



Surveillance de la qualité de l'air ambiant

Guide d'élaboration de plans d'échantillonnage temporel et de reconstitution de données



DONNÉES ET RÉFÉRENCES

ADEME



Agence de l'Environnement
et de la Maîtrise de l'Energie

Surveillance de la qualité de l'air ambiant

**Guide d'élaboration de plans
d'échantillonnage temporel et
de reconstitution de données**



► Auteurs

Michel Bobbia, AIR Normand
Fabrice Caïni, ATMO Poitou-Charentes
Tiphaine Delaunay, ATMO Nord Pas-de-Calais
Jean-Luc Houdret, LCSQA/EMD
Laure Malherbe, LCSQA/INERIS
Vladislav Navel, ATMO Poitou-Charentes
Frédéric Penven, AIR Pays de la Loire
Céline Phillips, ADEME
Arnaud Rebours, AIR Pays de la Loire
Pierre-Yves Robic, ORAMIP
Lionel Rosset, ATMO Auvergne

► Préface

La surveillance de la qualité de l'air ambiant repose en grande partie sur des stations fixes équipées d'analyseurs automatiques permettant une mesure en continue, et tout au long de l'année, de divers polluants atmosphériques.

D'autres approches reposant sur des campagnes de mesures périodiques peuvent toutefois également être mises en œuvre. Ce peut être le cas pour des polluants dont les niveaux et les risques n'imposent pas une mesure continue, pour des polluants pour lesquels l'on ne dispose pas de technique de mesures in situ et en temps réel, ou pour enrichir les informations issues de stations fixes.

C'est la raison pour laquelle, en France, les Associations Agréées de Surveillance de la Qualité de l'Air (AASQA) conduisent des campagnes de mesures « discontinues » au moyen de systèmes de prélèvement et de stations mobiles.

Ces campagnes peuvent être menées pour répondre à des objectifs de surveillance réglementaires ou à des fins d'études spécifiques. Elles comportent souvent un objectif d'évaluation des concentrations de polluants sur des échelles de temps (l'année par exemple) différentes de celles de la durée ou de la périodicité des campagnes. Ces campagnes imposent de définir des stratégies d'échantillonnage et de reconstitution de données permettant de s'assurer de leur représentativité, notamment temporelle.

Afin d'harmoniser les pratiques d'échantillonnage temporel et de reconstitution de données, le Ministère de l'écologie, de l'énergie, du développement durable et de la mer (MEEDDM) a souhaité, en 2005, la mise en place d'un groupe de travail visant à élaborer des démarches de référence et applicables par l'ensemble des AASQA.

Ce groupe de travail, coordonné par l'Agence de l'Environnement et de la Maîtrise de l'Énergie (ADEME), comportait des experts de six AASQA (Air Normand, ATMO Poitou-Charentes, ATMO Nord Pas-de-Calais, Air Pays de la Loire, ORAMIP et ATMO Auvergne). Il a bénéficié du support scientifique du Laboratoire Central de Surveillance de la Qualité de l'Air (LCSQA - INERIS et École des Mines de Douai) ainsi que du secrétariat de l'INERIS.

Les travaux du groupe de travail ont permis de dresser un état des lieux des pratiques des AASQA et de leurs homologues en Europe et, partant des méthodes employées, de rédiger ce guide et de développer des outils numériques associés. Ce guide, qui a été validé par le Ministère en charge du développement durable, constitue un document de référence lorsque l'on souhaite s'assurer de la représentativité des campagnes et des indicateurs qui en résultent.

Christian Elichegaray

Chef du département de Surveillance de la Qualité de l'Air de l'ADEME

► Sommaire

Introduction générale.....	6
1. Élaboration d'un plan d'échantillonnage.....	9
1.1 Introduction.....	10
1.2 Terminologie.....	12
1.2.1 Qu'est-ce qu'un échantillonnage ?.....	12
1.2.2 Qu'est-ce qu'un plan d'échantillonnage ?.....	13
1.2.3 Comment caractérise-t-on un plan d'échantillonnage ?.....	18
1.3 Étapes de l'élaboration d'un plan d'échantillonnage.....	20
1.3.1 Introduction.....	20
1.3.2 Étude des contraintes.....	20
1.3.3 Analyse de la variabilité temporelle des concentrations.....	21
1.3.4 Stratification temporelle.....	24
1.3.5 Dimensionnement de l'échantillonnage.....	26
1.3.6 Contrôle de la faisabilité du plan.....	30
1.3.7 Détermination des dates de mesure.....	31
1.4 Comment établir un plan d'échantillonnage en l'absence de données de référence ?.....	31
1.5 Évaluation des ressources.....	32
1.5.1 Planification des unités d'œuvre nécessaires pour une campagne à l'aide de moyens mobiles.....	32
1.5.2 Évaluation des coûts liés à des campagnes de prélèvement. Exemple : coûts de l'évaluation préliminaire des métaux lourds.....	33
1.5.3 Évaluation des coûts associés à un fonctionnement alterné des stations du réseau fixe.....	35
1.6 Conclusion.....	37

2. Reconstitution des paramètres statistiques	39
2.1 Introduction	40
2.1.1 Objectif	40
2.1.2 Mise en œuvre de la reconstitution	41
2.2 Les méthodes	42
2.2.1 Présentation des méthodes	42
2.2.2 Méthode des plans de sondage	42
2.2.3 Méthode « ISO » (méthode issue de la norme ISO 9359)	51
2.2.4 La régression linéaire	58
2.3 Choix d'une méthode	66
2.4 Conclusion	67
Références	68

Liste des annexes

1. Bilan des pratiques en Europe	69
2. Bilan des pratiques dans les AASQA	75
3. Fiche descriptive de la méthode des plans de sondage	85
4. Fiche descriptive de la méthode issue de la norme ISO	89
5. Fiche descriptive de la régression	93
6. Note théorique sur la corrélation	97
7. Définition de l'intervalle de confiance	104
8. Formulaire	106
9. Étude d'impact économique réalisée par AIR Pays de Loire	112
10. Cas d'application : échantillonnage temporel des HAP pour l'estimation d'une moyenne annuelle	117

► Introduction générale

L'objectif de ce document est de fournir un guide pratique pour toute Association Agréée de Surveillance de la Qualité de l'Air (AASQA) souhaitant élaborer une stratégie de surveillance temporelle et reconstituer des moyennes annuelles ainsi que des nombres de dépassements de seuils donnés.

Le champ d'application du guide concerne principalement la surveillance de polluants réglementés par modélisation et estimation objective. En effet, ces deux modes d'évaluation de la qualité de l'air autorisent le recours à la mesure discontinue, sans exigence particulière sur la nature du plan d'échantillonnage, et l'utilisation de méthodes d'estimation statistiques pour exploiter les données recueillies.

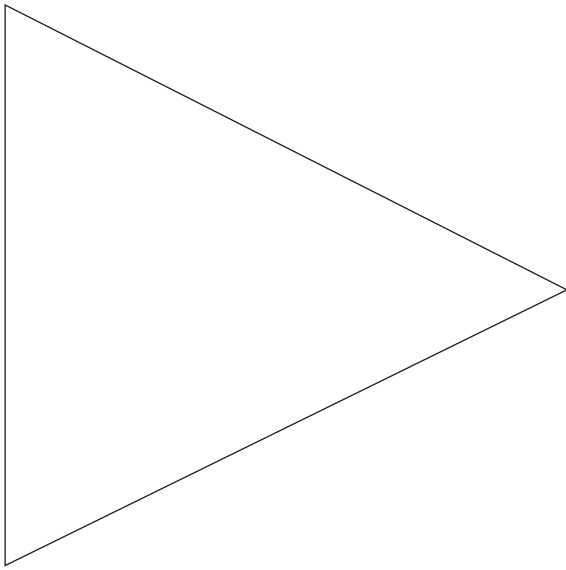
La mesure indicative et la mesure fixe (pour certains polluants précisés dans les directives européennes : particules, benzène, HAP, plomb et autres métaux lourds) constituent d'autres contextes de mise en œuvre de mesures discontinues. Dans ces cas cependant, le plan d'échantillonnage est fortement contraint par les directives européennes et aucune reconstitution statistique à des fins de surveillance réglementaire n'est possible. Toutefois, dans un souci de qualité, on pourra se servir de la première partie du guide pour s'assurer que la durée minimale imposée par la réglementation (ex : 14 % de l'année) suffit à un résultat de bonne précision et comparer l'efficacité de différents types de répartition uniforme (ex : une mesure aléatoire par semaine ou huit semaines réparties uniformément sur l'année).

Les campagnes ponctuelles destinées à des études spécifiques ne sont pas traitées en particulier dans ce document. La surveillance de la pollution de type industriel n'est pas non plus abordée, à cause de la grande variabilité temporelle des émissions, qui empêche de définir une stratégie générique d'échantillonnage temporel.

Ce guide comprend deux parties. Le chapitre 1 a pour objet de fournir les bases statistiques permettant à toute AASQA de planifier la répartition de ses mesures dans le temps, en fonction de contraintes relatives à la fois à la qualité de l'indicateur final et aux ressources disponibles. Il propose une méthode d'élaboration d'un plan d'échantillonnage temporel. Cette méthode a été développée par le groupe de travail en combinant différentes approches utilisées à ce jour par la communauté française de surveillance de la qualité de l'air et en s'appuyant sur les bases scientifiques de la théorie des sondages.

Le chapitre 2 présente trois méthodes de reconstitution de données : la méthode dite des « Plans de Sondage », la méthode « ISO » issue de la norme ISO 9359 et la méthode de régression linéaire. Ces méthodes optimisent l'exploitation des concentrations mesurées et permettent d'obtenir des indicateurs fiables, grâce à l'utilisation de variables auxiliaires. Les principes de ces méthodes sont décrits et illustrés par un exemple d'application. Les contraintes et performances comparées de ces méthodes sont également exposées.

En annexe, le bilan des pratiques françaises en matière d'échantillonnage temporel et de reconstitution des données est présenté ainsi qu'une synthèse des pratiques dans d'autres États membres. Afin de faciliter la compréhension des principes statistiques qui régissent les méthodes proposées dans ce guide, plusieurs documents en annexe expliquent les termes employés (coefficient de corrélation, intervalle de confiance...), synthétisent les trois méthodes de reconstitution, et explicitent les principales formules de calcul.



1

Élaboration d'un plan d'échantillonnage

► 1.1 Introduction

On souhaite établir des indicateurs de la qualité de l'air sur une période donnée (ex. : moyenne annuelle, nombre annuel de dépassements de seuils horaires ou journaliers) sans avoir recours à des mesures en continu sur l'ensemble de cette période.

Pourvu que la sélection des données dans le temps remplisse certaines conditions, il est possible de calculer des indicateurs d'une bonne précision tout en restreignant le nombre de mesures.

Quelles sont ces conditions ? Comment peut-on concevoir une sélection de données adaptée aux polluants considérés et à l'objectif de surveillance ?

En ce qui concerne la mesure fixe ou indicative, la Directive 2008/50/CE du Parlement européen et du Conseil du 21 mai 2008 et la directive 2004/107/CE du Parlement européen et du Conseil du 15 décembre 2004 contiennent un certain nombre d'exigences sur la période minimale à couvrir par des mesures et la répartition de ces dernières dans le temps (Tableau 1). L'utilisateur a peu de latitude dans la définition du plan d'échantillonnage mais il peut en ajuster le dimensionnement afin d'évaluer des indicateurs les plus fiables possible. Pour ce qui est de la modélisation et de l'estimation objective ou de toute étude sortant du cadre réglementaire, l'utilisateur est libre de choisir le plan d'échantillonnage qui correspond le mieux à ses contraintes (qualité souhaitée du résultat final, ressources disponibles).

Ce chapitre propose une démarche scientifique permettant d'élaborer une stratégie d'échantillonnage conforme aux contraintes existantes. La méthodologie développée s'inscrit dans le cadre scientifique de la théorie des sondages dont l'un des principaux aspects est la planification de la collecte des données. Cette théorie statistique est présentée en détail dans l'ouvrage de Tillé (2001). Saporta (2006, chapitre 20) en reprend les principaux aspects et les formules essentielles. La possibilité d'appliquer la théorie des sondages à la pollution atmosphérique a fait l'objet d'études antérieures (Lavancier et al., 2003 ; Houdret et Malherbe, 2005).

Polluant	SO ₂ NO _x NO ₂ CO	Benzène	PM ₁₀ PM _{2,5} Plomb	O ₃ NO NO ₂	B(a)P	As Ni Cd
Période minimale à prendre en compte pour la mesure fixe	Mesure en continu	35 % pour les sites de fond urbain et de proximité automobile 90 % pour les sites industriels ou par dérogation 14 % pour tous les types de sites*	Mesure en continu ou par dérogation 14 %*	Mesure en continu	33 % ou par dérogation 14 %*	50 % ou par dérogation 14 %*
Période minimale à prendre en compte pour la mesure indicative	14 %	14 %	14 %	10 % de l'été	14 % ou par dérogation 6 %*	14 % ou par dérogation 6 %*
Répartition des mesures sur la période d'étude pour la mesure fixe	Pour le benzène, en site de fond urbain ou de proximité automobile, égale répartition sur l'année				Égale répartition sur les jours de la semaine et sur l'année	
Répartition des mesures sur la période d'étude pour la mesure indicative	Une mesure aléatoire par semaine ou 8 semaines uniformément réparties sur l'année				Égale répartition sur les jours de la semaine et sur l'année	

* Si l'on peut démontrer que les objectifs de qualité sont atteints, c'est-à-dire que l'incertitude de mesure, augmentée de l'incertitude liée à la couverture temporelle incomplète, respecte la précision minimale demandée.

Tableau 1 - Surveillance réglementaire de la qualité de l'air : indications des Directives Européennes relatives à l'échantillonnage temporel, pour la mesure fixe et la mesure indicative.

► 1.2 Terminologie

1.2.1 Qu'est-ce qu'un échantillonnage ?

► 1.2.1.1 Échantillonnage

Un échantillonnage est un tirage d'un certain nombre d'**individus** (ou **unités**) dans une **population**.

► 1.2.1.2 Échantillonnage systématique

Un échantillonnage est **systématique** si les individus sont sélectionnés à intervalles réguliers (*ex : une mesure journalière tous les six jours*).

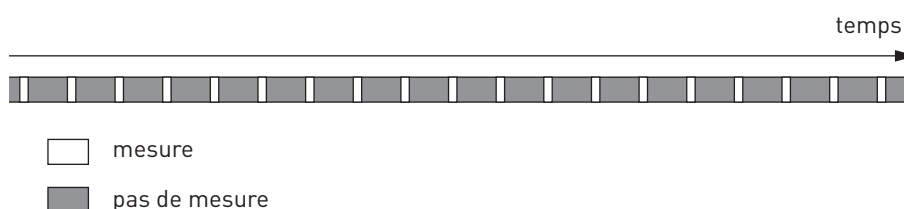


Figure 1 - Illustration d'un échantillonnage systématique.

► 1.2.1.3 Échantillonnage aléatoire

Un échantillonnage est **aléatoire** si les individus sont sélectionnés au hasard et de façon indépendante (*ex. : tirage au sort d'un certain nombre d'heures de mesure dans l'année*).

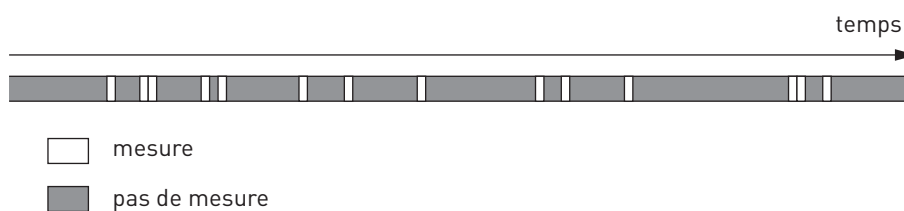


Figure 2 - Illustration d'un échantillonnage aléatoire simple.

► 1.2.1.4 Objet de l'échantillonnage

À l'aide de l'**échantillon d'individus**, on cherche à **reconstituer une donnée** relative à l'ensemble de la population (*ex. : la moyenne sur l'année de la concentration d'un polluant*).

1.2.2 Qu'est-ce qu'un plan d'échantillonnage ?

Afin d'estimer plus efficacement l'indicateur étudié (*ex. : afin de mieux estimer la moyenne avec un échantillon plus petit*), il convient de planifier la collecte des données.

► 1.2.2.1 Plan d'échantillonnage

Un plan d'échantillonnage définit combien d'individus seront sélectionnés et de quelle façon sera opérée cette sélection.

► 1.2.2.2 Plan d'échantillonnage stratifié

Dans un plan d'échantillonnage *stratifié*,

- la population est divisée en strates, c'est-à-dire en sous-ensembles distincts *a priori* plus homogènes que la population dans son entier ;
- un échantillon est prélevé au sein de chacune des strates, de façon systématique ou aléatoire.

La stratification d'une population est réalisée au moyen d'une **variable auxiliaire** ou d'une combinaison de variables auxiliaires dont les valeurs sont **disponibles pour tous les individus qui composent la population**.

En qualité de l'air, les variables auxiliaires considérées sont des variables liées aux concentrations, soit qu'elles influencent celles-ci directement (vent, température, émissions...) soit qu'elles traduisent indirectement ces influences (saison représentant les effets climatiques, jour de la semaine traduisant les quantités d'émissions, concentrations d'un autre polluant subissant les mêmes influences ou du même polluant sur un autre site...).

Ainsi, s'il est avéré que la concentration du polluant présente des variations saisonnières, du fait de conditions d'émission et de dispersion différentes, on pourra utiliser la variable auxiliaire « saison » comme variable de stratification. L'année étudiée sera alors découpée en strates saisonnières – les quatre trimestres par exemple – et chacune de ces strates fera l'objet d'un échantillonnage.

Par la suite, on désignera par **stratification temporelle** un découpage de l'année selon des variables de temps. **Les strates temporelles peuvent être continues dans le temps** (exemple des quatre trimestres cité ci-dessus) **ou discontinues** (*ex. 1 : on regroupe dans une même strate des mois non consécutifs tels que des mois de printemps et d'automne ; ex. 2 : on regroupe dans une même strate les jours ouvrés d'hiver, et dans une autre strate, les week-ends d'hiver.*) Elles peuvent être **de durées égales ou inégales**.

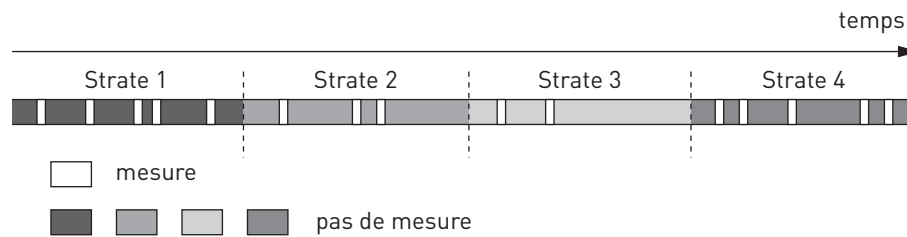


Figure 3 - Illustration d'un plan d'échantillonnage aléatoire stratifié, où les strates sont des périodes de temps continues et de même longueur.

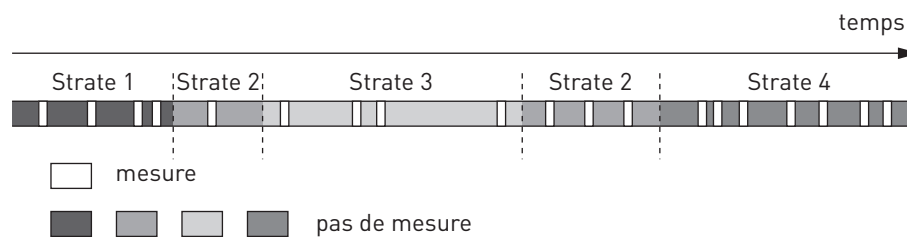


Figure 4 - Illustration d'un plan d'échantillonnage aléatoire stratifié, où les strates sont des périodes de temps continues ou discontinues et de longueurs différentes.

Quand il peut être mis en œuvre, l'échantillonnage stratifié permet :

- de réduire la variance de l'estimateur.
Cela signifie qu'à nombre de mesures égal, l'échantillonnage aléatoire stratifié permet d'estimer la moyenne annuelle plus précisément qu'un échantillonnage non stratifié.
- de réduire le nombre minimal de mesures requis afin d'estimer une donnée selon une précision recherchée.
En effet, soit une variable de concentration. Si à l'intérieur d'une même strate, les valeurs de concentration sont similaires, il suffit d'en tirer un petit nombre pour estimer précisément la concentration moyenne de la strate. Ainsi, en subdivisant judicieusement l'année, un effectif limité de mesures par strate est suffisant pour obtenir une estimation précise de la concentration moyenne annuelle. Sans stratification, il faudrait, pour atteindre cette même précision, que le nombre total de mesures dans l'année soit plus élevé.

► 1.2.2.3 Plan d'échantillonnage par grappes

S'il est plus pratique ou moins coûteux de sélectionner des groupes d'individus plutôt que des individus disséminés dans la population, on utilise un **plan par grappes** : la population est découpée en groupes d'individus, les **grappes**, dont un certain nombre est tiré dans chaque strate. On englobe dans l'échantillon tous les individus inclus dans les grappes sélectionnées.

La **taille d'une grappe** est le nombre d'individus qu'elle contient.

Un exemple type est celui de campagnes de mesures effectuées à l'aide d'un camion laboratoire : des périodes de mesure de plusieurs jours ou de plusieurs semaines, réalisées en nombre limité, induiront des temps de déplacement et des coûts moindres qu'un grand nombre de mesures horaires isolées. Dans ce cas, les grappes sont les périodes de mesure. Chacune d'elles est entièrement définie par son début et par sa taille. En pratique, on choisit souvent des grappes de même taille mais il est possible de considérer des grappes de tailles différentes.

Conséquence : En qualité de l'air, du fait de la corrélation temporelle des concentrations, les mesures individuelles à l'intérieur d'une grappe de données consécutives ne sont pas indépendantes.

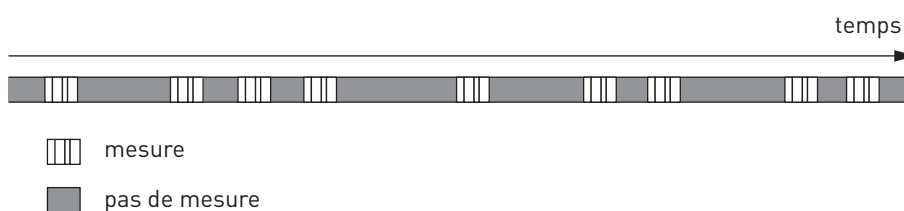


Figure 5 - Illustration d'un plan d'échantillonnage par grappes.

► 1.2.2.4 Plan d'échantillonnage par grappes stratifié

Un plan d'échantillonnage par grappes stratifié est un plan qui combine stratification temporelle et tirage de grappes.

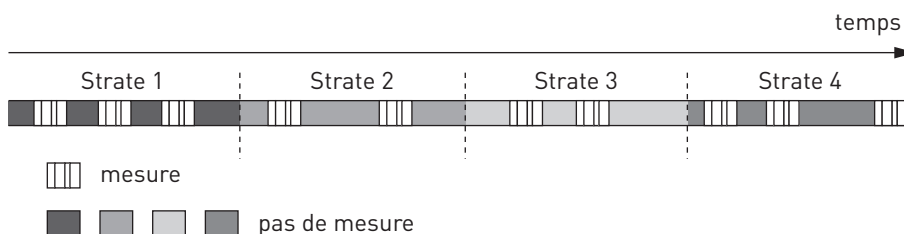


Figure 6 - Illustration d'un plan d'échantillonnage par grappes stratifié.

► 1.2.2.5 Remarque sur l'échantillonnage systématique

Dans un échantillonnage systématique (au sein d'une population entière ou d'une strate), seule la position du premier individu à tirer est sélectionnée aléatoirement parmi plusieurs débuts possibles. Les positions de tous les autres individus sont parfaitement déterminées par l'intervalle d'échantillonnage. Ainsi, d'un point de vue théorique, un tirage systématique peut être considéré comme un tirage d'une grappe unique.

Ex. : soit la réalisation d'une mesure individuelle tous les six jours durant une année. Cela équivaut à tirer une grappe de 60 ou 61 jours, parmi six grappes possibles. Dans cet exemple, les individus qui composent la grappe sélectionnée sont des mesures journalières espacées de cinq jours.



Figure 7 – Représentation d'un échantillonnage systématique. Tirage d'une mesure tous les six jours = sélection d'une grappe parmi six.

Cette représentation de l'échantillonnage systématique a son importance pour l'évaluation *a priori* de la qualité de la reconstitution. Dans la suite cependant, sauf mention du contraire, le terme de grappe désignera un ensemble de mesures consécutives.

Contexte statistique :	Exemple 1 : mesures automatiques	Exemple 2 : prélèvements journaliers	Exemple 3 : prélèvements hebdomadaires ou bihebdomadaires
Un échantillonnage est un tirage d'un certain nombre d'individus dans une population. À l'aide de cet échantillon d'individus, on cherche à estimer une donnée relative à l'ensemble de la population.	individus : mesures horaires population : l'ensemble des mesures sur une année donnée à reconstituer : • moyenne annuelle • ou pourcentage annuel de dépassement de seuil (pourcentage de 8 760 ou de 8 784 heures s'il s'agit d'un seuil horaire ; de 365 ou 366 jours s'il s'agit d'un seuil journalier). Ce pourcentage peut être ensuite traduit sous la forme d'un nombre de dépassements.	individus : prélèvements sur 24 heures population : l'ensemble des mesures sur une année donnée à reconstituer : moyenne annuelle	individus : prélèvements sur 1 ou 2 semaines population : l'ensemble des mesures sur une année donnée à reconstituer : moyenne annuelle
Échantillonnage systématique	Périodes de mesure effectuées à intervalles réguliers, par exemple fonctionnement cyclique de stations automatiques à raison de 6 semaines de mesure espacées de 6 semaines.	Prélèvements effectués à intervalles réguliers, par exemple un prélèvement journalier tous les 6 jours.	Prélèvements effectués à intervalles réguliers, par exemple des prélèvements bihebdomadaires espacés de 4 semaines.
Échantillonnage aléatoire	Périodes de mesure sélectionnées au hasard dans l'année ou dans les strates définies ci-dessous.	Prélèvements ou groupes de prélèvements consécutifs sélectionnés au hasard dans l'année ou dans les strates définies ci-dessous.	Prélèvements ou groupes de prélèvements consécutifs sélectionnés au hasard dans l'année ou dans les strates définies ci-dessous.
Répartition de la population en strates temporelles	strates : des parties de l'année , par exemple les saisons, délimitées selon les variations temporelles du polluant (observées sur des années antérieures et des sites similaires).	strates : des parties de l'année , par exemple les saisons, délimitées selon les variations temporelles du polluant ou d'espèces qui lui sont potentiellement corrélées.	strates : des parties de l'année , par exemple les saisons, délimitées selon les variations temporelles du polluant ou d'espèces qui lui sont potentiellement corrélées.
Détermination des grappes	grappes : mesures consécutives sur plusieurs jours ou plusieurs semaines.	absence de grappes : chaque période de mesure se réduit à un prélèvement individuel journalier grappes : prélèvements consécutifs sur plusieurs jours	absence de grappes : chaque période de mesure se réduit à un prélèvement individuel hebdomadaire ou bihebdomadaire grappes : prélèvements consécutifs sur plusieurs semaines

Tableau 2 - Illustration des termes statistiques pour des applications dans le domaine de la surveillance de la qualité de l'air.

1.2.3 Comment caractérise-t-on un plan d'échantillonnage ?

Le plan d'échantillonnage est établi :

- pour une période donnée,
- à l'aide d'une ou plusieurs variables auxiliaires (séries de données de concentration d'années antérieures ; si besoin, autres variables).

Il est caractérisé

- par des **paramètres de dimensionnement**,
- un mode de tirage,
- et une évaluation *a priori* de la qualité de la donnée à reconstituer.

► 1.2.3.1 Dimensionnement

- la proportion de la période considérée à couvrir par des mesures (couverture temporelle) ;
- le nombre de strates temporelles
(= 1 : pas de stratification ; > 1 : stratification) ;
- dans le cas de mesures individuelles : le nombre de mesures dans chaque strate ;
- dans le cas de mesures par grappes :
 - la taille des grappes : nombre de mesures horaires ou journalières composant les grappes ;
 - le nombre total de grappes sur la période considérée ;
 - le nombre de grappes dans chaque strate.

► 1.2.3.2 Mode de tirage

- tirage systématique ou aléatoire.

► 1.2.3.3 Évaluation *a priori* de la qualité de la donnée reconstituée

Au chapitre 2, nous verrons comment, une fois que l'échantillonnage est réalisé, les différentes méthodes de reconstitution permettent d'estimer une moyenne et son incertitude. Toutefois, avant même la mise en œuvre du plan d'échantillonnage, il est possible d'évaluer la précision qu'on peut attendre de ce plan, si la moyenne annuelle est estimée par une simple moyenne pondérée des données expérimentales.

Les formules de calcul appropriées sont fournies par la théorie statistique appelée « théorie des sondages », dans le cas d'un échantillonnage aléatoire stratifié ou dans le cas particulier d'un échantillonnage systématique. Elles permettent d'associer aux paramètres de dimensionnement une **variance théorique d'estimation** (V). Cette variance V est calculée sur une série complète de données, antérieure à l'année d'échantillonnage. Elle dépend des choix

d'échantillonnage (stratification choisie, grappes, mode de tirage), sans tenir compte d'un possible usage de variables auxiliaires au moment de la reconstitution (variables météorologiques, mesures en continu de concentration – du polluant étudié ou d'un ou plusieurs autres polluants – à une station de référence, etc.), mais ne nécessite aucun tirage de données (la formule de V est donnée page 29).

Le paramètre de qualité du plan d'échantillonnage Δ_Q est une *évaluation a priori de la précision d'estimation*. Pour une précision exprimée relativement à la moyenne $\bar{x}$, Δ_Q se déduit de V par la relation :

$$\Delta_Q = \frac{1}{2} \cdot L \cdot \frac{100}{\bar{x}} = t_{v,1-\alpha} \cdot \sqrt{V} \cdot \frac{100}{\bar{x}}$$

L : longueur de l'intervalle de confiance

$t_{v,1-\alpha}$: coefficient de Student d'ordre $1-\alpha$ à v degrés de liberté

Si la taille de l'échantillon est grande (supérieure à 30), on peut remplacer le coefficient de Student par le quantile d'ordre $1-(\alpha/2)$ de la loi normale. Pour un taux de confiance de 95 %, ce quantile vaut 1,96. Si ce paramètre constitue un critère d'appréciation, sinon de choix, dans la planification de l'échantillonnage, il ne permet pas de préjuger de la précision de la reconstitution finale.

Exemples :

Un plan d'échantillonnage répondant aux indications des directives européennes peut se caractériser par :

- *une couverture temporelle de **14 % de l'année civile**,*
- *un découpage de l'année en **4 strates de 3 mois**,*
- *des individus qui sont des mesures horaires,*
- *des **grappes d'une semaine**,*
- ***8 grappes** à répartir dans l'année civile,*
- *des grappes réparties dans l'année à raison de **deux par trimestre**,*
- *un tirage **aléatoire** des dates de début des grappes dans chaque trimestre,*
- *une **variance théorique V**, où V est évaluée sur des séries antérieures de concentration ;*

Ou encore par :

- *une couverture temporelle de **14 % de l'année civile**,*
- *un découpage de l'année en **4 strates de 3 mois**,*
- *des individus qui sont des prélèvements intégrés sur 7 jours,*
- ***8 prélèvements individuels** à répartir dans l'année civile,*
- *des individus répartis dans l'année à raison de **deux par trimestre**,*
- *un tirage **aléatoire** des dates de début des prélèvements dans chaque trimestre,*
- *une **variance théorique V**, où V est évaluée sur des séries antérieures de concentration.*

► 1.3 Étapes de l'élaboration d'un plan d'échantillonnage

1.3.1 Introduction

La définition de l'échantillonnage repose **sur l'analyse de séries annuelles de données antérieures à l'année d'échantillonnage, et qui proviennent d'un ou plusieurs sites jugés similaires au site d'étude**. Si l'historique de mesures le permet, des séries moyennes sur plusieurs années pourront être considérées, afin de lisser les variations de concentration imputables à des événements exceptionnels et à la variabilité météorologique.

La démarche proposée pour élaborer un plan d'échantillonnage se compose des étapes suivantes :

- étude des contraintes ;
- analyse de la variabilité temporelle des concentrations et détermination de la stratification temporelle ;
- dimensionnement de l'échantillonnage selon la précision souhaitée ; choix d'une taille de grappe ;
- détermination des dates de mesure ;
- contrôle de la qualité du plan et de sa faisabilité.

1.3.2 Étude des contraintes

Dans les applications de la surveillance réglementaire, le plan sera généralement défini en ayant pour contrainte initiale la qualité de l'indicateur final. **Il convient donc, en théorie, d'établir un plan d'échantillonnage pour chaque polluant et chaque indicateur à reconstituer.**

Il se peut néanmoins que le plan d'échantillonnage soit fixé ou contraint par la disponibilité des ressources (équipe, matériel, site...). Dans ce cas, il s'agira de contrôler l'impact du plan adopté sur la qualité finale des estimateurs.

Contrainte	Contraintes de qualité (% d'incertitude sur l'estimation) Maîtrise de la précision de l'estimateur	Contraintes de ressources (humaines et matérielles par exemple) Moins de souplesse sur la durée, le nombre et la répartition des grappes de mesures
Marge de manœuvre	Définir la durée, le nombre et la répartition des grappes de mesures	Accepter que l'estimateur soit moins précis
Calcul	Détermination du plan d'échantillonnage	Calcul de l'impact du plan d'échantillonnage sur la précision de l'estimateur
À contrôler	Stratégie acceptable du point de vue des ressources ?	Acceptable du point de vue de la qualité ?

Tableau 3 - Comparaison des démarches à suivre pour élaborer un plan d'échantillonnage en cas de contraintes de qualité ou de ressources.

Les paragraphes suivants décrivent l'élaboration d'un plan d'échantillonnage **lorsque la qualité de l'estimation finale représente la principale contrainte**. Cependant, en tant que facteur limitant, la question des ressources ne peut être ignorée. **Ainsi, certaines étapes constituent plutôt des procédés itératifs dont tout nouveau résultat doit être mis en rapport avec les ressources disponibles.**

1.3.3 Analyse de la variabilité temporelle des concentrations

Afin d'étudier la variabilité temporelle des concentrations sur l'année, nous proposons l'utilisation des représentations graphiques décrites ci-après.

a) Étude des profils annuels

Tracé des boîtes à moustaches :

Il s'agit de représenter les intervalles de variation des concentrations de polluants, selon les trimestres, pour un premier aperçu, et selon les mois de l'année, pour une analyse plus détaillée. Les statistiques figurées par une boîte à moustaches sont la médiane pour le point central et les quantiles 25 et 75 pour la dispersion. Cette représentation peut faire également ressortir les valeurs atypiques.

