Plan de sondage pour mesures mobiles de la pollution atmosphérique

Sampling plan for moving measures in atmospheric pollution

Frédéric LAVANCIER*, Fabrice CAINI**, Alain GAZEAU**

Résumé

Le but de cet article est de répondre à la problématique suivante : comment effectuer des mesures durant l'année afin d'estimer au mieux la moyenne d'un polluant (typiquement le NO_2) sans se contraindre d'un relevé permanent ? Nous nous appuyons sur la théorie des sondages pour proposer un protocole de mise en œuvre d'une campagne de mesures mobiles sur l'année selon la précision d'estimation souhaitée. Nous proposons alors différentes méthodes d'estimation de la moyenne annuelle du polluant et de son taux de dépassement de seuil. Ces estimations peuvent être effectuées uniquement à partir des mesures réalisées ou s'appuyer plus subtilement sur un site fixe auxiliaire supposé avoir un comportement assez proche. Dans tous les cas, nous calculons un intervalle de confiance autour de la valeur estimée avec un taux de confiance préalablement choisi. Cette méthode, pour son application, ne nécessite que les mesures du polluant sur le site à explorer grâce à des tubes passifs ou un camion-laboratoire. Aucune hypothèse sur le type de polluant ou sur le site à étudier n'est faite.

Abstract

In this paper we deal with the problem of moving measures. How can we estimate the annual average of a substance (e.g. NO_2) only thanks to a few measures throughout the year ? We present the random sampling theory which gives us a way to handle a sampling plan depending on the estimation accuracy that we wish to obtain. Then we estimate the annual average or the proportion of measures greater than a certain value, by different ways : either only thanks to the sample, or involving an auxiliary site. In every case, a confidence interval is implemented, whose confidence rate can be previously chosen. This method only requires the substance measures on the site, and no other hypothesis are done on the kind of substance or site we work with.

Introduction

La surveillance de la qualité de l'air s'est largement développée sur les grandes agglomérations françaises depuis la loi sur l'air et l'utilisation rationnelle de l'énergie (LAURE) de 1996. La plupart de ces agglomérations répondent à présent favorablement aux exigences de surveillance tant par le nombre de points de mesures que de polluants qui y sont suivis.

L'article 1 de la LAURE souligne le droit à chacun de respirer un air qui ne nuise pas à sa santé ; deux problématiques se posent pour la mise en place de stratégies de surveillance :

- affiner la connaissance des niveaux de concentrations sur certaines zones particulières d'une agglomération, proches de grandes voies de circulation par exemple ;
- fournir des informations sur la qualité de l'air sur des zones non couvertes par le réseau de mesures automatiques.

Devant le coût d'investissement et de fonctionnement des stations de mesures automatiques, il

* *Laboratoire de Statistique et Probabilités, Université des Sciences et Technologies de Lille, Villeneuve-d'Ascq.*
E-mail : lavancier@lps.univ-lille1.fr

** *ATMO Poitou-Charentes, surveillance de la qualité de l'air, La Rochelle.*

n'est pas possible d'en multiplier le nombre. À titre d'exemple, en Poitou-Charentes, seuls 27 % de la population ont accès à une information de la qualité de l'air relative à leur lieu de résidence.

La modélisation déterministe (CHIMÈRE, MOCAGE, par exemple) est dans certains cas une solution pour la connaissance de la qualité de l'air en zone non couverte. Mais l'application de ces modèles nécessite de gros moyens de mise en place, et une bonne connaissance préalable du site d'étude.

L'approche développée ici a été initiée lors d'un travail sur les mesures mobiles effectué pour un stage au cours de l'été 2001 à ATMO Poitou-Charentes. Elle consiste à utiliser les résultats issus de campagnes de mesures mobiles par camion-laboratoire ou tubes passifs. Cette approche, basée sur la théorie des sondages aléatoires, permet à la fois de planifier la campagne de mesures (durée et fréquence de passage) en fonction de la précision souhaitée, et de faire l'estimation des valeurs réglementaires avec son incertitude.

Elle est une alternative à l'effort de modélisation lorsque l'on souhaite éviter des relevés par station fixe, car elle permet une estimation de la qualité de l'air sans faire appel à une connaissance pointue du site à explorer ni du polluant à mesurer.

Avec cette méthode, et en attribuant une unité mobile de mesures à la surveillance des villes de plus de 10 000 habitants pendant deux ans (équivalent en coût à une station automatique de mesures), il serait possible d'accroître de 40 % le nombre de personnes ayant accès à une information sur la qualité de leur air en Poitou-Charentes (soit 38 % de la population totale).

Nous présentons dans une première partie à visée pédagogique les différentes notions couramment manipulées en sondage et le principe de fonctionnement d'un sondage aléatoire.

