Elaboration de plans d'échantillonnage temporel et reconstitution de données

Guide pratique

Août 2008

Elaboration de plans d'échantillonnage temporel et reconstitution de données

Guide pratique

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

Août 2008

Préambule

Traditionnellement constitué de stations fixes équipées d'analyseurs automatiques qui mesurent en continu la qualité de l'air, le dispositif français de surveillance fait de plus en plus appel à la mesure « discontinue » à l'aide de systèmes de prélèvement et de stations mobiles. Que ce soit dans le cadre de la surveillance réglementaire ou d'études, les Associations Agréées de Surveillance de la Qualité de l'Air sont conduites à planifier des campagnes de mesure, puis à exploiter les données pour déterminer les indicateurs de la qualité de l'air d'une zone.

Afin d'harmoniser les pratiques d'échantillonnage temporel et de reconstitution de données des AASQA, le Comité de Programmation Technique a proposé en 2005 de mettre en place un groupe de travail. Ce groupe avait pour mission de proposer des démarches d'échantillonnage temporel et de reconstitution des données applicables dans l'ensemble des AASQA.

Le groupe a fait un état des lieux des pratiques qui existent dans les AASQA et dans d'autres Etats Membres et, partant des méthodes déjà employées, a préparé ce guide et développé des outils numériques associés.

En 2008, une série de formations permettra de diffuser les méthodes et outils aux AASQA et d'identifier leurs éventuels besoins supplémentaires.

Le groupe de travail est constitué de représentants de six AASQA (Air Normand, ATMO Poitou-Charentes, ATMO Nord–Pas de Calais, Air Pays de la Loire, ORAMIP et ATMO Auvergne) et bénéficie du support scientifique du LCSQA (INERIS et Ecole des Mines de Douai). Le secrétariat est assuré par l'INERIS. Je remercie l'ensemble des membres du groupe pour leurs contributions.

Céline Phillips
ADEME
Animatrice du Groupe de travail « Plans d'échantillonnage et reconstitution de données »

Table des matières

Liste des annexes

1. Etat des pratiques en Europe
2. Etat des pratiques en France
3. Fiche descriptive de la méthode des plans de sondage
4. Fiche descriptive de la méthode issue de la norme ISO
5. Fiche descriptive de la régression
6. Note théorique sur la corrélation
7. Définition de l'intervalle de confiance
8. Formulaire
9. Etude d'impact économique réalisée par AIR Pays de Loire
10. Application : ***Echantillonnage temporel des HAP pour l'estimation d'une moyenne annuelle***

Introduction Générale

L'objectif de ce document est de fournir un guide pratique pour toute AASQA souhaitant, à l'occasion de la réalisation de mesures non permanentes, 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 par modélisation et estimation objective de polluants réglementés. 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, pour certains polluants précisés dans les directives européennes (particules, benzène, HAP, plomb et autres métaux lourds), la mesure fixe 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 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 national 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 Etats 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. ELABORATION D'UN PLAN D'ECHANTILLONNAGE

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).

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.

Polluant	SO_2 NOx 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				Egale 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				Egale 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.

1.2 Terminologie

1.2.1 Qu'est-ce qu'un échantillonnage ?

1.2.1.1 Echantillonnage

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

1.2.1.2 Echantillonnage systématique

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

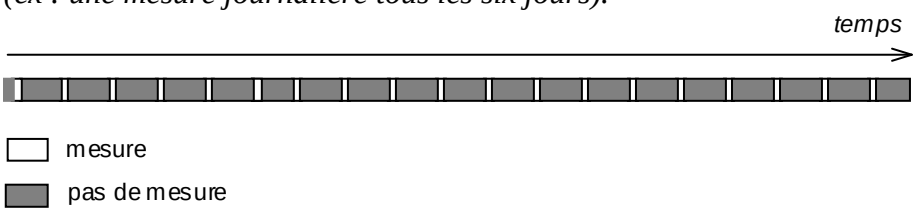


Figure 1– Illustration d'un échantillonnage systématique

1.2.1.3 Echantillonnage 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).*

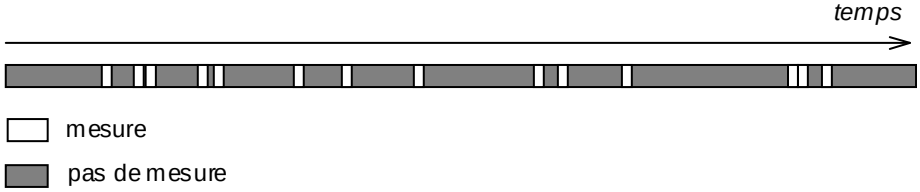


Figure 2 – Illustration d'un échantillonnage aléatoire simple

1.2.1.4 Objet de l'échantillonnage

A 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**.

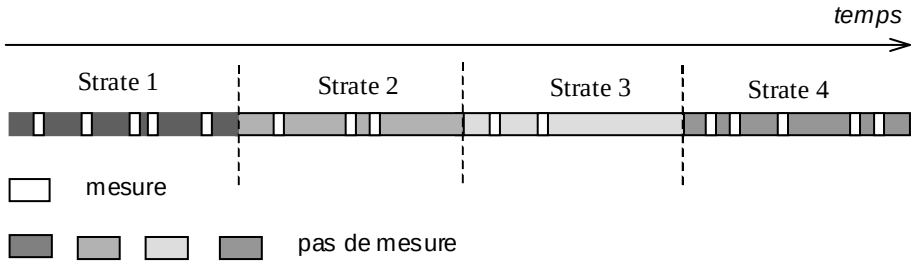


Figure 3 – Illustration d'un plan d'échantillonnage aléatoire stratifié, où les strates sont des périodes de temps continues et de même longueur.

