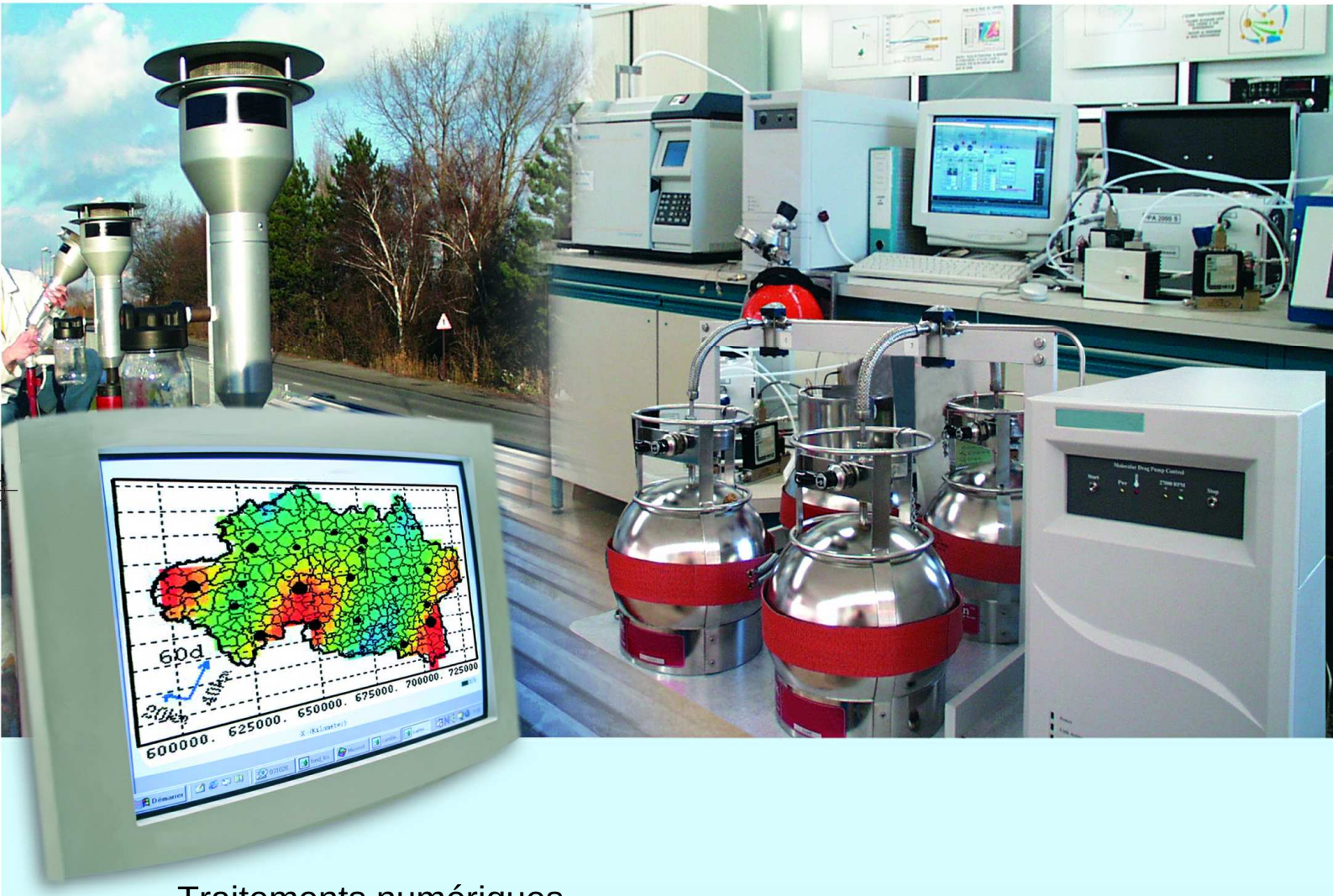
LCSQA

Laboratoire Central
de Surveillance de la Qualité de l'Air




Traitements numériques

# Assistance à l'exploitation de données de campagnes et à la réalisation de cartographies (1/2)

## Organisation d'une formation en statistique

**Décembre 2010**

Programme 2010

**L. MALHERBE**

INERIS

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Ministère
de l'Écologie,
de l'Énergie,
du Développement
durable
et de la Mer