Tracé des moyennes :

Il s'agit de représenter les moyennes des concentrations des polluants selon une périodicité mensuelle. La moyenne est indiquée par un point, et son erreur

type par deux barres. L'erreur type est fonction de l'écart-type σ_1 de l'échantillon de données et de la taille n_1 de cet échantillon :

$$\text{erreur type (moyenne)} = \frac{\sigma_1}{\sqrt{n_1}}$$

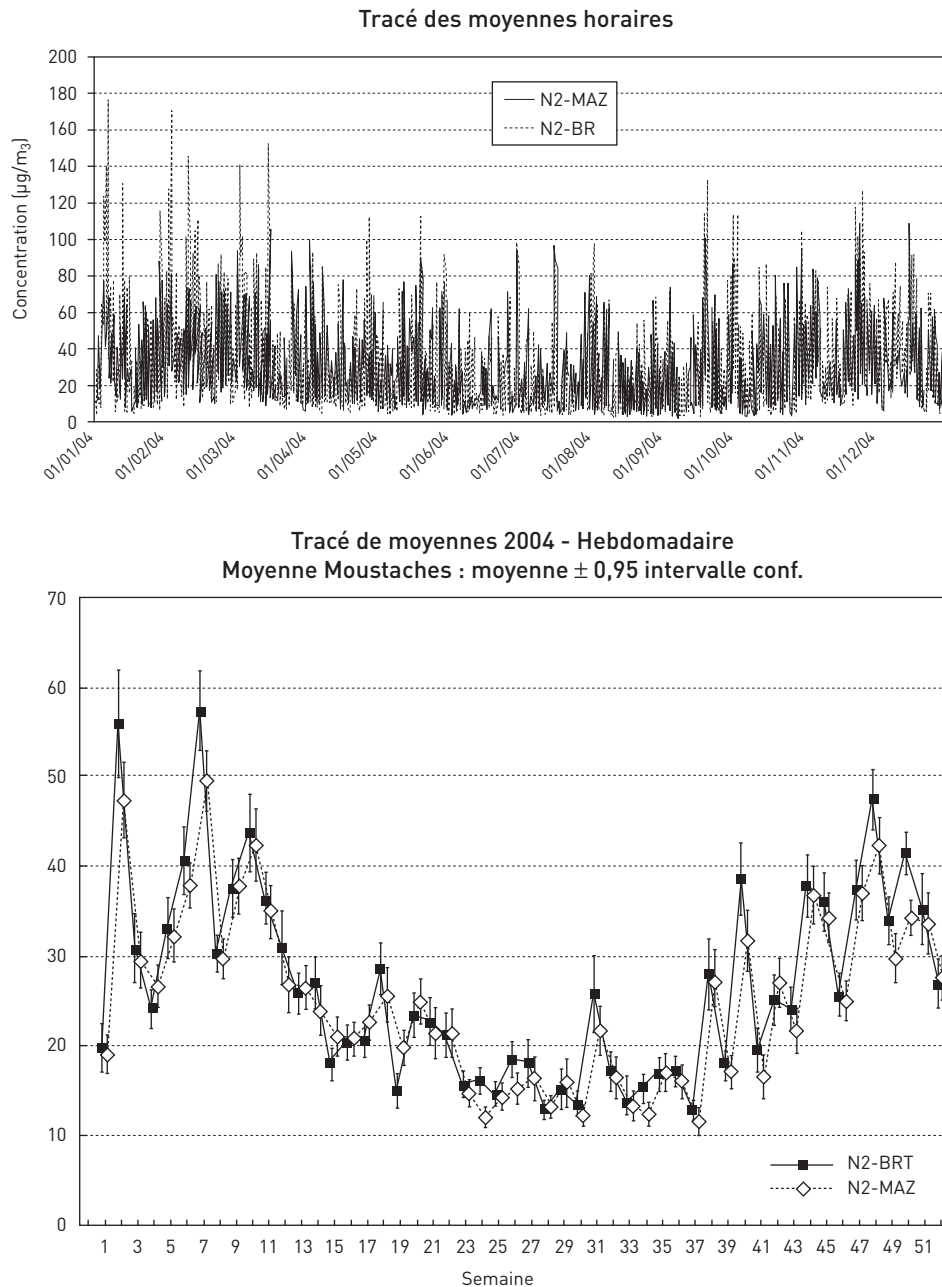


Figure 8 – Séries des moyennes horaires (graphique du haut) et hebdomadaires (graphique du bas). Les valeurs en ordonnée sont des concentrations en $\mu\text{g}/\text{m}^3$.

Exemple

Considérons le cas fictif suivant : en 2005, il est prévu d'effectuer une campagne d'échantillonnage du dioxyde d'azote en un site urbain de fond de Toulouse (le site Mazades - MAZ). Afin de définir le plan échantillonnage, la station urbaine de fond Berthelot (BRT), pour laquelle une série annuelle complète de données est disponible (année 2004), est choisie comme station de référence. La Figure 8 présente les séries des données horaires et, pour plus de lisibilité, des moyennes hebdomadaires. Sur la Figure 9 sont représentées les boîtes à moustaches trimestrielles et mensuelles et les moyennes correspondantes. (À titre d'information, on a fait également figurer sur ces graphiques les concentrations, supposées inconnues, de la station Mazades.)

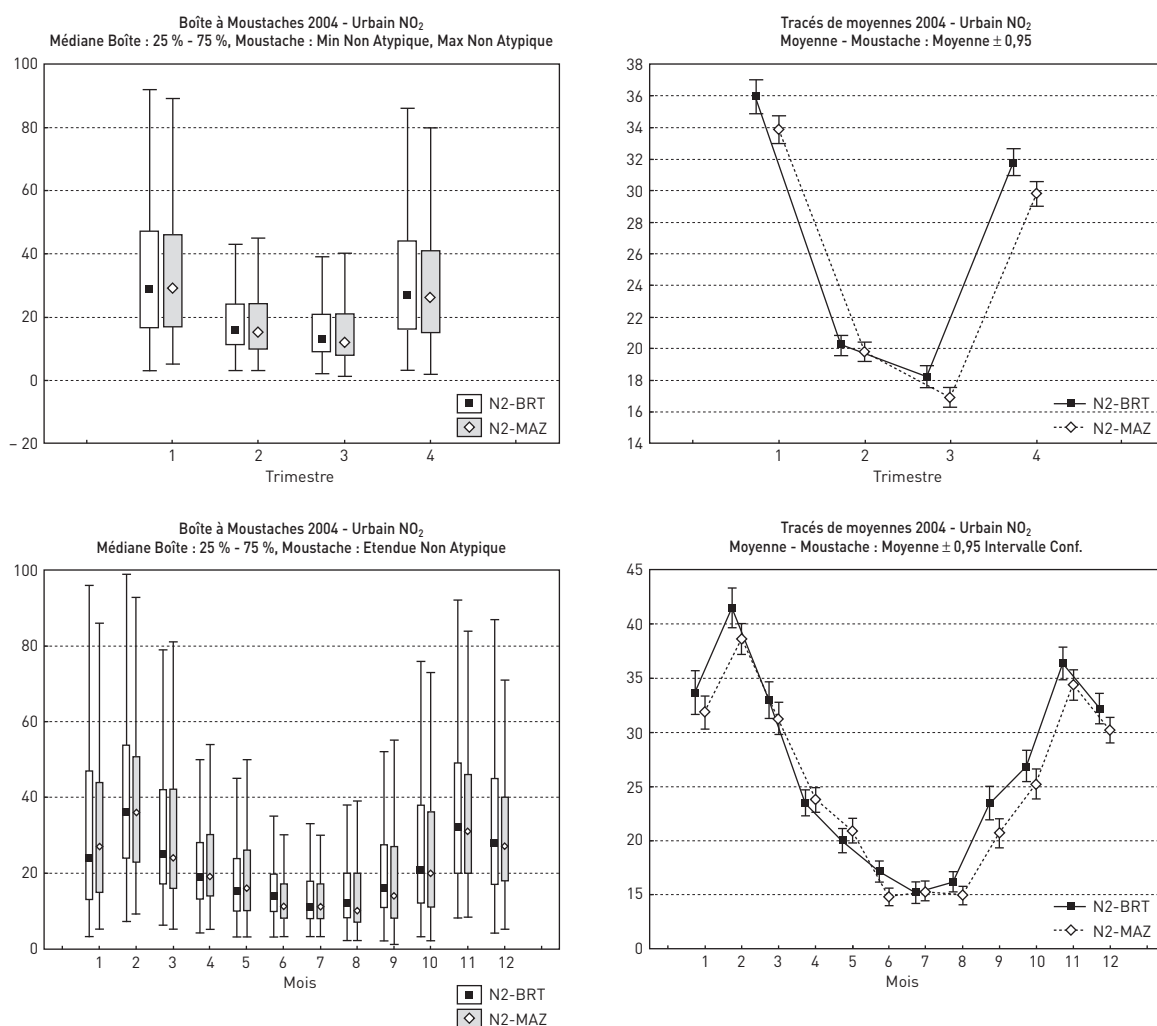


Figure 9 – Tracé des boîtes à moustaches (graphiques de gauche) et des moyennes (graphique de droite) trimestrielles et mensuelles. Les valeurs en ordonnée sont des concentrations en $\mu\text{g}/\text{m}^3$.

Les profils par trimestre montrent un net contraste entre les saisons hivernale et estivale. Les concentrations hivernales, environ deux fois supérieures en moyenne aux concentrations estivales, varient dans des plages de valeurs plus étendues. Au début et à la fin de la saison estivale (avril-mai et septembre), les profils par mois font apparaître une transition entre l'hiver et la période creuse de l'été.

b) Étude des profils hebdomadaires

En complément des graphiques précédents, les profils hebdomadaires (boîtes à moustaches et moyennes) permettent d'apprécier la variabilité des concentrations par jour de semaine et de détecter d'éventuels cycles hebdomadaires.

En hiver, les concentrations mesurées les jours ouvrés sont supérieures d'environ 40 % aux concentrations mesurées pendant les week-ends et varient dans de plus larges intervalles de valeurs. En été, le profil hebdomadaire est moins marqué ; seul le dimanche se distingue des autres jours de semaine.

1.3.4 Stratification temporelle

La stratification temporelle doit permettre à l'échantillonnage de saisir correctement la variabilité temporelle des concentrations avec un nombre limité de mesures. Le choix des strates découle donc directement de l'analyse qui précède.

a) Stratification selon le profil annuel

Dans une première étape, on s'appuiera sur les profils annuels pour diviser l'année en grandes périodes relativement homogènes du point de vue des niveaux et de la variabilité des concentrations mesurées.

Un découpage en deux grandes strates est possible si le profil annuel se caractérise par un simple contraste entre la saison estivale et la saison hivernale. Afin de répartir les mesures sur l'année, comme le demandent les directives, un découpage en 4 strates pourra être éventuellement préféré. Un découpage plus fin que 6 strates n'apporte pas nécessairement de gain sensible de précision (la variance à l'intérieur des strates diminue plus lentement au fur et à mesure que le nombre de strates augmente).

Lorsque les concentrations ne présentent aucun profil annuel (exemples de certaines séries de métaux), l'année peut être arbitrairement stratifiée selon les quatre trimestres.

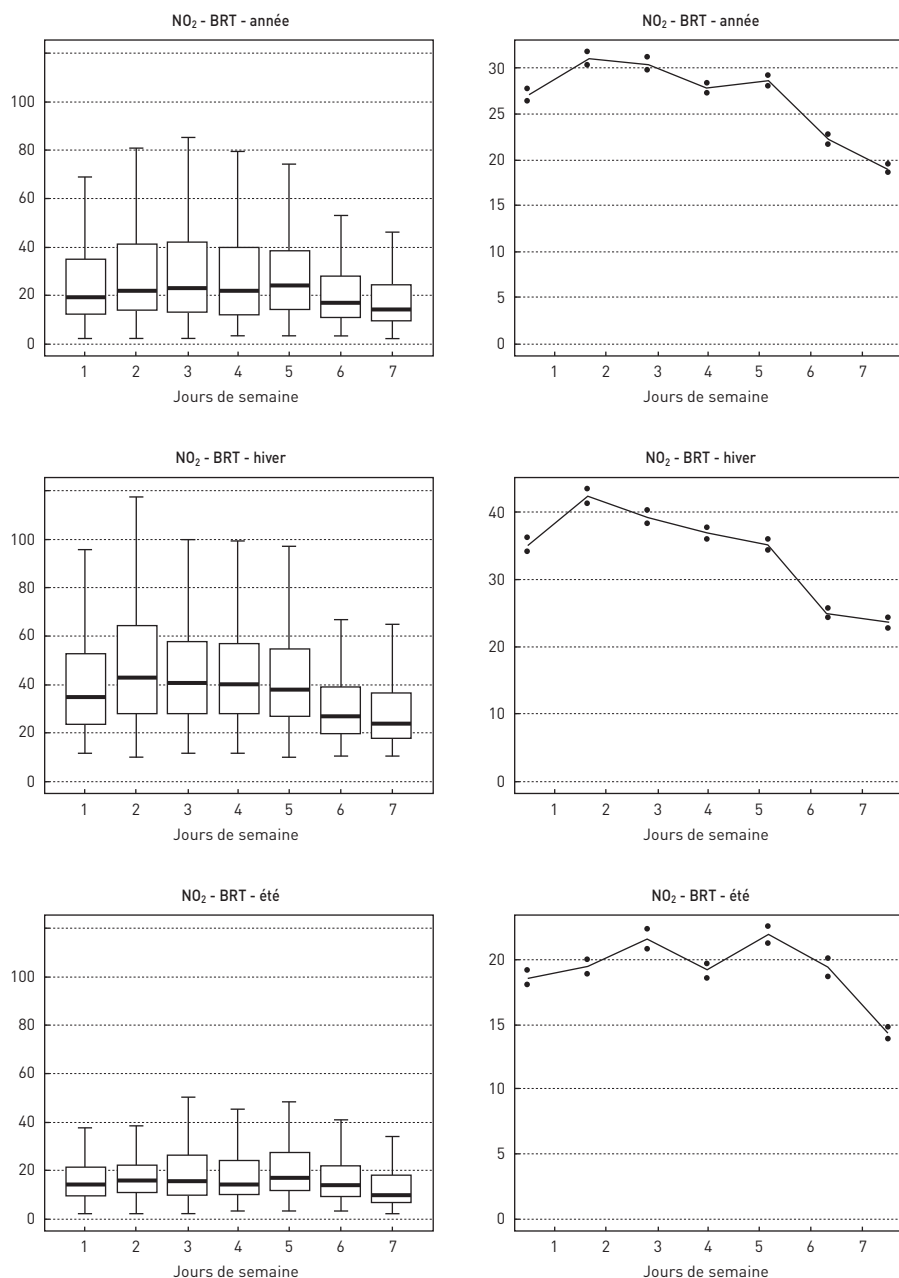


Figure 10 - Tracé des boîtes à moustaches (graphiques de gauche, hors valeurs atypiques) et des moyennes (graphique de droite) par jour de semaine pour l'année complète, la saison hivernale et la saison estivale. Les numéros en abscisse sont les jours de la semaine (6 et 7 = samedi et dimanche). Les valeurs en ordonnée sont des concentrations en µg/m³.

Reprenons l'exemple du site urbain de Toulouse (page 23)

La première stratification qui s'impose est un découpage hiver/été (saisons au sens large). En s'aidant du profil par mois, chacune de ces deux périodes est subdivisée en deux strates :

- pour la saison hivernale, on sépare le premier et le quatrième trimestre ;
- pour la saison estivale, on distingue les mois de transition et le creux de l'été.

La stratification suivante est ainsi proposée :

janvier-février-mars / avril-mai et septembre / juin-juillet-août /
octobre-novembre-décembre.

b) Stratification selon le profil hebdomadaire

Cette stratification peut éventuellement se combiner à la précédente si les jours de la semaine présentent de nettes différences de concentration.

Dans l'exemple précédent, on pourrait ainsi considérer les huit strates suivantes :
hiver1-jours ouvrés ; hiver1-week end ; transition-jours ouvrés ;
transition-week end ; été-jours ouvrés ; été-week end ;
hiver2-jours ouvrés ; hiver2-week end.

Cela suppose de dimensionner l'échantillonnage pour chacune de ces strates puis d'y sélectionner le nombre voulu de mesures. Ce type de stratification n'est donc applicable que pour des mesures individuelles journalières. Il n'est pas adapté à la réalisation de grappes de mesures sur une ou plusieurs semaines.

1.3.5 Dimensionnement de l'échantillonnage

La méthode de dimensionnement diffère selon le mode de tirage des données : elle est directe pour un échantillonnage aléatoire, itérative pour un échantillonnage systématique. En revanche, la démarche est identique, que l'on souhaite estimer une concentration moyenne annuelle ou un nombre annuel de dépassements de seuil. Dans ce dernier cas, les données de concentration sont remplacées par 1, si elles dépassent le seuil, et par 0 si elles lui sont inférieures.

► 1.3.5.1 Cas d'un échantillonnage aléatoire

Le dimensionnement de l'échantillonnage s'effectue à l'aide de la théorie statistique des sondages mentionnée en introduction. **Il exploite la stratification temporelle.**

1. On fixe une **précision d'estimation**. Cette précision correspond à la demi-largeur de l'intervalle de confiance associé à la valeur estimée.

Dans l'exemple de Toulouse, fixer une précision de $4 \mu\text{g}/\text{m}^3$, respectivement de 10 %, à un niveau de confiance de 95 % signifie que si $\hat{y}$ est la concentration moyenne annuelle de NO_2 reconstituée au site urbain Mazades et $\text{IC}_{95\%}$, l'intervalle de confiance à 95 % autour de $\hat{y}$, on souhaite avoir $\text{IC}_{95\%} = \hat{y} \pm 4$, respectivement $\text{IC}_{95\%} = \hat{y} \pm 0,1 \cdot \hat{y}$. Cette précision est reliée à la variance V de l'estimateur de la moyenne (estimation en l'absence de variables auxiliaires) par la relation : $4 = 1,96 \cdot \sqrt{V}$ ou $0,1 = 1,96 \cdot \sqrt{V / \hat{y}}$

2. On calcule, pour différentes tailles de grappe, le nombre minimal de grappes à prélever dans l'année afin d'atteindre cette précision.

Ce nombre dépend de la stratification temporelle adoptée et de la variabilité des concentrations moyennes par grappe à l'intérieur de chaque strate (Tillé, 2001, p.134) :

Soit un découpage de l'année en H strates ; soient h les indices des strates.

Le nombre minimal de grappes à tirer dans l'année pour obtenir une variance d'estimation V est :

$$m^* = \frac{\left(\sum_1^H M_h \cdot S_h \right)^2}{M^2 V + \sum_1^H M_h \cdot S_h^2}$$

$$S_h^2 = \frac{1}{M_h - 1} \sum_{j=1}^{M_h} (y_{h,j} - \bar{y}_h)^2 : \text{variance intra-strate corrigée.}$$

M : nombre total de grappes dans l'année

M_h : nombre total de grappes dans la strate h

$y_{h,j}$: moyennes par grappe dans la strate h

$\bar{y}_h$: moyenne de la strate h

$$S_h = \sqrt{S_h^2}$$

À une même précision correspondent plusieurs couples (**taille de grappe, nombre de grappes**). Il suffit de retenir, parmi ces propositions, celle qui s'accorde le mieux avec la disponibilité des ressources. S'il est nécessaire, on augmentera le nombre de grappes de façon que la couverture temporelle minimale exigée par la réglementation soit respectée.

Soit l'exemple du point 1 et la stratification temporelle précédemment définie. Les nombres de grappes requis pour une précision de 10 % et pour différentes tailles de grappes sont (nombres arrondis à l'unité supérieure) :

Taille de grappe	7 jours	14 jours	21 jours
Nombre de grappes requis dans l'année	16	7	4

Remarque : Supposons qu'au lieu d'estimer la moyenne annuelle, l'objectif ait été d'estimer le **nombre annuel de dépassements d'une ou plusieurs valeurs seuils**. Les données de NO₂ de la station Berthelot sont remplacées par 1 si la concentration est supérieure au seuil considéré et par 0 dans le cas contraire. Le nombre de grappes requis pour une précision de 10 % devient :

Taille de grappe	7 jours	14 jours	21 jours
Nombre requis Seuil horaire = 50 µg/m ³ (dépassé 13 % du temps à la station Berthelot, en 2004)	33	14	8
Nombre requis Seuil horaire = 70 µg/m ³ (dépassé 5,5 % du temps à la station Berthelot, en 2004)	35	14	10

3. À chaque couple (taille de grappes, nombre de grappes) est associée une répartition optimale des grappes entre les strates temporelles (Tillé, 2001, p. 133) :

$$m_h = \frac{m \cdot M_h \cdot S_h}{\sum_{l=1}^H M_l \cdot S_l}$$

m_h : nombre de grappes à tirer dans la strate h

Exemple :

L'année a été divisée en 4 strates temporelles : hiver 1^{er} trimestre, transition, été, hiver 4^e trimestre.

Après calcul d'optimisation, la répartition des grappes entre ces quatre strates est :

- si l'on retient la proposition 16 x 7 jours : 5 ; 3 ; 2 ; 6
- si l'on retient la proposition 7 x 14 jours : 3 ; 1 ; 1 ; 2.

Remarque :

Comme c'était le cas pour l'effectif total de grappes, les nombres théoriques de grappes par strate temporelle sont rarement des entiers et doivent être arrondis. À cause de cette approximation, la répartition des grappes entre les strates n'est plus tout à fait optimale et par suite, la précision d'estimation associée au plan d'échantillonnage n'est plus exactement égale à la valeur attendue. Il convient donc une nouvelle fois de l'évaluer (cf. étape suivante).

4. On contrôle la précision d'estimation associée au plan finalement choisi en calculant la variance théorique de l'estimateur de la moyenne (Tillé, 2001, p. 131) :

$$Var(\hat{\bar{y}}) = \frac{1}{M^2} \sum_{h=1}^H M_h \cdot \frac{M_h - m_h}{m_h} \cdot S_h^2$$

$$précision [\%] = 100 \cdot \frac{1,96 \cdot Var(\hat{\bar{y}})}{\bar{y}}$$

$Var(\hat{\bar{y}})$: variance de l'estimateur de la moyenne

$\bar{y}$: moyenne de la série de référence considérée

La précision attendue pour le plan [5 ; 3 ; 2 ; 6] x 7 jours est de 9,8 %.

La précision attendue pour le plan [3 ; 1 ; 1 ; 2] x 14 jours est de 9,9 %.

Ce résultat est une **précision estimée a priori**. C'est la précision que l'on obtiendrait par calcul à l'issue de l'échantillonnage si les concentrations de 2005 au site Mazades étaient identiques à celles de 2004 au site Berthelot et que les données d'échantillonnage permettent d'estimer parfaitement la variance dans chaque strate.

► 1.3.5.2 Cas d'un échantillonnage systématique

Si pour des raisons pratiques, un échantillonnage systématique, c'est-à-dire répété à intervalle régulier, est préféré à un échantillonnage aléatoire, **les formules de dimensionnement précédentes ne s'appliquent plus**.

D'un point de vue théorique en effet, comme il est expliqué en page 15, un échantillonnage systématique équivaut au tirage d'une grappe unique dont la taille est donnée par le nombre de mesures (Tillé, 2001, p. 170). Soit donc une précision d'estimation fixée. Dimensionner l'échantillonnage revient à rechercher la taille optimale des grappes à prélever dans chaque strate, ce qu'on ne peut obtenir par les formules précédentes.

Une approche itérative est proposée :

1. Un dimensionnement est défini pour chaque strate.
2. On calcule la variance théorique associée et la précision d'estimation correspondante :

$$Var(\hat{y}) = \frac{1}{N^2} \sum_{h=1}^H N_h^2 \cdot Var(\hat{y}_h)$$

$$\text{avec } Var(\hat{y}_h) = K_h \cdot \sum_{j=1}^{K_h} \left(\frac{\bar{y}_{h,j} \cdot n_{h,j}}{N_h} - \frac{\bar{y}_h}{K_h} \right)^2$$

K_h : nombre d'échantillons systématiques possibles dans la strate h

$n_{h,j}$: nombre de mesures* dans l'échantillon j

N_h : nombre de mesures* dans la strate h

N : nombre de mesures* dans l'année

$\bar{y}_{h,j}$: moyenne de l'échantillon j appartenant à la strate h

$\bar{y}_h$: moyenne de la strate h

mesure = une mesure individuelle ou une grappe de mesures consécutives représentée par sa moyenne.

3. Si cette précision ne convient pas, le dimensionnement est réajusté.
4. On s'assure que la fréquence d'échantillonnage déterminée par le dimensionnement ne coïncide pas avec un cycle saisonnier ou hebdomadaire (cf. analyse de la variabilité temporelle).

1.3.6 Contrôle de la faisabilité du plan

Les ressources nécessaires pour la mise en œuvre du plan établi sont évaluées (cf. page 32). Si elles sont compatibles avec les ressources disponibles, le plan d'échantillonnage est accepté. Si elles ne le sont pas :

- soit la précision d'estimation est revue à la baisse et l'on répète l'étape de dimensionnement ;
- soit le dimensionnement est directement réajusté en fonction des ressources et l'on contrôle la précision associée au nouveau plan.

Exemple 1 :

Les ressources autorisent le tirage de 16 grappes d'une semaine mais non pas la répartition [5 ; 3 ; 2 ; 6]. On décide de modifier cette répartition qui devient : [5 ; 3 ; 3 ; 5]. La précision attendue, qui était de 9,8 %, passe à 10,0 %.

Exemple 2 :

Les ressources fixent le plan d'échantillonnage à 12 grappes d'une semaine, selon la répartition [3 ; 3 ; 3 ; 3]. La précision attendue vaut à présent 13,4 %.

Cette étape d'évaluation est illustrée par des exemples (cf. page 32).

1.3.7 Détermination des dates de mesure

Une fois que le plan d'échantillonnage est compatible avec les ressources, il reste à déterminer les dates de mesure.

Au sein de chaque strate temporelle, on tire **aléatoirement** :

- cas d'un échantillonnage aléatoire : les dates de début des grappes ;
- cas d'un échantillonnage systématique : une date de début d'échantillonnage (compatible avec le nombre total de mesures à répartir sur l'année.)

On s'assure que les dates ainsi sélectionnées s'accordent avec la disponibilité des moyens humains ou matériels. S'il est nécessaire, un nouveau tirage de dates est effectué.

► 1.4 Comment établir un plan d'échantillonnage en l'absence de données de référence ?

Il existe toujours des situations dans lesquelles la méthodologie décrite précédemment ne peut être appliquée pour construire un plan d'échantillonnage, faute de données pouvant raisonnablement servir de référence. Dans ces cas-là, il faut garder à l'esprit le principe suivant :

Pour qu'un échantillon soit représentatif d'une période donnée, plus les concentrations pendant cette période sont variables, plus il faut de mesures.

Le premier objectif sera donc d'estimer, même de façon grossière, les périodes de grande variabilité – où l'on prélèvera un large échantillon de mesures – et les périodes de faible variabilité – où l'on se permettra de tirer un plus petit échantillon –.

Pour cela, toute information doit être considérée : d'autres types de données réputées suffisamment corrélées, d'autres paramètres (météorologiques ou autre), ou encore d'autres études faites par ailleurs (bibliographie).

Dans tous les cas, il est vivement conseillé, lorsque c'est possible, de réaliser de nombreuses mesures la première année – quitte à en faire plus qu'il ne paraît nécessaire –, afin de pouvoir optimiser par la suite le plan d'échantillonnage.

► 1.5 Évaluation des ressources

Si elle est généralement guidée par un objectif de qualité, la planification de l'échantillonnage doit s'accorder avec un autre aspect de la surveillance : la gestion des ressources, et notamment, des unités d'œuvre.

Quelles sont les ressources mobilisées par un plan d'échantillonnage ? Trois exemples de réponses sont ici proposés. Ils correspondent à des usages différents de la mesure discontinue :

- campagnes de mesure à l'aide de moyens mobiles (ex : camions laboratoires, cabines mobiles) ou de tubes à échantillonnage passif ;
- campagnes de prélèvement des polluants de la quatrième directive fille (HAP, métaux lourds) ;
- fonctionnement alterné des stations du réseau fixe.

Une méthodologie plus générale d'évaluation n'est pas développée à ce jour, étant donné la diversité des situations et des pratiques.

1.5.1 Planification des unités d'œuvre nécessaires pour une campagne à l'aide de moyens mobiles

La planification des unités d'œuvre nécessite de calculer :

1. le nombre de jours d'installation et de désinstallation, en incluant, pour les moyens mobiles, les travaux d'étalonnage des analyseurs ;
2. le nombre de jours de maintenance préventive ou de remplacement des tubes (ex. une maintenance d'un analyseur automatique se fait tous les 15 jours, la durée d'exposition de tubes peut varier de 24 heures à 15 jours) ;
3. les temps de déplacement vers le ou les sites de mesure et la durée des trajets de retour à l'AASQA.

Exemple :

Soit un plan d'échantillonnage constitué de 9 grappes de 2 semaines.

On suppose que pour chaque grappe de mesure, les temps (exprimés en jours travaillés) nécessaires aux travaux et aux déplacements s'élèvent à :

- déplacement (aller-retour) : 0,5 JT,
- installation : 0,5 JT,
- désinstallation : 0,5 JT,
- maintenance préventive : fonction du plan d'échantillonnage.

Généralement, les visites sur site pour maintenance préventive sont effectuées tous les 15 jours. Dans ce plan d'échantillonnage, des grappes isolées de deux semaines ont été choisies, c'est pourquoi de telles visites ne figurent pas dans l'évaluation.

La mise en œuvre du plan requiert donc les temps suivants (en nombre de JT) :

% d'utilisation de la cabine sur la période (ici l'année)	34 %
Déplacement des techniciens	9 pour les 18 déplacements
Temps d'installation	4,5
Temps de désinstallation	4,5
Temps de maintenance préventive (tous les 15 jours)	0*
Total	18

Remarque :

Le calcul du nombre de jours travaillés a été réalisé pour un site unique, en supposant que le moyen mobile effectue un aller-retour entre l'AASQA et le point d'échantillonnage. Si plusieurs sites sont surveillés pendant la même période, l'optimisation du parcours du moyen mobile peut éventuellement abaisser ce nombre.

1.5.2 Évaluation des coûts liés à des campagnes de prélèvement. Exemple : coûts de l'évaluation préliminaire des métaux lourds

L'évaluation suivante, réalisée par l'ORAMIP, porte sur la surveillance des métaux lourds en des sites sous influence industrielle mais elle peut être transposée à d'autres types de sites. Elle compare les coûts entre une stratégie de 8 x 1 semaine et une stratégie de 4 x 2 semaines. Deux situations sont décrites : un seul site de suivi et deux sites éloignés. Pour ces deux exemples, l'approche de 4 x 2 semaines apparaît nettement plus rentable.

Avec 1 seul site :

Pour rejoindre un site éloigné	
Trajet A/R	
Distance de Colomiers	340 km
Durée Transport	4 h 45
Temps technicien nécessaire	1 journée x 2 personnes
Évaluation préliminaire : 8 x 1 semaine pendant 3 années consécutives (15 %)	
Nbre de déplacements sur site	16
Installation + retrait du Partisol	32 journées techn. = 10067,2 euros
Transport	2 176 euros (0,4 €/km)
Analyse	708 euros
Pour 3 années	38 854 € + 1 Partisol utilisé pour chaque série de 2 semaines
Évaluation préliminaire : 4 x 2 semaines pendant 3 années consécutives (15 %)	
Nbre de déplacements sur site	8
Installation + retrait du Partisol	16 journées techn. = 5033,6 euros
Transport	1088 euros (0,4 €/km)
Analyse	354 euros
Pour 3 années	19 427 € + 1 Partisol utilisé pour chaque série de 2 semaines

Avec 2 sites :

Pour rejoindre deux sites éloignés	
Trajet A/R	
Distance de Colomiers	370 km
Durée Transport	5 h 20
Temps technicien nécessaire	1 journée x 2 personnes
Évaluation préliminaire : 8 x 1 semaine pendant 3 années consécutives (15 %)	
Nbre de déplacements sur site	16
Installation + retrait des 2 Partisols	32 journées techn. = 10067,2 euros
Transport	2368 euros (0,4 €/km)
Analyse	1416 euros
Pour 3 années	41 554 € + 2 Partisols utilisés pour chaque série de 2 semaines
Évaluation préliminaire : 4 x 2 semaines pendant 3 années consécutives (15 %)	
Nbre de déplacements sur site	8
Installation + retrait des 2 Partisols	16 journées techn. = 5033,6 euros
Transport	1 184 euros (0,4 €/km)
Analyse	708 euros
Pour 3 années	20 777 € + 2 Partisols utilisés pour chaque série de 2 semaines

1.5.3 Évaluation des coûts associés à un fonctionnement alterné des stations du réseau fixe

L'association Air Pays de la Loire a engagé une réflexion sur un fonctionnement cyclique des stations urbaines de son réseau de surveillance. À l'occasion de son Programme de Surveillance de la Qualité de l'Air (PSQA), elle a simulé différents cycles de mesure sur des séries de données existantes et comparé les variations de coûts que la mise en place de ces cycles engendrerait. Les principaux résultats de cette évaluation (jointe intégralement en annexe 9) sont fournis ci-après.

Description des simulations

Pour chaque agglomération considérée, la station de centre-ville conserve son caractère permanent tandis que les autres stations, groupées par paires, fonctionnent en alternance. Les simulations réalisées concernent quatre couples de stations : deux à Nantes, un à Angers et un au Mans, et quatre types de cycles sur un an :

- 8 x 1,5 mois : la première station du couple est instrumentée durant 1,5 mois, la deuxième est équipée durant la période suivante de 1,5 mois (4 cycles dans l'année)
- 4 x 3 mois : 2 cycles dans l'année
- 2 x 6 mois : 1 cycle dans l'année
- 1 x 12 mois : la première station est équipée l'année n , la deuxième l'année $n + 1$

Le dernier cycle constitue un cas particulier d'échantillonnage puisqu'aucune mesure de la station au repos n'est disponible pour estimer les indicateurs de l'année : cette estimation nécessite des informations de l'année passée ou des années précédentes, en complément des données de la station de référence permanente ; elle sort du contexte de surveillance considéré dans le présent guide. Ce cas particulier doit être vu comme une approche d'estimation objective (calcul de concentrations à partir de valeurs mesurées en d'autres lieux et/ou à d'autres périodes⁽¹⁾).

Impact des cycles sur stations urbaines

Le bilan économique de la mise en place des cycles de mesures sur les différents couples de sites urbains est présenté ci-dessous. La situation de base considérée dans le Tableau 4 est l'arrêt définitif d'une station par couple et le fonctionnement permanent de la station restante. La situation qui lui est comparée est le maintien de toutes les stations avec, pour chaque couple, un fonctionnement alterné des deux stations. Les coûts de déplacements

1. Guidance on Assessment under the EU Air Quality Directives – Final Draft.

(carburant et péage) tiennent compte de la localisation du Service Métrologie (près de Nantes) par rapport aux lieux d'intervention (Nantes, Angers à 91 km dont 81 km de voie à péage, Le Mans à 185 km dont 170 km de voie à péage).

	UO terrain (en jours)	Déplacement (en euros)
8 x 1,5 mois	+ 70	+ 700
4 x 3 mois	+ 33	+ 360
2 x 6 mois	+ 15	+ 200
1 x 12 mois	+ 6	+ 100

Tableau 4 – Variation entre 2004 et 2006 des unités d'œuvre (UO) et des coûts de déplacements liés à différents scénarios de mesure cyclique sur les sites urbains des Pays de la Loire.