Bien que très utilisés en tant que techniques d'enquêtes (à l'INSEE notamment), les sondages aléatoires sont inédits dans le domaine des mesures de pollution atmosphérique. Ainsi, dans une deuxième partie, nous expliquerons les choix que nous avons effectués pour notre stratégie de sondage : ils répondent à des contraintes spécifiques liées à ce domaine, mais aussi à des exigences de simplicité et de large utilisation.

Nous présenterons ensuite les plans de sondages retenus, les plans par grappes stratifiés, en expliquant leur mise en œuvre pratique. Nous verrons alors comment estimer au mieux la moyenne d'un polluant avec un intervalle de confiance, en se servant éventuellement d'informations auxiliaires. Nous aborderons de même l'estimation de la proportion de données au-dessus d'un certain seuil.

Nous terminerons par des simulations et la présentation d'un cas pratique réalisé à ATMO Poitou-Charentes, qui illustrent la démarche et ses mises en garde.

Présentation des sondages aléatoires

Cette partie a pour objectif de familiariser le lecteur non spécialiste avec les sondages aléatoires. Nous y présentons la démarche générale d'un sondage, son enjeu et les paramètres sur lesquels on peut jouer pour optimiser l'estimation de la moyenne annuelle. Puis nous expliquons comment on peut utiliser les données pour proposer un intervalle de confiance contenant très probablement la vraie moyenne. Enfin, nous donnons un exemple simple de plan de sondage pour illustrer nos propos.

Estimateur et caractère aléatoire

Lorsque l'on fait une série de mesures de NO_2 durant l'année, on obtient un échantillon de mesures à partir duquel on peut estimer la moyenne annuelle. Il n'est question que d'estimation, la vraie moyenne est, elle, inaccessible si on n'a pas effectué de mesures exhaustives. Pour entreprendre une estimation, on utilise un estimateur, c'est-à-dire une quantité calculée à partir des données de l'échantillon qui estime le caractère d'intérêt. Dans notre cas, l'estimateur est la moyenne des données de l'échantillon. C'est cet estimateur qui estime au mieux la vraie moyenne annuelle. Ce n'est pas toujours aussi simple : par exemple, si on veut estimer la variance annuelle, l'estimateur le plus judicieux n'est pas de calculer la variance sur les données de l'échantillon (nous reviendrons sur l'estimateur de la variance ultérieurement, nous en aurons besoin).

On réalise un bon sondage lorsque l'estimateur est proche de la vraie valeur. Dans notre cas, nous voulons que l'estimateur de la moyenne soit le plus proche possible de la vraie moyenne annuelle. Mais le plus important est de contrôler l'erreur que l'on a pu commettre en estimant cette moyenne, c'est-à-dire de pouvoir donner un intervalle de confiance autour de la valeur que l'on avance.

Cette exigence ne peut être réalisée que dans le cadre de la théorie des sondages aléatoires, c'est-à-dire lorsque l'échantillon de mesures est tiré de façon aléatoire (mais pas n'importe comment, nous y reviendrons). Une autre façon de concevoir un échantillon aurait été de fixer par avance les dates de mesures, de façon déterministe, en utilisant son bon sens et son expérience du milieu d'étude (par la méthode des « quotas » par exemple). Cette approche n'est pas retenue ici car aucun cadre théorique ne la légitime, et, en outre, on ne peut pas contrôler l'erreur que l'on a pu commettre en estimant la moyenne à partir de l'échantillon de mesures choisi. Ceci dit, il est clair que la connaissance du milieu sera nécessaire pour optimiser la façon de tirer aléatoirement les données et de les exploiter.

Ainsi, toutes les méthodes exposées ici, et toutes les formules utilisées découlent de la théorie des sondages et de son approche probabiliste. L'objectif n'étant pas d'entrer dans les détails des probabilités, nous nous contenterons de donner les formules utiles.

Le lecteur pourra trouver les justifications théoriques des résultats dans l'ouvrage de Y. Tillé [1] et celui de J.J. Droesbeke, B. Fichet et P. Tassi [2].

Paramètres à optimiser pour avoir un bon plan de sondage

Le plan de sondage spécifie la façon dont on tire les données. C'est un choix complexe dont l'enjeu principal est de fournir un bon échantillon de données tout en tenant compte des contraintes techniques du milieu (nombre de relevés possibles, coût…).

La quantité qui nous intéresse est la moyenne annuelle de NO_2 notée $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ où les y_i sont les mesures de NO_2 durant l'année ; il y en a *N*. « $\bar{y}$ » est une notation courante pour la moyenne des y_i.