# 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. Ces travaux en matière de pollution atmosphérique supportés financièrement par la Direction Générale de l'Energie et du Climat du Ministère de l'Ecologie, de l'Energie, du Développement Durable et de la Mer sont réalisés avec le souci constant d'améliorer le dispositif de surveillance de la qualité de l'air en France, coordonné au plan technique par l'ADEME, en apportant un appui scientifique et technique 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.**

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Ministère de l'Écologie, de l'Énergie, du Développement durable et de la Mer

# Assistance à l'exploitation de données de campagnes et à la réalisation de cartographies

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**

Ce document comporte 12 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)<br>Direction des Risques Chroniques | Responsable de l'unité Modélisation Atmosphérique et Cartographie Environnementale (MOCA)<br>Direction des Risques Chroniques | Responsable LCSQA/INERIS<br>Direction des Risques Chroniques |
| Visa | | | |

# TABLE DES MATIÈRES

## RESUME

Le LCSQA a été chargé d'organiser pour 2010 une formation en statistique à l'intention des AASQA. Par un questionnaire qui leur était destiné sur le site Internet du LCSQA, ces dernières ont été invitées à exprimer leurs attentes.

27 réponses issues de 22 AASQA ont été recueillies. Les besoins des AASQA relèvent à la fois de la statistique descriptive élémentaire et de la statistique multivariée plus avancée. On note également un intérêt pour le logiciel libre R. Une première trame de formation susceptible de satisfaire à ces besoins a été élaborée. Elle se compose de trois parties : statistique descriptive mono et bivariée ; initiation à R ; analyse multivariée et régression.

Deux sessions de formation identiques de 3 jours chacune auront lieu en 2010. Deux sessions supplémentaires sont prévues pour le début de l'année 2011.

# 1. INTRODUCTION

Pour satisfaire à la demande exprimée en CPT, le LCSQA organisera en 2010 une formation en statistique à l'intention des AASQA. Ces dernières ont été invitées à répondre à un questionnaire sur le site Internet du LCSQA et à y préciser leurs attentes.

La présente note synthétise les réponses obtenues. Une première trame de formation susceptible de convenir au plus grand nombre est ensuite proposée.

# 2. BILAN DES REPONSES AU QUESTIONNAIRE

## 2.1 REPONSES OBTENUES

22 AASQA sont représentées sur un total de 27 réponses (selon les AASQA, réponses globales ou individuelles) :

- AIR APS
- AIR BREIZH
- AIR COM
- AIR Languedoc-Roussillon
- AIR Normand
- AIRAQ
- AIRLOR
- AIRPARIF
- ASPA
- ATMO Auvergne
- ATMO Champagne-Ardenne
- ATMO Franche-Comté
- ATMO Lorraine Nord
- ATMO NPDC
- ATMO Poitou-Charentes
- ATMO Rhône-Alpes
- ATMOSF'Air Bourgogne
- GWADAIR
- LIG'AIR
- LIMAIR
- Madininair
- SCAL'AIR

## 2.2 BILAN DES PRATIQUES

### 2.2.1 METHODES ET CONTEXTES D'UTILISATION

Les AASQA traitent statistiquement leurs données à l'occasion d'études diverses, notamment : élaboration de plans d'échantillonnage, exploitation de données de campagnes (le plus souvent), bilans annuels, prévision statistique, cartographie, amélioration des résultats de modélisations.