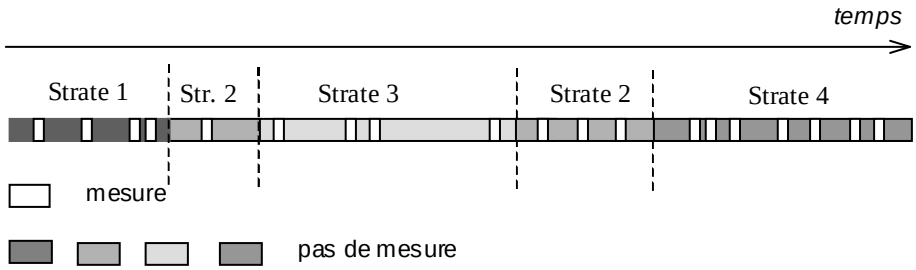


Figure 4 – Illustration d'un plan d'échantillonnage aléatoire stratifié, où les strates sont des périodes de temps continues ou discontinues et de longueurs différentes.

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.

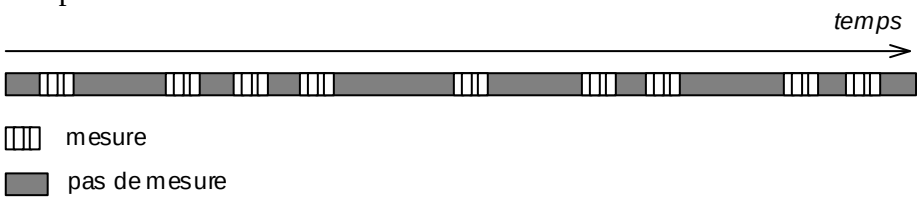


Figure 5 – Illustration d'un plan d'échantillonnage par grappes.

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.

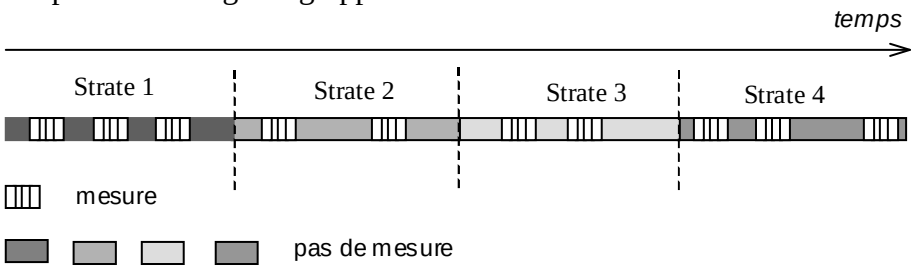


Figure 6– Illustration d'un plan d'échantillonnage par grappes stratifié.

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.

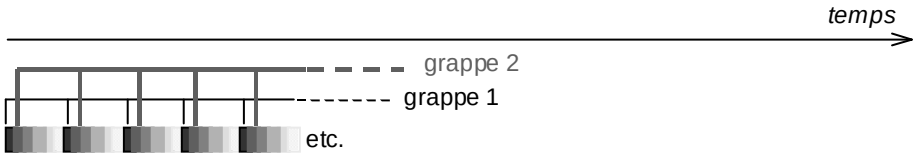


Figure 7 – Représentation d'un échantillonnage systématique. Tirage d'une mesure tous les six jours = sélection d'une grappe parmi six.

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.

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

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**. A 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 8760 ou de 8784 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**
Echantillonnage 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.
Echantillonnage 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

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 a 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 Evaluation *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 au point 4 du paragraphe 1.3.5.1.).

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

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

L : longueur de l'intervalle de confiance

$t_{\nu,1-\alpha}$: coefficient de Student d'ordre $1-\alpha$ à ν 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.

<u>*Exemples :*</u>

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 Etapes 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 Etude 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.

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

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
A contrôler	Stratégie acceptable du point de vue des ressources ?	Acceptable du point de vue de la qualité ?

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) Etude 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 :

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

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. (A titre d'information, on a fait également figurer sur ces graphiques les concentrations, supposées inconnues, de la station Mazades.)