En les considérant isolément, ces actions ont un impact économique négatif puisque, par rapport au maintien d'une seule station par couple, elles augmentent le temps d'intervention, en particulier pour les deux premiers scénarios. En revanche, si l'on établit le bilan global de **l'ensemble des actions programmées par le PSQA entre 2004 et 2006** et qui ont un impact positif ou négatif sur les temps d'intervention, les coûts d'exploitation et de maintenance préventive et les coûts d'analyse chimique, des gains sont possibles (Tableau 5). En particulier, la réduction du nombre d'analyseurs permanents, prévue dans le PSQA, compense assez largement les temps d'intervention des techniciens engendrés par la mise en place des cycles. Un fonctionnement cyclique des stations urbaines est donc envisageable; par rapport au scénario de base, cette approche a l'avantage de préserver la couverture spatiale des villes concernées.

	UO terrain (en jours)	Déplacement (en euros)	Analyses chimiques (en euros)	Exploitation (en euros)
8 x 1,5 mois	+ 3	+ 800	- 7 000	- 1 700
4 x 3 mois	- 34	+ 500	- 7 000	- 1 700
2 x 6 mois	- 52	+ 300	- 7 000	- 1 700
1 x 12 mois	- 61	+ 200	- 7 000	- 1 700

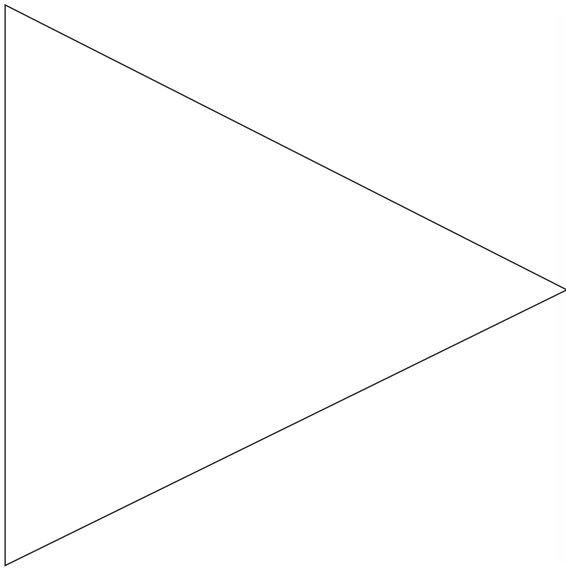
Tableau 5 – Variation entre 2004 et 2006 des unités d'œuvre (UO) et des coûts liés à l'ensemble du programme de surveillance dans les Pays de la Loire.

► 1.6 Conclusion

Nous avons vu dans ce chapitre comment, à partir d'informations existantes, il était possible de construire un plan d'échantillonnage qui permette de saisir au mieux la variabilité temporelle des concentrations. La méthodologie développée repose sur l'exploitation de données caractéristiques du polluant et du type de site étudiés et fait appel aux principes statistiques de la théorie des sondages. Elle comprend plusieurs étapes :

- analyse des contraintes de qualité et de ressources ;
- analyse de la variabilité temporelle des concentrations et découpage de l'année en grandes périodes, les strates temporelles ;
- détermination des caractéristiques de l'échantillonnage et du nombre de mesures à effectuer dans chaque strate temporelle ;
- évaluation des ressources nécessaires à la mise en œuvre du plan choisi et contrôle de la faisabilité de ce plan.

Dans le chapitre suivant, nous montrons comment les données d'échantillonnage recueillies peuvent être exploitées par différentes approches afin de reconstituer des valeurs moyennes annuelles.



2

Reconstitution des paramètres statistiques

Avertissement : ce chapitre ne concerne pas l'exploitation de données de mesure fixe ou indicative ; dans ce cas, la seule méthode autorisée pour estimer une moyenne est la moyenne arithmétique des données expérimentales.

► 2.1 Introduction

2.1.1 Objectif

Les plans d'échantillonnage décrits dans le chapitre précédent permettent d'obtenir des mesures à plusieurs époques de la période d'étude. Cette période est celle qui est prise en compte dans la définition des valeurs réglementaires. Il s'agit presque toujours de l'année, sauf dans le cas du dioxyde de soufre et de l'ozone pour lesquels la saison respectivement hivernale (valeur limite pour la protection des écosystèmes) et estivale (AOT40) peut être aussi distinguée. Les travaux du GT se sont jusqu'à présent attachés à des **périodes d'étude annuelles**, aussi est-ce toujours de l'année dont il sera question dans ce chapitre. Mais les méthodes présentées peuvent être appliquées à des périodes plus courtes.

À partir de l'échantillon de données recueilli, l'objectif est de reconstituer des indicateurs de la qualité de l'air sur toute la période. Dans les situations qui nous intéressent, ces indicateurs peuvent être :

- une valeur de **concentration moyenne annuelle** et son **incertitude** associée ;
- un **nombre annuel de dépassements de seuils** horaires ou journaliers et son **incertitude** associée.

Dans ce guide, la reconstitution de données individuelles (concentrations horaires, journalières) n'est pas considérée comme une fin en soi mais comme une possible étape intermédiaire pour estimer les indicateurs cités ci-dessus.

Remarque :

À la reconstitution d'un nombre de dépassements de seuil, on peut, dans certains cas, préférer l'estimation de centiles horaires ou journaliers (d'ordre 90, 95, 98...), qui sont comparés au(x) seuil(s) réglementaire(s) horaire(s) ou journalier(s) considéré(s).

Exemple :

On réalise une campagne de mesure pendant laquelle on s'attend à enregistrer plusieurs dépassements d'un certain seuil. Le caractère aléatoire de l'échan-

tillonnage fait qu'aucun dépassement ne se produit pendant la campagne ou qu'il n'en survient qu'un très petit nombre. Le nombre estimé de dépassements sur l'année est nul ou très faible, mais il se révèle aussi très imprécis : on ne peut rien conclure quant au respect de la réglementation. Supposons que le seuil soit une concentration journalière et qu'il soit permis de le dépasser 35 fois par an ($\approx 10\%$ de l'année). Une autre approche sera de reconstituer non pas le nombre de dépassements mais le centile 90 de la série des données journalières : on considérera que la réglementation est respectée si celui-ci, assorti de son incertitude, est significativement inférieur au seuil.

Ce guide se limite à des indications sur la façon d'estimer ou d'approcher des centiles à l'aide des méthodes de reconstitution étudiées. Une difficulté notable est d'assortir ces estimations d'une incertitude.

2.1.2 Mise en œuvre de la reconstitution

Trois méthodes de reconstitution sont décrites dans ce chapitre. La mise en œuvre de chacune est illustrée par un exemple, toujours le même, appelé exemple « fil rouge » :

Description de l'exemple :

Après planification de l'échantillonnage (voir l'exemple du chapitre 1), une campagne de mesure^[2] du NO_2 par moyen mobile a été réalisée en 2005 sur un site urbain de Toulouse (site Mazades). Les mesures se répartissent comme suit* :

- 3 fois 1 semaine pendant la période janvier-février-mars
- 3 fois 1 semaine pendant la période avril-mai-septembre
- 3 fois 1 semaine pendant la période juin-juillet-août
- 3 fois 1 semaine pendant la période octobre-novembre-décembre

L'objectif est de reconstituer la concentration moyenne annuelle de NO_2 en ce site.

Pour cette reconstitution, on dispose également des données d'une station fixe de référence, la station Berthelot, et de données météorologiques.

*NB : Il s'agit d'une campagne simulée, la station Mazades fonctionnant en réalité toute l'année.

2. Dans toute la suite, on désignera par campagne de mesure l'ensemble des mesures réalisées sur la période d'étude, que ces mesures soient ou non consécutives.

► 2.2 Les méthodes

2.2.1 Présentation des méthodes

Ce chapitre décrit trois méthodes de reconstitution de données. Deux d'entre elles relèvent de la théorie statistique des sondages. Il s'agit de :

- la **méthode des plans de sondage** qui estime directement une moyenne annuelle ou un nombre annuel de dépassements de seuil, en tenant compte de la stratification temporelle et des caractéristiques du plan d'échantillonnage ;
- la **méthode issue de la norme ISO 9359** qui estime directement une moyenne annuelle ou un nombre annuel de dépassements de seuil, en stratifiant les données recueillies en fonction de variables influentes, et non plus sur une base de répartition exclusivement temporelle ;

La troisième est la **régression linéaire** qui, à partir de variables auxiliaires, reconstitue d'abord des séries annuelles de données horaires, journalières ou hebdomadaires, avant d'en déduire les indicateurs étudiés.

Les observations et avertissements émis pour chacune sont fondés sur la simulation d'un grand nombre de plans d'échantillonnage à partir d'une base de données annuelle fournie par l'ORAMIP (travaux internes aux GT)^[3].

2.2.2 Méthode des plans de sondage

► 2.2.2.1 Principe

La méthode des plans de sondage s'applique dans le cadre d'un **échantillonnage aléatoire stratifié** tel qu'il a été décrit au chapitre 1. Pour la définition des termes qui la caractérisent (strate, grappe, taille de grappe), nous renvoyons à la lecture de ce chapitre.

3. Les simulations d'échantillonnage réalisées portent sur les situations suivantes : surveillance des polluants NO₂, O₃, PM₁₀, PM_{2,5} et BTEX en grande et petite agglomération, surveillance de l'ozone en milieu rural. Les plans d'échantillonnage testés tiennent compte des pratiques répandues dans les AASQA. Ils couvrent 6 % à 50 % environ de l'année. Ces tests étaient destinés à examiner la précision des méthodes, à vérifier que les intervalles de confiance étaient correctement estimés, et à déterminer les conditions d'application les plus appropriées pour chacune des méthodes. L'essentiel des simulations a porté sur la reconstitution de moyennes annuelles. Des travaux complémentaires consacrés à l'estimation de nombres de dépassements de seuils sont prévus.

Dans sa plus simple application, cette méthode se borne à estimer la concentration moyenne inconnue par une **moyenne pondérée des données expérimentales**. Ces dernières sont :

- soit les mesures individuelles tirées de façon indépendante dans les strates temporelles
- soit les grappes réduites à leur valeur moyenne, si l'échantillon se compose de grappes.

Le poids de chaque donnée dépend des caractéristiques du plan d'échantillonnage, c'est-à-dire de la stratification temporelle et du dimensionnement. Il est d'autant plus grand que le nombre de données tirées dans la même strate est plus petit et que cette strate représente une plus large part de la période d'étude.

Soit un tirage de grappes. La reconstitution peut être décomposée en deux étapes. Elle consiste :

- à estimer la moyenne dans chaque strate temporelle :

$$\hat{y}_{pi\ s} = \sum_{g \in G_s} p_{s,g} \cdot \bar{y}_{s,g}$$

Y : variable d'intérêt

G_s : ensemble des grappes prélevées dans la strate s

$\bar{y}_{s,g}$: moyenne des mesures individuelles qui constituent la grappe g

$p_{s,g}$: poids de la grappe g

Si toutes les grappes ont exactement la même taille : $p_{s,g} = \frac{1}{m_s}$ où m_s est le nombre de grappes tirées dans la strate s.

- puis à en déduire la moyenne sur la période entière :

$$\hat{y}_{pi} = \sum_{s \in S_p} \frac{N_s}{N} \cdot \hat{y}_{pi\ s}$$

S_p : ensemble des strates dont se compose la période d'étude

N_s : nombre total de données dans la strate s

N : nombre total de données sur la période d'étude

$\hat{y}_{pi}$ est appelé le **π -estimateur⁽⁴⁾** (par facilité d'écriture, il sera désigné dans la suite par estimateur PI).

Si des variables auxiliaires corrélées à la variable d'intérêt (en particulier, des mesures d'un site fixe de référence) sont disponibles sur l'année entière, les moyennes par strate peuvent être corrigées à l'aide de ces variables : c'est le

4. Cette appellation, propre à la théorie des sondages, vient du fait que cette théorie fait appel à des notions de probabilité, désignant par π_k la probabilité d'appartenance d'un individu y_k à un échantillon.

redressement décrit ci-dessous. Il existe plusieurs méthodes de redressement et par suite, plusieurs estimateurs redressés.

À chaque estimateur (PI ou redressé) est associée une variance, que l'on estime à partir des données de mesure. Cela permet d'encadrer la moyenne estimée par un intervalle de confiance. Le mode de calcul de l'incertitude est présenté page 46.

Estimation d'un nombre de dépassements de seuil

Pour estimer le nombre de dépassements d'un seuil horaire ou journalier, on évalue d'abord un taux annuel de dépassement. La série des données d'échantillonnage et, pour un redressement, la série auxiliaire sont préalablement transformées en séries de 0 (pas de dépassement) et de 1 (dépassement). Tout se passe ensuite exactement comme un calcul de moyenne. Le résultat obtenu est un réel compris entre 0 et 1 : c'est le pourcentage de dépassement estimé sur l'année, que l'on peut transformer en nombre, connaissant le nombre d'heures (ou de jours) dans l'année. Pour un seuil horaire :

$$\hat{N}_{>s} = \hat{T}_{>s} \cdot 8\,760$$

$\hat{T}_{>s}$: taux estimé de dépassement du seuil horaire s ($0 \leq \hat{T}_{>s} \leq 1$)

$\hat{N}_{>s}$: nombre estimé de dépassements

Estimation d'un centile

La méthode des plans de sondage ne permet pas d'estimer directement une valeur de centile.

Une approche, proposée comme une piste possible, est de procéder par dichotomie. Exemple : on souhaite estimer le centile 98 des concentrations horaires, c'est-à-dire la concentration horaire dépassée moins de 2 % du temps.

Soient s_1 et s_1' , $s_1 < s_1'$, deux valeurs seuils telles que le taux de dépassement de chacune (noté T_{s_1} et $T_{s_1'}$) est respectivement supérieur et inférieur à 2 %. On augmente ensuite progressivement s_1 tandis qu'on diminue progressivement s_1' . Si s_n et s_n' sont chacune la plus grande et la plus petite valeur possible telle que $s_n < s_n'$ et $T_{s_n} < 2\% < T_{s_n'}$, alors l'intervalle $[s_n; s_n']$ représente une approximation du centile 98.

► 2.2.2.2 Redressement

Le redressement consiste à corriger la moyenne estimée dans chacune des strates par un facteur multiplicatif ou additif fonction d'une variable auxiliaire X . Le plus souvent, X est une variable de concentration mesurée en un site fixe.

La variable X doit remplir deux critères :

- être mesurée sur toute la période d'étude (l'année) ;
- présenter une bonne corrélation avec la variable d'intérêt. L'annexe 6 décrit la façon d'apprécier la corrélation linéaire entre deux variables.

Dans ces conditions, le redressement permet d'accroître sensiblement la précision de l'estimation.

Il existe plusieurs types de redressement (Tillé, 2001, p. 199 à 204) :

- Le **redressement par le quotient** : dans chaque strate, on multiplie l'estimation PI par le rapport entre la moyenne de X mesurée sur la strate entière et la moyenne de X estimée à partir des périodes de campagne ;

$$\hat{y}_{quot} = \sum_{s \in S_p} \cdot \frac{N_s}{N} \cdot \left(\frac{\hat{y}_{pi\ s}}{\hat{x}_{pi\ s}} \cdot \bar{x}_s \right)$$

- Le **redressement par la différence** : dans chaque strate, on ajoute à l'estimation PI l'écart entre la moyenne de X mesurée sur la strate entière et la moyenne de X estimée à partir des périodes de campagne ;

$$\hat{y}_{diff} = \sum_{s \in S_p} \cdot \frac{N_s}{N} \cdot \left(\hat{y}_{pi\ s} - \hat{x}_{pi\ s} + \bar{x}_s \right)$$

- Le **redressement par la régression** : dans chaque strate, on ajoute à l'estimation PI une fonction linéaire de l'écart entre la moyenne de X mesurée sur la strate entière et la moyenne de X estimée à partir des périodes de campagne ;

$$\hat{y}_{reg} = \sum_{s \in S_p} \cdot \frac{N_s}{N} \cdot \left(\hat{y}_{pi\ s} + b \cdot (\bar{x}_s - \hat{x}_{pi\ s}) \right)$$

Les deux premiers modes de redressement fournissent généralement des résultats comparables **mais attention** : le redressement par la différence considère que dans chaque strate, l'écart entre les moyennes vraie et estimée de X peut être directement reporté au point d'échantillonnage. Lorsque les mesures fixes sont d'un autre type que les mesures temporaires (ex : on redresse une concentration de benzène avec des données de NO₂) ce type de redressement n'est pas utilisable.

Le redressement par la régression ne se pratique qu'avec un nombre suffisant de grappes par strate (3 est un strict minimum). Un moyen d'accroître ce nombre est de regrouper les strates qui se ressemblent : au moment de l'estimation, on considérera par exemple que les deux strates hivernales

(janvier-mars et octobre-décembre), respectivement estivales (avril-juin et mai-septembre) ne font qu'une. En admettant que le lien entre la moyenne du site d'échantillonnage et la moyenne du site auxiliaire reste identique au cours de l'année, il est même envisageable de regrouper l'ensemble des strates.

► 2.2.2.3 Données d'entrée

La reconstitution par la méthode des plans de sondage requiert :

- dans tous les cas,
les données du polluant dont on souhaite estimer la moyenne annuelle et qui ont été recueillies pendant la campagne de mesure (variable Y).
Exemple : les mesures de NO₂ enregistrées pendant 12 semaines de l'année 2005 au site Mazades (MAZ) à Toulouse.
- lorsqu'on veut redresser l'estimation,
les données **sur l'année entière**, et au même pas de temps que les données d'échantillonnage, d'une ou plusieurs variables de redressement. Il s'agit le plus souvent d'une série annuelle de station fixe.
Exemple : X = mesures de NO₂ du site urbain Berthelot (BRT).

► 2.2.2.4 Incertitude

Pour évaluer l'incertitude associée à la moyenne estimée (concentration moyenne annuelle ou taux annuel de dépassement de seuil), il faut calculer la variance d'estimation V.

Celle-ci dépend de la variance estimée de Y et, dans le cas d'un redressement, de la variance estimée de X et de la covariance entre X et Y à l'intérieur de chaque strate. Les formules correspondant à l'estimateur PI et aux estimateurs redressés par la différence, le quotient et la régression sont fournis en annexe 8.

L'intervalle de confiance à 95 % autour de la moyenne estimée

vaut $\left[\hat{y} - 1,96\sqrt{V} ; \hat{y} + 1,96\sqrt{V} \right]$,

ce qui représente une incertitude relative de $\frac{100 \cdot 1,96 \times \sqrt{V}}{\hat{y}}$.

Attention : il s'agit d'un intervalle **estimé** à partir des données d'échantillonnage. Il quantifiera d'autant plus sûrement l'incertitude que les variances par strate auront été mieux estimées.

Le nombre de grappes par strate est parfois insuffisant (ex : ≤ 2) pour estimer correctement la variance dans ces strates. Par suite, l'intervalle de confiance est peu fiable : on n'est pas sûr à 95 % qu'il contient la vraie moyenne. Ce problème se pose quel que soit l'estimateur et de façon plus sensible encore pour l'estimateur redressé par la régression.

Afin de mieux approcher l'incertitude, une solution est de calculer l'intervalle de confiance en regroupant tout ou partie des strates, comme il a été indiqué à propos du redressement par la régression. Le taux de confiance s'en trouve augmenté.

► 2.2.2.5 Application à l'exemple « fil rouge »

Y représente la série des mesures horaires de NO_2 effectuées au site urbain Mazades pendant les 12 semaines de campagne (3 fois 1 semaine par trimestre).

La Figure 11 et le Tableau 6 illustrent pas à pas l'application de la méthode et récapitulent les résultats de mesure et d'estimation.

Grappes			Moyennes mesurées ⁽²⁾ ($\mu\text{g}/\text{m}^3$)		Strates	Variable d'intérêt (Y) mesure est	
Début	Fin	Poids ⁽¹⁾	Variable d'intérêt	Variable auxiliaire	Poids ⁽³⁾	Moyenne réelle ⁽⁴⁾	Moyenne estimée ⁽⁵⁾
07/01/05	14/01/05	0,37	37,8	43,9	0,25	32,2	32,8
17/01/05	24/01/05	0,38	23,2	19,9			
27/02/05	06/03/05	0,25	39,6	35,6			
03/04/05	10/04/05	0,31	18,4	19,4	0,25	18,1	19,7
19/04/05	26/04/05	0,35	20,9	18,7			
14/05/05	21/05/05	0,35	19,6	20,5			
04/07/05	11/07/05	0,34	12,5	11,7	0,25	17,4	16,2
12/07/05	19/07/05	0,31	16,2	18,7			
23/08/05	30/08/05	0,35	19,7	16,7			
23/10/05	30/10/05	0,30	13,2	13,6	0,25	30,5	33,5
03/12/05	10/12/05	0,35	32,3	27,2			
20/12/05	27/12/05	0,35	52,3	52,7			

1. Poids de chaque grappe, fonction de sa longueur (une grappe a moins de poids si elle compte des données manquantes).

2. Moyennes par grappe.

3. Poids de la strate, fonction de sa longueur. On considère ici quatre strates de même taille.

4. Moyenne réelle sur la strate entière (il s'agit de la valeur supposée ici inconnue $\bar{y}_s$).

5. Moyenne sur la strate estimée à partir des 3 moyennes de grappes (il s'agit de $\bar{y}_{pi\ s}$).

Tableau 6 – Application de la méthode des plans de sondage à l'exemple « fil rouge ». Estimation de la moyenne annuelle de NO_2 . Moyennes par grappe et par strate temporelle sans redressement. Résultats numériques.

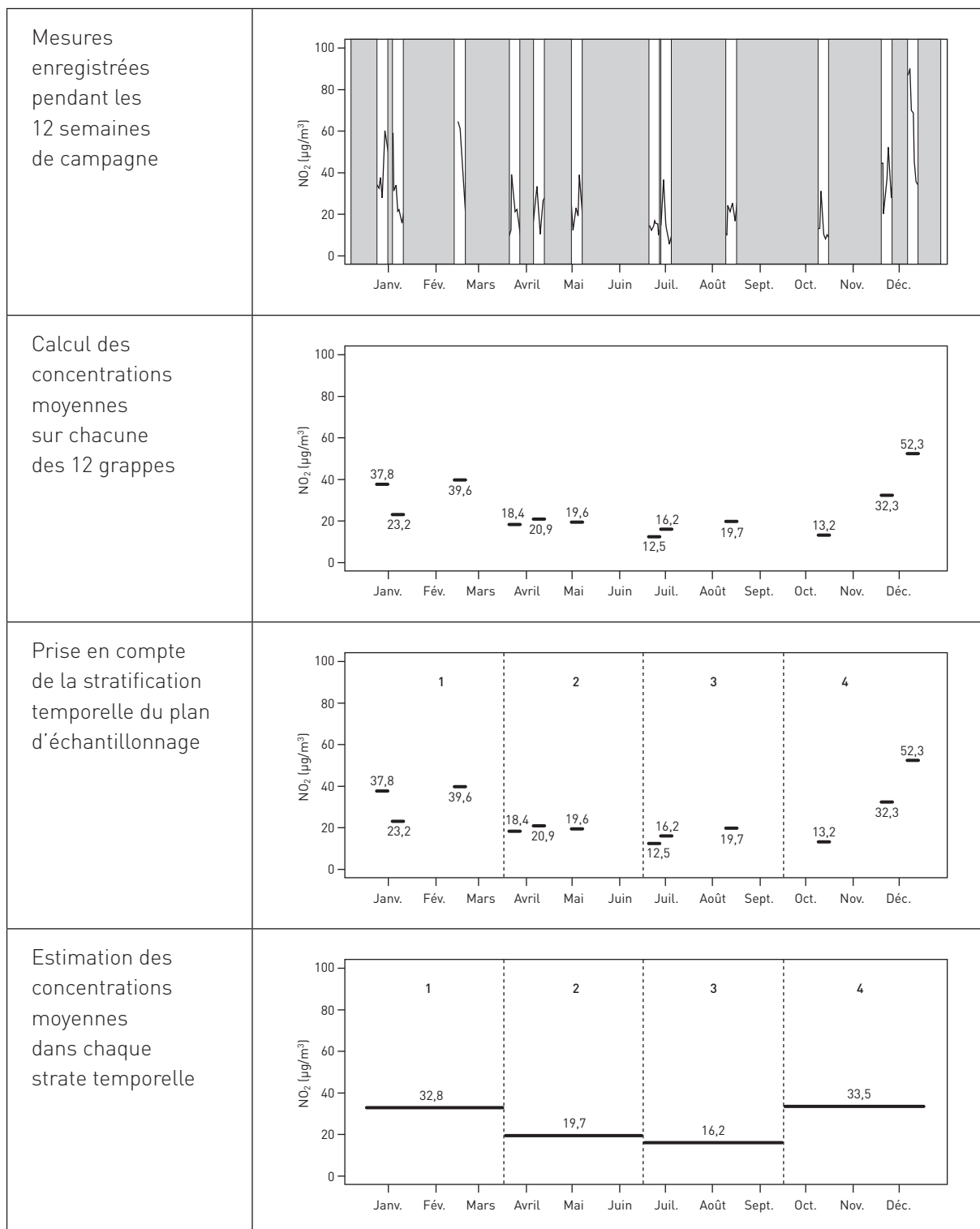


Figure 11 – Application de la méthode des plans de sondage à l'exemple « fil rouge ». Estimation de la moyenne annuelle de NO_2 . Moyennes par grappe et par strate temporelle sans redressement. Illustration.

X_1 est la série des mesures horaires de NO_2 enregistrées par la station fixe Berthelot pendant toute l'année 2005. Compte tenu de la bonne corrélation entre les deux sites (Figure 12, Figure 13), cette station est choisie comme station de redressement.

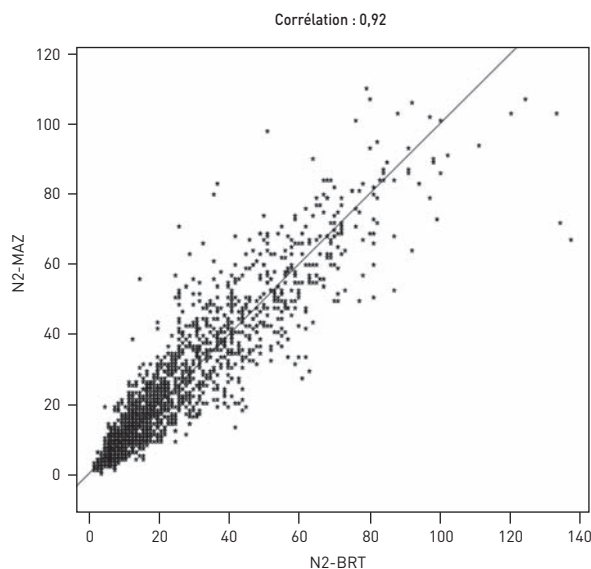


Figure 12 – Nuage de corrélation entre les concentrations mesurées sur le site de Mazades et les concentrations mesurées sur le site de Berthelot pendant la campagne de mesure.

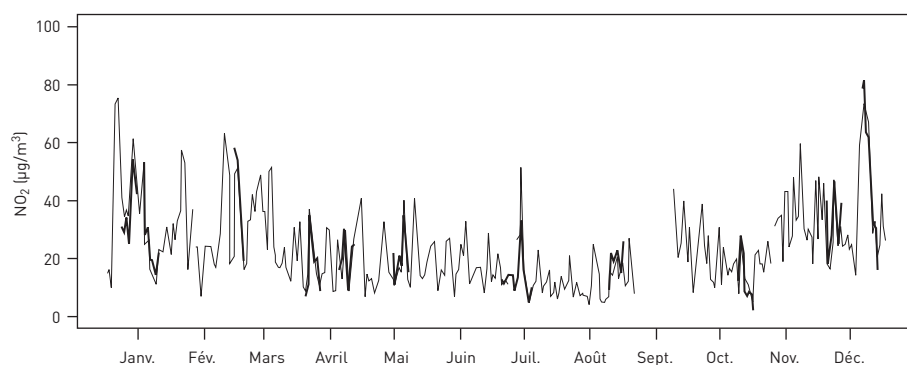


Figure 13 – Concentrations mesurées sur le site de Mazades (tracé épais) et sur le site fixe de Berthelot (tracé fin).

Pour chaque strate temporelle, le Tableau 7 présente la moyenne de X_1 sur la période d'échantillonnage – cette moyenne est proche de celle de Y – et la moyenne de X_1 sur la strate entière.

Grappes		Variable auxiliaire (X ₁)		Variable d'intérêt (Y)
Début	Fin	Moyenne réelle ⁽⁴⁾	Moyenne estimée ⁽⁵⁾	Moyenne estimée ⁽⁵⁾
07/01/05	14/01/05	31,9	32,8	32,8
17/01/05	24/01/05			
27/02/05	06/03/05			
03/04/05	10/04/05	18,7	19,5	19,7
19/04/05	26/04/05			
14/05/05	21/05/05			
04/07/05	11/07/05	14,8	16,4	16,2
12/07/05	19/07/05			
23/08/05	30/08/05			
23/10/05	30/10/05	27,8	31,1	33,5
03/12/05	10/12/05			
20/12/05	27/12/05			

4. Moyenne réelle sur la strate entière.

5. Moyenne sur la strate estimée à partir des 3 moyennes de grappes (il s'agit de $\hat{X}_{1pi s}$ et $\hat{Y}_{pi s}$).

Tableau 7 – Moyennes trimestrielles au site Berthelot (X₁) calculées respectivement sur trois mois complets de données et sur trois grappes hebdomadaires. Comparaison avec le site Mazades (Y).

Après redressement, les moyennes par strate estimées pour Y sont plus proches des valeurs réelles. Il en va de même pour la moyenne annuelle (égale à 24,4 µg/m³). Le gain de précision lié au redressement se traduit par une diminution de l'incertitude estimée (Tableau 8).

Remarque : l'incertitude associée au redressement par la régression est la plus faible. Si l'estimation correspondante est bien la plus proche de la moyenne réelle, l'expérience montre qu'avec un nombre limité de grappes par strate, l'incertitude obtenue par ce mode de redressement a tendance à être sous-estimée.

Estimation de la moyenne de la variable d'intérêt par strate et sur la population										
Sans redressement			Redressement par la différence			Redressement par le quotient			Redressement par la régression	
Moyenne		Incertitude estimée	Moyenne		Incertitude estimée	Moyenne		Incertitude estimée	Moyenne	Incertitude estimée
32,8	25,5	23,3 %	31,9	23,9	10,1 %	31,9	23,8	10,5 %	32,2	24,5
19,7			18,8			18,8			20,2	
16,2			14,5			14,5			15,1	
33,5			30,3			30,0			30,4	

Tableau 8 – Application de la méthode des plans de sondage à l'exemple « fil rouge ». Estimation de la moyenne annuelle de NO₂. Moyennes par strate temporelle sans et avec redressement et moyennes annuelles. Résultats numériques.

► 2.2.2.6 Avantages et limites

Avantages :

- La méthode des plans de sondage est aisée et rapide à mettre en œuvre.
- Elle peut être utilisée en l'absence de toute variable auxiliaire.
- Si les données d'échantillonnage sont bien corrélées à celles d'un site fixe (qui mesure le même polluant ou un polluant de même origine), la précision de l'estimation peut être sensiblement accrue par un redressement.

Limites :

- La méthode est exigeante en ce qui concerne le plan d'échantillonnage :
 - En théorie, les données doivent être tirées aléatoirement dans les strates ;
 - Pour une estimation fiable de l'intervalle de confiance, des grappes courtes et nombreuses (au moins 3 par strate) sont préférables à des grappes longues et peu nombreuses.
- L'estimation d'un nombre de dépassements de seuil est moins précise que celle de la moyenne annuelle.
- Les centiles peuvent être seulement approchés.

2.2.3 Méthode « ISO » (méthode issue de la norme ISO 9359)

La méthode ISO, ainsi nommée parce qu'elle a été adaptée de la norme ISO 9359 (Houdret, 2002 à 2004), procède de la théorie des sondages, comme la méthode des plans de sondage décrite précédemment. Dans son principe, elle est similaire à cette dernière ; la différence essentielle ne réside que dans la façon de grouper les données de mesure.

Dans la méthode des plans de sondage, l'estimation tient compte de la répartition des données en grappes et en strates telle qu'elle a été définie par le plan d'échantillonnage.

La méthode ISO, à l'issue la campagne, et avant de réaliser l'estimation, procède à un nouveau regroupement des données en fonction de variables auxiliaires (on parle de post-stratification pour distinguer celle-ci de la stratification temporelle du plan d'échantillonnage).

► 2.2.3.1 Principe

La méthode ISO s'appuie sur la norme « *ISO 9359 – Qualité de l'air – Échantillonnage aléatoire stratifié pour l'évaluation de la qualité de l'air ambiant* » qui considère l'influence des variables météorologiques et/ou temporelles suivantes :

- directions de vent sélectionnées par rapport à un émetteur particulier s'il est l'objet de la campagne,
- classes de valeurs de paramètres météorologiques : température et vitesse du vent,
- heures du jour et jours de semaine.

La limitation volontaire à ces variables disponibles dans toutes les AASQA rend cette méthode aisément praticable par tous mais d'autres paramètres pourraient être aussi utilisés : pression, humidité relative, radiations solaires (pour O₃)...

La méthode ISO a également été appliquée à des mesures journalières en utilisant des variables pertinentes pour de tels pas de temps. En revanche, cette méthode est peu utilisable pour des données hebdomadaires (HAP, ML...), car des données météorologiques n'ont pas de sens sur une base hebdomadaire.

La mise en œuvre de la méthode ISO peut être décomposée en trois étapes.

1. Définition des strates paramétriques

Cette étape préalable a pour objet d'évaluer l'impact des paramètres influents sur les concentrations du ou des polluants d'intérêt puis à délimiter, pour chaque paramètre, des classes de valeurs plus ou moins propices à des concentrations élevées (Tableau 9). Elle est conduite sur des séries annuelles de données propres à la zone géographique étudiée (année de la campagne et/ou années antérieures) ou, à défaut de telles séries, sur les données collectées pendant la campagne de mesure.

Polluants	Moyennes				Dépassements de seuils			
	VV m.s ⁻¹	T en °C	Périodes en heures	Jours	VV m.s ⁻¹	T en °C	Périodes en heures	Jours
NO	< 2	< 12	7-11, 16-21	1-5	< 2	> 18	7-11, 16-21	2-4
NO ₂	< 2	< 12	7-11, 16-21	1-5	< 2	> 18	7-11, 16-21	2-4
O ₃	> 2	> 20	10-19	1-7	< 2	> 27	11-19	1-3
CO	< 1,5	< 12	7-11, 16-21	2-6	< 1,5	< 18	7-10, 16-21	1-4
PM	< 2	< 6, > 20	7-11, 19-22	1-5	< 2	< 6	7-11, 19-22	1-4

Tableau 9 – Exemple des classes de paramètres d'influence favorisant des valeurs horaires élevées pour plusieurs polluants.

À partir de cette analyse, des strates paramétriques sont définies selon la méthodologie suivante :

- Pour chaque mesure horaire d'un polluant donné, un paramètre influent est affecté de la valeur de classe « 2 » s'il correspond à une concentration élevée de ce polluant selon les conditions fixées dans le Tableau 9, et de la valeur « 1 » pour toutes les autres.

- Puis, pour chaque mesure horaire, on regarde combien de paramètres valent simultanément 2. Si quatre paramètres sont pris en compte, ce nombre « Σ » est compris entre 0 et 4 : $\Sigma = 0$ à 4.
- On définit ensuite 4 strates dans lesquelles sont ventilées les données horaires :
Strate 1 : $\Sigma = 0$, Strate 2 : $\Sigma = 1$, Strate 3 : $\Sigma = 2$ et Strate 4 : $\Sigma = 3$ et 4.