L'estimateur de la moyenne est le suivant :

$\hat{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n} y_i$ où les y_i sont les données de l'échantillon tiré aléatoirement selon le plan de sondage et *n* le nombre de mesures relevées (i.e. la taille de l'échantillon). De façon générale, il est courant de noter avec un « chapeau » l'estimateur associé au caractère d'intérêt.

L'échantillon est tiré de façon aléatoire, ainsi les y_i intervenant dans l'estimateur sont aléatoires, c'est-à-dire qu'ils dépendent de l'échantillon tiré et donc l'estimateur lui-même est aléatoire. Le caractère aléatoire signifie simplement que la valeur de l'estimateur n'est pas fixée, il renvoie une valeur par échantillon. Le problème majeur d'un sondage est précisément cette variabilité de l'estimateur selon les échantillons. Et comme on ne sait pas *a priori* quel est l'échantillon de données qui nous fournira la meilleure estimation de la moyenne, il convient de tenir compte du risque de tirer un échantillon pour lequel l'estimation de la moyenne est mauvaise. Ceci est d'autant plus important que l'on n'a aucun moyen de contrôle une fois les données tirées : comment savoir si l'estimation à partir de ces dernières est bonne puisqu'on n'a pas la vraie moyenne sur l'année ?

Pour contrôler ce risque, il faut contrôler la variabilité de l'estimateur selon les échantillons tirés, autrement dit sa dispersion. Si l'estimateur varie peu selon les échantillons, on pourra se fier à la valeur qu'il nous fournit quel que soit l'échantillon tiré.

Cela se traduit essentiellement à travers deux caractéristiques qui résument le comportement aléatoire de l'estimateur ; il s'agit de son espérance et de sa variance. On cherche généralement à les connaître lorsque l'on est en présence d'une quantité aléatoire. L'espérance traduit le comportement moyen de la quantité aléatoire selon les échantillons. La variance traduit la variabilité de la quantité aléatoire autour de son espérance selon les échantillons. Concrètement, si on tire un très grand nombre d'échantillons et que pour chacun d'entre eux on calcule la valeur de l'estimateur, on peut retrouver l'espérance de l'estimateur en faisant la moyenne de ces valeurs. De même, la façon dont toutes les valeurs se dispersent autour de l'espérance est directement liée à la variance : en gros 95 % des valeurs seront comprises entre l'espérance plus deux fois la racine carrée de sa variance et l'espérance moins deux fois la racine carrée de sa variance.

Dans notre cas, pour l'estimateur de la moyenne, il a été montré que son espérance est toujours égale à la vraie moyenne. Ceci est rassurant : en moyenne, notre estimateur renvoie la vraie valeur. La question est de savoir de combien il peut s'éloigner de la vraie moyenne annuelle, en somme de combien peut-on se tromper en estimant sur un échantillon particulier la moyenne : cela est quantifié par sa variance. Celle-ci dépend de la façon dont on a tiré les données (de façon indépendante ou par blocs par exemple).

Pour le choix d'un bon plan de sondage, il conviendra donc de calculer la variance de l'estimateur associée, et de choisir celui qui la minimise. C'est la démarche que nous adopterons dans la partie suivante pour proposer des plans de sondage aussi pertinents que possible.

L'intervalle de confiance

Intervalle de confiance théorique

L'avantage d'un sondage aléatoire réside dans la possibilité de donner un intervalle de confiance autour de la moyenne estimée.

On a introduit la variance de l'estimateur en indiquant que cette caractéristique quantifie sa dispersion. Le résultat exact qui justifie cette affirmation est le suivant : il y a 95 % de chances pour que l'estimation associée à un échantillon tiré de façon indépendante soit à moins de $1{,}96 * \sqrt{\text{var}(\hat{\bar{y}})}$ de la vraie moyenne sur la population, où $\hat{\bar{y}}$ est l'estimateur de la moyenne annuelle et var ($\hat{\bar{y}}$) sa variance, ce qui s'écrit :

$\hat{\bar{y}} \in \left[\bar{y} - 1{,}96 * \sqrt{\text{var}(\hat{\bar{y}})} \; ; \; \bar{y} + 1{,}96 * \sqrt{\text{var}(\hat{\bar{y}})}\right]$ avec une probabilité de 95 %.

Inversement, nous pouvons avancer le résultat suivant. Il propose, à partir de l'échantillon, un intervalle contenant la vraie moyenne avec 95 % de chances :

$\bar{y} \in \left[\hat{\bar{y}} - 1{,}96 * \sqrt{\text{var}(\hat{\bar{y}})} \; ; \; \hat{\bar{y}} + 1{,}96 * \sqrt{\text{var}(\hat{\bar{y}})}\right]$ avec une probabilité de 95 %.