Sont cités comme types de traitements: invalidation ou correction de données, analyse de séries temporelles, reconstitution de moyennes annuelles, analyse et modélisation des relations entre variables, interpolation

Ces traitements s'appuient notamment sur les techniques suivantes (entre parenthèses, le nombre de réponses mentionnant la technique) :

- Calcul de statistiques élémentaires (tous)
- Tests statistiques (1)
- Analyse en composantes principales (ACP) (2)
- Classification (CAH), arbres de décision, CART (3)
- Régression multiple (3)
- Reconstitution de données par stratification (4)

### 2.2.2 LOGICIELS UTILISES

Ont été cités les outils de calcul suivants (entre parenthèses, le nombre de réponses mentionnant l'outil)

- Excel (14)
- Statistica (1)
- R (9)
- OpenCalc (1)
- Python (1)
- Fortran (1)
- Xair (2)
- Isatis (11)
- ArcGis (2)

## 2.3 BESOINS

Pour 3 réponses : aucun besoin n'a été exprimé.

### 2.3.1 BESOINS METHODOLOGIQUES

- Connaissances de base : 14 réponses positives
- Statistiques multidimensionnelles : 15 réponses positives
- Modélisation statistique (régression) : 12 réponses positives
- Autres :
    - Tests statistiques et conditions d'utilisation (1 réponse)
    - Evaluation de la validité d'un échantillonnage ; Estimation d'une moyenne annuelle et de son intervalle de confiance ; Calcul des incertitudes de mesure sur divers pas de temps (1 réponse)
    - Etude de tendance (1 réponse)
    - Logique floue, réseaux de neurones (1 réponse)
    - Informations théoriques sur la géostatistique (1 réponse)
    - Géostatistique : simulations conditionnelles (1 réponse)

Pour 4 réponses, les besoins se limitent à des connaissances de base.

Pour 5 réponses, les besoins portent sur l'analyse multidimensionnelle et /ou sur la régression mais non sur les connaissances de base.

Pour 10 réponses, les besoins portent à la fois sur les connaissances de base et sur les méthodes d'analyse multidimensionnelle/méthodes de régression.

### 2.3.2 BESOINS PRATIQUES

Les réponses montrent un intérêt général pour R (cité 16 fois) et le souhait de se former à l'utilisation de ce logiciel. Les besoins ne sont cependant pas tous les mêmes :

- Utilisation de R pour des calculs simples,
- Formation avancée : statistiques élémentaires, représentations graphiques, régression, géostatistique,
- Utilisation de R si des bibliothèques sont spécifiquement développées pour la qualité de l'air ; utilisation de R en sortie de Xair ; utilisation de modules tels que Qair (développé par ATMO PC),
- Utilisation de R pour l'automatisation de tâches, notamment pour la cartographie,
- Utilisation de R via une interface logicielle.

Autre : intérêt pour XLstat (1 réponse)

Notons que pour 4 réponses, les besoins se limitent à la théorie. Pour 1 réponse, ils se limitent à la pratique.

# 3. PROPOSITIONS

## 3.1 CONTENU DE LA FORMATION

La formation s'organisera de manière progressive selon les trois parties suivantes:

1. Exploration des données à l'aide de graphes, interprétations graphiques : statistiques descriptives univariées (distribution d'une variable), bivariées (boîte à moustaches selon un facteur, nuage de corrélation) et multivariées (ACP, classification).
2. Tests statistiques (égalité de moyennes, significativité de la corrélation) et modélisation (ANOVA à un facteur, régression linéaire).
3. Traitement des dynamiques temporelles, approfondissement de la régression.

*NB : Cette trame n'est pas figée.*

Tous les exemples et travaux pratiques seront élaborés à partir de jeux de données préalablement fournis par les AASQA. Le module R-Commander, version interfacée de R, sera principalement utilisé. Des compléments sur les fonctions utiles de R, en particulier sur les fonctions spécifiques aux données de pollution de l'air, seront fournis dans la troisième partie.