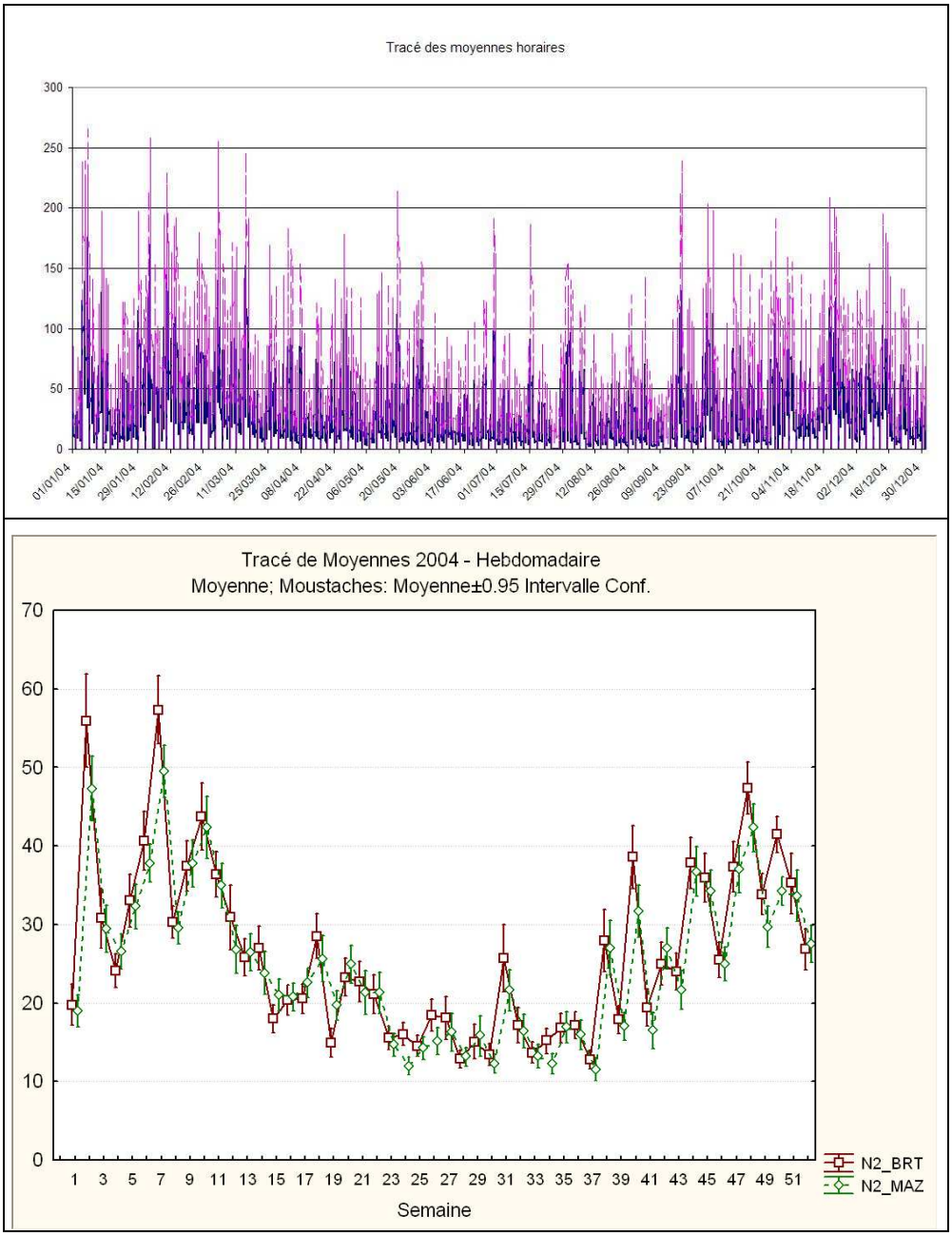


Figure 8 – Séries des moyennes horaires (graphique du haut) et hebdomadaires (graphique du bas. Les valeurs en ordonnée sont des concentrations en µg/m^3.

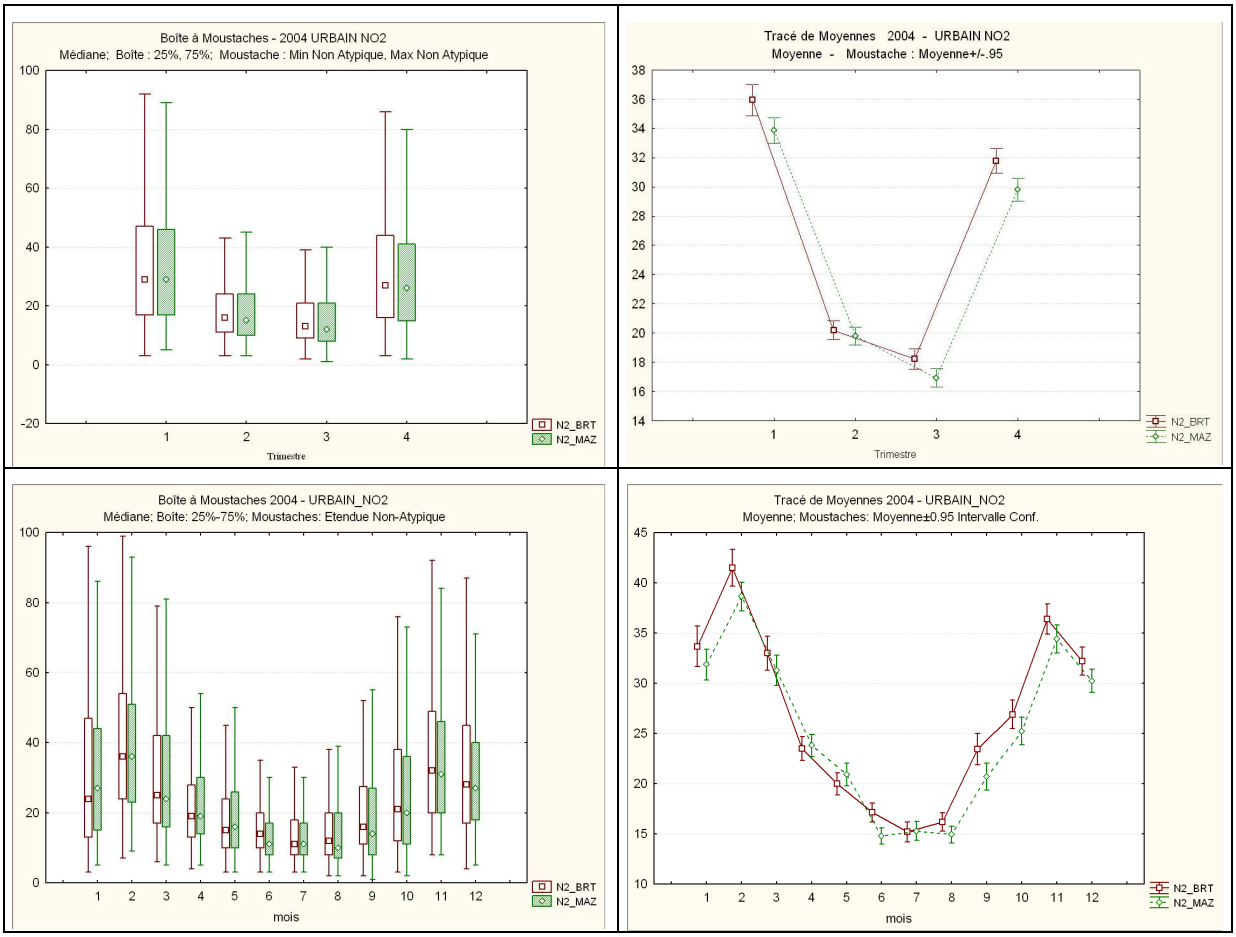


Figure 9 –Tracé des boîtes à moustaches (graphiques de gauche) et des moyennes (graphique de droite) trimestrielles et mensuelles. Les valeurs en ordonnée sont des concentrations en µg/m^3.

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) Etude 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.