Ainsi, la strate 1 réunit les concentrations les plus faibles, la strate 4 les concentrations les plus élevées, et les strates 2 et 3 les situations intermédiaires.

Note 1 : pour simplifier la méthode, on a attribué une importance égale à chacun des paramètres d'influence pris en compte ; la stratification pourrait être affinée en pondérant les influences des divers paramètres. On peut en effet imaginer une hiérarchie des influences dans cet ordre décroissant : les pointes horaires, la température, la vitesse du vent, les jours ouvrés.

Note 2 : ce choix permet d'obtenir des nombres de données par strate les plus voisins possibles. La norme ISO 9359 recommande de limiter à 3 ou 4 le nombre de strates ; en effet, une stratification plus fine n'abaisse pas nécessairement les variances dans les strates, et risque de réduire trop fortement les effectifs par strate (notamment si elle est utilisée pour des campagnes de courte durée).

Exemple :

Le Tableau 10 a montré que les concentrations en NO_2 sont plus élevées :

- lorsque la vitesse du vent est $< 2 \text{ m/s}$,
- lorsque la température est $< 12^\circ\text{C}$,
- de 7 à 11 h et de 16 à 21 h,
- du lundi au vendredi.

Soit une heure pour laquelle les paramètres de vent, température, heure du jour et jour de semaine sont simultanément en classe 2. La donnée de concentration correspondante sera donc affectée à la strate 4. Si cette mesure a lieu un dimanche, elle se trouvera encore dans la strate 4. Mais elle passera dans la strate 3 si de plus, le vent a une vitesse supérieure à 2 m.s^{-1} ; dans la strate 2, si de plus, la température est supérieure à 10°C , et dans la strate 1, si de plus elle intervient en dehors des heures de pointe (Tableau 10).

[NO ₂]	$v < 2 \text{ m.s}^{-1}$	$T < 10^\circ\text{C}$	Lundi au vendredi	6 h à 9 h et 15 h à 21 h	Σ	N° de strate i
élevée	✓	✓	✓	✓	4	4
	✓	✓		✓	3	4
		✓		✓	2	3
				✓	1	2
faible					0	1

Tableau 10 – Correspondance entre variables d'influence et numéros de strates. Exemple pour le NO_2 . La condition indiquée en première ligne du tableau est vérifiée si la case est cochée.

Note 3 : Il ressort des pré-études réalisées sur des données de plusieurs années consécutives dans plusieurs villes, que les roses des vents et les roses de pollution sont dans l'ensemble **reproductibles d'une année sur l'autre**, et que par suite, il n'y a pas nécessité de les refaire. Elles sont néanmoins très utiles lorsque l'on ne connaît pas exactement les limites des classes de vitesse de vent et de température.

Même si elle peut être affinée au cours du temps, la définition des strates n'est donc pas nécessairement spécifique à une campagne de mesure. Si elle se fonde sur un solide historique de données, **elle peut être utilisée pour différentes campagnes conduites dans une même zone géographique (dont il conviendra d'apprécier l'étendue)** et en des sites de même typologie.

2. Répartition des heures dans les strates paramétriques

Cette étape requiert les séries complètes des paramètres d'influence sur l'année étudiée. À chaque heure de l'année est associé un numéro de strate, selon les valeurs de ces paramètres. On détermine alors les N_i , nombres totaux d'heures sur l'année, et les n_i , nombres d'heures sur la campagne de mesure, qui appartiennent aux strates i .

Désormais, les données de mesure ne sont plus groupées dans le temps selon les grappes et strates temporelles qui caractérisent le plan d'échantillonnage mais elles sont dispersées entre les strates paramétriques.

Soit par exemple un échantillonnage composé de grappes hebdomadaires. D'après la stratification précédemment définie, les données d'une grappe prélevées un lundi aux heures de pointe par vent calme ne se retrouveront pas dans la même strate que les données de cette même grappe prélevées un dimanche en milieu de journée par vent fort : les premières appartiendront à la strate 4 et les secondes, à la strate 1 ou 2. Selon la variabilité des conditions météorologiques, une grappe hebdomadaire se verra ainsi complètement ou partiellement éclatée entre plusieurs strates paramétriques.

En conséquence, une strate paramétrique pourra contenir aussi bien des mesures individuelles isolées que des séquences de deux à plusieurs heures consécutives.

3. Reconstitution

Hormis le fait que les strates et les grappes ne sont plus celles du plan d'échantillonnage, la reconstitution se déroule de la même façon que pour les plans de sondage :

- on calcule tout d'abord les moyennes par strate.

$$\hat{\bar{y}}_i = \frac{1}{n_i} \sum_j y_{ij}$$

Y : variable d'intérêt

y_{ij} : ensemble des données horaires d'échantillonnage appartenant à la strate i

- on émet l'hypothèse que le résultat obtenu pour les n_i heures de la strate i pendant la campagne est identique, aux incertitudes près, à celui que donneraient les N_i heures de la strate i pendant l'année.
- on estime la moyenne annuelle en pondérant les moyennes par strate la probabilité d'occurrence de ces strates :

$$\hat{\bar{y}} = \sum_i \cdot \frac{N_i}{N} \cdot \hat{y}_i \quad (\text{estimateur PI des plans de sondage})$$

N : nombre total d'heures sur l'année

Estimation d'un nombre de dépassements de seuil

L'estimation d'un nombre de dépassements de seuil se passe comme pour la moyenne annuelle : il suffit de remplacer préalablement les concentrations individuelles par 0 si le seuil n'est pas dépassé et par 1 s'il est franchi.

Estimation d'un centile

Une approche fondée sur le redressement est actuellement à l'étude. La principale difficulté réside dans le calcul de l'incertitude.

► 2.2.3.2 Redressement

Les moyennes estimées dans chaque strate paramétrique peuvent être redressées à l'aide d'une variable auxiliaire X par la méthode du quotient, de la différence ou de la régression (page 44).

► 2.2.3.3 Données d'entrée

La reconstitution par la méthode ISO requiert :

- dans tous les cas,
 - les données du polluant dont on souhaite estimer la moyenne annuelle et qui ont été recueillies pendant la campagne de mesure (variable Y).
Exemple : les mesures de NO₂ enregistrées pendant 12 semaines de l'année 2005 au site Mazades (MAZ) à Toulouse.
 - des données **sur l'année entière** de vent, de température, ou de toute autre variable auxiliaire qu'il a été décidé de prendre en compte.
- lorsqu'on veut redresser l'estimation,
 - les données **sur l'année entière**, et au même pas de temps que les données d'échantillonnage, d'une ou plusieurs variables de redressement. Il s'agit le plus souvent d'une série annuelle de station fixe.
Exemple : X = mesures de NO₂ du site urbain Berthelot (BRT).

► 2.2.3.4 Incertitude

L'estimation de l'incertitude est fondée sur le calcul de la variance d'estimation V , qui dépend elle-même des variances (et des covariances en cas de redressement) dans chaque strate paramétrique (page 44).

Les formules correspondant à l'estimateur PI et aux estimateurs redressés par la différence, le quotient et la régression sont fournis en annexe 8.

L'intervalle de confiance à 95 % autour de la moyenne estimée

$$\text{vaut } \left[\hat{y} - 1,96\sqrt{V} ; \hat{y} + 1,96\sqrt{V} \right],$$

ce qui représente une incertitude relative de $\frac{100 \cdot 1,96 \times \sqrt{V}}{\hat{y}}$.

Attention : il s'agit d'un intervalle estimé à partir des données d'échantillonnage. Il quantifiera d'autant plus sûrement l'incertitude que les variances par strate auront été mieux estimées.

Or cette estimation repose sur l'hypothèse que les données sont tirées aléatoirement dans les **strates paramétriques** et qu'elles sont indépendantes (absence de corrélation temporelle). On peut admettre que la première condition est assurée par le caractère aléatoire de la météorologie. En revanche, la seconde condition est mise en défaut si des séquences de plusieurs heures consécutives se retrouvent dans une même strate. C'est pourquoi, dans une strate paramétrique donnée, toutes les mesures issues d'une même grappe d'échantillonnage restent encore assimilées à une grappe, représentée par sa valeur moyenne.

► 2.2.3.5 Application à l'exemple « fil rouge »

Y représente la série des mesures horaires de NO_2 effectuées au site urbain Mazades pendant les 12 semaines de campagne. X_1 est la série des mesures horaires de NO_2 enregistrées par la station fixe Berthelot pendant toute l'année 2005. Compte tenu de la bonne corrélation entre les deux sites (voir le nuage de corrélation présenté page 42), cette station est choisie comme station de redressement. Quatre strates paramétriques sont considérées, ainsi qu'il a été défini page 51.

En fonction des conditions météorologiques, du jour et de l'heure, les données des grappes sont réparties entre ces quatre strates. Les portions d'une grappe affectées à des strates paramétriques différentes deviennent de nouvelles grappes de taille variable dont la Figure 14 indique les valeurs moyennes.



Strates	Variable auxiliaire		Variable d'intérêt	
Poids	Moyenne réelle	Moyenne estimée	Moyenne réelle	Moyenne estimée
0,04	11,6	10,3	12,6	11,6
0,19	19,5	18,6	19,5	19,1
0,40	22,5	25,6	22,7	25,3
0,37	28,6	30,1	29,6	30,4

Tableau 11 – Application de la méthode ISO à l'exemple « fil rouge ». Estimation de la moyenne annuelle de NO_2 . Moyennes par strate paramétrique sans redressement.

Estimation de la moyenne de la variable d'intérêt par strate et sur la population											
Sans redressement			Redressement par la différence			Redressement par le quotient			Redressement par la régression		
Moyenne		Incertitude	Moyenne		Incertitude	Moyenne		Incertitude	Moyenne		Incertitude
11,6	25,4	18,4 %	12,9	23,9	5,8 %	13,1	23,9	5,8 %	13,4	24,1	4,7 %
19,1			19,9			20,0			19,9		
25,3			22,2			22,2			22,4		
30,4			28,9			28,9			28,9		

Tableau 12 – Application de la méthode ISO à l'exemple « fil rouge ». Estimation de la moyenne annuelle de NO_2 . Moyennes par strate paramétrique sans et avec redressement et moyennes annuelles.

► 2.2.3.6 Avantages et limites

Avantages :

- Une fois que la stratification paramétrique est définie, la méthode est aisée et rapide à mettre en œuvre.
- Les fluctuations de la météorologie assurant le caractère aléatoire de l'échantillonnage dans les strates paramétriques, la méthode peut s'appliquer avec un échantillonnage aléatoire ou systématique dans le temps.
- Par la stratification paramétrique, la méthode permet de prendre en compte des variables influentes pour grouper les données en ensembles relativement homogènes.
- Si les données d'échantillonnage sont bien corrélées à celles d'un site fixe (qui mesure le même polluant ou un polluant de même origine), la précision de l'estimation peut être sensiblement accrue par un redressement.

Limites :

- La méthode ne s'applique pas ou s'applique plus difficilement à des prélèvements de longue durée.
- Elle requiert des données météorologiques ; toutefois les données de vent et de température sont généralement accessibles.
- Elle exige d'examiner, polluant par polluant et pour des zones géographiques à délimiter, l'influence de variables auxiliaires sur les concentrations. Cette étape préalable peut être assez longue.
- Pour une stratification paramétrique donnée, il est nécessaire que toutes les strates paramétriques soient représentées dans l'échantillonnage.
- L'estimation d'un nombre de dépassements de seuil est moins précise que celle de la moyenne annuelle et les centiles peuvent être seulement approchés.

2.2.4 La régression linéaire

► 2.2.4.1 Principe

La modélisation par régression consiste à établir une relation statistique entre les mesures de concentration réalisées lors d'une campagne (variable à expliquer) et un ensemble de variables auxiliaires (variables explicatives). Ces dernières décrivent un impact direct et avéré sur les concentrations (émissions, météorologie) ou reflètent indirectement ces influences (concentrations du même polluant mesurées ailleurs, concentrations d'autres polluants issus de sources communes, etc.). Leurs données doivent être disponibles sur toute la période de reconstitution (l'année).

Pour des explications détaillées sur la régression, on pourra se référer aux ouvrages de Saporta (2006, chapitres 16 et 17), Tomassone et al. (1992), Lebart et al. (1982).

Il existe une multitude de modèles dont le choix dépend du type de relation envisagé. Le plus simple est le modèle linéaire, qui s'écrit traditionnellement :

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \dots + \beta_k \cdot X_k + \varepsilon$$

Soit sous forme matricielle : $Y = X \cdot \beta + \varepsilon$.

X : matrice des variables explicatives (incluant la constante) ;

β : vecteur des paramètres $(\beta_i)_{0 \leq i \leq k}$ du modèle.

Il est ajusté sur les données collectées durant la campagne de mesure : cet ajustement comprend la sélection des variables explicatives et l'estimation des paramètres $(\beta_i)_{0 \leq i \leq k}$, où k est le nombre de variables explicatives retenues.

L'hypothèse essentielle pour la reconstitution est que le modèle obtenu s'applique à tout moment de l'année. En dehors de la campagne de mesure, les valeurs inconnues de Y sont donc estimées par l'opération : $\hat{Y} = X_p \cdot \hat{\beta}$, où X_p représente la matrice des variables d'influence (constante incluse) mesurées hors campagne et $\hat{\beta}$, le vecteur des paramètres estimés du modèle.

La série finale, notée encore $Y = (y_i)_{1 \leq i \leq N}$ se compose des données mesurées pendant la campagne, $(y_i)_{i \in \text{campagne}}$, et des concentrations individuelles reconstituées sur le reste de l'année, $(\hat{y}_i)_{i \notin \text{campagne}}$.

L'estimation de la moyenne annuelle est égale à la moyenne de cette série :

$$\hat{\bar{y}} = \frac{1}{N} \left(\sum_{i \in \text{campagne}} y_i + \sum_{i \notin \text{campagne}} \hat{y}_i \right)$$

Remarque : Lorsque le modèle de régression possède un terme constant β_0 , les séries mesurée et reconstituée ont même moyenne sur la période d'échantillonnage. La valeur de $\hat{\bar{y}}$ n'est donc quasiment pas modifiée si sur la période de campagne, les prédictions $(\hat{y}_i)_{i \in \text{campagne}}$ du modèle sont substituées aux valeurs individuelles $(y_i)_{i \in \text{campagne}}$.

Estimation d'un nombre de dépassements de seuil et d'un centile

Puisque la série annuelle complète est reconstituée, il est possible d'en déduire une estimation de n'importe quel indicateur annuel, à condition que ce dernier se réfère au même pas de temps que la série (seuil horaire et centile des valeurs horaires pour une série de concentrations horaires, seuil journalier et centile des valeurs journalières pour une série de concentrations journalières). La précision de ces estimations dépend toutefois de l'aptitude du modèle à restituer correctement les concentrations plus élevées.

Les différentes étapes de la reconstitution sont récapitulées et illustrées ci-après (Tableau 13, Figure 15).

Sélection des variables explicatives :

Dans le cas d'un modèle linéaire, on effectue un premier choix de variables corrélées linéairement à la variable d'intérêt (calcul des corrélations et examen des nuages de corrélation, cf. annexe 6). Des algorithmes de sélection automatique développés dans les logiciels de statistique (tel l'algorithme stepwise, Besse, 2006) permettent de n'en conserver qu'un certain nombre.

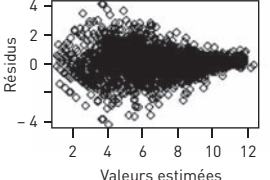
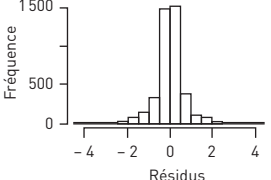
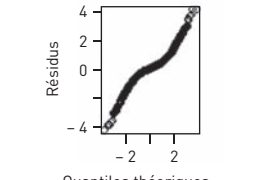
Estimation des paramètres du modèle :

Les coefficients $(\beta_i)_{0 \leq i \leq k}$ sont généralement estimés par la méthode des moindres carrés.

Contrôle du modèle :

Les résidus de la régression, c'est-à-dire les écarts entre les valeurs de Y mesurées et estimées par régression sur la période de la campagne, sont analysés. En examinant les graphiques du type de ceux qui sont présentés ci-dessous, et à l'aide de tests statistiques, on s'assure qu'ils satisfont aux hypothèses théoriques :

- d'*homoscédasticité* : la variance des résidus est constante sur tout le domaine des variables explicatives (on parle d'*hétéroscédasticité* dans le cas contraire),
- de *normalité* : les résidus se distribuent selon une loi normale.

Nuages résidus / valeurs estimées 	Histogramme des résidus 	Graphique quantiles Quantiles pour la loi normale 
Résidus en fonction des valeurs prédites. Si le critère d'homoscédasticité est vérifié, la dispersion du nuage est indépendante des valeurs prédites.	Histogramme des résidus	Relation entre les résidus et les valeurs qu'ils prendraient s'ils se distribuaient parfaitement selon une loi normale. Si le critère de normalité est vérifié, la courbe expérimentale suit la bissectrice.

Application du modèle :

Reconstitution des données en dehors des périodes de mesure.
Calcul du ou des indicateurs d'intérêt : moyenne annuelle, nombre de dépassements d'un seuil donné, centile.

Tableau 13 – Étapes d'une reconstitution par régression.

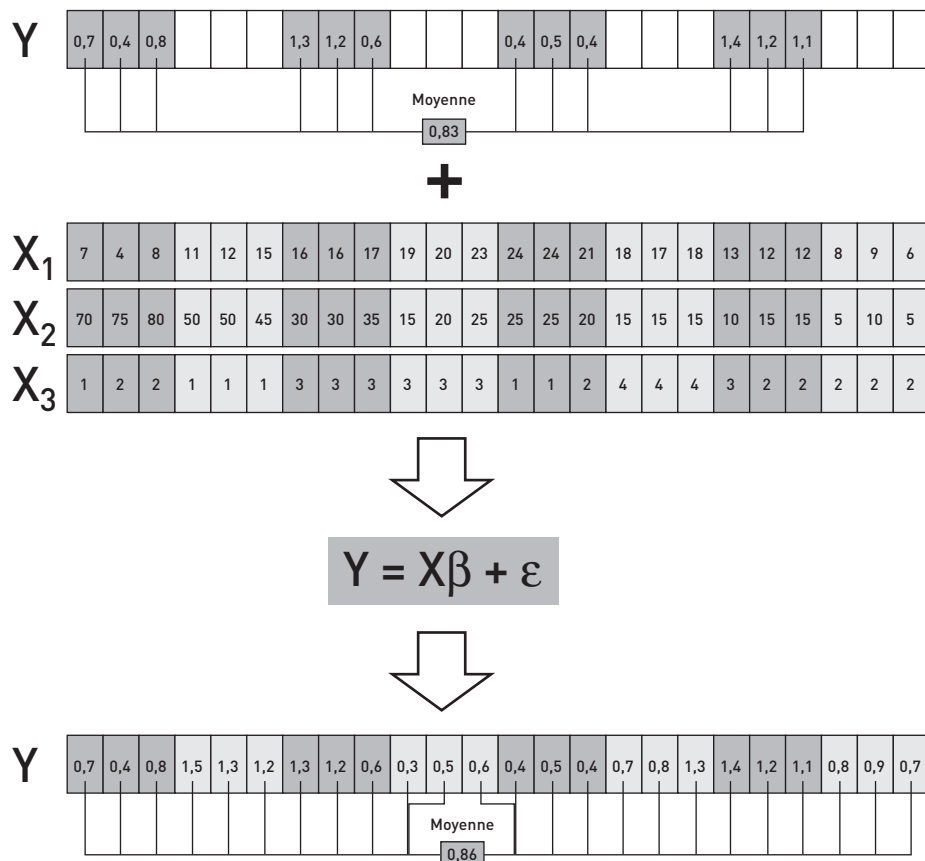


Figure 15 – Principe de la reconstitution par régression. *Y* est la variable à expliquer. *X* est la matrice des variables explicatives (*X*₁, *X*₂ et *X*₃ dans cet exemple). *β* est le vecteur des paramètres du modèle de régression et *ε* est un résidu aléatoire.

► 2.2.4.2 Redressement

Il n'y a pas lieu de parler de redressement, c'est-à-dire de correction de l'estimation par des variables auxiliaires, puisque par construction même du modèle de régression, les données reconstituées s'expriment à l'aide de telles variables.

► 2.2.4.3 Données d'entrée

La reconstitution par régression requiert :

- les données du polluant dont on souhaite estimer la moyenne annuelle et qui ont été recueillies pendant la campagne de mesure (variable *Y*).

Exemple : les mesures de NO₂ enregistrées pendant 12 semaines de l'année 2005 au site Mazades (MAZ) à Toulouse.

- les données sur l'année entière, et au même pas de temps que les données d'échantillonnage, de toutes les variables explicatives $\{X_i\}_{1 \leq i \leq k}$.

Exemple : X_1 = mesures de NO_2 du site urbain Berthelot (BRT), X_2 = mesures d'ozone de ce même site, X_3 = mesures de température...

Les données concomitantes de Y et des X_i servent à caler le modèle. Les données des X_i disponibles en dehors de la campagne de mesure sont employées pour la reconstitution.

► 2.2.4.4 Incertitude

Incertainitude sur la moyenne annuelle reconstituée

Comme la moyenne annuelle estimée découle de la série reconstituée sur l'année entière, son incertitude ne reflète pas le caractère partiel de l'échantillonnage, mais elle traduit à la fois l'erreur de modélisation et la variabilité due à l'imprécision des estimations du modèle.

On a vu que la moyenne annuelle était donnée par :

$$\hat{\bar{y}} = \frac{1}{N} \cdot \sum_{i=1}^N \hat{y}_i = \frac{1}{N} \cdot \sum_{i=1}^N x'_{p,i} \cdot \hat{\beta}$$

$x'_{p,i} = (1, x_{1,i}, x_{2,i}, \dots, x_{k,i})$: vecteur ligne des valeurs au temps i des variables explicatives

$\hat{\beta}$: vecteur colonne des paramètres estimés du modèle

N : nombre total de données dans la série reconstituée

Il en résulte que : $\hat{\bar{y}} = \bar{x}_p \cdot \hat{\beta}$

$\bar{x}_p = (1, \bar{x}_1, \bar{x}_2, \dots, \bar{x}_k)$: vecteur ligne des valeurs moyennes annuelles des variables explicatives

Pour évaluer l'incertitude sur la moyenne annuelle, cette seconde formulation est utilisée. On calcule à cette fin la variance V de l'erreur de reconstitution, qui dépend des deux sources de variabilité citées plus haut (Dégerine, 2002) :

$$V = u^2(\hat{\bar{y}}) = \sigma^2 \cdot \left(\frac{1}{N} + \bar{x}_p' (X'X)^{-1} \bar{x}_p \right)$$

$$\text{avec : } \sigma^2 = \frac{1}{n - k - 1} \sum_{j=1}^n (y_j - \hat{y}_j)^2$$

X : matrice des variables du modèle sur la période de la campagne

n : nombre de mesures individuelles collectées durant la campagne

k : nombre de variables explicatives dans le modèle

L'intervalle de confiance à 95 % autour de la moyenne annuelle estimée est obtenu avec un facteur d'élargissement $k = 2$:

$$\left[\hat{y} - 2 \times \sqrt{V} ; \hat{y} + 2 \times \sqrt{V} \right]$$

Soit une incertitude relative de $\frac{100 \cdot 2 \times \sqrt{V}}{\hat{y}}$.

Attention : Cette incertitude est d'autant mieux estimée que l'hypothèse d'un modèle linéaire est bien vérifiée. Plus la réalité est éloignée d'un tel modèle, moins l'intervalle de confiance est fiable.

Incertitude sur le nombre de dépassements de seuil ou sur un centile

La méthode ne permet pas d'associer aisément une incertitude à ce type de donnée reconstituée. Ce problème n'a pas été examiné par le GT.

► 2.2.4.5 Application à l'exemple « fil rouge »

Y représente la série des mesures horaires de NO_2 effectuées au site urbain Mazades pendant les 12 semaines de campagne. X_1 est la série des mesures horaires de NO_2 enregistrées par la station fixe Berthelot pendant toute l'année 2005. Compte tenu de la bonne corrélation entre les deux sites (voir le nuage de corrélation présenté page 42), X_1 est retenue comme variable explicative.

L'étape suivante est de déterminer la relation entre les deux variables. On part du postulat que cette relation est de la forme $Y = \beta_0 + \beta_1 \cdot X_1$. Les paramètres β_0 et β_1 sont estimés par moindres carrés. Leurs estimations valent $b_0 = 2,881$ et $b_1 = 0,914$.

Les résidus ne satisfont pas rigoureusement aux hypothèses de normalité et d'homoscédasticité ; leur variance croît notamment avec la valeur de X_1 (Figure 16). Pour une première estimation, la reconstitution est cependant poursuivie.

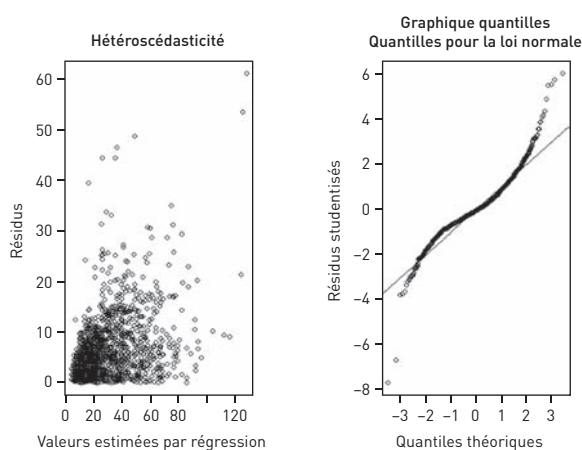


Figure 16 – Étude de l'homoscédasticité et de la normalité des résidus.

Dans une troisième étape, la série annuelle est reconstituée heure par heure. Pour chaque heure h de l'année 2005, si h appartient à la campagne de mesure, alors $Y(h)$ = valeur mesurée au site de campagne, sinon $Y(h) = 2,881 + 0,914 \cdot x_1$, où x_1 est la concentration de NO_2 mesurée à la station fixe.

Enfin, on estime la moyenne annuelle de Y en faisant la moyenne de toutes les données horaires de la série (mesurées sur la durée de la campagne, reconstituées pour le reste de l'année).

Cette moyenne vaut **24,49 $\mu\text{g}/\text{m}^3$** . Elle est proche de la moyenne réelle qui vaut 24,40 $\mu\text{g}/\text{m}^3$. L'intervalle de confiance à 95 % autour de la moyenne vaut [24,07 $\mu\text{g}/\text{m}^3$; 24,92 $\mu\text{g}/\text{m}^3$], ce qui représente une incertitude de 1,7 %. Cette incertitude est peut-être sous-estimée du fait que les résidus ne satisfont pas entièrement aux hypothèses théoriques.

Le modèle de régression peut être complété par d'autres variables auxiliaires. Une sélection automatique pas à pas (algorithme *stepwise*) conduit à retenir les mesures de monoxyde d'azote, d'ozone et de dioxyde de soufre à la station urbaine Berthelot ($X_2 = \text{NO_BRT}$, $X_3 = \text{O3_BRT}$, $X_4 = \text{S2_BRT}$), la direction de vent, la vitesse du vent, les précipitations et la température à la station périurbaine de Colomiers ($X_5 = \text{DV_COL}$, $X_6 = \text{VV_COL}$, $X_7 = \text{PR_COL}$, $X_8 = \text{T_COL}$). Le modèle de régression est de la forme :

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta_4 \cdot X_4 + \beta_5 \cdot X_5 + \beta_6 \cdot X_6 + \beta_7 \cdot X_7 + \beta_8 \cdot X_8.$$

Les valeurs estimées par moindres carrés des coefficients β_i sont :

$$b_0 = 300,338; b_1 = 0,766; b_2 = 0,027; b_3 = -0,072; b_4 = 0,383; b_5 = 0,007; b_6 = -0,563; b_7 = -0,288 \text{ et } b_8 = -0,175.$$

L'ajout de variables explicatives améliore les statistiques d'erreur et accroît légèrement la corrélation ($r = 0,89$) entre les concentrations mesurées et estimées. Néanmoins, les résidus ne satisfont toujours pas totalement aux hypothèses théoriques. Comme le modèle s'ajuste correctement sur les données expérimentales, on décide de continuer la reconstitution, tout en sachant que l'incertitude pourra être sous-estimée.

Les concentrations horaires de NO_2 à la station Mazades en dehors des 12 semaines de mesure sont reconstituées par la relation :

$$\text{N2_MAZ} = 300,338 + 0,766 \cdot \text{NO2_BRT} + 0,027 \cdot \text{NO_BRT} - 0,072 \cdot \text{O3_BRT} + 0,383 \cdot \text{S2_BRT} + 0,007 \cdot \text{DV_COL} - 0,563 \cdot \text{VV_COL} - 0,288 \cdot \text{PR_COL} - 0,175 \cdot \text{T_COL}$$

L'estimation de la moyenne annuelle reste proche de la valeur réelle : elle est maintenant de **25,07 $\mu\text{g}/\text{m}^3$** . Son intervalle de confiance à 95 % a pour bornes [24,62 $\mu\text{g}/\text{m}^3$; 25,52 $\mu\text{g}/\text{m}^3$], ce qui représente une incertitude de 1,8 %. Cette incertitude est effectivement sous-estimée : l'intervalle de confiance ne contient

pas la vraie moyenne. Si la régression linéaire est capable d'estimer précisément une moyenne annuelle, le respect des hypothèses théoriques demeure une condition nécessaire à une évaluation fiable de l'incertitude.

► 2.2.4.6 Avantages et limites

Avantages :

- Même si des précautions d'usage s'imposent, aussi bien avant la construction du modèle (étude des corrélations avec les variables auxiliaires) qu'après l'estimation (analyse des résidus), la régression linéaire est une méthode statistique aisée à mettre en œuvre.
- Elle s'applique sans contrainte théorique sur le plan d'échantillonnage (tirage aléatoire ou systématique, grappes courtes et nombreuses ou longues et peu nombreuses, etc.).
- Pourvu que l'on trouve des variables explicatives bien corrélées avec la concentration du polluant d'intérêt, la régression permet d'estimer la moyenne annuelle avec une bonne précision.
- Le fait de reconstituer entièrement la série permet de calculer de façon immédiate différents indicateurs annuels (nombres de dépassements de seuil, centiles).

Limites :

- L'efficacité de la méthode dépend de la disponibilité des variables auxiliaires et du soin apporté à la construction du modèle (sélection des variables explicatives, contrôle des hypothèses, évaluation...).
- Le domaine de validité du modèle est fonction de la gamme des valeurs de concentration et de variables auxiliaires enregistrées pendant la campagne. (Aussi, bien que la méthode puisse être mise en œuvre quel que soit le plan d'échantillonnage, deux ou plusieurs périodes de mesure dans l'année sont préférables à une seule longue période.)
- Comme l'ajustement d'un modèle de régression tend à lisser la réalité, celui-ci ne permet pas d'estimer un nombre de dépassements d'un seuil élevé ou des centiles d'ordre élevé (qui représentent des situations à caractère exceptionnel) aussi précisément que la moyenne annuelle.
- La fiabilité de l'intervalle de confiance autour de la moyenne reconstituée est liée à la qualité du modèle ajusté et au respect des hypothèses sous-jacentes (sur les résidus).

► 2.3 Choix d'une méthode

Dans certaines situations, l'utilisateur pourra utiliser indifféremment l'une, l'autre ou les trois méthodes ; tout dépendra de son expérience et du temps dont il dispose. Dans d'autres cas, les circonstances et les données disponibles orienteront son choix, comme il est indiqué dans le Tableau 14.

Données disponibles →	station fixe de référence*	station fixe de référence*	pas de station fixe de référence	pas de station fixe de référence ni de station météorologique
Échantillonnage ↓	station météorologique	pas de station météorologique	station météorologique	
Une seule grande grappe	(Rég.)	(Rég.)	(Rég.)	
2 grandes grappes	Rég.	(Rég.)	(Rég.)	(PS-PI)
Échantillonnage aléatoire et nombre limité de grappes	Rég. (PS-PI) PS-Diff PS-Quot (ISO-PI) ISO-Diff ISO-Quot	(Rég.) (PS-PI) PS-Diff PS-Quot	(Rég.) (PS-PI) (ISO-PI)	(PS-PI)
Échantillonnage aléatoire et nombreuses grappes*	Rég. PS-PI PS-Diff PS-Quot PS-Reg ISO-PI ISO-Diff ISO-Quot	Rég. PS-PI PS-Diff PS-Quot PS-Reg	(Rég.) PS-PI ISO-PI	PS-PI
Échantillonnage systématique et nombre limité de grappes	Rég. (ISO-PI) ISO-Diff ISO-Quot	(Rég.)	(Rég.) (ISO-PI)	
Échantillonnage systématique et nombreuses grappes*	Rég. ISO-PI ISO-Diff ISO-Quot	Rég.	(Rég.) ISO-PI	

* cas où il existe une bonne corrélation entre les données du site d'échantillonnage et celles d'une station de référence.

Rég. : régression

PS : méthode des plans de sondage

ISO : méthode ISO

-PI : estimateur PI, -Diff : redressement par la différence, -Quot : redressement par le quotient, -Reg : redressement par la régression

Tableau 14 – Méthodes possibles selon le plan d'échantillonnage et les données disponibles. Une méthode est indiquée entre parenthèses si la situation permet l'utilisation de la méthode mais ne favorise pas nécessairement une bonne précision.

D'autre part, parmi les méthodes possibles, pourront être préférés :

- la régression et les estimateurs PS et ISO redressés, si l'objectif est d'atteindre une bonne précision d'estimation ;
- la régression, s'il s'agit de reconstituer divers centiles (avec les réserves qui s'imposent quant à la précision obtenue) ;
- la méthode des plans de sondage ou la méthode ISO, si l'échantillonnage se compose de nombreuses grappes et qu'une estimation fiable de l'intervalle de confiance soit recherchée. Pour un calcul précis de l'incertitude, la régression linéaire ne pourra servir que si les hypothèses d'un modèle linéaire sont bien vérifiées.

► 2.4 Conclusion

Trois méthodes statistiques ont été proposées afin d'exploiter les données d'échantillonnage et d'estimer une moyenne annuelle et son incertitude.

La méthode des plans de sondage, dans sa plus simple application, consiste à prendre la moyenne expérimentale des mesures mais pour éviter tout biais dans le résultat, la pondération des données tient compte de la stratification temporelle du plan d'échantillonnage. Lorsque de telles variables auxiliaires se présentent, la précision de l'estimation peut être sensiblement accrue à l'aide de variables bien corrélées aux concentrations et mesurées en continu sur l'année (une division par 2 ou 3 de l'incertitude peut être constatée).

La méthode ISO obéit au même principe que la méthode des plans de sondage. La différence réside dans le fait qu'à l'issue de l'échantillonnage, on procède à un nouveau regroupement des données en fonction de paramètres influents (post-stratification). Ainsi, dans le calcul de la moyenne, les données ne sont plus pondérées selon leur répartition dans les strates temporelles mais selon leur répartition dans les strates paramétriques. Comme dans la méthode des plans de sondage, l'estimation peut être corrigée à l'aide de variables auxiliaires.

La régression linéaire s'appuie entièrement sur l'existence de liens statistiques entre la concentration du polluant étudié et des variables explicatives afin de reconstituer toute la série de données et d'en déduire la moyenne annuelle.