**Remarques**

- Les questions théoriques ou pratiques liées à la géostatistique, abordées partiellement lors d'une formation LCSQA 2009, pourront faire l'objet d'une formation ultérieure (2011) si le besoin est avéré.
- Les questions liées à l'échantillonnage et à la reconstitution de données relèvent de la formation dispensée aux AASQA par le GT *Plans d'échantillonnage et reconstitution de données* (décembre 2008 et janvier 2009). Elles ne seront pas traitées de manière spécifique dans la formation mais pourront être mentionnées en même temps que certaines méthodes (ex : régression et reconstitution). S'il reste du temps, des rappels sur l'utilisation du logiciel pourront être éventuellement proposés à la fin de la formation.
- Un point d'information sur les modules de R spécifiques à l'analyse de données de qualité de l'air (tel le module Qair, développé par ATMO Poitou-Charentes) est prévu. Le temps disponible ne permettra cependant pas d'en faire une présentation détaillée.
- Le calcul des incertitudes de mesure, qui a fait l'objet de formations de la part du GT Incertitudes, ne sera pas inclus dans le programme de la formation.
- La logique floue et les réseaux de neurones (une seule demande) ne seront pas mis à l'ordre du jour. Ces méthodes auraient plutôt leur place dans une formation consacrée à la prévision statistique.

## 3.2 ORGANISATION

La formation s'étalera sur trois jours. La capacité d'accueil de la salle informatique limite le nombre de participants à 14 par session.

Deux sessions sont programmées en 2010. Compte tenu de l'intérêt manifesté pour cette formation, deux sessions supplémentaires sont prévues pour 2011.

Lieu : INERIS, site de Verneuil-en-Halatte, afin de bénéficier des installations informatiques et faciliter l'organisation pratique.

Période : entre la fin octobre et début décembre

*Tableau 1 – Déroulé possible de la formation*

| | | |
|---|---|---|
| 1) Première partie<br>Représentations et interprétations graphiques | Introduction à R et R-commander<br>Statistiques descriptives univariées et bivariées | Matin 1 |
| | Statistiques descriptives multivariées | A-midi 1 |
| 2) Deuxième partie<br>Tests et modélisation | Tests statistiques<br>Analyse de variance à un facteur | Matin 2 |
| | Régression | A-midi 2 |
| 3) Troisième partie<br>Compléments théoriques et pratiques | Traitement des dynamiques temporelles<br>Approfondissement de la régression<br>Fonctions utiles de R | Matin 3 |
| | Retour sur certains points<br>Application des notions abordées pendant la formation aux données apportées par chacun<br>Questions complémentaires | A-midi 3 |

## 3.3 INTERVENANTS

L'INERIS a pris contact avec plusieurs intervenants possibles. Deux prestataires ont été retenus.

La formation sera animée par le LCSQA, avec la participation de Frédéric Lavancier, maître de conférences à l'université de Nantes. F. Lavancier apportera son expertise théorique et pratique en statistique appliquée à la qualité de l'air.

Le LCSQA fera également appel à la société StatConsult (www.stat-consult.fr) pour la préparation des cas d'étude.

## 4. SUITES POSSIBLES POUR 2011

Deux sessions de formation supplémentaires sont prévues en janvier 2011. La nécessité de reconduire cette formation sera évaluée selon les demandes d'inscription.

Parmi les besoins exprimés par les AASQA, figure la possibilité de réaliser des calculs statistiques simples avec R par l'intermédiaire du site Internet du LCSQA, de la même façon que pour la planification de l'échantillonnage et la reconstitution de données. D'après l'expérience acquise par le LCSQA, de tels développements demanderaient un travail important. A la différence des programmes consacrés à l'échantillonnage et à la reconstitution de données de pollution de l'air, ce nouvel outil n'apporterait pas nécessairement de valeur ajoutée par rapport aux logiciels de statistique existants. Il semble préférable, et il pourrait être proposé, d'explorer les possibilités déjà offertes par certaines applications interfacées de R et de rédiger une note pratique à l'usage des AASQA.

## 5. LISTE DES ANNEXES

| Repère | Désignation | Nombre de pages |
|---|---|---|
| Annexe 1 | Fiche descriptive de l'étude | 2 |

**Annexe 1**

**Fiche descriptive de l'étude**

## *THEME 6 : Modélisation et 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 pas, 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<br>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 | | | |