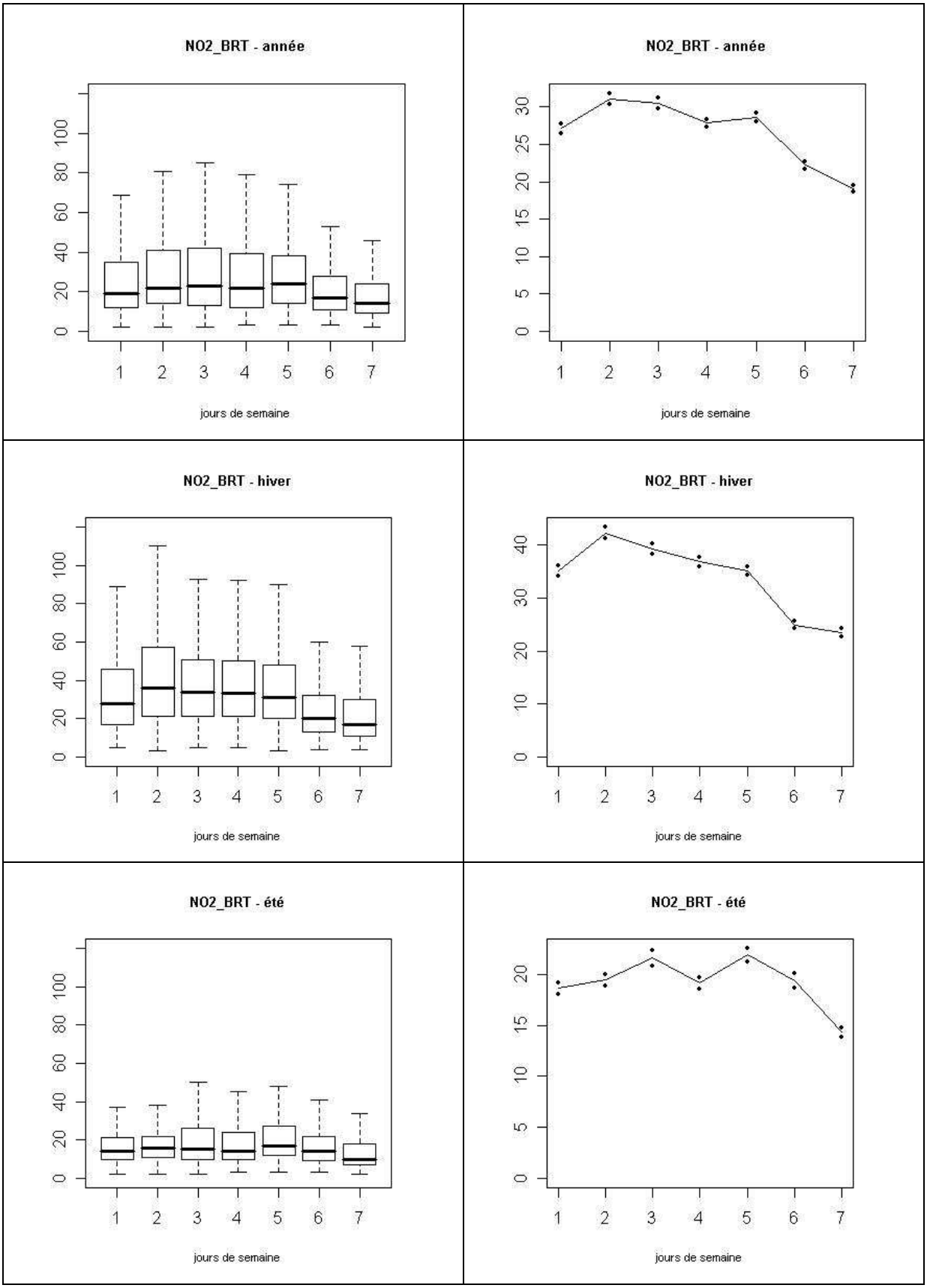


Figure10 - Tracé des boîtes à moustaches (graphiques de gauche, hors valeurs atypiques) et des moyennes (graphique de droite) par jour de semaine pour l'année complète, la saison hivernale et la saison estivale. Les numéros en abscisse sont les jours de la semaine (6 et 7 = samedi et dimanche). Les valeurs en ordonnée sont des concentrations en µg/m^3.

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.

Reprenons l'exemple du site urbain de Toulouse (§ 1.3.3**.)**
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 µg/m^3, respectivement de 10%, à un niveau de confiance de 95% signifie que si $\hat{\bar{y}}$ est la concentration moyenne annuelle de NO_2 reconstituée au site urbain Mazades et $IC_{95\%}$, l'intervalle de confiance à 95% autour de $\hat{\bar{y}}$, on souhaite avoir $IC_{95\%} = \hat{\bar{y}} \pm 4$, respectivement $IC_{95\%} = \hat{\bar{y}} \pm 0{,}1.\ \hat{\bar{y}}$. 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 $0.1 = 1.96 \cdot \sqrt{V} / \hat{\bar{y}}$

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 :

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

   $S_h^2 = \frac{1}{M_h - 1} \sum_{j=1}^{M_h} (y_{h,j} - \bar{y}_h)^2$: 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

   $S_h = \sqrt{S_h^2}$

A 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

Rem. : 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 NO_2 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. A 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 1er trimestre, transition, été, hiver 4e 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. A 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) :

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

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

$Var(\hat{\bar{y}})$: 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, i.e. 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 1.2.2.5, 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 :

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

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

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. §1.5). 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%.

Le paragraphe 1.5 illustre sur des exemples cette étape d'évaluation.

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 Evaluation 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 tout 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 Evaluation 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	4h45				
Temps technicien nécessaire	1	journée	x	2	personnes

EVALUATION PRELIMINAIRE :	**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	2176	euros (0.4€/km)			
Analyse	708	euros			
Pour 3 années	**38 854 €**	+ 1 partisol utilisé pour chaque série de 2 semaines			

EVALUATION PRELIMINAIRE :	**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	5h20				
Temps technicien nécessaire	1	journée	x	2	Personnes

EVALUATION PRELIMINAIRE :	**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			

EVALUATION PRELIMINAIRE :	**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	1184	euros (0.4€/km)			
Analyse	708	euros			
Pour 3 années	**20 777 €**	+ 2 partisols utilisés pour chaque série de 2 semaines			

1.5.3 Evaluation 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. A 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'autre 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 (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).