Quelle que soit la méthode employée, son efficacité dépend du plan d'échantillonnage choisi et du fait que ce plan représente correctement l'étendue et la variabilité des concentrations et des variables auxiliaires utilisées. D'autre part, l'intervalle de confiance autour de la moyenne

reconstituée est d'autant mieux estimé que pour chaque méthode, les hypothèses théoriques sous-jacentes sont mieux vérifiées.

Il existe d'autres techniques de reconstitution non abordées dans ce guide. Les trois méthodes retenues ont l'avantage d'être compréhensibles dans leurs principes et aisément applicables et automatisables.

► Références

Directive 2008/50/CE du Parlement européen et du Conseil du 21 mai 2008 concernant la qualité de l'air ambiant et un air pur pour l'Europe.

Directive 2004/107/CE du Parlement européen et du Conseil du 15 décembre 2004 concernant l'arsenic, le cadmium, le mercure, le nickel et les hydrocarbures aromatiques polycycliques dans l'air ambiant.

Houdret J.-L., 2002, 2003, 2004. *Influence des paramètres météorologiques sur la stratégie de mesure à l'aide de moyens mobiles*. Rapports LCSQA disponibles à l'adresse www.lcsqa.org.

Houdret J.-L., Malherbe L., 2005. *Méthodes de reconstitution de moyennes et de nombres de dépassements de seuils à partir de données de campagnes*. Rapport LCSQA disponible à l'adresse www.lcsqa.org.

Lavancier F., Caïni F., Gazeau A., 2003. *Plan de sondage pour mesures mobiles de la pollution atmosphérique*. Pollution atmosphérique, N° 180, octobre-décembre.

Lebart L., Morineau A., Fenelon J.-P. *Traitement des données statistiques*, 510 pages, seconde édition, Dunod, 1982.

Norme ISO 9359:1989. Qualité de l'air - Échantillonnage stratifié pour l'estimation de la qualité de l'air ambiant.

Saporta G. *Probabilités, analyse des données et statistique*, 656 pages, 2^e édition révisée et augmentée, Technip, 2006.

Tillé Y. *Théorie des sondages. Échantillonnage et estimation en populations finies*. Cours et exercices avec solutions, 284 pages, Éditions Dunod, Paris, 2001.

Tomassone R., Audrain S., Lesquoy-Deturckheim E., Millier C. *La régression : nouveau regard sur une ancienne méthode statistique*, 190 pages, seconde édition, Masson, 1992.

Annexe 1

Bilan des pratiques
en Europe

► Bilan du questionnaire sur l'utilisation de la mesure non permanente en Europe

Un questionnaire a été envoyé durant l'été 2006 à une dizaine de correspondants européens rencontrés régulièrement lors de GT CEN. Seulement quatre réponses ont été obtenues après un délai conséquent :

- CH** Robert Gehrig (EMPA),
- D** Dieter Siegel (UBW),
- UK** Paul Quincey (NPL),
- B** Sébastien Fays (Issep).

Réponse globale en Suisse : tous les paramètres soumis à réglementation font l'objet de mesures permanentes exclusivement, à la seule exception des métaux lourds analysés dans les PM₁₀ prélevées tous les 4 jours, soit 90 échantillons journaliers par an.

► 1. Moyens utilisés pour la mesure non permanente

• Tubes à échantillonnage passif

- D =** Des mesures passives de NO₂ sont réalisées en hiver sur environ 100 sites.
- UK =** Des mesures passives sont réalisées pour le 1,3-butadiène. Des tubes de Palmes sont utilisés pour NO₂, mais pas dans un contexte de réseau National.
- B =** ?

• Autres préleveurs (pour les métaux par exemple ou les HAP)

- D =** Métaux lourds, HAP... sur filtres journaliers 180 jours par an. Des mesures de suies (soots) comme indicateur de PM par périodes de 2 semaines. À partir de janvier 2007 aux 25 sites les plus exposés, des mesures seront réalisées comme indiqué ci-après, ainsi que les PM₁₀ en quotidien.
- UK =** Seulement en sites permanents.
- B =** VOC, Heavy metals, PAH... par préleveurs isolés.

- **Moyens mobiles (camions labos, analyseurs embarqués)**

- D =** Les mesures mobiles selon une grille préétablie ont cessé depuis l'an 2000.
- UK =** Non systématique, seulement pour des campagnes limitées dans le temps
- B =** ISSeP a 5 stations semi-mobile (remorques) équipées pour : des mesures automatiques de SO₂, H₂S, NO_x, CO, CH₄, BTEX, PM_x...

- **Station fixe fonctionnant partiellement dans l'année**

- D =** ?
- UK =** Non
- B =** ?

► **2. Contexte d'utilisation de la mesure non permanente**

- D =** Les sites sont choisis en fonction de la densité de trafic et la densité de l'habitat.
- UK =** Les sites sont choisis en fonction de l'historique et la modélisation ayant montré des concentrations de fond importantes.
- B =** Le programme de mesure des polluants et prélèvement pour analyse différée en laboratoire est défini au cas par cas.

► **3. Périodes de prélèvement ou de mesure**

- **Durée des périodes**

- D =** Les périodes de mesure sont fixes et régulièrement espacées sur l'année.
Des mesures à temps partiel sont réalisées en sites fixes pour NO₂ et PM₁₀.
- UK =** 14 jours.
- B =** En général, les stations sont installées pour 2 campagnes de mesure pour une durée minimale de 6 semaines chacune.

- **Fréquence des périodes**

- D =** ?
- UK =** 14 jours ?
- B =** 2 par an.

- Réparties régulièrement ou aléatoirement dans le temps

D = ?

UK = Continues ?

B = Une campagne durant la période printemps/été et une en automne/hiver.

► 4. Estimation d'indicateurs annuels

- Moyennes annuelles

D = oui

UK = oui

B = oui

- Nombre de dépassements

D = oui pour PM₁₀

UK = oui

B = oui

- Leur incertitude est-elle estimée aussi ?

D = ?

UK = oui

B = ?

- Par quelle méthode ?

Simple calcul sur les données de l'échantillon

D = ?

UK = oui

B = Une extrapolation linéaire sur l'année des données obtenues pendant les campagnes de mesure permet d'obtenir certaines conclusions sur base annuelle.

Méthode statistique : ISO (météo), plans de sondage (redressement avec des données auxiliaires), modélisation par régression, ISO 11222 (pour l'estimation de l'incertitude)...

D = ?

UK = ?

B = ?

► 5. Reconstitution de séries temporelles

• Cherche-t-on à reconstituer des séries temporelles ?

D = oui

UK = oui, avec les données de 14 jours

B = oui et comparées ensuite aux données des réseaux fixes.

• Par quelle méthode ? Avec quelles variables ?

D = ?

UK = ?

B = ?

► 6. Statut des mesures non-permanentes

• Comment les résultats sont-ils incorporés dans le contrôle qualité ?

D = Totalement intégrées.

UK = Des essais de validation sont réalisés et les données sont redressées.

B = Une validation de tous les résultats est nécessaire avant incorporation.

• Quel est leur statut au niveau national ou européen ?

D = Pour les PM₁₀, action consécutive selon les plans et programmes appliqués aux valeurs limites.

UK = Niveau national uniquement.

B = Le rapport est publié au niveau régional wallon.

• Sont-elles totalement intégrées dans les bases de données ?

D = Yes

UK = Yes

B = Yes

• Sont-elles considérées comme estimation objective ou modélisation ?

D = ?

UK = Estimation objective.

B = ?

Annexe 2

Bilan des pratiques dans les AASQA



© AIR Pays de la Loire

I Objectif

Ce document a pour objectif de synthétiser les pratiques des AASQA en matière de plan d'échantillonnage et de reconstitution des données. Il s'appuie sur les réponses des AASQA au questionnaire du GT.

II Bilan des retours

Au 11 janvier 2007, douze AASQA ont répondu : ATMO CA, ATMO PACA, GIERSA, ASPA, LIG'AIR, AIRCOM, AIR PL, LIM'AIR, AIRAQ, ATMO AUVERGNE, ATMO PC, ORAMIP. Le taux de réponse est relativement faible par rapport au nombre d'AASQA dans le dispositif mais permet un premier bilan intéressant. Ce bilan pourra être complété ultérieurement.

III Méthode de traitement des réponses

Le traitement des réponses au questionnaire a été réalisé à partir des catégories de méthodes de mesure mises en œuvre (analyseurs automatiques, tubes passifs, préleveurs séquentiels) qui induisent des plans d'échantillonnage différents ou qui correspondent parfois à des objectifs distincts de surveillance.

IV Analyse des réponses

► Objectifs de surveillance et polluants

Analyseurs automatiques

- l'extension de la couverture à l'ensemble du territoire d'agrément et la validation de sites fixes (problématique similaire) sont des objectifs suivis par les 12 AASQA (NO₂, O₃, PM, SO₂, CO) ;
- le fonctionnement en cycles de stations urbaines est proposé par AIRPL (mais ORAMIP envisage également cette stratégie) (gaz et particules) ;
- le maintien de l'historique est mis en œuvre par ATMO PACA (gaz et particules) ;
- d'autres objectifs correspondant principalement à l'étude de phénomènes (couverture saisonnière ozone, étude ozone et brise : ATMO PACA et AIRCOM) ou des études en proximité industrielle ou de surveillance du métro (ORAMIP et ATMO PC) sont également poursuivis. Ces stratégies ne seront pas détaillées plus en avant puisqu'il s'agit principalement d'études ponctuelles, pour lesquelles la reconstitution des données n'est pas encore envisagée.

Tubes passifs

- l'objectif numéro un est la réalisation de cartographies urbaines et régionales du NO₂, des BTX et de l'O₃ (5 AASQA) ;
- le suivi réglementaire est réalisé par AIRPL et ORAMIP ;
- l'évaluation préliminaire des aldéhydes est mise en œuvre par ATMO CA ;
- des études du NO₂ en proximité industrielle sont réalisées par ATMO PC

Préleveurs séquentiels

Les espèces chimiques surveillées par ces systèmes sont le plus souvent des « nouveaux » polluants (HAP, métaux, pesticides, COV, aldéhydes, dioxines). Seuls les deux premières catégories de polluants sont réglementées dans l'air ambiant. Les objectifs poursuivis correspondent soit à de l'évaluation préliminaire, soit à un suivi réglementaire.

► Stratégie de surveillance

Élaboration de la stratégie

Ce point n'était pas directement demandé dans le questionnaire mais quelques AASQA l'ont présenté. Il ressort que cette élaboration découle :

- des Directives Européennes (période minimale de prise en compte) : AIRPL, GIERSA ;
- du guide JRC de l'Agence Européenne : ASPA ;
- de calculs statistiques sur l'historique existant : ATMO PACA.

Plans d'échantillonnage et couverture annuelle

Analyseurs automatiques :

On peut distinguer deux types d'approche :

- un échantillonnage saisonnier : soit été et hiver (ATMO CA, ATMO PACA, ASPA, AIRCOM, AIRPL, AIRAQ, ORAMIP) ou bien une campagne par trimestre (GIERSA, LIG'AIR) ;
- un échantillonnage plus régulier : 8 x 2 semaines (ASPA), 8 à 9 campagnes de 2 semaines sur l'année (ATMO PC).

La première catégorie est moins contraignante pour une mise en œuvre opérationnelle et moins coûteuse en unités d'œuvre (ATMO CA, AIRPL). AIRAQ met ainsi en œuvre des campagnes d'une durée minimale de 4 semaines.

Pour cette catégorie de méthodes de mesure, la couverture temporelle annuelle est variable et inférieure ou égale à 50 % : 8 %, 16 %, 30 %, 40 %, 50 %.

Tubes passifs :

Les plans d'échantillonnage sont là aussi variables :

- approche en continu : LIG'AIR ;

- approche saisonnière : été et hiver (ATMO CA, AIRCOM, ORAMIP), par trimestre (ATMO CA, AIRPL) ;
- approche plus régulière : 6 x 2 semaines, 1 semaine par mois entre avril et septembre – suivi ozone (GIERSA), 8 à 9 campagnes de 2 semaines sur l'année – suivi NO₂ (ATMO PC).

La couverture temporelle annuelle est variable et inférieure à 40 % : 8 %, 11 %, 15 %, 23 %, 30 %, 38 %.

Préleveurs séquentiels :

On distingue :

- un suivi réglementaire en continu des métaux en site industriel : AIRPL ;
- un échantillonnage régulier : 8 x 2 semaines ou 2 x 8 semaines – COV/aldéhydes (ASPA), 1 mois sur 2 – métaux en site urbain (AIRPL) ;
- des études ponctuelles de quelques semaines à quelques mois (pesticides, dioxines) : ATMO CA, AIRPL, AIRCOM, ORAMIP.

La couverture temporelle est là aussi inférieure à 50 % mais est plus importante que pour les autres techniques : 25 %, 30 %, 50 %.

Reconstitution des données

Les AASQA qui mettent en œuvre des méthodes de reconstitution :

- AIRPL (reconstitution par régression avec pollution – NO₂, O₃ et météo – VV, DV, HR, T, pluvio) ;
- ATMO PC et GIERSA (plans de sondage) ;
- ATMO PACA (reconstitution de moyennes annuelles par redressement à partir d'une station fixe) ;
- AIRCOM (réseaux de neurones).

L'AASQA qui utilise des données de comparaison :

- LIG'AIR (écart stations fixes-tubes, en réflexion pour la mise en œuvre de méthodes de reconstitution) ;

Les AASQA qui ne pratiquent pas de reconstitution :

- ATMO CA, ASPA, LIM'AIR, AIRAQ, ATMO AUVERGNE, ORAMIP.

V Conclusions

- les objectifs prioritaires : extension de la couverture par analyseurs automatiques, suivi réglementaire, cartographies, fonctionnement en cycles de stations automatiques ;
- les objectifs secondaires : maintien de l'historique, études de phénomènes, campagnes ou évaluation préliminaire « nouveaux » polluants ;
- les polluants prioritaires : NO_x, O₃, PM₁₀, benzène, PM_{2,5}, métaux, HAP ;
- les polluants secondaires : SO₂, CO, aldéhydes, pesticides, dioxines ;
- les plans d'échantillonnage prioritaires : campagnes été-hiver ou par saison, approche plus régulière ;
- les plans d'échantillonnage secondaires : études ponctuelles non réparties sur l'année
- méthodes de reconstitution prioritaires : régression, plans de sondage, méthode ISO (non citée dans les réponses).

Résultat du traitement des réponses par catégories de méthodes de mesure

Analyseurs automatiques

Méthode de mesure	Objectif de surveillance	Polluants	Plan d'échantillonnage	Couverture annuelle	
Analyseurs automatiques	Couverture territoire, validation site fixe	NO ₂ , O ₃ , PM, SO ₂ , (CO ?)	1 mois été + 1 mois hiver	16 %	
			1-2 mois été et/ou 1-2 mois hiver	16 % - 33 %	
			2 semaines par trimestre	15 %	
			8 x 2 semaines ou 2 x 8 semaines (été-hiver)	30 %	
			3 semaines par trimestre	23 %	
			15 à 60 jours en hiver et en été	8 % - 32 %	
			3 mois été et 2 mois hiver (fond)	41 %	
			En continu sur l'année (trafic)	100 %	
			8 à 9 campagnes de 2 semaines sur l'année	30 %	
			3 semaines en été et en hiver	11 %	
			? empirique		
			4 campagnes sur l'année couvrant hiver et été	?	
			? empirique		
	Fonctionnement en cycles de stations		1 an sur chaque site urbain en alternance – tests sur des cycles + courts	100 % sur un an – 50 % sur 2 ans	
	Maintien historique	NO ₂ , O ₃	1-2 mois été et/ou 1-2 mois hiver	16 % - 33 %	
	Proximité industrielle	En fonction des émissions du site	Une campagne de 2 à 6 mois	16 à 50 %	
		Polluants réglementés + spécifiques	3 semaines en été et en hiver	11 %	
	Surveillance du métro	Polluants réglementés + spécifiques	2 fois 15 jours par an	5 %	
	Couverture saisonnière	O ₃	6 mois d'avril à septembre	50 %	
	Reconstitution spatiale	O ₃	2 mois par an	16 %	
	Étude de phénomènes	O ₃	4 à 5 mois	33 %	

	Répartition sur l'année	Reconstitution	AASQA
	oui	non	ATMO CA
	oui	Si corrélation avec station fixe auxiliaire, reconstitution d'une moyenne annuelle par redressement	ATMO PACA
	oui	Si écart / station de référence > 10 %, redressement par plans de sondage	GIERSA
	oui	non	ASPA
	oui	Calcul envisagé des percentiles	LIG'AIR
	oui	Étude de corrélation/régression avec les stations fixes et/ou reconstitution en fonction de la rose des vents	AIRCOM
	oui	Non : évaluation des risques de dépassement par comparaison avec d'autres agglomérations	AIRPL
		inutile	AIRPL
	oui	Reconstitution par plans de sondage pour la moyenne et les percentiles	ATMOPC
	oui	non	ORAMIP
	?	non	LIM'AIR
	oui	non	AIRAQ
	?	non	ATMO AUVERGNE
		Reconstitution par régression avec pollution et météo (VV, DV, HR)	AIRPL
	oui	Si corrélation avec station fixe auxiliaire, reconstitution d'une moyenne annuelle par redressement	ATMO PACA
	non	non	ATMO PC
	oui	non	ORAMIP
	non	non	ORAMIP
	non	non	ATMO PACA
	non	non	AIRCOM
	non	non	AIRCOM

Tubes passifs

Méthode de mesure	Objectif de surveillance	Polluants	Plan d'échantillonnage	Couverture annuelle	
Tubes passifs	Cartographie urbaine ou régionale	NO ₂ , BTX, O ₃	NO ₂ , BTX : 4 x 2 semaines en été et en hiver	30 %	
			NO ₂ : 6 x 2 semaines sur l'année O ₃ : 1 semaine par mois entre avril et septembre	23 % 11 %	
			NO ₂ , BTX : en continu avec des périodes de 1 mois	100 %	
			NO ₂ : 8 à 9 campagnes de 2 semaines sur l'année	30 %	
			BTX : 15 jours en été et en hiver	8 %	
	Évaluation préliminaire	Aldéhydes	4 x 2 semaines sur l'année	15 %	
	Suivi réglementaire annuel	BTX	4 x 1 mois en urbain 4 x 3 semaines ou 4 x 5 semaines en trafic	33 % 23 % 38 %	
		NO ₂ , BTX	Suivi été/hiver Plusieurs périodes de 15 jours (de 3 à 7 selon budget)	12 à 28 %	
	Proximité industrielle	En fonction des émissions du site	Une campagne de 2 à 6 mois	16 à 50 %	
Préleveurs séquentiels (partisols, DA 80, canisters...)	Surveillance par campagnes	HAP	2 périodes dans l'année	?	
		Pesticides	En fonction des traitements	?	
		Métaux	2 x 2 mois sur l'année (2 saisons)	33 %	
		Pesticides	Prélèvements journaliers ou hebdomadaires pendant les périodes de traitement		
		COV, aldéhydes	8 x 2 semaines ou 2 x 8 semaines (été-hiver)	30 %	
		Pesticides, HAP, dioxines, métaux	3 à 6 mois de mesure	25 % 50 %	
		Métaux	1 mois sur 2 en site urbain 52 semaines sur sites industriels	50 % 100 %	

	Répartition sur l'année	Reconstitution	AASQA
	oui	non	ATMO CA
	oui	Non : moyenne des périodes ; erreur relative contrôlée sur sites fixes avant les campagnes	GIERSA
		Non : calcul de la moyenne annuelle : incertitude estimée par calcul de l'écart entre tubes et stations fixes	LIG'AIR
	oui	Reconstitution par plans de sondage pour la moyenne	ATMO PC
	oui		AIRCOM
	oui	non	ATMO CA
	oui	Reconstitution par régression avec pollution (NO _x) et météo (VV, T)	AIRPL
	oui	non	ORAMIP
	non	non	ATMO PC
	oui	non	ATMO CA
	?	non	ATMO CA
	oui	non	ATMO CA
	non	non	ORAMIP
	oui	non	ASPA
	non	non	AIRCOM
	oui	Reconstitution par régression avec pollution (O ₃ , NO ₂) et météo (Pluvio)	AIRPL

Annexe 3

Fiche descriptive
de la méthode
des plans de sondage

Méthode des plans de sondages (PS)

Description
Cette méthode statistique, fondée sur la théorie des sondages, comprend deux aspects : • la planification de la collecte des données en fonction de la précision d'estimation souhaitée ; • l'estimation des caractères d'intérêt et de leur incertitude par un ou plusieurs estimateurs. La mise en œuvre du premier point conduit à la définition d'un plan d'échantillonnage aléatoire stratifié dans le temps. À cause des contraintes de mesure, il s'agit généralement d'un plan par grappes. Le second point concerne la reconstitution proprement dite, après que l'échantillon de données a été collecté. L'opération est la suivante : • estimation des moyennes ou des nombres de dépassements de seuil dans les strates temporelles ; • pondération de ces estimations par la fréquence d'occurrence de chaque strate et reconstitution des moyennes annuelles ou des nombres annuels de dépassements. <i>Type d'échantillonnage requis</i> : échantillonnage aléatoire stratifié constitué de nombreuses mesures indépendantes ou de multiples grappes de mesures.

Indicateur recherché		Pas de temps maximal des mesures
☒	Moyenne annuelle	1 mois
☒	Incertitude associée	
☒	Nombre de dépassements d'une moyenne horaire	1 heure
~	Incertitude associée	
☒	Nombre de dépassements d'une moyenne journalière	1 jour
~	Incertitude associée	
☒	Nombre de dépassements d'une moyenne sur 8 heures (glissante ou non)	1 heure (la taille minimale des grappes est 8 heures)
~	Incertitude associée	
☐	Reconstitution d'une série chronologique	
☐	Incertitude associée	

Variables complémentaires

Pour une simple reconstitution : **pas de variable complémentaire requise**

Pour effectuer un redressement sur les indicateurs estimés : la mesure du même polluant sur un site permanent continu (une station fixe par exemple). Le pas de temps des données explicatives doit être un diviseur de celui de la mesure.

Estimation d'une incertitude d'estimation

L'estimation d'indicateurs (moyenne, nombre de dépassements) sur une série temporelle incomplète induit une incertitude. La méthode PS fournit une estimation de l'incertitude liée au plan de sondage sous la forme d'intervalles de confiance à 95 %. Ceux-ci sont d'autant mieux estimés que le nombre de grappes est plus grand (au minimum une dizaine de grappes environ). La fiabilité des intervalles de confiance peut être contrôlée en simulant un grand nombre de fois le plan de sondage sur une station fixe annexe.

Outil nécessaire

Excel ☐	Logiciel statistique : R (freeware) disponible à l'adresse http://cran.cict.fr	Autres :
--------------------------------	---	----------

Références bibliographiques

- Lavancier F., Caïni F., Gazeau A., *Plan de sondage pour mesures mobiles de la pollution atmosphérique*, Pollution atmosphérique, N° 180, octobre-décembre, 2003.
- Tillé Yves, *Théorie des sondages. Échantillonnage et estimation en populations finies*, Cours et exercices, Dunod, Paris, 2001.
- Dreesbeke J.-J., Fichet B., Tassi P., *Les Sondages*, Economica 1987.
- Houdret J.-L., Malherbe L., *Méthodes de reconstitution temporelle de moyennes et de dépassements de seuils à partir de données de campagnes*, Rapport LCSQA, 2005. www.lcsqa.org

Annexe 4

Fiche descriptive
de la méthode issue
de la norme « ISO »

Méthode ISO (issue de la norme ISO 9359)

Description
La méthode ISO s'appuie sur la norme « <i>ISO 9359 – Qualité de l'air – Échantillonnage aléatoire stratifié pour l'évaluation de la qualité de l'air ambiant</i> » qui considère l'influence des variables météorologiques et/ou temporelles sur les concentrations de polluants. Son application comprend plusieurs étapes : a) Sur des séries de données issues d'un ou plusieurs sites de référence : • Étude des influences des paramètres météorologiques et temporels sur les niveaux de pollution mesurés ; en général, ces influences peuvent être considérées comme reproductibles d'une année sur l'autre et représentatives d'une zone géographique relativement vaste. • Définition de strates paramétriques : chaque strate paramétrique correspond à une ou plusieurs combinaisons de valeurs de paramètres concourant à des situations de pollution homogènes. b) Sur les données d'échantillonnage : • Répartition des données d'échantillonnage dans les strates paramétriques selon les valeurs correspondantes des paramètres influents ; • Estimation des moyennes ou des nombres de dépassements de seuil dans les strates paramétriques ; • Pondération de ces estimations par la fréquence d'occurrence de chaque strate et reconstitution des moyennes annuelles ou des nombres annuels de dépassements. <i>Type d'échantillonnage requis</i> : échantillonnage aléatoire ou systématique constitué de nombreuses mesures indépendantes ou de plusieurs grappes de mesure réparties sur l'année.

Indicateur recherché		Pas de temps maximal des mesures
☒	Moyenne annuelle	1 mois
☒	Incertitude associée	
☒	Nombre de dépassements d'une moyenne horaire	1 heure
~	Incertitude associée	
☒	Nombre de dépassements d'une moyenne journalière	1 jour
~	Incertitude associée	
☒	Nombre de dépassements d'une moyenne sur 8 heures (glissante ou non)	1 heure (la taille minimale des grappes est 8 heures)
~	Incertitude associée	
☐	Reconstitution d'une série chronologique	
☐	Incertitude associée	

Variables complémentaires

Données météorologiques permettant la définition des strates

paramétriques (données relatives aux vents et à la température lorsque la mesure est horaire, ensoleillement total sur une journée, etc.). Le pas de temps des données individuelles doit être égal à celui de la mesure.

Pour effectuer un redressement sur les indicateurs estimés : la mesure du même polluant sur un site permanent continu (une station fixe par exemple). Le pas de temps des données auxiliaires doit être un diviseur de celui de la mesure.

Incertitude d'estimation

L'estimation d'indicateurs (moyenne, nombre de dépassements) sur une série temporelle incomplète induit une incertitude. La méthode ISO fournit une **estimation** de cette incertitude sous la forme d'intervalles de confiance à 95 %. Ceux-ci sont d'autant mieux estimés que le nombre de mesurages par strate est plus grand et que l'estimation prend en compte l'éventuelle dépendance temporelle entre les données. La fiabilité des intervalles de confiance peut être contrôlée en simulant un grand nombre de fois l'échantillonnage sur une station fixe annexe.

Outil nécessaire

Excel ☒ Disponible sur le site LCSQA	Logiciel statistique : R (freeware) disponible à l'adresse http://cran.cict.fr	Autres :
---	---	----------

Références bibliographiques

- Norme ISO 9359 – Qualité de l'air – Échantillonnage aléatoire stratifié pour l'évaluation de la qualité de l'air ambiant.
- Houdret J.-L., Malherbe L., Méthodes de reconstitution temporelle de moyennes et de dépassements de seuils à partir de données de campagnes, Rapport LCSQA, 2005. www.lcsqa.org

Annexe 5

Fiche descriptive
de la régression

Reconstitution par modèle de régression

Description
Le modèle de régression consiste à déterminer, au moyen d'un échantillon d'apprentissage, la relation qui existe entre une variable « à expliquer » Y et n variables « explicatives » ($X_1, X_2, \dots, X_n$) qui peuvent être les concentrations d'autres polluants ou des paramètres météorologiques. Dans le cas de la mesure non permanente, l'échantillon d'apprentissage est constitué par les données recueillies au cours des périodes de mesure. Il existe une multitude de modèles. Dans le cas d'un modèle linéaire, la relation s'écrit : $Y = a_1.X_1 + a_2.X_2 + \dots + a_n.X_n + \epsilon$ où les coefficients ($a_1, a_2, \dots, a_n$) sont calculés sur l'échantillon d'apprentissage par la méthode des moindres carrés. En utilisant cette relation, il est ensuite possible de prédire, à partir des valeurs des variables explicatives ($X_1, X_2, \dots, X_n$), les valeurs individuelles de la variable Y sur les périodes pendant lesquelles elle n'a pas été mesurée, et ainsi, de reconstituer différents indicateurs annuels. <i>Hypothèses à vérifier :</i> • Indépendance des variables explicatives ; • Indépendance et normalité des résidus ; • Homoscédasticité (la variance des résidus est la même pour toutes les valeurs des variables explicatives) ; • Diagnostic du modèle (coefficient de détermination). <i>Type d'échantillonnage requis :</i> n'importe quel type d'échantillonnage. Cependant, pour une reconstitution plus précise, un minimum de deux périodes de mesure est préférable à une unique période plus longue.

Indicateur recherché		Pas de temps maximal des mesures
☒	Moyenne annuelle	1 mois
☒	Incertitude associée	
☒	Nombre de dépassements d'une moyenne horaire	1 heure
~	Incertitude associée	
☒	Nombre de dépassements d'une moyenne journalière	1 jour
~	Incertitude associée	
☒	Nombre de dépassements d'une moyenne sur 8 heures (glissante ou non)	1 heure (la taille minimale des grappes est 8 heures)
~	Incertitude associée	
☒	Reconstitution d'une série chronologique	Identique à celui de la série à reconstituer
☒	Incertitude associée	

Variables complémentaires

Au moins une variable bien corrélée avec la variable à expliquer (il reste néanmoins préférable de disposer de plusieurs variables pour la construction du modèle).

Le pas de temps des mesures individuelles correspondant aux variables complémentaires doit être un diviseur du pas de temps de la variable à expliquer.

Estimation d'une incertitude d'estimation

L'estimation donnée par le modèle peut être accompagnée d'un intervalle de confiance.

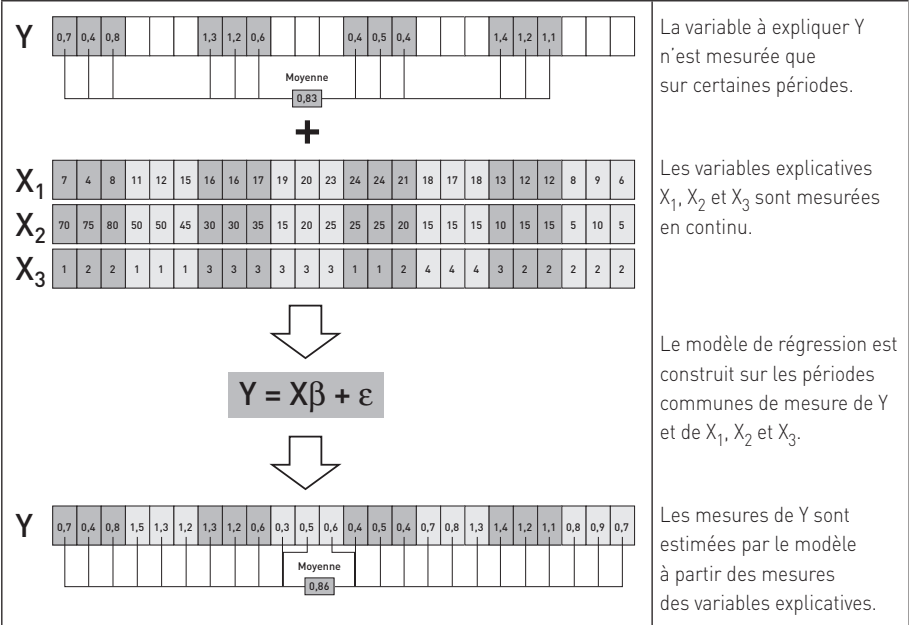
Outil nécessaire

Excel	Logiciel statistique R (freeware disponible à l'adresse http://cran.cict.fr)	Autres :
-------	---	----------

Références bibliographiques

- Moilime D., Penven F., *Reconstitution de données. Mise en œuvre de modèle de régression dans le cadre de la mesure cyclique*, Rapport AIRPL, 2005.

Principe



Annexe 6

Note théorique
sur la corrélation

► Le coefficient de corrélation

On dispose de deux ensembles de valeurs conjointes (des concentrations concomitantes) : $\{x_i\}_{1 \leq i \leq n}$ et $\{y_i\}_{1 \leq i \leq n}$, que l'on considère comme les observations de deux variables aléatoires X et Y dont les réalisations x_i et y_i sont les concentrations **mesurées aux mêmes instants i** .

Étudier la corrélation entre deux variables aléatoires (ou statistiques), c'est étudier l'**intensité** de la liaison qui peut exister entre ces variables. La liaison recherchée est une relation **affine**.

1. Étude graphique de la corrélation

Afin d'examiner s'il existe une liaison entre X et Y on représente chaque observation i comme un point de coordonnées $\{x_i, y_i\}$ dans un repère cartésien. La forme du nuage de points ainsi tracé est fondamentale pour la suite : ainsi la figure 1 montre :

- a) une absence de liaison ;
- b) une absence de liaison en moyenne mais pas en dispersion ;
- c) une corrélation linéaire positive ;
- d) une corrélation non linéaire.

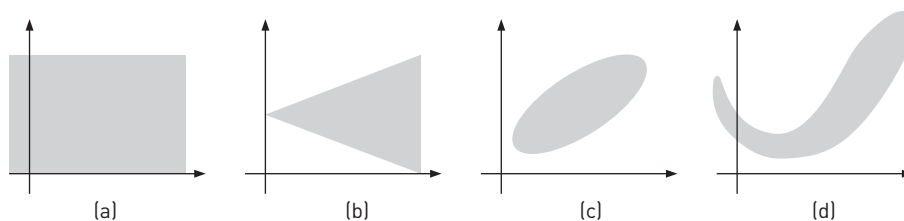


Figure 1

On dit qu'il y a corrélation s'il y a dépendance en moyenne : à $X = x$ fixé, la moyenne $\bar{Y}$ est fonction de x . Si cette liaison est approximativement linéaire on se trouve dans le cas de la corrélation linéaire.

2. Le coefficient de corrélation linéaire

Ce coefficient mesure **exclusivement** le caractère plus ou moins **linéaire** du nuage de points, autrement dit la qualité de la relation linéaire ou le degré de dépendance linéaire entre les deux variables.

► **A. Définition**

$$r = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

où s_x et s_y sont les écarts-types de x et y :

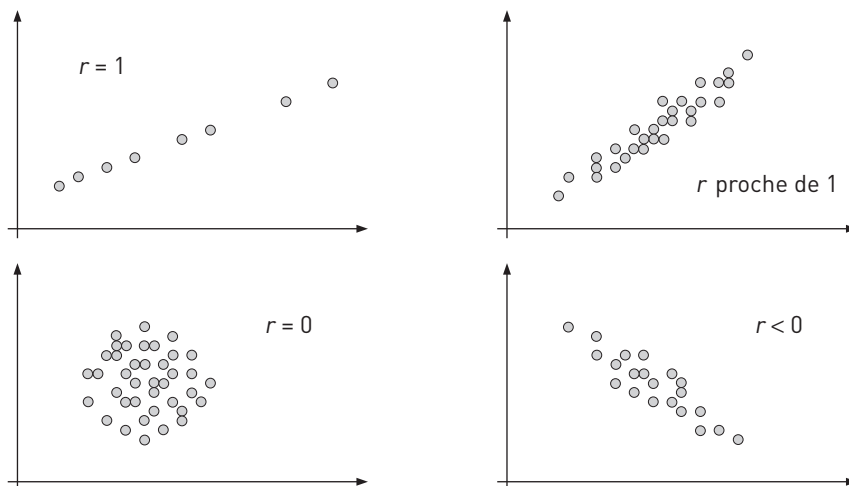
$$s_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad s_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$$

Remarque : le numérateur $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ est la covariance observée.

