantillonnage d'importance (voir [3] pour une revue).

Une autre famille de méthodes bien connue pour générer (approximativement) des simulations selon une loi π inaccessible sont les méthodes de Monte-Carlo par chaînes de Markov (voir [14]). Elles consistent à simuler une chaîne de Markov de loi stationnaire π .

L'idée de l'algorithme population Monte-Carlo (PMC) est la suivante : si l'on considère à l'itération t d'un algorithme MCMC un échantillon de N points $\mathbf{x}_t = (x_t^{(1)}, \dots, x_t^{(N)})$, on produira l'itération suivante de $\mathbf{x}_{t+1}$ en générant les composantes $x_{t+1}^{(i)}$ selon un noyau propositionnel $q(x|x_t^{(i)})$. Cependant, plutôt que d'accepter chacun des nouveaux états $x_{t+1}^{(i)}$ (ce qui reviendrait à utiliser N chaînes en parallèle) ou bien l'ensemble de l'échantillon $\mathbf{x}_{t+1}$ globalement (ce qui engendrerait des taux d'acceptation très faibles en raison de la dimension), l'algorithme PMC permet de s'affranchir des justifications de convergence de la méthode MCMC, en corrigeant à chaque itération l'emploi d'une mauvaise densité par l'introduction de poids d'importance. Ainsi, si l'échantillon $\mathbf{x}_t$ est généré en simulant $x_t^{(i)}$ selon $q_t^{(i)}$, indépendamment les uns des autres, en associant à chaque point $x_t^{(i)}$ le poids d'importance

$$\varrho_t^{(i)} = \pi(x_t^{(i)})/q_t^{(i)}(x_t^{(i)}, x_{t-1}^{(i)}), \quad i = 1, \dots, N,$$

on peut construire des estimateurs sans biais

$$\widehat{\mathcal{J}}_t = \frac{1}{N} \sum_{i=1}^N \varrho_t^{(i)} f(x_t^{(i)}),$$

pour toute fonction intégrable. Evidemment, dans la plupart des cas, la méconnaissance de π (de sa constante de normalisation) implique l'emploi des poids suivants :

$$\varrho_t^{(i)} \propto \pi(x_t^{(i)})/q_t^{(i)}(x_t^{(i)}, x_{t-1}^{(i)}), \quad i = 1, \dots, N,$$

normalisés. Afin d'éviter des problèmes de dégénérescence au cours de l'algorithme, plutôt que de mettre à jour les poids à chaque itération, il est préférable de rééchantillonner parmi les $(x_t^{(1)}, \dots, x_t^{(N)})$ avec remplacement selon les poids $\varrho_t^{(i)}$, auquel cas, l'échantillon ainsi construit est équivalent à un échantillon i.i.d. suivant la loi empirique pondérée associée à π .

Voici le schéma de l'algorithme :

Algorithm 2 Population Monte-Carlo

```
for  $t = 1 \dots T$  do
  for  $i = 1 \dots N$  do
    Sélectionner le noyau propositionnel  $q_t^{(i)}(\cdot, x_{t-1}^{(i)})$ 
    Générer  $x_t^{(i)} \sim q_t^{(i)}(\cdot, x_{t-1}^{(i)})$ 
    Calculer  $\varrho_t^{(i)} = \pi(x_t^{(i)})/q_t^{(i)}(x_t(i), x_{t-1}^{(i)})$ 
  end for
  Normaliser les  $\varrho_t^{(i)}$  de sorte que  $\sum_{i=1}^N \varrho_t^{(i)} = 1$ 
  Rééchantillonner  $N$  valeurs parmi les  $x_t^{(i)}$  avec replacement, en utilisant les poids  $\varrho_t^{(i)}$ 
end for
```

L'avantage majeur de cet algorithme est la possibilité d'employer des noyaux $q_t^{(i)}$ arbitraires, tout en maintenant la validité de la méthode. Les noyaux propositionnels peuvent ainsi être choisis en fonction des performances des noyaux précédents $q_{t-1}^{(i)}$ et notamment en fonction des échantillons déjà générés. Le cadre de cet algorithme nous autorise donc à construire des schémas adaptatifs en fonction de nos besoins, la validité étant assurée par l'étape d'échantillonnage d'importance.