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

	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**

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.

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

	UO terrain (en jours)	**Déplacement (en euros)**	**Analyses chimiques (en euros)**	**Exploitation (en euros)**
8 x 1.5 mois	+3	+ 800	- 7000	**- 1700**
4 x 3 mois	- 34	+ 500	- 7000	**- 1700**
2 x 6 mois	- 52	+ 300	- 7000	**- 1700**
1 x 12 mois	- 61	+200	- 7000	**- 1700**

[1] *Guidance on Assessment under the EU Air Quality Directives – Final Draft -*

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 PARAMETRES 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.

A 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 :
A 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'échantillonnage 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 (≈ 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 considèrera 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.*

2.2 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] **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.**

2.2.2 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.
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 :

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

  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 :

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

  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

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

---

[3] Les simulations d'échantillonnage réalisées portent sur les situations suivantes : surveillance des polluants NO2, O3, PM10, PM2.5 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.

[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.

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 **redressement** décrit au paragraphe 2.2.2.2. Il existe plusieurs méthodes de redressement et par suite, plusieurs estimateurs redressés.

A 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 au paragraphe 2.2.2.4.

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 :

$\hat{N}_{>s} = \hat{T}_{>s} \cdot 8760$

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

$\hat{N}_{>s}$: 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, i.e., 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 Ts_1 et Ts_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 Ts_n'$<2\%<Ts_n$, 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;

$$\hat{\bar{y}}_{quot} = \sum_{s \in S_P} \cdot \frac{N_S}{N} \cdot \left(\frac{\hat{\bar{y}}_{pi\ s}}{\hat{\bar{x}}_{pi\ s}} \cdot \bar{x}_S \right)$$

- 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;

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

- 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;

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

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 NO_2) 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 considèrera 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 NO_2 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 NO_2 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 fournies en annexe 8.

L'intervalle de confiance à 95% autour de la moyenne estimée vaut $\left[\hat{\bar{y}} - 1{,}96\sqrt{V}; \hat{\bar{y}} + 1{,}96\sqrt{V}\right]$, ce qui représente une incertitude relative de $\frac{100 \cdot 1.96 \times \sqrt{V}}{\hat{\bar{y}}}$.

<u>Attention</u> : il s'agit d'un intervalle <u>estimé</u> 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.

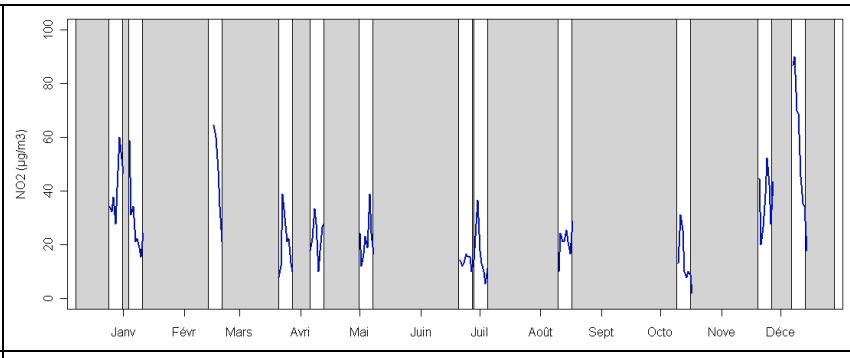
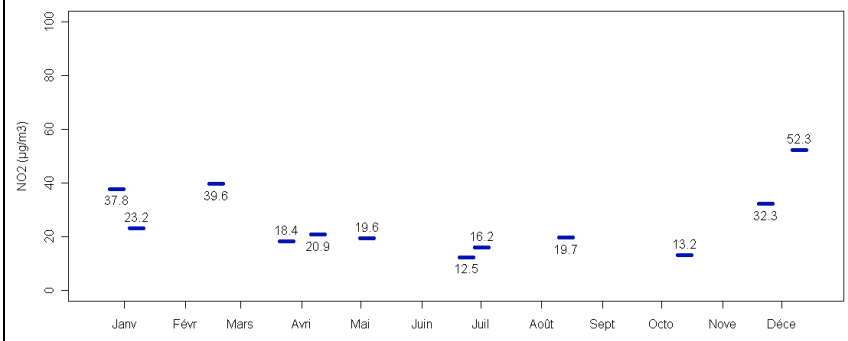
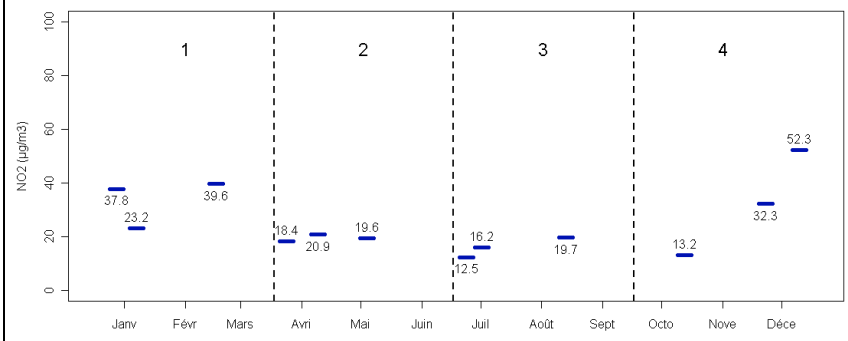
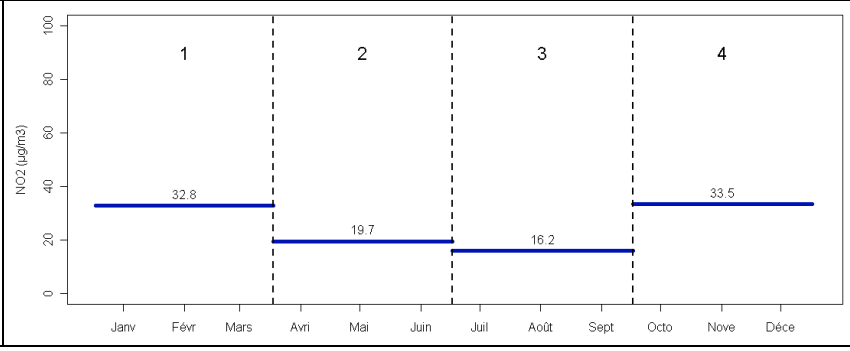
Mesures enregistrées pendant les 12 semaines de campagne	
Calcul des concentrations moyennes sur chacune des 12 grappes	
Prise en compte de la stratification temporelle du plan d'échantillonnage	
Estimation des concentrations moyennes dans chaque strate temporelle	