Propriétés :

- $-1 \leq r \leq 1$ (ou bien $|r| \leq 1$)
- r est un indicateur sans dimension ; il n'est pas sensible aux unités de chacune des variables. Ainsi par exemple, le coefficient de corrélation linéaire entre 2 mesures sera identique que l'une des mesures soit exprimée en ppb ou en $\mu\text{g}/\text{m}^3$.
- $|r| = 1$ est équivalent à l'existence d'une relation linéaire exacte (tous les points sont alignés sur une droite) : $ax_i + by_i + c = 0 \quad \forall i$.
- X et Y indépendantes $\Rightarrow r = 0$, mais la réciproque n'est pas vraie en général : **la non corrélation n'est pas nécessairement l'indépendance**, alors que l'indépendance entraîne forcément une non corrélation.
- Notons pour finir que la corrélation n'est pas transitive : X très corrélée avec Y , Y très corrélée avec Z , n'implique nullement que X soit corrélée avec Z .

Exemples :



► B. Du bon usage du coefficient r

Attention, il est toujours possible de calculer un coefficient de corrélation (sauf cas très particulier) mais un tel coefficient de corrélation n'arrive pas toujours à rendre compte de la relation qui existe en réalité entre les variables étudiées. En effet, il suppose que l'on essaye de juger de l'existence d'une relation linéaire entre nos variables. Il n'est donc pas adapté pour juger de corrélations qui ne seraient pas linéaires et non linéarisables. Il perd également de son intérêt lorsque les données étudiées sont très hétérogènes puisqu'il représente **une relation moyenne** et que l'on sait que la moyenne n'a pas toujours un sens, notamment si la distribution des données est multi modale ou très asymétrique (c'est typiquement le cas pour la répartition du SO_2 par exemple).

Le coefficient r ne mesure que le caractère linéaire d'une liaison et son usage doit être réservé à des nuages où les points sont répartis de part et d'autre d'une tendance linéaire (figure 1.c).

La figure 2 montre les risques d'un usage inconsidéré du coefficient de corrélation linéaire r . On notera en particulier que r est très sensible à la présence de valeurs extrêmes et/ou aberrantes (valeurs très éloignées de la majorité des autres, pouvant être considérées comme des exceptions), et n'est donc pas « robuste ».

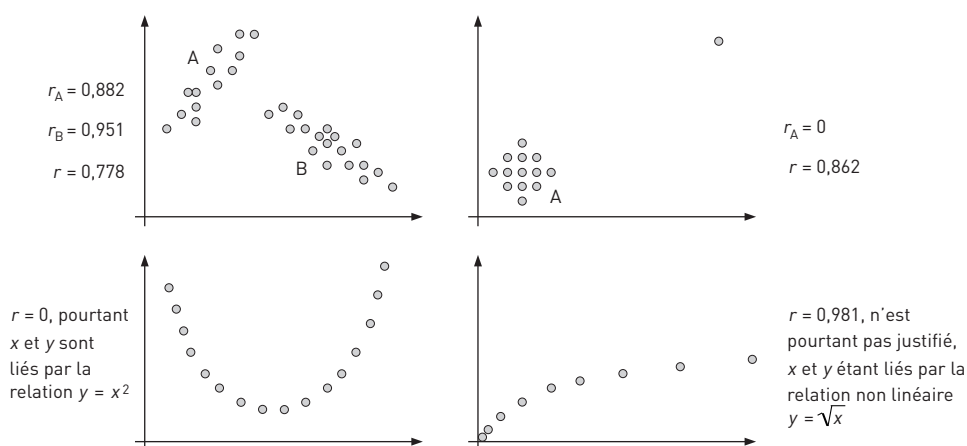


Figure 2

Dans l'exemple suivant, les 4 nuages de la figure 3 ont mêmes moyennes, mêmes variances et même coefficient de corrélation :

$$\begin{aligned} \bar{x} &= 9 & \bar{y} &= 7,5 \\ s_x^2 &= 10,0 & s_y^2 &= 3,75 \\ r &= 0,82 \end{aligned}$$

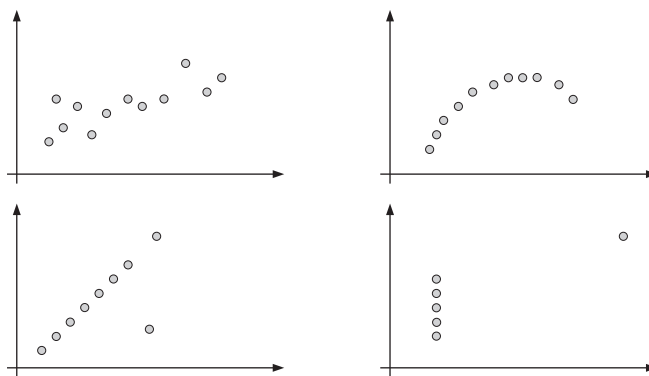


Figure 3

Seul le premier nuage justifie l'usage de r .

D'une manière générale, l'étude de la relation entre des variables, quelles qu'elles soient, doit s'accompagner de graphiques descriptifs, exhaustifs ou non dans l'appréhension des données à notre disposition, pour éviter de subir les limites purement techniques des calculs que nous utilisons.

3. Caractère significatif d'un coefficient de corrélation

En admettant que l'on se trouve dans le cas où l'usage de r est justifié, à partir de quelle valeur la liaison est-elle significative ?

On se fixe un risque acceptable α (en général 5 %)

et on calcule la quantité r_0 suivante :
$$r_0 = \frac{t}{\sqrt{n - 2 + t^2}}$$

où n est le nombre d'observations et t le fractile d'ordre $\alpha/2$ de la loi de Student à $n - 2$ degrés de liberté (donné par la table *ad hoc*).

On rejette l'hypothèse que le coefficient de corrélation est nul, au risque α de se tromper, dès que $|r| \geq r_0$. On pourra alors dire, avec une probabilité α de se tromper, que le coefficient de corrélation est significativement différent de 0.

Par exemple, au risque de 5 % on déclarera qu'une liaison est significative sur un échantillon de 30 observations si $|r| > 0,36$ ($t = 2,048$).

Attention : on remarquera que le seuil de signification (r_0) décroît quand n croît ; **pour les grands échantillons, le fait de trouver que r diffère significativement de 0 ne garantit nullement que la liaison soit forte**. De plus, rappelons que *le fait de trouver que r n'est pas significativement différent de 0 n'entraîne pas nécessairement l'indépendance*.

On peut aussi déterminer un **intervalle de confiance** sur le coefficient de corrélation (afin de quantifier la qualité de la régression) grâce à l'introduction de la transformation suivante :

$$z_r = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) \quad \text{et} \quad r = \frac{e^{2z_r} - 1}{e^{2z_r} + 1}$$

L'intervalle de confiance est défini par :

$$P \left(z_r - z_{\alpha/2} \sqrt{\frac{1}{n-3}} < Z < z_r + z_{\alpha/2} \sqrt{\frac{1}{n-3}} \right) = 1 - \alpha$$

où $z_{\alpha/2}$ est tel que $P(Y < z_{\alpha/2}) = 1 - \alpha/2$ avec Y une variable aléatoire normale centrée-réduite.

Application : en clair, pour un coefficient de corrélation calculé r :

- on se fixe un risque acceptable, en général $\alpha = 0,05$ (5 %) ;
- on détermine $z_{\alpha/2}$ avec la table de la loi normale centrée-réduite, en général $z_{\alpha/2} = 1,96$;
- on calcule z_r à partir de la valeur r : $z_r = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right)$

- on calcule les bornes de l'intervalle de confiance sur Z , soit :

$$z_{inf} = z_r - z_{\alpha/2} \sqrt{\frac{1}{n-3}} \quad \text{et} \quad z_{sup} = z_r + z_{\alpha/2} \sqrt{\frac{1}{n-3}} ,$$

- puis celles de l'intervalle de confiance sur R , à savoir :

$$r_{inf} = \frac{e^{2z_{inf}} - 1}{e^{2z_{inf}} + 1} \quad \text{et} \quad r_{sup} = \frac{e^{2z_{sup}} - 1}{e^{2z_{sup}} + 1} .$$

=> Le coefficient de corrélation, estimé par r , appartient à l'intervalle $[r_{inf} ; r_{sup}]$ avec une probabilité de $1 - \alpha$.

Exemple : Soit $r = 0,54$ obtenu sur un échantillon de taille $n = 69$. On souhaite construire l'intervalle de confiance à 99 % autour de cette valeur. On obtient successivement $z_r = 0,604$. Dans la table de la loi normale, on lit $z_{0,005} = 2,575$ et donc $P(0,293 < Z < 0,927) = 0,99$. Par inversion, on obtient l'intervalle de confiance sur l'estimation du coefficient de corrélation : $P(0,285 < R < 0,729) = 0,99$.

4. Corrélation partielle *(pour information)*

Une erreur courante est de croire qu'un coefficient de corrélation élevé induit une relation de cause à effet entre les deux phénomènes mesurés. En réalité, les deux phénomènes peuvent être corrélés à un même « phénomène-source » : une troisième variable non mesurée, et dont dépendent les deux autres. Autrement dit, il arrive fréquemment que la dépendance apparente entre deux variables soit due en réalité aux variations d'une 3^e variable.

La littérature statistique abonde en exemples de fausses corrélations surprenantes entre phénomènes variés qui disparaissent lorsque l'on fixe une 3^e variable (souvent non aléatoire comme le temps); ainsi de la corrélation entre le nombre de maladies mentales déclarées chaque année et le nombre de postes de radio installés, ou encore le nombre de coups de soleil observés dans une station balnéaire, qui est fortement corrélé au nombre de lunettes de soleil vendues; pourtant aucun des deux phénomènes n'est bien sûr la cause de l'autre... On cite souvent aussi l'exemple plus morbide de la consommation de pétrole et de la mortalité des personnes âgées.

Les coefficients de corrélation partielle constituent un moyen d'éliminer l'influence d'une ou plusieurs variables.

Le coefficient de corrélation partielle entre X et Y conditionnellement

à une 3^e variable Z est donné par :
$$r_{XY|Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$$

où r_{AB} est le coefficient de corrélation entre les variables A et B.

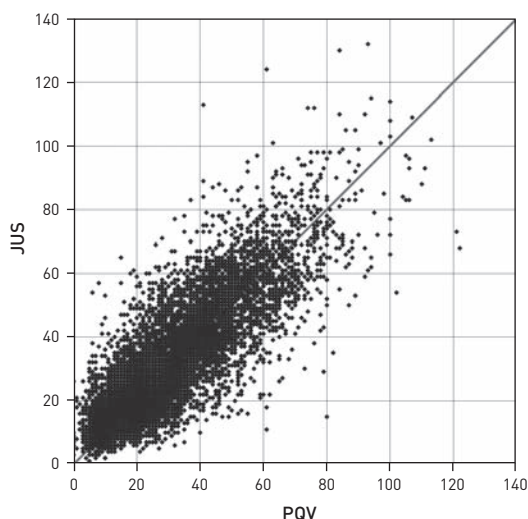
5. Problèmes d'interprétation :

corrélation entre deux stations urbaines pour le NO₂

Un coefficient qui peut sembler plutôt élevé (ici 83 %) peut conduire à des défauts d'interprétation...

En fait, il y a un assez grand nombre de points éloignés de la bissectrice, et notamment dans les concentrations les plus fortes, il n'y a **aucun** point sur la bissectrice. En fait, si on s'intéresse aux valeurs supérieures à 40, le coefficient de corrélation tombe à 0,56 et si on ne garde que les valeurs > 80, il tombe à 0,066...

Notons aussi que dans de nombreux cas, la différence de concentrations peut aller « du simple au double » : quand PQV enregistre 60 µg/m³, JUS enregistre dans le même temps une valeur pouvant aller de 20 à 100 µg/m³...



Annexe 7

Définition de l'intervalle de confiance

► Incertitude et intervalle de confiance

L'incertitude doit être comprise comme un paramètre permettant de dire quelles sont les valeurs possibles de la concentration réelle autour de la concentration annoncée.

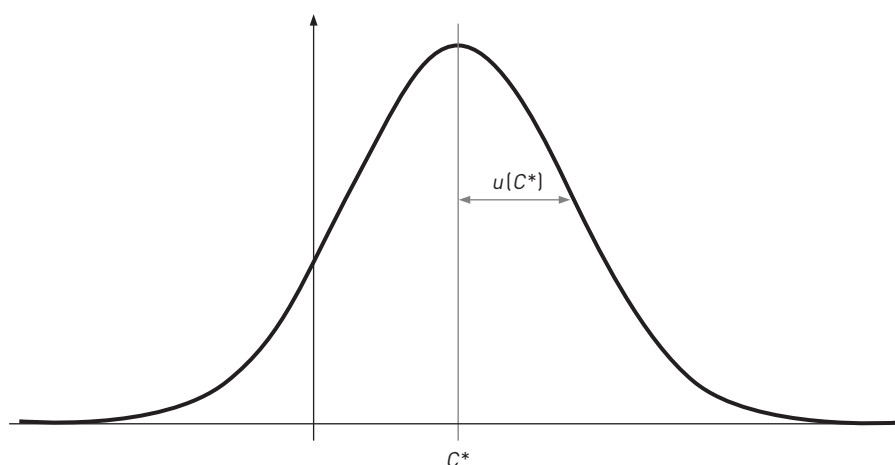
Elle est généralement traduite sous la forme d'un **intervalle de confiance**.

Un intervalle de confiance est un intervalle supposé contenir, avec un certain degré de confiance, la valeur à estimer. Par exemple, un intervalle de confiance à 95 % (ou au seuil de risque de 5 %) a 95 % de chances de contenir la valeur du paramètre que l'on cherche à estimer ; autrement dit, la probabilité pour que cet intervalle contienne la valeur à estimer est de 0,95.

Exemple :

Soit C une concentration moyenne annuelle à estimer. Soient C^* la valeur estimée de C et $u(C^*)$ son écart-type associé.

En admettant que les valeurs possibles de C se distribuent selon la loi normale de moyenne C^* et d'écart-type $u(C^*)$:



alors l'intervalle $[C^* - 2.u(C^*); C^* + 2.u(C^*)]$ est un intervalle de confiance à environ 95 %.

Ex : Si $C^* = 80 \mu\text{g}/\text{m}^3$ et $u(C^*) = 5 \mu\text{g}/\text{m}^3$, il y aura 95 % de chances pour que la vraie valeur appartienne à l'intervalle $[70 \mu\text{g}/\text{m}^3; 90 \mu\text{g}/\text{m}^3]$, avec peu de chances pour qu'elle se trouve proche des extrêmes et plus de chances pour qu'elle soit proche du milieu, c'est-à-dire $80 \mu\text{g}/\text{m}^3$.

Annexe 8

Formulaire

► Formulaire

► Notations

Individu : concentration quart-horaire, horaire, journalière, etc. selon le pas de temps de la base

Grappe : groupe de données généralement consécutives (ex : mesures horaires sur une semaine, un jour, etc.)

Strate : sous-ensemble de la période d'étude au sein duquel le polluant se comporte de façon plus homogène que sur la période entière

N : nombre total d'individus sur la période d'étude

S : ensemble des strates sur la période

N_s : nombre total d'individus dans la strate s

m_s : nombre de grappes prélevées dans la strate s

$\hat{X}_0$: estimateur de la moyenne de X_0 sur la période d'étude

$\hat{\text{var}}(\hat{X}_0)$: variance estimée de $\hat{X}_0$

$\hat{X}_{s,0}$: estimateur de la moyenne de X_0 sur la strate s

$\bar{X}_{g,0}$: moyenne de la grappe g dans la strate s

p_g : poids de la grappe g dans la strate s

n_g : nombre d'individus composant la grappe g dans la strate s

$X_0(t_{g,i})$: valeur prise par l'individu i dans la grappe g

Plan de sondage par grappes avec stratification

► Sans redressement

$$\hat{\bar{X}}_0 = \sum_s \frac{N_s}{N} \hat{\bar{X}}_{s,0}$$

$$\text{avec} \left[\begin{array}{l} \hat{\bar{X}}_{s,0} = \sum_{g=1}^{m_s} p_g \bar{X}_{g,0} \\ \sum_{g=1}^{m_s} p_g = 1 \\ \forall g \bar{X}_{g,0} = \frac{1}{n_g} \sum_i X_0(t_{g,i}) \end{array} \right.$$

population infinie

$$\hat{var}[\hat{\bar{X}}_0] = \sum_s \left(\left(\frac{N_s}{N} \right)^2 \frac{m_s \sum_{g=1}^{m_s} p_g^2}{m_s + m_s \sum_{g=1}^{m_s} p_g^2 - 2} \sum_{g=1}^{m_s} (\bar{X}_{g,0} - \hat{\bar{X}}_{s,0})^2 \right)$$

population finie

$$\hat{var}[\hat{\bar{X}}_0] = \sum_s \left(\left(\frac{N_s}{N} \right)^2 \frac{m_s}{M_s} \frac{M_s - m_s}{M_s - 1} \sum_{g=1}^{m_s} \left(p_g \bar{X}_{g,0} - \frac{1}{m_s} \hat{\bar{X}}_{s,0} \right)^2 \right)$$

► Redressement par la différence

$$\hat{\bar{X}}_0 = \sum_s \frac{N_s}{N} \left(\hat{\bar{X}}_{s,0} - \hat{\bar{X}}_{s,1} + \bar{X}_{s,1} \right)$$

population infinie

$$\begin{aligned} \hat{var} \left[\hat{\bar{X}}_0 \right] = & \sum_s \left(\left(\frac{N_s}{N} \right)^2 \frac{m_s \sum_{g=1}^{m_s} p_g^2}{m_s + m_s \sum_{g=1}^{m_s} p_g^2 - 2} * \right. \\ & \left. \sum_{g=1}^{m_s} \left(\left(\bar{X}_{g,0} - \hat{\bar{X}}_{s,0} \right)^2 + \left(\bar{X}_{g,1} - \hat{\bar{X}}_{s,1} \right)^2 - 2 \left(\bar{X}_{g,0} - \hat{\bar{X}}_{s,0} \right) \left(\bar{X}_{g,1} - \hat{\bar{X}}_{s,1} \right) \right) \right) \end{aligned}$$

population finie

$$\begin{aligned} \hat{var} \left[\hat{\bar{X}}_0 \right] = & \sum_s \left(\left(\frac{N_s}{N} \right)^2 \frac{m_s}{M_s} \frac{M_s - m_s}{M_s - 1} * \right. \\ & \left. \sum_{g=1}^{m_s} \left(\left(p_g \bar{X}_{g,0} - \frac{1}{m_s} \hat{\bar{X}}_{s,0} \right)^2 + \left(p_g \bar{X}_{g,1} - \frac{1}{m_s} \hat{\bar{X}}_{s,1} \right)^2 - 2 \left(p_g \bar{X}_{g,0} - \frac{1}{m_s} \hat{\bar{X}}_{s,0} \right) \left(p_g \bar{X}_{g,1} - \frac{1}{m_s} \hat{\bar{X}}_{s,1} \right) \right) \right) \end{aligned}$$

► Redressement par le quotient

$$\hat{X}_0 = \sum_s \frac{N_s}{N} \left(\frac{\hat{X}_{s,0}}{\hat{X}_{s,1}} \bar{X}_{s,1} \right)$$

population infinie

$$\hat{var} \left[\hat{X}_0 \right] = \sum_s \left(\left(\frac{N_s}{N} \right)^2 \frac{m_s \sum_{g=1}^{m_s} p_g^2}{m_s + m_s \sum_{g=1}^{m_s} p_g^2 - 2} * \right. \\ \left. \sum_{g=1}^{m_s} \left(\left(\bar{X}_{g,0} - \hat{X}_{s,0} \right)^2 + r_s^2 \left(\bar{X}_{g,1} - \hat{X}_{s,1} \right)^2 - 2 r_s \left(\bar{X}_{g,0} - \hat{X}_{s,0} \right) \left(\bar{X}_{g,1} - \hat{X}_{s,1} \right) \right) \right)$$

$$\text{avec } r_s = \frac{\hat{X}_{s,0}}{\hat{X}_{s,1}}$$

population finie

$$\hat{var} \left[\hat{X}_0 \right] = \sum_s \left(\frac{N_s}{N} \right)^2 \frac{m_s}{M_s} \frac{M_s - m_s}{M_s - 1} * \\ \sum_{g=1}^{m_s} \left(\left(p_g \bar{X}_{g,0} - \frac{1}{m_s} \hat{X}_{s,0} \right)^2 + r_s^2 \left(p_g \bar{X}_{g,1} - \frac{1}{m_s} \hat{X}_{s,1} \right)^2 - 2 r_s \left(p_g \bar{X}_{g,0} - \frac{1}{m_s} \hat{X}_{s,0} \right) \left(p_g \bar{X}_{g,1} - \frac{1}{m_s} \hat{X}_{s,1} \right) \right)$$

$$\text{avec } r_s = \frac{\hat{X}_{s,0}}{\hat{X}_{s,1}}$$

► Redressement par la régression

$$\hat{\bar{X}}_0 = \sum_s \frac{N_s}{N} \left(\hat{\bar{X}}_{s,0} + b_s (\bar{X}_{s,1} - \hat{\bar{X}}_{s,1}) \right)$$

$$\text{avec } b_s = \frac{\sum_{g=1}^{m_s} \left(\rho_g \bar{X}_{g,0} - \frac{1}{m_s} \hat{\bar{X}}_{s,0} \right) \left(\rho_g \bar{X}_{g,1} - \frac{1}{m_s} \hat{\bar{X}}_{s,1} \right)}{\sum_{g=1}^{m_s} \left(\rho_g \bar{X}_{g,1} - \frac{1}{m_s} \hat{\bar{X}}_{s,1} \right)^2}$$

population infinie

$$\hat{\text{var}} \left[\hat{\bar{X}}_0 \right] = \sum_s \left(\left(\frac{N_s}{N} \right)^2 \frac{m_s \sum_{g=1}^{m_s} \rho_g^2}{m_s + m_s \sum_{g=1}^{m_s} \rho_g^2 - 2} \sum_{g=1}^{m_s} \left(\bar{X}_{g,0} - \hat{\bar{X}}_{s,0} \right)^2 (1 - \hat{\rho}_s^2) \right)$$

$$\text{avec } \hat{\rho}_s^2 = \frac{\left(\sum_{g=1}^{m_s} (\bar{X}_{g,0} - \hat{\bar{X}}_{s,0}) (\bar{X}_{g,1} - \hat{\bar{X}}_{s,1}) \right)^2}{\sum_{g=1}^{m_s} (\bar{X}_{g,0} - \hat{\bar{X}}_{s,0})^2 \sum_{g=1}^{m_s} (\bar{X}_{g,1} - \hat{\bar{X}}_{s,1})^2}$$

population finie

$$\hat{\text{var}} \left[\hat{\bar{X}}_0 \right] = \sum_s \left(\left(\frac{N_s}{N} \right)^2 \frac{m_s}{M_s} \frac{M_s - m_s}{M_s - 1} \sum_{g=1}^{m_s} \left(\rho_g \bar{X}_{g,0} - \frac{1}{m_s} \hat{\bar{X}}_{s,0} \right)^2 (1 - \hat{\rho}_s^2) \right)$$

$$\text{avec } \hat{\rho}_s^2 = \frac{\left(\sum_{g=1}^{m_s} \left(\rho_g \bar{X}_{g,0} - \frac{1}{m_s} \hat{\bar{X}}_{s,0} \right) \left(\rho_g \bar{X}_{g,1} - \frac{1}{m_s} \hat{\bar{X}}_{s,1} \right) \right)^2}{\sum_{g=1}^{m_s} \left(\rho_g \bar{X}_{g,0} - \frac{1}{m_s} \hat{\bar{X}}_{s,0} \right)^2 \sum_{g=1}^{m_s} \left(\rho_g \bar{X}_{g,1} - \frac{1}{m_s} \hat{\bar{X}}_{s,1} \right)^2}$$

Annexe 9

Étude d'impact
économique réalisée
par AIR Pays de Loire



► Cycles de mesure sur les sites urbains des Pays de la Loire

Impact économique

► Introduction

Dans le cadre de l'élaboration de son PSQA, Air Pays de la Loire a engagé une réflexion sur la mise en place de cycles de mesures sur les stations urbaines de son réseau de surveillance. L'objectif est de pouvoir réduire le nombre d'analyseurs, donc de limiter l'effort de renouvellement tout en permettant de conserver un nombre d'indicateurs équivalent en associant au plan d'échantillonnage mis en place des techniques de reconstitution de données.

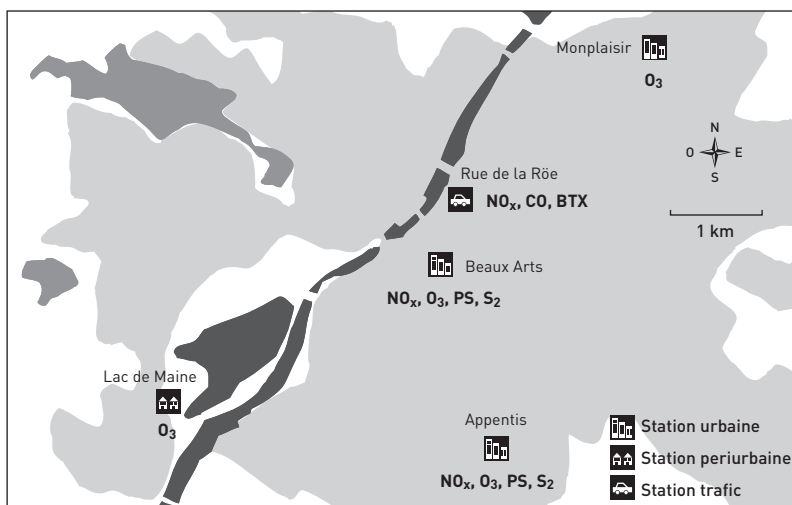
Ce document a pour objectif d'évaluer l'impact économique de ces actions. Les variations du nombre d'unités d'œuvre, des coûts de déplacement (carburant et péage) et des coûts d'analyse sont évaluées entre 2004 (état initial) et 2006 (état final).

Dans un premier temps l'impact spécifique du développement des cycles de mesure sur les sites urbains est considéré. Dans un deuxième temps, ce bilan est corrigé du gain obtenu par la mise en place de l'ensemble du programme de surveillance, notamment la réduction du réseau permanent.

► Description des simulations

Dans les agglomérations de plus de 100 000 habitants, Air Pays de la Loire disposait en 2004 d'informations provenant de 5 stations urbaines à Nantes, de 3 stations urbaines à Angers et au Mans et de 2 stations urbaines à Saint-Nazaire. Air Pays de la Loire a proposé dans son PSQA de réaliser à partir de 2006 des cycles de mesures sur ces stations en conservant le caractère permanent de la station de centre-ville pour référence. Saint-Nazaire n'a pas fait l'objet de cette étude en raison du nombre de sites limité à deux.

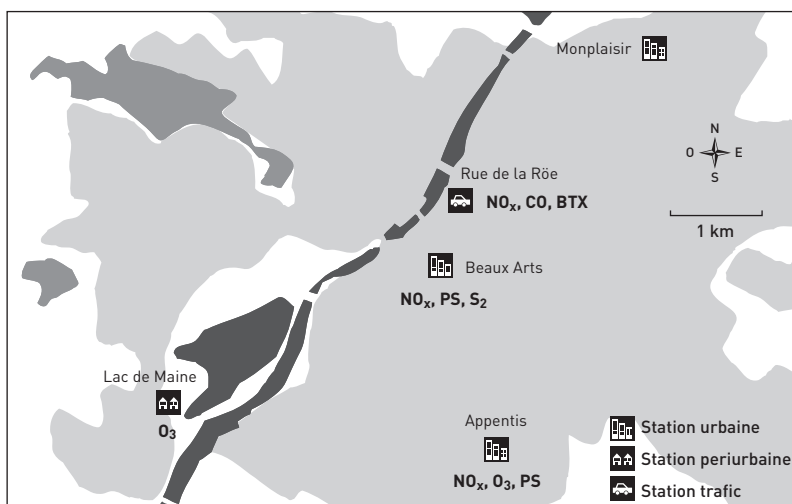
Ainsi, les cycles de mesure concernent 4 couples de stations: deux à Nantes et un à Angers et un au Mans. L'exemple de l'approche sur l'agglomération angevine est donné dans la page suivante (le couple est formé des stations Monplaisir et Appentis, la station de référence du centre-ville est Beaux Arts).



Angers

Fin 2004

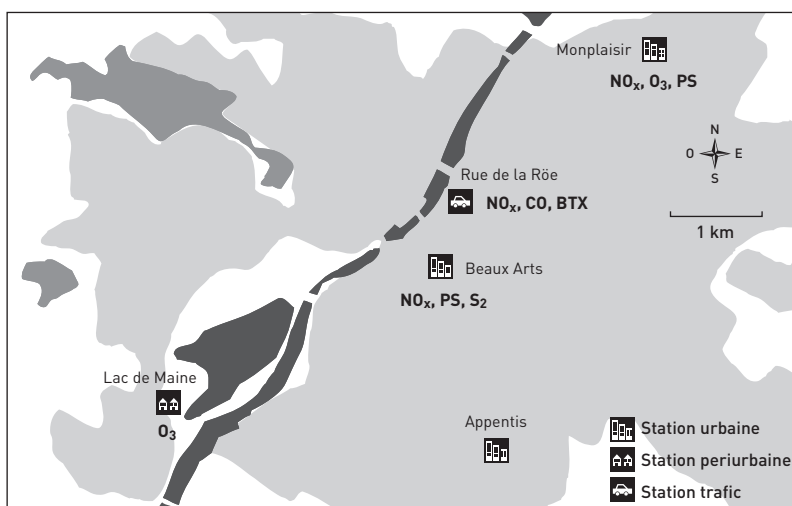
- 5 sites
- 13 points mesure



Angers

Période cyclique 1
à compter de 2006

- 4 sites
- 10 points mesure



Angers

Période cyclique 2
à compter de 2006

- 4 sites
- 10 points mesure

Les simulations de l'impact économique concernent au total 4 types de cycles sur un an :

- 8 x 1,5 mois : la première station du couple est instrumentée durant 1,5 mois, la deuxième est équipée durant la période suivante de 1,5 mois (4 cycles dans l'année)
- 4 x 3 mois : 2 cycles dans l'année
- 2 x 6 mois : 1 cycle dans l'année
- 1 x 12 mois : la première station est équipée l'année n, la deuxième l'année n + 1

Le dernier cycle constitue un cas particulier d'échantillonnage et de reconstitution puisqu'il nécessite l'utilisation des informations de l'année passée ou des années précédentes en complément des données de la station de référence permanente pour reconstruire les indicateurs : aucune mesure sur la station cyclique n'est disponible pour l'année à reconstituer. Les performances de la méthode de régression, seule méthode permettant ce type de calcul parmi les trois techniques étudiées, ne sont pas traitées dans ce guide qui considère que l'année à reconstituer est en partie échantillonnée. Ce cas particulier doit être considéré comme une approche d'estimation objective (calcul de concentrations à partir de valeurs mesurées en d'autres lieux et/ou à d'autres périodes⁽¹⁾).

► Impact spécifique des cycles sur stations urbaines

Le bilan économique de l'impact de la mise en place des cycles de mesures sur les sites urbains est présenté ci-dessous. Les coûts de déplacements (carburant et péage) tiennent de la localisation du Service Métrologie (près de Nantes) par rapport aux lieux d'intervention (Nantes, Angers à 91 km dont 81 km de voie à péage, Le Mans à 185 km dont 170 km de voie à péage).

	UO terrain (en jours)	Déplacement (en euros)
8 x 1,5 mois	+ 70	+ 700
4 x 3 mois	+ 33	+ 360
2 x 6 mois	+ 15	+ 200
1 x 12 mois	+ 6	+ 100

Tableau 1 – Variation entre 2004 et 2006 des unités d'œuvre (UO) et des coûts de déplacements liés à différents scénarii d'approche cyclique sur les sites urbains des Pays de la Loire.

En les considérant isolément, ces actions ont un impact économique négatif puisqu'elles génèrent logiquement plus de temps d'intervention, en particulier pour les deux premiers types de cycles.

1. Guidance on Assessment under the EU Air Quality Directives – Final Draft.

► Impact de l'ensemble des actions du programme PSQA

Le bilan précédent doit être corrigé avec l'ensemble des actions programmées par le PSQA entre 2004 et 2006 qui ont un impact, positif ou négatif, sur les temps d'intervention, les coûts d'exploitation et de maintenance préventive et les coûts d'analyse chimique à savoir :

- la réduction de 15 % du nombre d'analyseurs de sites permanents (dioxyde de soufre, notamment mais aussi ozone et dioxyde d'azote...);
- le programme cyclique annuel de surveillance du dioxyde de soufre dans les 3 villes de 50 000 habitants (La Roche sur Yon, Cholet, Laval);
- la surveillance des PM_{2,5} par période d'un an dans les 3 villes de 50 000 habitants et par période d'un an dans 3 agglomérations de plus de 100 000 habitants (Saint-Nazaire, Angers, Le Mans);
- la diminution de la couverture temporelle de la mesure du benzène par tubes passifs dans les 4 agglomérations de plus de 100 000 habitants (passage de 100 % à 50 %) mais légère augmentation de la couverture temporelle dans les 3 villes de 50 000 habitants (passage de 33 % à 50 %);
- la diminution de la couverture temporelle de la mesure des métaux toxiques en milieu urbain (passage de 100 % à 50 %).

Le bilan global de l'impact économique du PSQA des Pays de la Loire suivant 4 scénarii de cycles sur les stations urbaines est donné ci-dessous :

	UO terrain (en jours)	Déplacement (en euros)	Analyses chimiques (en euros)	Exploitation (en euros)
8 x 1,5 mois	+ 3	+ 800	- 7 000	- 1 700
4 x 3 mois	- 34	+ 500	- 7 000	- 1 700
2 x 6 mois	- 52	+ 300	- 7 000	- 1 700
1 x 12 mois	- 61	+ 200	- 7 000	- 1 700

Tableau 2 – Variation entre 2004 et 2006 des unités d'œuvre (UO) et des coûts liés à l'ensemble du programme de surveillance dans les Pays de la Loire.

L'approche de 8 x 1,5 mois est, pour Air Pays de la Loire, à écarter puisqu'elle ne génère pas de gains d'unités d'œuvre par rapport à 2004. En revanche, les deux cycles de 2 x 6 mois et de 1 x 12 mois permettent des gains appréciables, en lien principalement avec la réduction du nombre de d'analyseurs permanents qui compense les temps d'intervention des techniciens liés à la mise en place de cycles sur stations urbaines. C'est le cycle de 1 x 12 mois qui a été retenu en 2006 dans l'attente des recommandations du guide du GT en matière de choix des plans d'échantillonnage et des performances des méthodes de reconstitution.

Annexe 10

Cas d'application :
échantillonnage temporel
des HAP pour l'estimation
d'une moyenne annuelle

► 1. Résumé

Au moment où se développe la surveillance des HAP dans l'air ambiant, il convient de s'interroger sur les stratégies d'échantillonnage les plus efficaces pour estimer une concentration moyenne annuelle avec une bonne précision. L'historique de données disponibles pour une telle analyse est à ce jour limité. De février à décembre 1998 cependant, des prélèvements quotidiens de HAP ont été réalisés en un site urbain de Paris (sous influence du trafic). AIRPARIF a mis les séries de mesures à la disposition du GT *Plans d'échantillonnage et reconstitution de données*. La présente note synthétise les résultats et recommandations d'échantillonnage issus de l'exploitation de ces données. Le cas du BaP est traité plus particulièrement.

Les profils annuels des concentrations de HAP montrent une saisonnalité marquée, avec des valeurs et une variabilité sensiblement plus élevées en hiver (mois de février, mars, novembre et décembre). Les profils hebdomadaires moyens présentent un maximum en semaine (le mercredi pour le BaP) et un minimum le samedi.

Différents plans d'échantillonnage sont simulés et comparés : des plans d'échantillonnage systématiques, qui consistent à effectuer des mesures à intervalles de temps réguliers ; des plans d'échantillonnage aléatoires stratifiés, qui consistent à découper l'année en grandes périodes, les strates, puis à y sélectionner aléatoirement un nombre approprié de mesures individuelles (c'est-à-dire de prélèvements journaliers) ou de groupes de mesures consécutives (les grappes).

D'après les statistiques d'erreur, et pour une même couverture annuelle, la réalisation de prélèvements isolés permet d'estimer plus précisément la moyenne annuelle que des prélèvements groupés par semaines. Pour ce qui est du BaP, un échantillonnage systématique d'un jour sur six, soit une soixantaine de mesures également réparties sur l'année, conduit à une erreur d'estimation inférieure à 12 % pour cinq des six échantillons possibles et voisine de 33 % pour le sixième échantillon. Un plan aléatoire stratifié constitué de quinze mesures individuelles dans chaque trimestre conduit à une précision comparable : 200 échantillons ont été simulés selon un tel plan ; sur ces 200 tirages, la moyenne, la médiane et le centile 95 de l'erreur relative sont respectivement de 15 %, 13 % et 35 %. Ces valeurs deviennent respectivement 24 %, 22 % et 52 % lorsque les mesures sont effectuées par grappes hebdomadaires.

Dans les tests décrits ci-dessus, l'année est découpée en strates de même taille, les trimestres, et le nombre de mesures par trimestre est identique. La théorie des sondages offre un cadre rigoureux pour concevoir d'autres plans d'échantillonnage dans lesquels les strates et/ou les échantillons extraits de ces dernières sont de taille différente. Les strates sont de préférence des parties de l'année relativement homogènes en ce qui concerne les valeurs et la variabilité des concentrations (pour le BaP on a considéré deux strates hivernales, une strate estivale et une strate de transition) ; le nombre, s'il n'est pas fixé par des contraintes externes, et la répartition des grappes entre les strates sont calculés en tenant compte de cette variabilité.

D'autre part, il s'avère que les concentrations de NO mesurées par une station urbaine voisine ont une variabilité temporelle très similaire à celle du BaP. Dans ce cas, l'élaboration d'un plan d'échantillonnage du BaP peut s'appuyer sur les mesures de NO. À ce jour, les données disponibles ne permettent pas de généraliser ce résultat.

Du point de vue de la réglementation, la mesure fixe ou indicative des HAP n'admet qu'un échantillonnage régulièrement réparti sur l'année et une moyenne annuelle estimée comme la moyenne arithmétique des mesures. Lorsque le suivi des HAP peut s'effectuer par modélisation ou estimation objective, ou dans tout contexte autre que la surveillance réglementaire, l'utilisateur a le choix de la stratégie d'échantillonnage (systématique, aléatoire stratifié). Selon les données dont il dispose, il peut alors appliquer une ou plusieurs méthodes de reconstitution afin d'estimer une concentration moyenne annuelle et son incertitude associée.

► 2. Introduction

2.1 Objectifs

En exploitant une série annuelle de prélèvements de HAP, le GT *Plans d'échantillonnage et reconstitution de données* a cherché à déterminer, indépendamment de toute contrainte technique, quels plans d'échantillonnage convenaient le mieux à l'estimation d'une concentration moyenne annuelle. Ce document synthétise les principaux résultats de cette étude.

2.2 Contraintes réglementaires

Afin que les résultats de cette étude soient utilisés à bon escient, il convient de rappeler que :

- dans le cadre de la mesure fixe et indicative, l'échantillonnage des HAP doit couvrir respectivement 33 % et 14 % de l'année au minimum et se répartir régulièrement sur l'année et les jours de semaine. Seule la moyenne arithmétique est autorisée pour estimer la moyenne annuelle. Il est demandé aux AASQA de pratiquer ce type de suivi pour l'évaluation préliminaire actuellement en cours. L'incertitude sur la moyenne annuelle ne doit pas dépasser 50 %.
- dans le cadre de l'estimation objective et de la modélisation, la réglementation n'impose pas de contrainte sur la stratégie d'échantillonnage. L'incertitude sur la moyenne annuelle ne doit pas dépasser 60 % en modélisation.

2.3 Données utilisées

Les données traitées ont été fournies par AIRPARIF. Elles sont issues d'une campagne de mesure des HAP conduite du 10 février au 31 décembre 1998 selon un **pas de temps journalier**.

Le site de mesure est un site urbain du 13^e arrondissement de Paris soumis à l'influence du trafic. **La série de mesure couvre au total 75 % de l'année 1998 ;** elle comprend deux grandes plages de données manquantes : du 1^{er} janvier au 9 février et du 1^{er} juillet au 27 juillet.

L'étude se limite aux HAP cités dans la quatrième Directive fille et pour lesquels il existe des données (Tableau 1) :

- benzo(a)anthracène (B[a]A),
- benzo(b)fluoranthène (B[b]F),
- benzo(k)fluoranthène (B[k]F),
- benzo(a)pyrène (B[a]P),
- indénopyrène (IP).

La base ne contient pas de données de benzo(j)fluoranthène ni de données de dibenzo(ah)anthracène.

	Effectif (nombre de prélèvements journaliers valides)	Couverture annuelle [%]	Moyenne (m) [pg/m ³]	Écart-type (σ) [pg/m ³]	Coefficient de variation = σ/m [%]
BaP	273	74,8	290,5	481,4	165,7
BaA	273	74,8	354,5	377,7	106,5
BbF	271	74,8	586,3	735,9	125,5
BkF	273	74,8	189,1	215,2	113,9
IP	273	74,8	458,9	535,5	116,7

Tableau 1 – Caractéristiques des données de HAP étudiées.

► 3. Étude de la variabilité temporelle

3.1 Profils annuels de concentration

Les séries des mesures journalières (Figure 1) montrent des pics de concentration relativement élevés en février (début des courbes) et en novembre.

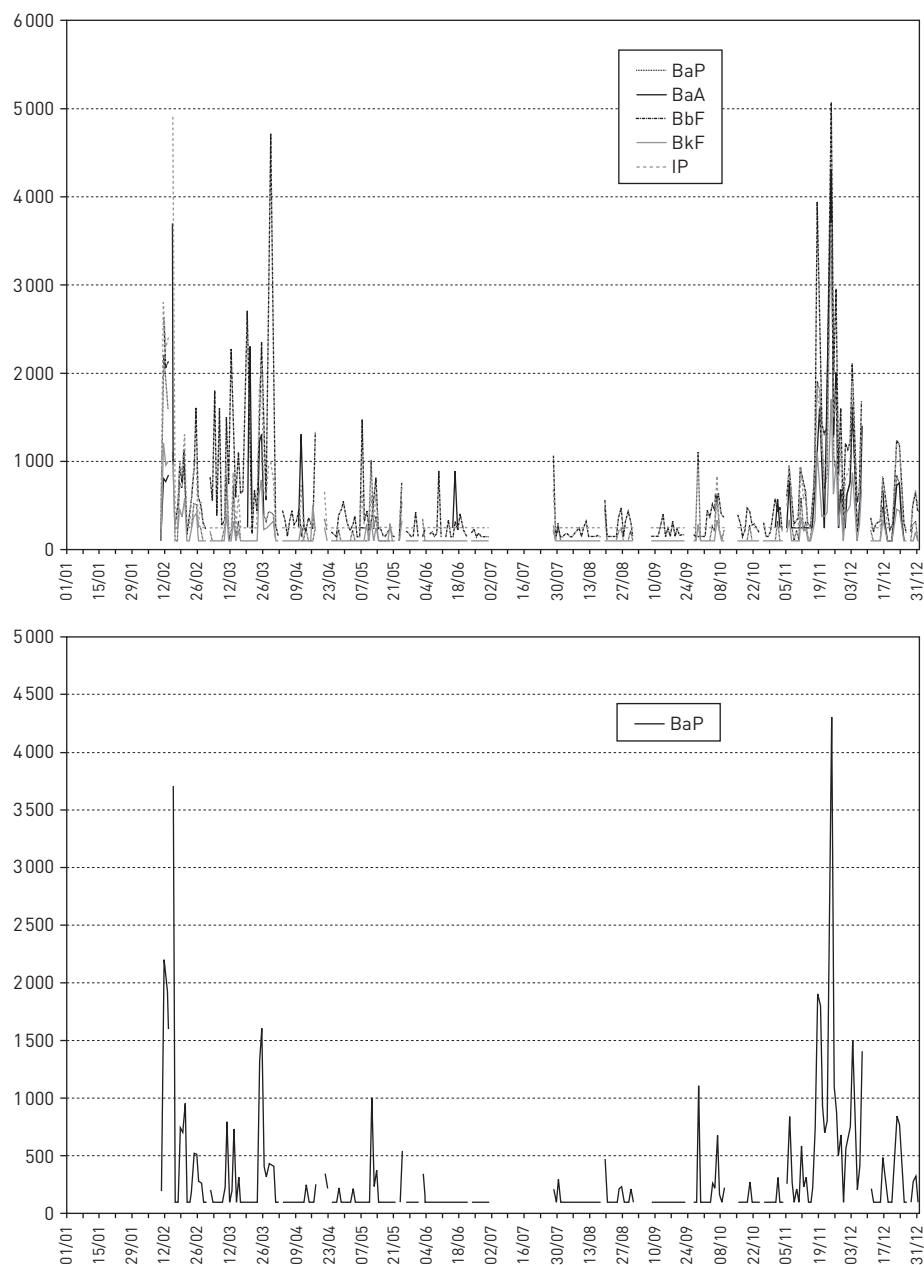


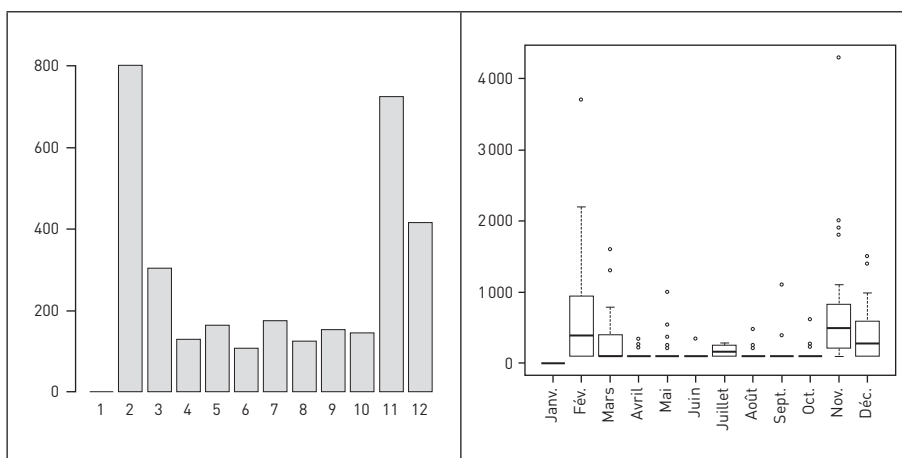
Figure 1 – Séries temporelles des concentrations de 5 HAP (figure du haut), du BaP (figure du bas), sur la période allant du 10 février au 31 décembre 1998. Concentrations en pg/m^3 .

Pour chacun des HAP, les figures suivantes représentent la moyenne et la dispersion mensuelle des données.

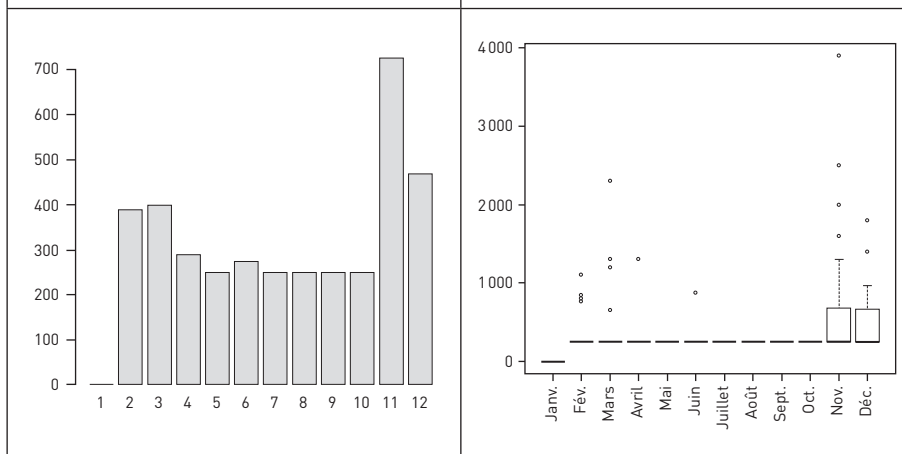
Elles font voir clairement la saisonnalité des concentrations, avec des **valeurs et une variabilité sensiblement plus élevées en hiver** (mois de février, mars, novembre et décembre ; on peut supposer qu'il en est de même en janvier bien

qu'aucune mesure ne soit disponible pour l'attester). L'augmentation des concentrations de HAP en juillet n'est pas significative compte tenu du petit nombre de données valides (moins de 5 valeurs).

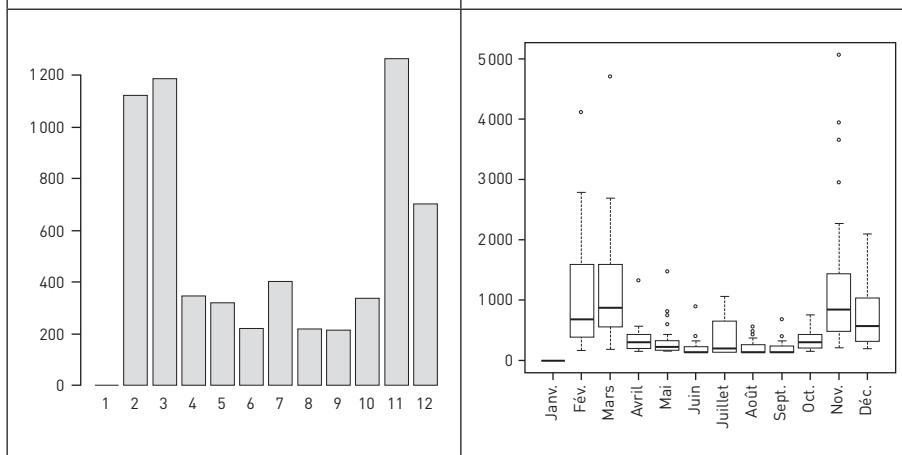
B[a]P



B[a]A



B[b]F



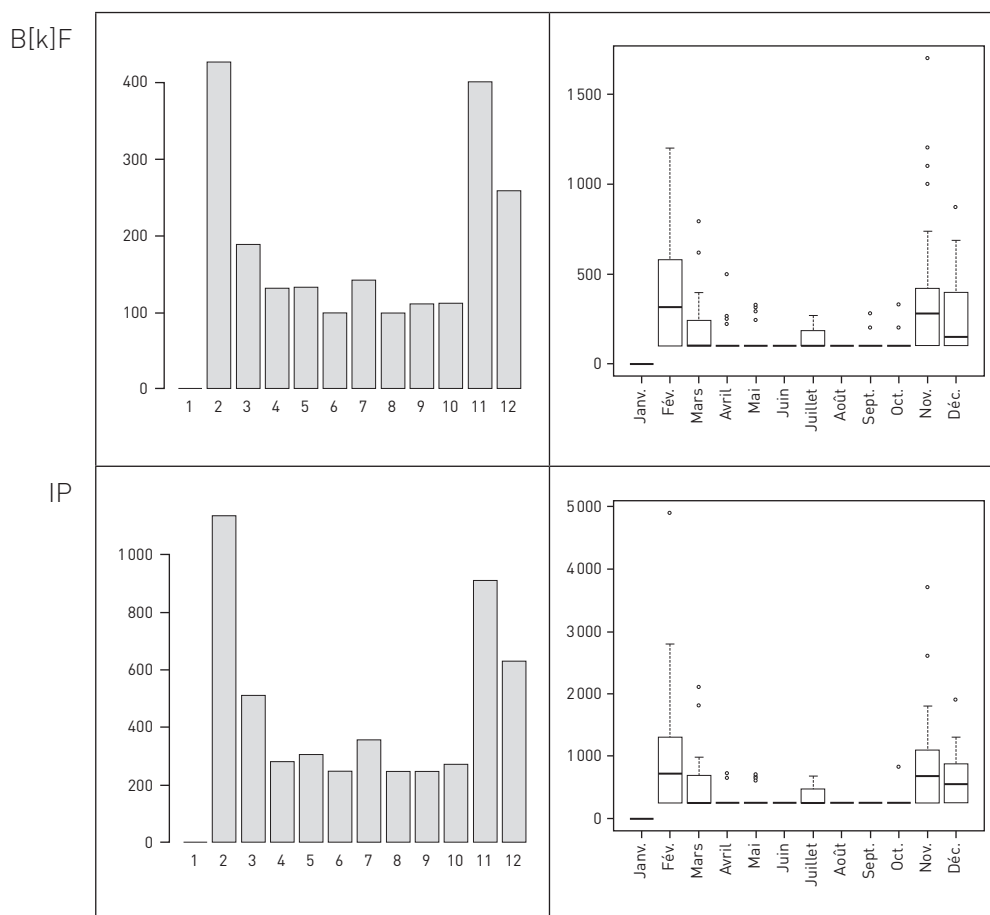


Figure 2 – Colonne de gauche : moyennes mensuelles [pg/m^3] des cinq HAP étudiés. Colonne de droite : représentation sous forme de boîtes à moustaches de la dispersion mensuelle des concentrations. Les traits horizontaux d'une boîte à moustaches matérialisent le minimum (hors valeurs atypiques), le centile 25, la médiane, le centile 75, le maximum (hors valeurs atypiques); les points isolés représentent les données atypiques.

3.2 Profils hebdomadaires

- B[k]F, B[a]P et IP montrent des profils hebdomadaires similaires, avec, en moyenne sur l'année, une concentration maximale le mercredi et minimale le samedi (Figure 3). Les valeurs mesurées le dimanche sont en moyenne supérieures à celles du samedi; pour l'IP, elles sont même comparables aux concentrations du lundi.
- Le profil hebdomadaire du B[b]F ressemble à celui de l'IP à ceci près que le maximum de concentration s'observe le jeudi et que la décroissance du jeudi au vendredi est plus brutale.

- En ce qui concerne le B[a]A, les écarts entre jours ouvrés sont moindres. La concentration est maximale le mardi. Les valeurs mesurées le week end sont plus faibles qu'en semaine ; on ne note pas de différence entre le samedi et le dimanche.

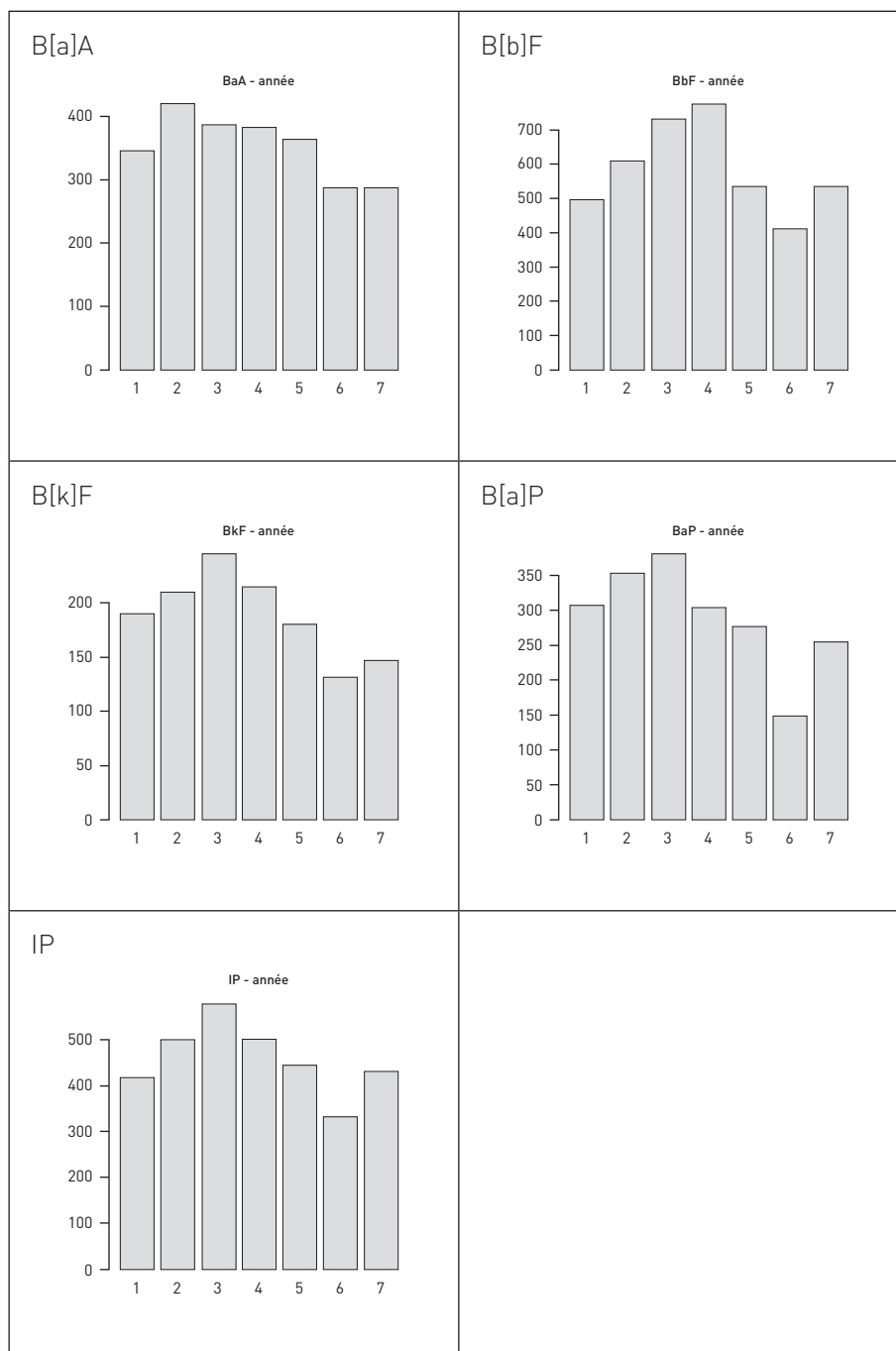


Figure 3 – Profils hebdomadaires de concentration pour les cinq HAP étudiés [pg/m³].
1 : lundi... 7 : dimanche.

► 4. Comparaison de différents plans d'échantillonnage temporel

Dans cette partie, nous comparons plusieurs plans d'échantillonnage qui couvrent environ **14 % de l'année** et comptent le même nombre de mesures dans chaque trimestre. Nous examinons quelles stratégies d'échantillonnage conduisent aux estimations les plus précises. Les principaux termes de statistique relatifs à l'échantillonnage sont décrits dans le guide *Élaboration d'un plan d'échantillonnage et reconstitution de données* rédigé par le GT. Sont rappelées ci-après quelques définitions utiles à la compréhension de cette note.

*Dans un échantillonnage **systématique**, les mesures sont réalisées à intervalles de temps réguliers tandis que dans un échantillonnage **aléatoire**, les dates de mesure sont sélectionnées au hasard.*

Un plan d'échantillonnage systématique ou aléatoire est dit :

- **par grappes**, si au lieu de mesures isolées, on effectue des groupes de mesures consécutives, par exemple des groupes de sept prélèvements successifs ;
- **stratifié**, si l'année est préalablement découpée en grandes périodes, les strates (de taille égale, comme les saisons, ou inégale), dans chacune desquelles est tiré un échantillon convenablement dimensionné.

Les plans d'échantillonnage systématiques suivants sont simulés sur les données d'AIRPARIF :

- Mesures **un jour sur six** : à cause de longues périodes de données manquantes (janvier, partie de février, juillet), stratégie testée séparément sur le deuxième et le quatrième trimestres et, à titre indicatif, sur l'année entière ;
- Mesures **un jour sur trois dans chaque première moitié de trimestre** : pour la même raison que précédemment (peu de données disponibles dans les premières moitiés du premier et du troisième trimestre), stratégie testée uniquement pour le deuxième et le quatrième trimestres ;
- Mesures **un jour sur trois dans chaque seconde moitié de trimestre** : stratégie testée pour le deuxième et le quatrième trimestres et pour l'année entière.

En étendant la notion d'*égale répartition des mesures sur l'année* à celle, plus souple, d'*effectif égal par saison*, deux plans d'échantillonnage aléatoires stratifiés sont également considérés :

- **15 mesures journalières sélectionnées aléatoirement dans chaque trimestre** : cette stratégie a été testée en réalisant 200 simulations d'échantillonnage ;

- **deux grappes de 7 prélèvements journaliers sélectionnées aléatoirement dans chaque trimestre** : cette stratégie a été testée en réalisant 200 simulations d'échantillonnage.

Avertissement : À cause des données manquantes, certains échantillons peuvent comporter moins de mesures ou de grappes que ce qui a été planifié. Si individuellement, les résultats de certains tirages doivent être considérés prudemment, les statistiques sur les 200 simulations fournissent des informations exploitables.

Dans ces différents plans, qu'ils reposent sur un échantillonnage systématique ou aléatoire, les trimestres jouent le rôle de strates. Celles-ci étant toutes de même taille (trois mois), l'estimation de la concentration moyenne annuelle d'un HAP est la moyenne arithmétique des concentrations moyennes par strate. Comme en outre chaque strate compte le même nombre de prélèvements (une quinzaine), cette estimation correspond (sans considération des données manquantes) à la moyenne arithmétique de l'ensemble des mesures.

4.1 Échantillonnage systématique

► 4.1.1 Comparaison de trois stratégies sur les deuxième et quatrième trimestres

Pour chaque stratégie testée :

1. mesures 1 jour sur 6 pendant tout le trimestre
2. mesures 1 jour sur 3 pendant chaque 1^{ère} moitié de trimestre
3. mesures 1 jour sur 3 pendant chaque 2^{nde} moitié de trimestre

On effectue respectivement six, trois et trois simulations d'échantillonnage, décalées chacune d'une journée. À partir de chaque échantillon, on estime la concentration du trimestre. Comme on pouvait s'y attendre au vu de la variabilité temporelle des concentrations (Figure 1, Figure 2), les estimations du quatrième trimestre sont sensibles à l'échantillonnage, en particulier si la stratégie d'échantillonnage consiste à effectuer une mesure tous les trois jours pendant une moitié de trimestre. Entre la première et la seconde moitié du trimestre, les estimations peuvent différer d'un facteur 3 à 4 à cause des pics de novembre (Tableau 4, Tableau 5). En ce qui concerne le BaP, l'écart relatif à la moyenne mesurée est compris entre 20 % et 100 %. Les estimations relatives à la stratégie « une mesure tous les six jours » sont moins variables, bien qu'elles dépendent de la séquence de mesure. Un décalage d'une journée peut modifier sensiblement les résultats (Tableau 2 à Tableau 5). Pour le B[a]P au 4^e trimestre, les moyennes estimées s'écartent de 4 à 50 % de la moyenne mesurée. Cette erreur relative est moindre pour le B[a]A : de 2 à 22 %.

Fréquence d'échantillonnage (freq)	Numéro d'échantillon (n°)	Nombre de prélèvements valides (N)	Moyenne des prélèvements (m) [pg/m³]	Erreur relative* (err.r.) [%]
Tous les jours		83	134,7	

Mesures un jour sur 6 pendant tout le trimestre

1 jour/6	1	15	128,0	- 5
1 jour/6	2	12	100,0	- 25,8
1 jour/6	3	15	147,3	9,4
1 jour/6	4	14	108,6	- 19,4
1 jour/6	5	13	180,8	34,2
1 jour/6	6	14	141,4	5

Mesures un jour sur 3 pendant la 1^{ère} moitié du trimestre

1 jour/3	1	15	136,0	1
1 jour/3	2	12	187,5	39,2
1 jour/3	3	14	143,6	6,6

Mesures un jour sur 3 pendant la 2^{nde} moitié du trimestre

1 jour/3	1	14	100,0	- 25,8
1 jour/3	2	13	100,0	- 25,8
1 jour/3	3	15	145,3	7,9

* L'erreur relative désigne l'écart relatif entre la moyenne de l'échantillon et la moyenne de la série entière.

Tableau 2 – Simulation de différents plans d'échantillonnage sur le 2^e trimestre.

Concentration moyenne de **BaP** [pg/m³] pour chaque échantillon tiré.

freq	n°	B[a]A			B[b]F			B[k]F			IP		
		N	moy	err.r.	N	moy	err.r.	N	moy	err.r.	N	moy	err.r.
tlj		83	270,2		81	298,8		83	121,4		83	279,6	

Mesures un jour sur 6 pendant tout le trimestre

1/6	1	15	250	- 7,5	15	351,3	17,6	15	122	0,5	15	302,7	8,3
1/6	2	12	250	- 7,5	12	236,7	- 20,8	12	100	- 17,6	12	250	- 10,6
1/6	3	15	250	- 7,5	15	275,3	- 7,9	15	122,7	1,1	15	276,7	- 1
1/6	4	14	250	- 7,5	14	302,9	1,4	14	128,6	5,9	14	250	- 10,6
1/6	5	13	379,2	40,3	12	305,8	2,3	13	137,7	13,4	13	313,1	12
1/6	6	14	250	- 7,5	13	311,5	4,3	14	115,7	- 4,7	14	282,1	0,9

Mesures un jour sur 3 pendant la 1^{ère} moitié du trimestre

1/3	1	15	250	- 7,5	15	407,3	36,3	15	136	12	15	302,7	8,3
1/3	2	12	337,5	24,9	12	347,5	16,3	12	140,8	16	12	318,3	13,8
1/3	3	14	250	- 7,5	14	347,1	16,2	14	124,3	2,4	14	278,6	- 0,4

Mesures un jour sur 3 pendant la 2^{nde} moitié du trimestre

1/3	1	14	250	- 7,5	14	242,9	- 18,7	14	113,6	- 6,4	14	250	- 10,6
1/3	2	13	298,5	10,5	12	195	- 34,7	13	100	- 17,6	13	250	- 10,6
1/3	3	15	250	- 7,5	14	237,1	- 20,6	15	14	- 5,5	15	280	0,1

Tableau 3 – Simulation de différents plans d'échantillonnage sur le 2^e trimestre. Résultats pour les autres HAP.

Fréquence d'échantillonnage (freq)	Numéro d'échantillon (n°)	Nombre de prélèvements valides (N)	Moyenne (m) [pg/m³]	Erreur relative (err.r.) [%]
Tous les jours		84	434,6	

Mesures un jour sur 6 pendant tout le trimestre

1 jour/6	1	15	654,7	50,6
1 jour/6	2	15	456	4,9
1 jour/6	3	14	272,1	- 37,4
1 jour/6	4	13	367,7	- 15,4
1 jour/6	5	13	378,5	- 12,9
1 jour/6	6	14	452,9	4,2

Mesures un jour sur 3 pendant la 1^{ère} moitié du trimestre

1 jour/3	1	14	189,3	- 56,4
1 jour/3	2	13	168,5	- 61,2
1 jour/3	3	14	190,7	- 56,1

Mesures un jour sur 3 pendant la 2^{nde} moitié du trimestre

1 jour/3	1	15	638	46,8
1 jour/3	2	14	534,3	22,9
1 jour/3	3	14	853,6	96,4

Tableau 4 – Simulation de différents plans d'échantillonnage sur le **4^e trimestre**.

Concentration moyenne de BaP [pg/m³] pour chaque échantillon tiré.

freq	n°	B[a]A			B[b]F			B[k]F			IP		
		N	moy	err.r.	N	moy	err.r.	N	moy	err.r.	N	moy	err.r.
tlj		84	486,7		84	779,6		84	261,0		84	611,4	

Mesures un jour sur 6 pendant tout le trimestre

1/6	1	15	592,7	21,8	15	1 100	41,1	15	352,7	35,1	15	779,3	27,5
1/6	2	15	535,3	10,0	15	766	- 1,7	15	248,7	- 4,7	15	597,3	- 2,3
1/6	3	14	466,4	- 4,2	14	644,3	- 17,4	14	212,9	- 18,4	14	473,6	- 22,5
1/6	4	13	393,8	- 19,1	13	628,5	- 19,4	13	226,2	- 13,3	13	541,5	- 11,4
1/6	5	13	433,1	- 11,0	13	681,5	- 12,6	13	246,2	- 5,7	13	586,2	- 4,1
1/6	6	14	477,1	- 2,0	14	817,9	4,9	14	270	3,4	14	672,9	10,1

Mesures un jour sur 3 pendant la 1^{ère} moitié du trimestre

1/3	1	14	281,4	- 42,2	14	397,1	- 49,1	14	131,4	- 49,7	14	323,6	- 47,1
1/3	2	13	273,8	- 43,7	13	369,2	- 52,6	13	110	- 57,9	13	303,1	- 50,4
1/3	3	14	250	- 48,6	14	407,1	- 47,8	14	142,1	- 45,6	14	357,9	- 41,5

Mesures un jour sur 3 pendant la 2^{nde} moitié du trimestre

1/3	1	15	673,3	38,3	15	1 036,7	33,0	15	366,7	40,5	15	842,7	37,8
1/3	2	14	693,6	42,5	14	1 055	35,3	14	340,7	30,5	14	788,6	29,0
1/3	3	14	719,3	47,8	14	1 365	75,1	14	456,4	74,9	14	1 014,3	65,9

Tableau 5 – Simulation de différents plans d'échantillonnage sur le **4^e trimestre**. Résultats pour les **autres HAP**.

► 4.1.2 Test sur l'année entière

Comme il a été mentionné, à cause des données manquantes, seule la troisième stratégie, à savoir un jour de mesure sur trois dans chaque seconde moitié de trimestre, peut être testée sur l'année entière de manière satisfaisante. Trois simulations d'échantillonnage sont réalisées, décalées d'une journée l'une par rapport à l'autre.

Les moyennes estimées à partir des deuxième et troisième séquences de mesure sont proches des moyennes annuelles mesurées. L'erreur relative, qui indique une légère surestimation de ces moyennes, est inférieure à 13 %. En revanche, la première séquence conduit à des écarts plus importants (de 20 à 45 % environ). Les surestimations constatées sont dues principalement à l'influence de deux valeurs extrêmes aux premier et dernier trimestres (quel que soit le HAP).