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.

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.

Grappes			Moyennes mesurées[2] (µg/m³)		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 $\hat{\bar{y}}_{pis}$)

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.

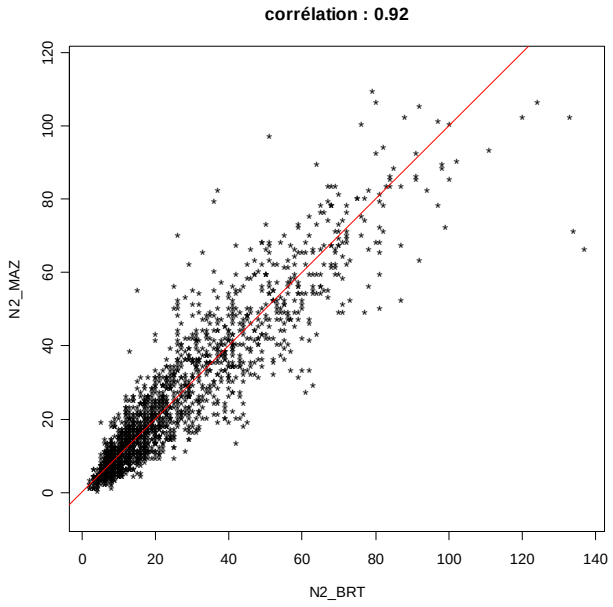


Figure 12 – Nuage de corrélation entre les concentrations mesurées sur le site de Mazades et les concentrations mesurées sur le site de Berthelot pendant la campagne de mesure

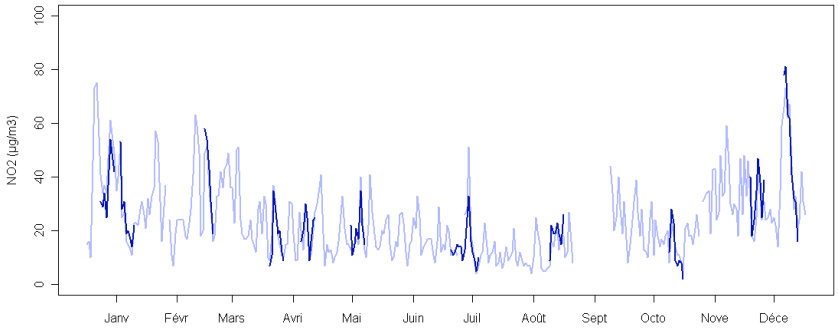


Figure 13 – Concentrations mesurées sur le site de Mazades (bleu foncé) et sur le site fixe de Berthelot (bleu clair)

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_1sur la strate entière.

Tableau 7 – Moyennes trimestrielles au site Berthelot (X_1) calculées respectivement sur trois mois complets de données et sur trois grappes hebdomadaires. Comparaison avec le site Mazades (Y).

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 $\hat{\bar{x}}_{1pis}$ et $\hat{\bar{y}}_{pis}$)

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/m^3). 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.

Tableau 8 – Application de la méthode des plans de sondage à l'exemple « fil rouge ». Estimation de la moyenne annuelle de NO_2. Moyennes par strate temporelle sans et avec redressement et moyennes annuelles. Résultats numériques.

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	5.4%
19.7			18.8			18.8			20.2		
16.2			14.5			14.5			15.1		
33.5			30.3			30.0			30.4		

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 – Echantillonnage 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.

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

	MOYENNES				**DEPASSEMENTS DE SEUILS**			
Polluants	**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

A 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 : **Σ = 0 à 4** .

- On définit ensuite 4 strates dans lesquelles sont ventilées les données horaires:

Strate 1 : **Σ = 0 ,** Strate 2 : **Σ = 1 ,** Strate 3 : **Σ = 2** et Strate 4 **: Σ = 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 m/s,
- lorsque la température est < 12°C,
- de 7 à 11h et de 16 à 21h,
- 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).

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.

[NO_2]	**v< 2 $m.s^{-1}$**	**T<10°C**	**Lundi au vendredi**	**6h à 9h et 15h à 21h**	**Σ**	**N° de strate i**
élevée	✓	✓	✓	✓	4	4
	✓	✓		✓	3	4
		✓		✓	2	3
				✓	1	2
faible					0	1

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. A 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 séquence 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.

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

  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:

  $$\hat{\bar{y}} = \sum_i \cdot \frac{N_i}{N} \cdot \hat{\bar{y}}_i$$ (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 (2.2.2.2).

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 NO_2 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 NO_2 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 (2.2.2.2).

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

L'intervalle de confiance à 95% autour de la moyenne estimée vaut $\left[\hat{\bar{y}} - 1{,}96\sqrt{V};\ \hat{\bar{y}} + 1{,}96\sqrt{V}\right]$, ce qui représente une incertitude relative de $\frac{100 \cdot 1.96 \times \sqrt{V}}{\hat{\bar{y}}}$.

Attention : il s'agit d'un intervalle <u>estimé</u> 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 <u>strates paramétriques</u> 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 au chapitre 2.2.2), cette station est choisie comme station de redressement. Quatre strates paramétriques sont considérées, ainsi qu'il a été défini au paragraphe 2.2.3.1.

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.

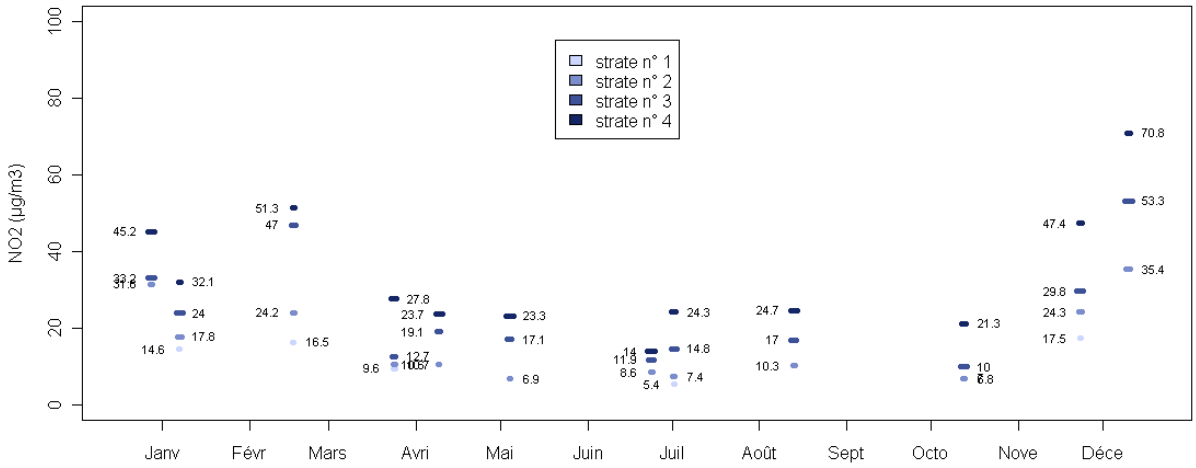


Figure 14 – Répartition des données de grappes entre les quatre strates paramétriques.

Les deux tableaux suivants présentent les résultats de mesure et d'estimation. Après redressement, les moyennes par strate sont plus proches des valeurs réelles. Il en va de même pour la moyenne annuelle (égale à 24,4 µg/m^3). Le gain de précision lié au redressement se traduit par une diminution de l'incertitude estimée.

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.

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 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.

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		

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 nécessaires 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 :

$$Y = \beta_0 + \beta_1.X_1 + \beta_2.X_2 + \ldots + \beta_k.X_k + \varepsilon$$

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

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 :

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

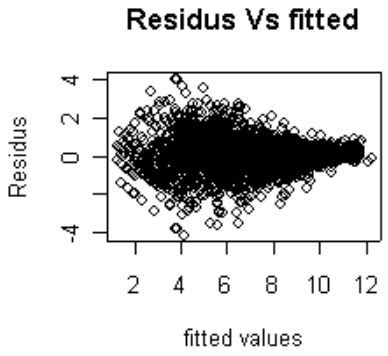
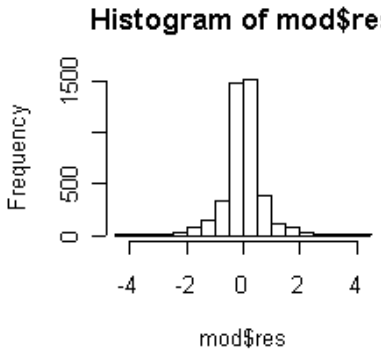
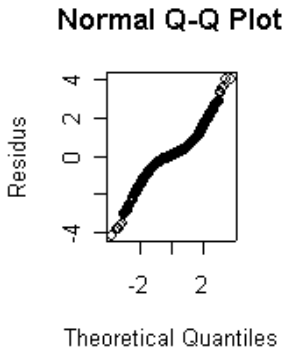
Remarque : Lorsque le modèle de régression possède un terme constant β_0, les séries mesurée et reconstituée ont même moyenne sur la période d'échantillonnage. La valeur de $\hat{\bar{y}}$ n'est donc quasiment pas modifiée si sur la période de campagne, les prédictions $(\hat{y}_i)_{i\in \text{campagne}}$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

<u>Estimation d'un nombre de dépassements de seuil et d'un centile</u>

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).

Tableau 13 – Etapes d'une reconstitution par régression

<table>
<tr><td colspan="3">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.</td></tr>
<tr><td colspan="3">Estimation des paramètres du modèle :
Les coefficients $(\beta_i)_{0\leq i\leq k}$ sont généralement estimés par la méthode des moindres carrés.</td></tr>
<tr><td colspan="3">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.</td></tr>
<tr><td> 
</td><td> 
</td><td> 
</td></tr>
<tr><td>Résidus en fonction des valeurs prédites. Si le critère d'homoscédasticité est vérifié, la dispersion du nuage est indépendante des valeurs prédites.</td><td>Histogramme des résidus</td><td>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.</td></tr>
<tr><td colspan="3">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.</td></tr>
</table>