Séquence 1

	Effectif	Couverture annuelle (%)	Moyenne de l'échantillon (pg/m ³)	Ecart-type	Biais (pg/m ³)	Erreur relative (%)	Moyenne de la série entière (pg/m ³)
BaP	54	14,8	422,4	794,1	131,9	45,4	290,5
BaA	54	14,8	425,4	617,3	70,9	20,0	354,5
BbF	54	14,8	815,7	1 134,7	229,4	39,1	586,3
BkF	54	14,8	235,2	296,7	46,1	24,4	189,1
IP	54	14,8	588,7	830,2	129,8	28,3	458,9

Séquence 2

	Effectif	Couverture annuelle (%)	Moyenne de l'échantillon (pg/m ³)	Ecart-type	Biais (pg/m ³)	Erreur relative (%)	Moyenne de la série entière (pg/m ³)
BaP	55	15,1	309,1	386,7	18,6	6,4	290,5
BaA	55	15,1	394,2	327,6	39,7	11,2	354,5
BbF	54	15,1	620,6	621,3	34,3	5,8	586,3
BkF	55	15,1	203,6	181,5	14,6	7,7	189,1
IP	55	15,1	487,3	414,5	28,4	6,2	458,9

Séquence 3

	Effectif	Couverture annuelle (%)	Moyenne de l'échantillon (pg/m ³)	Ecart-type	Biais (pg/m ³)	Erreur relative (%)	Moyenne de la série entière (pg/m ³)
BaP	59	16,2	296,3	382,5	5,8	2,0	290,5
BaA	59	16,2	380,0	408,5	25,5	7,2	354,5
BbF	58	15,9	661,6	767,8	75,2	12,8	586,3
BkF	59	16,2	197,5	221,7	8,4	4,5	189,1
IP	59	16,2	474,9	487,4	16,0	3,5	458,9

Tableau 6 – Estimation de la moyenne annuelle à partir de prélèvements journaliers réalisés un jour sur trois pendant chaque seconde moitié de trimestre.

À titre indicatif, nous présentons les résultats obtenus pour la stratégie « 1 jour sur 6 ».

Séquence 1

	Effectif	Couverture annuelle (%)	Moyenne de l'échantillon (pg/m ³)	Ecart-type	Biais (pg/m ³)	Erreur relative (%)	Moyenne de la série entière (pg/m ³)
BaP	46	12,6	287,8	386,6	- 2,7	- 0,9	290,5
BaA	46	12,6	346,5	381,5	- 8,0	- 2,3	354,5
BbF	46	12,6	507,2	492,1	- 79,1	- 13,5	586,3
BkF	46	12,6	184,1	184,1	- 4,9	- 2,6	189,1
IP	46	12,6	435,9	426,0	- 23,0	- 5,0	458,9

Séquence 2

	Effectif	Couverture annuelle (%)	Moyenne de l'échantillon (pg/m ³)	Ecart-type	Biais (pg/m ³)	Erreur relative (%)	Moyenne de la série entière (pg/m ³)
BaP	41	11,2	268,5	321,3	- 21,9	- 7,6	290,5
BaA	41	11,2	322,4	183,5	- 32,1	- 9,0	354,5
BbF	41	11,2	506,3	500,5	- 80,0	- 13,6	586,3
BkF	41	11,2	188,5	184,9	- 0,5	- 0,3	189,1
IP	41	11,2	459,8	435,3	0,9	0,2	458,9

Séquence 3

	Effectif	Couverture annuelle (%)	Moyenne de l'échantillon (pg/m ³)	Ecart-type	Biais (pg/m ³)	Erreur relative (%)	Moyenne de la série entière (pg/m ³)
BaP	47	12,9	274,0	339,4	- 16,4	- 5,7	290,5
BaA	47	12,9	326,4	342,3	- 28,1	- 7,9	354,5
BbF	47	12,9	539,1	601,5	- 47,2	- 8,0	586,3
BkF	47	12,9	177,9	189,1	- 11,2	- 5,9	189,1
IP	47	12,9	437,0	426,2	- 21,9	- 4,8	458,9

Séquence 4

	Effectif	Couverture annuelle (%)	Moyenne de l'échantillon (pg/m ³)	Ecart-type	Biais (pg/m ³)	Erreur relative (%)	Moyenne de la série entière (pg/m ³)
BaP	44	12,1	387,5	854,6	97,0	33,4	290,5
BaA	44	12,1	386,1	581,2	31,6	8,9	354,5
BbF	44	12,1	812,3	1 218,8	226,0	38,5	586,3
BkF	44	12,1	219,8	311,2	30,7	16,3	189,1
IP	44	12,1	559,3	887,6	100,4	21,9	458,9

Séquence 5

	Effectif	Couverture annuelle (%)	Moyenne de l'échantillon (pg/m ³)	Ecart-type	Biais (pg/m ³)	Erreur relative (%)	Moyenne de la série entière (pg/m ³)
BaP	48	13,2	271,7	387,0	- 18,8	- 6,5	290,5
BaA	48	13,2	394,0	355,1	39,5	11,1	354,5
BbF	47	12,9	558,5	601,6	- 27,8	- 4,7	586,3
BkF	48	13,2	178,3	167,9	- 10,7	- 5,7	189,1
IP	48	13,2	427,9	382,4	- 31,0	- 6,8	458,9

Séquence 6

	Effectif	Couverture annuelle (%)	Moyenne de l'échantillon (pg/m ³)	Ecart-type	Biais (pg/m ³)	Erreur relative (%)	Moyenne de la série entière (pg/m ³)
BaP	47	12,9	257,0	414,9	- 33,5	- 11,5	290,5
BaA	47	12,9	348,5	324,9	- 6,0	- 1,7	354,5
BbF	46	12,6	597,2	748,2	10,9	1,9	586,3
BkF	47	12,9	187,7	234,0	- 1,4	- 0,7	189,1
IP	47	12,9	440,2	523,1	- 18,7	- 4,1	458,9

Tableau 7 – Estimation de la moyenne annuelle à partir de prélèvements journaliers réalisés un jour sur six pendant toute l'année.

À l'exception de la quatrième séquence pour le BaP et le BbF, l'erreur sur la moyenne annuelle en révèle presque toujours une sous-estimation mais relativement faible.

► 4.1.3 Conclusions

- Dans les trimestres de plus grande variabilité interjournalière, l'erreur d'estimation commise avec un échantillonnage d'un jour sur six, uniformément réparti sur cette période, est généralement plus faible qu'avec un échantillonnage d'un jour sur trois, restreint à une moitié de trimestre. En effet, si comme l'automne 1998, le trimestre contient des plages de valeurs contrastées, cette seconde stratégie augmente le risque de biaiser l'estimation de la moyenne trimestrielle.
- En ce qui concerne le BaP, un échantillonnage d'un jour sur six sur l'année entière conduit à une erreur d'estimation sur la moyenne annuelle inférieure à 12 % dans cinq cas sur six et de 33 % pour l'un des échantillons.

			moy_vraie	moymoyest	etmoyest	minmoyest	P25moyest	P50moyest	P75moyest	maxmoyest	
BaP	1	15 x 4	290,5	309,2	53,4	168,9	272,4	308,1	349,7	475,1	
BaA	1	15 x 4	354,5	354,6	41,7	265,0	326,0	353,1	381,3	477,0	
BbF	1	15 x 4	586,3	637,1	79,5	465,3	574,4	635,8	690,9	858,5	
BkF	1	15 x 4	189,0	199,6	23,7	149,3	182,9	196,2	217,1	264,7	
IP	1	15 x 4	458,9	484,1	64,2	343,3	437,8	481,0	523,5	747,2	

4.2 Échantillonnage aléatoire stratifié

► 4.2.1 Comparaison de deux plans d'échantillonnage

Deux plans d'échantillonnage sont évalués :

- 15 prélèvements individuels effectués aléatoirement dans chaque trimestre ;
- 2 grappes hebdomadaires de mesures sélectionnées aléatoirement dans chaque trimestre.

200 tirages d'échantillons sont simulés pour chacun. À chaque tirage et pour chaque HAP, on calcule la différence entre la moyenne de l'échantillon et la vraie moyenne annuelle. Les tableaux ci-après fournissent des statistiques sur ces écarts.

Signification des variables des tableaux :

- moy_vraie : moyenne des mesures journalières sur l'année.
- moymoyest, etmoyest : moyenne et écart-type sur les 200 simulations de la moyenne annuelle estimée.
- minmoyest, P25moyest, P50moyest, P75moyest, maxmoyest : valeur minimale, centile 25, médiane, centile 75 et valeur maximale des moyennes annuelles estimées pour les 200 simulations.
- biais : moyenne annuelle estimée – moyenne annuelle réelle
- erreur relative : $100 \times (\text{moyenne annuelle estimée} - \text{moyenne annuelle réelle}) / \text{moyenne annuelle réelle}$
- biaismoy, biaiset : moyenne et écart-type du biais sur les 200 simulations.
- errelmoy, errelet : moyenne et écart-type de l'erreur relative sur les 200 simulations.
- P25errel, P50errel, P75errel, P90errel, P95errel, maxerrel : autres statistiques de l'erreur relative sur les 200 simulations (centile 25, médiane, centile 75 et valeur maximale.)

4.2.1.1 Quinze prélèvements journaliers répartis aléatoirement dans chaque trimestre

Les moyennes annuelles estimées s'écartent en moyenne de **10 à 15 %** environ de la moyenne annuelle mesurée (Tableau 8). Pour 95 % des simulations, cet écart est inférieur à 36 % (BaP), 22 % (BaA), 34 % (BbF), 26 % (BkF), 29 % (IP).

	biaismoy	biaiset	errelmoy	errelet	P25errel	P50errel	P75errel	P90errel	P95errel	maxerrel
	18,7	53,4	15,4	11,9	6,2	13,1	23,1	29,7	35,3	63,6
	0,1	41,7	9,5	6,9	4,1	7,9	13,8	19,6	21,1	34,5
	50,8	79,5	12,8	9,8	4,9	10,9	18,6	25,3	33,1	46,4
	10,6	23,7	10,9	8,3	3,5	9,8	16,2	23,5	25,9	40,0
	25,2	64,2	11,7	9,4	4,7	9,7	16,0	23,8	28,3	62,8

Tableau 8 – Estimation de la moyenne annuelle à partir de 15 prélèvements sélectionnés aléatoirement dans chaque trimestre (valeurs de concentration en pg/m^3).

4.2.1.2 Deux grappes de sept prélèvements journaliers consécutifs tirées aléatoirement dans chaque trimestre

Les estimations sont moins précises qu'avec 15 mesures journalières par trimestre. Les concentrations moyennes annuelles estimées s'écartent en moyenne de **15 % à 25 %** des valeurs mesurées. Par rapport au cas précédent, les centiles et l'écart-type des erreurs relatives ainsi que l'écart-type du biais augmentent.

			moy_vraie	moymoyest	etmoyest	minmoyest	P25moyest	P50moyest	P75moyest	maxmoyest	
BaP	7	2 x 4	290,5	286,1	84,1	127,5	219,1	279,4	350,2	536,1	
BaA	7	2 x 4	354,5	353,0	70,3	256,1	303,3	334,8	379,8	593,9	
BbF	7	2 x 4	586,3	622,9	132,1	404,5	523,0	595,6	702,3	1 070,4	
BkF	7	2 x 4	189,0	186,7	43,4	118,6	154,4	176,7	216,8	306,0	
IP	7	2 x 4	458,9	467,6	105,7	258,3	384,5	457,9	543,2	765,3	

Tableau 9 – Estimation de la moyenne annuelle à partir de 2 grappes hebdomadaires sélectionnées aléatoirement dans chaque trimestre (valeurs de concentration en $\mu\text{g}/\text{m}^3$).

► 4.2.2 Conclusions

D'après les statistiques d'erreur, et pour un même taux d'échantillonnage, la réalisation de prélèvements isolés fournit une estimation plus précise de la moyenne annuelle que des prélèvements groupés par semaines.

S'agissant du BaP, la moyenne et le centile 95 des erreurs relatives sont respectivement de 15 % et 35 % pour 15 prélèvements aléatoires par trimestre, et de 24 % et 52 % pour 2 grappes de 7 jours par trimestre.

	biaismoy	biaiset	errelmoy	errelet	P25errel	P50errel	P75errel	P90errel	P95errel	maxerrel
	- 4,4	84,1	24,0	16,1	10,4	22,3	34,1	44,3	51,9	84,6
	- 1,5	70,3	15,5	12,3	6,4	12,9	21,9	30,1	41,2	67,5
	36,6	132,1	17,6	15,3	6,1	12,8	24,2	40,0	50,9	82,6
	- 2,3	43,4	18,7	13,2	8,0	17,1	25,8	36,2	41,7	61,8
	8,7	105,7	19,2	12,8	9,4	17,5	26,3	35,4	43,9	66,8

► 5. Échantillonnage dimensionné en fonction de la variabilité temporelle des données

L'analyse des séries temporelles a révélé un net contraste entre la saison hivernale (concentrations et variabilité plus élevées) et la saison estivale. Celui-ci n'est pas pris en compte lorsqu'on tire un nombre égal de mesures par saison.

Or, comme les concentrations sont peu variables en été, la précision d'estimation de la concentration estivale ne devrait pas se dégrader si le nombre de prélèvements réalisés pendant cette saison est réduit dans une certaine limite. À l'inverse, en augmentant le nombre de mesures en hiver, on peut espérer obtenir une meilleure représentation de la variabilité des données sur cette saison et estimer plus précisément la concentration hivernale.

Cette démarche, qui consiste à adapter le dimensionnement de l'échantillonnage à la variabilité temporelle des données est l'approche développée par le GT dans le guide *Élaboration d'un plan d'échantillonnage et reconstitution de données* (et brièvement appelé *guide* dans la suite).

Elle n'est pas applicable dans le cadre de la mesure fixe ou indicative, puisque la moyenne arithmétique des données, seule formule autorisée dans ce contexte pour le calcul de la moyenne annuelle, risque alors de fournir une estimation biaisée (si le nombre de mesures est supérieur en hiver, la moyenne annuelle risque d'être surestimée).

En revanche, complétée par l'utilisation de méthodes de reconstitution, elle constitue une stratégie pertinente pour l'estimation objective ou la modélisation.

Une illustration en est présentée ci-après.

5.1 Élaboration d'un plan d'échantillonnage pour une précision d'estimation choisie

Dans cet exemple d'application, nous cherchons à atteindre une **précision d'estimation sur la moyenne annuelle de 25 %**.

L'examen du profil annuel et de la variabilité mensuelle des HAP, et notamment du BaP, nous conduit à découper l'année en quatre strates temporelles :

1. une première strate hivernale : (janvier)-février-mars (concentrations et variabilité relativement grandes) ;
2. une strate regroupant les périodes de transition : avril-mai et septembre-octobre (concentrations relativement basses avec quelques) ;
3. une strate estivale : juin-juillet-août (concentrations et variabilité faibles) ;
4. une seconde strate hivernale (concentrations et variabilité relativement grandes).

Trois tailles de grappes sont considérées :

- des grappes de taille unitaire, autrement dit des prélèvements journaliers isolés ;
- des grappes de trois prélèvements journaliers consécutifs ;
- des grappes de sept prélèvements journaliers consécutifs.

Pour atteindre la précision souhaitée, **l'application de la théorie des sondages** (voir le chapitre 1 du guide) fournit le nombre total de grappes à tirer dans l'année et la répartition optimale des grappes entre les strates. Elle aboutit aux dimensionnements suivants (on se limite pour cet exemple au BaP et au BbF) :

Précision fixée : 25 %		Nombre de grappes par strate	Nombre total de grappes à prélever dans l'année	Nombre total de jours de mesure dans l'année
Mesures individuelles	B[a]P	32 9 3 21	65	65
	B[b]F	19 5 3 13	40	40
Grappes de 3 jours	B[a]P	15 4 1 10	30	90
	B[b]F	7 3 1 6	17	51
Grappes de 7 jours	B[a]P	8 1 1 4	14	98
	B[b]F	3 1 1 4	9	63

Tableau 10 – Nombre de grappes à prélever dans chaque strate pour une précision attendue de 25 %.

Plus les grappes sont longues, plus la couverture temporelle requise est grande. D'autre part, pour une même précision souhaitée, le plan d'échantillonnage du BaP est plus contraignant que celui du BbF.

5.2 Évaluation de l'incertitude associée à un plan prédéfini

Reprenons l'exemple précédent. Supposons que des considérations pratiques conduisent au choix de grappes de 3 jours mais que les ressources disponibles ne permettent pas la réalisation du plan 15, 4, 1, 10 proposé pour le BaP. Supposons alors que le compromis trouvé entre les contraintes de ressources et l'intérêt d'augmenter le nombre de mesures en hiver aboutisse au plan 7, 4, 3, 7. D'après les formules de la théorie des sondages, l'incertitude d'estimation que l'on peut attendre de ce nouveau plan d'échantillonnage s'élève à 36 % pour le BaP et 22 % pour le BbF.

En ce qui concerne le BaP, cette incertitude est supérieure aux 25 % fixés initialement, ce qui est logique vu qu'on a réduit le nombre de grappes et modifié leur répartition. S'agissant du BbF, elle diminue puisque dans chaque strate, le plan d'échantillonnage comprend plus de grappes qu'il n'était requis pour ce polluant.

5.3 Échantillonnage systématique/ échantillonnage aléatoire stratifié. Comparaison de différents plans

Dans cette partie, nous comparons plusieurs plans d'échantillonnage qui totalisent chacun une soixantaine de jours de mesure.

Les plans aléatoires stratifiés ont été dimensionnés spécifiquement pour le **BaP** au moyen de la théorie des sondages et réajustés le cas échéant pour tenir compte de certaines contraintes (cf. commentaires du Tableau 12). Pour les cinq HAP, nous avons calculé l'incertitude d'estimation de la concentration moyenne annuelle associée a priori à chacun de ces plans. Les résultats relatifs au BaP et au BbF sont ici mentionnés.

À titre d'information, ce même calcul a été effectué pour le NO et le NO₂ mesurés en un site fixe proche d'AIRPARIF (station Paris 13^e). La variabilité temporelle des concentrations de NO se révèle tout à fait comparable à celle du BaP alors que le NO₂ est nettement moins variable (Tableau 11, Figure 4).

	Effectif	Couverture annuelle (%)	Moyenne annuelle	Ecart-type	Coefficient de variation (%)
BaP [$\mu\text{g}/\text{m}^3$]	273	74,8	290,5	481,4	165,7
NO [$\mu\text{g}/\text{m}^3$]	358	98,0	24,5	52,6	165,1
NO ₂ [$\mu\text{g}/\text{m}^3$]	358	98,0	40,5	20,5	38,9

Tableau 11 – Caractéristiques des données de BaP étudiées et des données de NO_x issues d'une station fixe AIRPARIF proche du site de prélèvement (Paris 13^e).

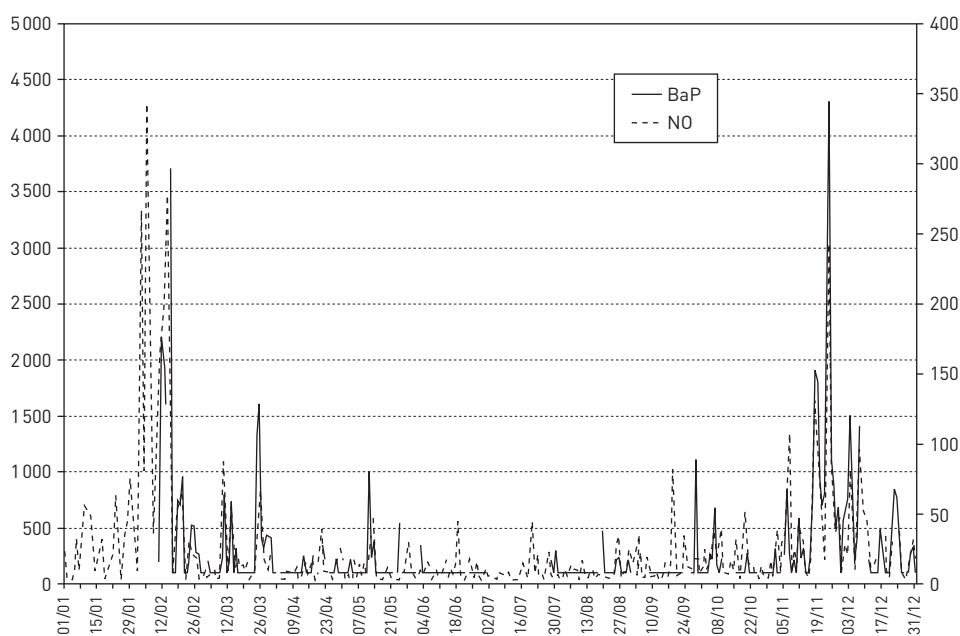


Figure 4 – Séries temporelles des concentrations de BaP [$\mu\text{g}/\text{m}^3$] et de NO [$\mu\text{g}/\text{m}^3$] sur la période allant du 10 février au 31 décembre 1998.

Type d'échantillonnage	Stratification	Taille des grappes	Nombre de grappes par strate	Nombre total de jours de mesure	Couverture annuelle	Évaluation a priori de l'incertitude
Systématique	/	1 jour	60 (1 jour sur 6)	60	16,4 %	BaP : 28 % BbF : 35 %** NO : 32 % NO ₂ : 8,7 %
Aléatoire stratifié	1. janv.-mars 2. avril-mai + sept.-oct. 3. juin-août 4. nov.-déc.	1 jour	29, 9, 3, 19	60	16,4 %	BaP : 26 % BbF : 20 % NO : 26 % NO ₂ : 14 %
Aléatoire stratifié	1. janv.-mars 2. avril-mai + sept.-oct. 3. juin-août 4. nov.-déc.	3 jours	(10, 2, 1, 7)* 9, 2, 2, 7	60	16,4 %	BaP : 34 % BbF : 23 % NO : 37 % NO ₂ : 13 %
Aléatoire stratifié	1. janv.-mars 2. avril-mai + sept.-oct. 3. juin-août 4. nov.-déc.	7 jours	(5, 1, (0,3), 3)* 3, 2, 2, 2	63	17,2 %	BaP : 49 % BbF : 29 % NO : 49 % NO ₂ : 12 %

* NB : La répartition théorique est légèrement modifiée afin d'avoir un minimum de 2 à 3 grappes par strate. En effet, la variance théorique, qui permet de déterminer la répartition optimale des grappes, est calculée sur la série de référence complète. En revanche, après que le plan d'échantillonnage aura été mis en œuvre, la variance qui lui correspondra réellement ne pourra qu'être qu'estimée sur les données recueillies. Pour que l'estimation de la variabilité dans chaque strate soit possible, il est préférable que chacune d'elles contienne plus d'une grappe.

** L'incertitude associée à un plan d'échantillonnage systématique se déduit de la variabilité, évaluée sur la série de référence, entre les différentes séquences d'échantillonnage possibles. La valeur obtenue pour le BbF, supérieure à celle du BaP, est due au fait que l'une des 6 séquences inclut les deux mesures de BbF les plus élevées, conduisant à une nette surestimation de la moyenne.

Tableau 12 – Évaluation a priori de l'incertitude associée à des plans d'échantillonnage systématiques ou aléatoires stratifiés qui totalisent une soixantaine de jours de mesure.

En ce qui concerne le BaP, **plus la taille de grappe augmente** et donc plus le nombre de grappes par strate diminue, **plus grande est l'incertitude associée au plan d'échantillonnage** (Tableau 12). Cette observation rejoint les résultats des simulations (Tableau 8 / Tableau 9). Les plans d'échantillonnage aléatoires stratifiés ayant été spécifiquement dimensionnés pour le BaP, cette remarque ne se vérifie pas nécessairement pour les autres composés. On constate cependant que les incertitudes calculées pour le BaP et pour le NO sont proches, ce qui s'accorde avec la variabilité temporelle similaire de ces deux polluants.

L'incertitude calculée pour le NO₂, polluant dont la concentration est sensiblement moins variable, est logiquement plus faible.

Pour des grappes unitaires, les incertitudes attendues d'un plan systématique et d'un plan aléatoire stratifié sont comparables. C'est bien ce que suggéraient les erreurs relatives obtenues dans les simulations d'échantillonnage (pages 134 et 137). Sachant que les incertitudes évaluées a priori ont une valeur **indicative** (voir la Figure 5 ci-après et les commentaires correspondants), les chiffres du Tableau 12 ne permettent pas de recommander une stratégie plutôt qu'une autre.

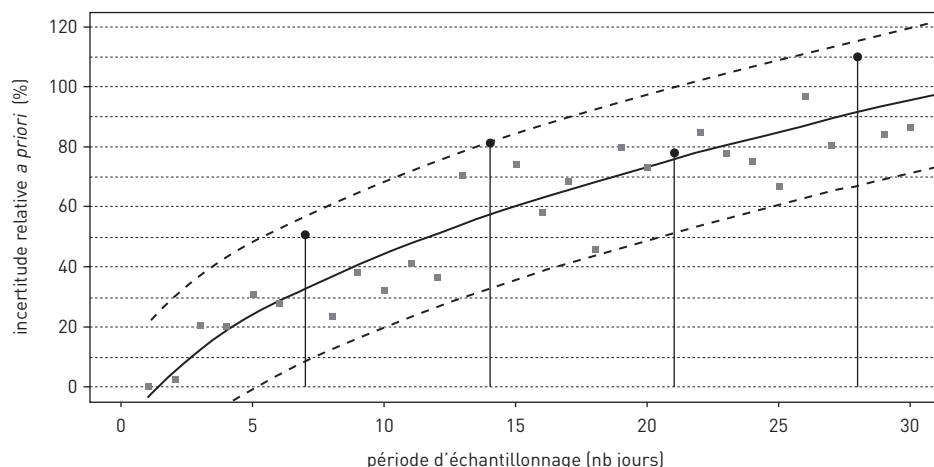


Figure 5 – Évaluation a priori de l'incertitude associée à un plan d'échantillonnage systématique. Évolution de l'incertitude en fonction de la période d'échantillonnage (période de 6 jours = une mesure tous les six jours).

- Courbe en trait continu : tendance (résultat de la régression linéaire entre l'incertitude a priori et la racine carrée de la période d'échantillonnage).
- Courbes en traits pointillés : intervalle de confiance à 95 % autour de la courbe de régression.
- Points noirs : valeurs d'incertitude correspondant à des périodes multiples de 7 jours.

La figure ci-dessus représente l'incertitude associée a priori à un plan d'échantillonnage systématique. Celle-ci a tendance à croître avec la période d'échantillonnage. Toutefois, la variabilité associée au calcul d'incertitude fait qu'une petite augmentation de la période d'échantillonnage peut s'accompagner d'une légère diminution de l'incertitude. D'autre part, il convient de noter l'augmentation quasi systématique de l'incertitude pour des périodes d'échantillonnage multiples de sept jours. Dans ce cas, l'incertitude exprime la variabilité entre les jours de semaine : si l'on considère les profils hebdomadaires, il est évident que la moyenne d'un échantillon constitué uniquement de lundis sera sensiblement différente de la moyenne d'un échantillon de même taille constitué uniquement de samedis.

► 6. Conclusions - Discussion

6.1 Conclusions sur l'échantillonnage

La base de données de HAP est incomplète mais offre un bon aperçu de la variabilité temporelle des concentrations pour les différents composés. Les nombres de grappes et les incertitudes calculés sont cohérents avec cette variabilité. Ainsi, bien que les données manquantes puissent affecter les valeurs numériques obtenues, elles ne semblent pas mettre en cause les indications générales que l'on peut tirer des résultats.

Pour **ce site urbain de mesure des HAP soumis à l'influence du trafic**, l'analyse des données fait ressortir les points suivants :

- Pour une meilleure précision d'estimation de la moyenne annuelle, des grappes courtes (1 à 3 jours) et nombreuses sont préférables à des grappes plus longues (hebdomadaires) et en effectif réduit.
- Dans le cas de prélèvements isolés (grappes de 1 jour), un échantillonnage systématique d'un jour sur six et un échantillonnage aléatoire stratifié conduisent à des précisions d'estimation comparables.
- Un échantillonnage d'un jour sur six uniformément réparti sur l'année est préférable à un échantillonnage d'un jour sur trois par moitié de trimestre. Certains trimestres sont en effet assez contrastés du point de vue de la variabilité des concentrations, notamment le quatrième trimestre qui oppose le mois d'octobre (faible variabilité) à novembre et décembre. Avec un échantillonnage d'un jour sur trois par moitié de trimestre, le risque de biaiser l'estimation de la moyenne annuelle est accru.
- Du fait que ces deux polluants présentent une variabilité temporelle similaire, l'élaboration d'un plan d'échantillonnage du BaP peut s'appuyer sur une série de NO issue d'un site proche et de mêmes caractéristiques.

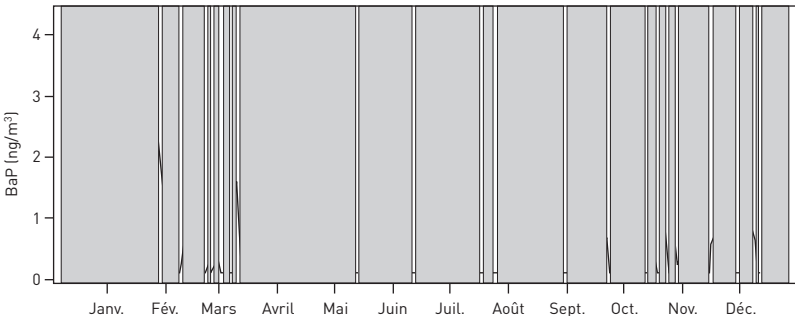
Il conviendrait d'examiner si cette corrélation entre le BaP et le NO se vérifie pour des sites de typologie différente et, si ce n'est pas le cas, déterminer quelles autres variables de pollution pourraient être éventuellement exploitées.

Notons que toute connaissance supplémentaire sur la variabilité temporelle des HAP et l'existence de variables corrélées avec ces HAP constituera une information utile pour étendre ou non à d'autres sites les conclusions précédentes.

6.2 Remarques sur l'estimation de la moyenne annuelle

Dans le cas d'un échantillonnage systématique, l'estimation de la moyenne annuelle est la moyenne arithmétique des mesures. Il en va de même dans celui d'un plan d'échantillonnage aléatoire stratifié comportant des **strates de même taille et un nombre identique de grappes par strate** : dans un tel plan, toutes les données ont le même poids.

Dans le cas d'un plan d'échantillonnage aléatoire stratifié, avec des **strates de taille différente et/ou des nombres différents de grappes par strate**, la **moyenne arithmétique est une estimation biaisée de la moyenne annuelle**. Il convient alors de mettre en œuvre une ou plusieurs **techniques de reconstitution**, comme celles qui sont décrites dans le guide publié par le GT. L'application de la théorie des sondages est illustrée dans les figures suivantes (Figure 6).

Plan d'échantillonnage aléatoire stratifié	4 strates : 1. janv.-mars 2. avril-mai + sept.-oct. 3. juin-août 4. nov.-déc. 21 grappes de 3 jours réparties dans les strates de la façon suivante : 7 - 4 - 3 - 7
Représentation des données d'échantillonnage	

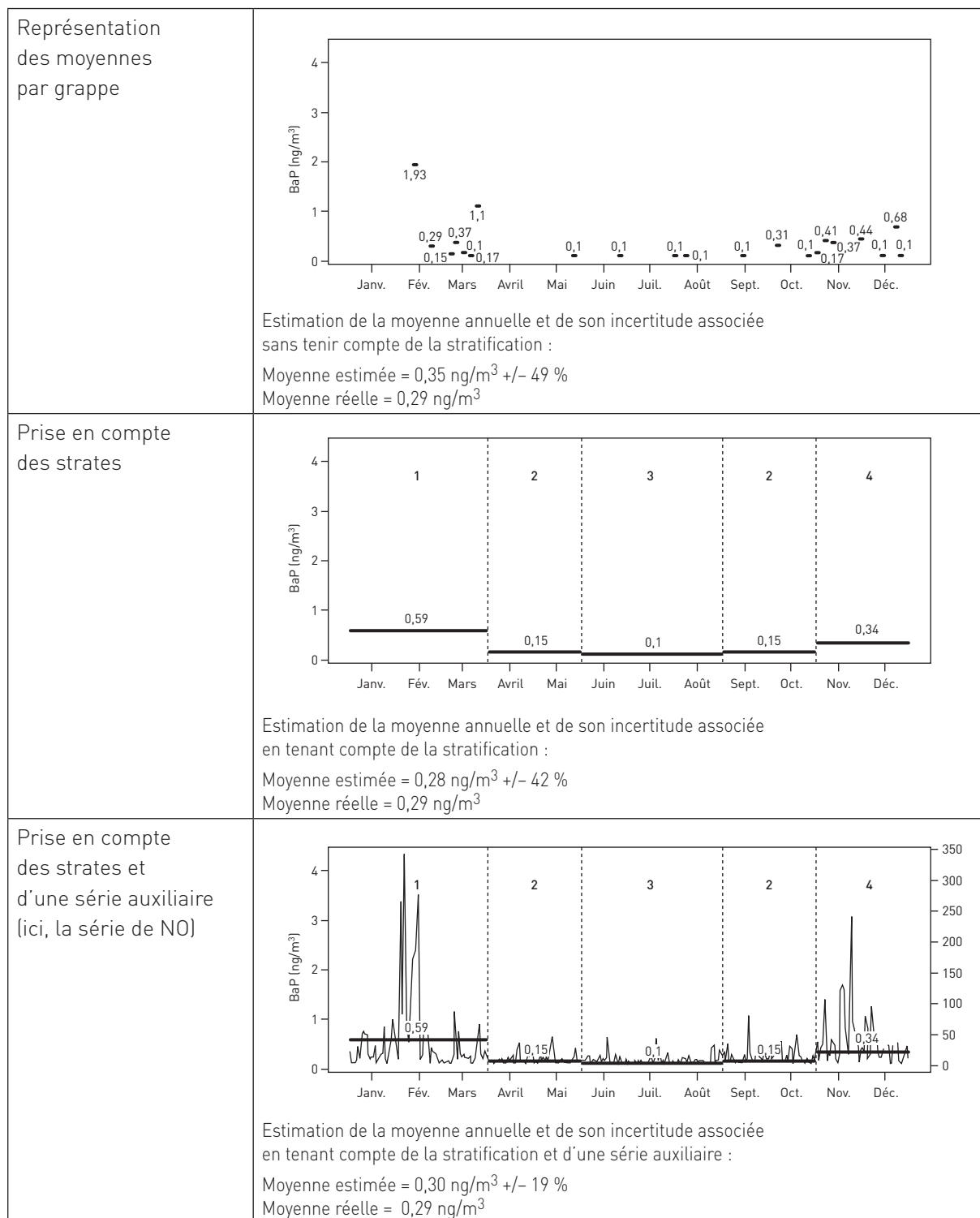


Figure 6 – Application de la théorie des sondages à l'estimation de la moyenne annuelle du BaP.
Exemple extrait des résultats présentés lors du séminaire LCSQA Surveillance dans l'air ambiant des HAP & des métaux, Paris, 30 novembre 2007.

Surveillance de la qualité de l'air ambiant

Guide d'élaboration de plans d'échantillonnage temporel et de reconstitution de données

En matière de surveillance de la qualité de l'air, des campagnes de mesure peuvent être menées par les Associations Agréées de Surveillance de la Qualité de l'Air (AASQA) pour répondre à des objectifs réglementaires ou à des fins d'études spécifiques.

Ces campagnes de mesure comportent souvent un objectif d'évaluation des concentrations de polluants sur des échelles de temps (l'année par exemple) différentes de celles de la durée ou de la périodicité des campagnes. Elles imposent de définir des stratégies d'échantillonnage et de reconstitution de données permettant de s'assurer de leur représentativité, notamment temporelle.

Ce guide a été élaboré par un collectif d'experts sous la coordination de l'ADEME pour harmoniser les pratiques d'échantillonnage temporel et de reconstitution de données. Il définit des démarches de référence applicables par l'ensemble des AASQA, dans le champ de la surveillance par modélisation et estimation objective de polluants réglementés.



Agence de l'Environnement et de la Maîtrise de l'Énergie
Direction de l'Air, du Bruit et de l'Efficacité Énergétique
Département de Surveillance de la Qualité de l'Air
Centre de Paris : 27, rue Louis Vicat - 75737 Paris Cedex 15
www.ademe.fr