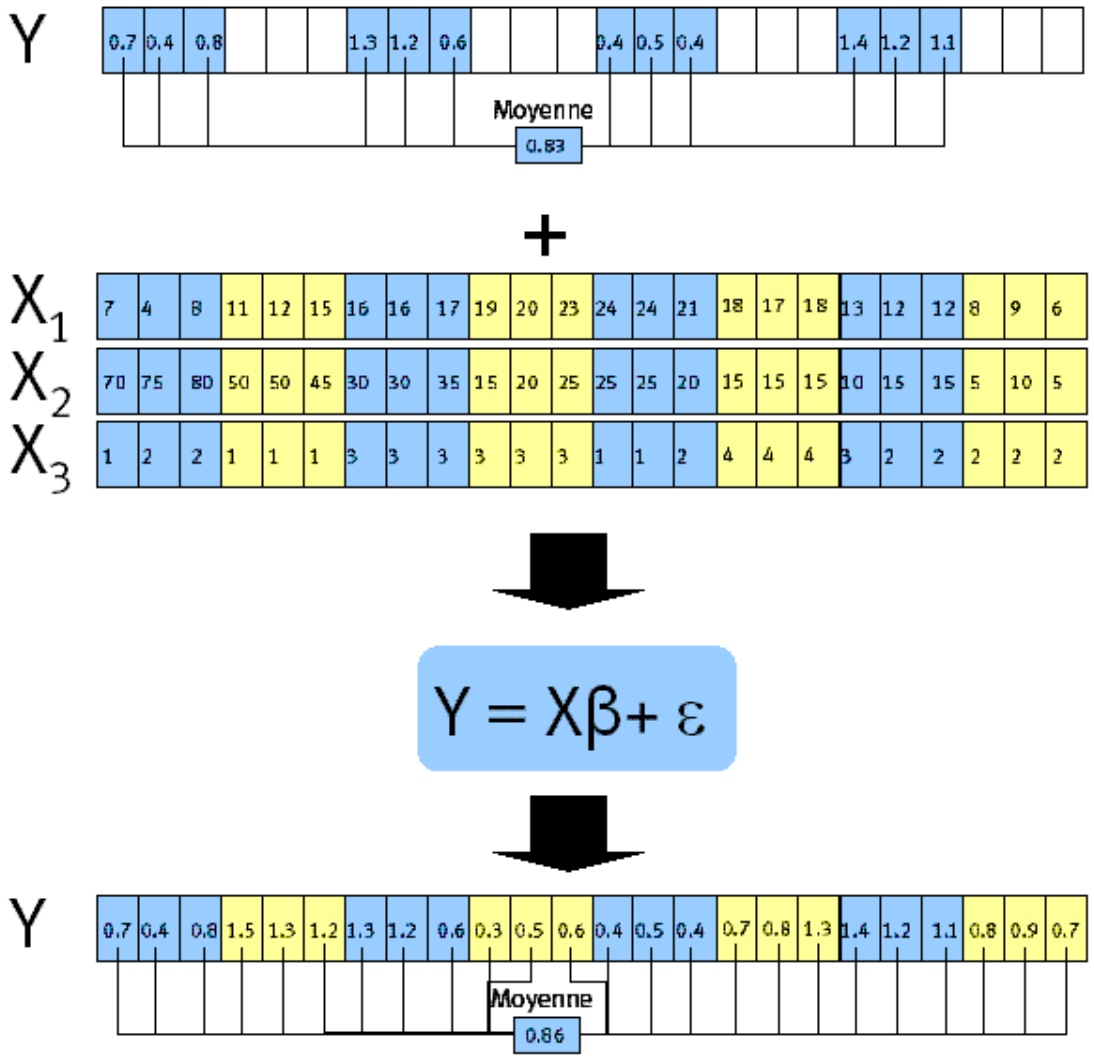


Figure 15 – Principe de la reconstitution par régression. Y est la variable à expliquer. X est la matrice des variables explicatives (X1, X2 et X3 dans cet exemple). β est le vecteur des paramètres du modèle de régression et ε est un résidu aléatoire.

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 NO_2 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

Incertitude 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 :

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

$x'_{p,i} = (1, x_{1,i}, x_{2,i}, ..., x_{k,i})$: 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 :

$$\hat{\bar{y}} = \bar{x}_p \cdot \hat{\beta}$$

$\bar{x}'_p = (1, \bar{x}_1, \bar{x}_2, ..., \bar{x}_k)$: 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) :

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

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

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 :

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

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

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 au chapitre 2.2.2), 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 $Y = \beta_0 + \beta_1.X_1$. 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.

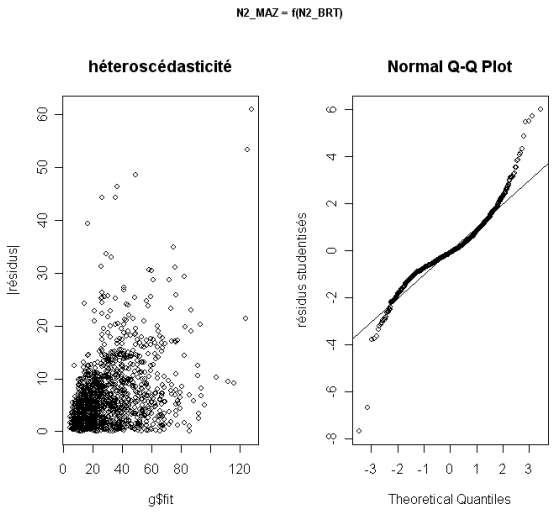


Figure 16 – Etude de l'homoscédasticité et de la normalité des résidus.

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 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 µg/m^3**. Elle est proche de la moyenne réelle qui vaut 24,40 µg/m^3. L'intervalle de confiance à 95% autour de la moyenne vaut [24.07 µg/m^3; 24.92 µg/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=NO_BRT, X_3=O3_BRT, X_4=S2_BRT), la direction de vent, la vitesse du vent, les précipitations et la température à la station périurbaine de Colomiers (X_5=DV_COL, X_6=VV_COL, X_7=PR_COL, X_8=T_COL). Le modèle de régression est de la forme :

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

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, 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 :

N2_MAZ = 300,338 + 0,766.NO2_BRT + 0,027.NO_BRT - 0,072.O3_BRT+ 0,383.S2_BRT + 0,007.DV_COL - 0,563.VV_COL -0,288.PR_COL - 0,175.T_COL

L'estimation de la moyenne annuelle reste proche de la valeur réelle : elle est maintenant de **25.07** µg/m^3. Son intervalle de confiance à 95% a pour bornes [24.62 µg/m^3; 25.52 µg/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

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.

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

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.

3. REFERENCES

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, 2ème édition révisée et augmentée, Technip, 2006.

Tillé Y. Théorie des sondages. Echantillonnage et estimation en populations finies. Cours et exercices avec solutions, 284 pages, Editions 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.

ANNEXES