C'est ce dernier intervalle qui est appelé intervalle de confiance à 95 %.

On peut comprendre intuitivement ce que signifie le « 95 % de chances ». Si on tirait un très grand nombre d'échantillons et que pour chacun d'entre eux on calculait l'intervalle de confiance associé, on constaterait que 95 % de ces intervalles contiennent la vraie moyenne annuelle. En tirant un seul échantillon, on a donc bien 95 % de chances que la moyenne soit dans l'intervalle de confiance proposé.

De façon plus générale, l'intervalle de confiance à (1–α)*100 %, avec α entre 0 et 1, s'écrit de la manière suivante :

$$IC(1-\alpha) = \left[\hat{\bar{y}} - z_{1-\alpha/2}\sqrt{\text{var}(\hat{\bar{y}})} \; ; \; \hat{\bar{y}} + z_{1-\alpha/2}\sqrt{\text{var}(\hat{\bar{y}})}\right]$$

Où $z_{1-\alpha/2}$ est le quantile d'ordre 1–α/2 d'une loi normale centrée réduite (on trouve ce quantile dans des tables). Par exemple pour α = 0,05 (i.e. un intervalle de confiance à 95 %) $z_{1-\alpha/2}$ =1,96, pour α = 0,1 (i.e. un intervalle de confiance à 99 %) $z_{1-\alpha/2}$ =1,64.

Ces résultats sont classiques en théorie des probabilités et sont valables pour un échantillon de grande taille tiré de façon indépendante. En pratique, dès que la taille de l'échantillon est raisonnablement grande (plus de 30 mesures indépendantes), on peut les utiliser. Sans cela, on risque de perdre de la précision, c'est-à-dire qu'un intervalle de confiance annoncé à 95 % ne sera peut être qu'à 90 %, mais l'intervalle ne perd pas toute signification pour autant.

Intervalle de confiance estimé

Il y a pourtant un problème majeur à l'intervalle de confiance que nous avons proposé : ce dernier fait intervenir la variance de l'estimateur. Or, nous le verrons par la suite, cette variance dépend des données sur toute la population, données dont nous ne disposons évidemment pas.

En pratique, pour pouvoir avancer un intervalle de confiance autour de la moyenne, il conviendra donc d'estimer également la variance de l'estimateur à l'aide de l'échantillon tiré. Le résultat que nous pourrons alors donner sera un intervalle de confiance estimé.

La démarche adoptée pour estimer la variance de l'estimateur est la même que celle expliquée pour estimer la moyenne : on propose un estimateur qui a de bonnes propriétés. Dans le cas de la moyenne, le meilleur estimateur est la moyenne sur l'échantillon, pour la variance d'un estimateur c'est plus compliqué. En effet, la variance de l'estimateur de la moyenne est différente selon le plan de sondage que l'on propose, son estimateur sera donc différent à chaque fois. On donnera par la suite, pour chaque estimateur de la moyenne proposée, sa variance théorique var($\hat{\bar{y}}$), ainsi que son estimateur $\hat{\text{var}}(\hat{\bar{y}})$, distingué par son « chapeau ».

L'intervalle de confiance estimé à (1–α)*100 % est donc :

$$IC(1-\alpha) = \left[\hat{\bar{y}} - z_{1-\alpha/2}\sqrt{\hat{\text{var}}(\hat{\bar{y}})} \; ; \; \hat{\bar{y}} + z_{1-\alpha/2}\sqrt{\hat{\text{var}}(\hat{\bar{y}})}\right]$$

Un exemple de plan de sondage simple

Présentation du plan de sondage simple

Nous allons dans cette partie simuler un exemple de plan de sondage afin de manipuler les différentes notions introduites précédemment.

Le plan de sondage que nous choisissons est le plus courant : tirer de façon indépendante *n* individus dans la population. Dans notre cas cela revient à tirer *n* quarts d'heure dans l'année de façon indépendante et d'y mesurer la concentration de NO_2, notée *y*. Comme nous nous intéressons à la moyenne annuelle, notée $\bar{y}$, l'estimateur utilisé sera de façon naturelle la moyenne des concentrations sur l'échantillon tiré, noté $\hat{\bar{y}}$.

Nous rappelons que $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ et $\hat{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n} y_i$, où *N* est la taille de la population, soit le nombre de quarts d'heure dans une année dans notre cas.

On a vu qu'il fallait s'intéresser à la variance de l'estimateur pour juger de la pertinence du plan de sondage.

Dans ce cas (tirage de *n* quarts d'heure indépendants) elle vaut :

$$\text{var}(\hat{\bar{y}}) = \frac{N-n}{n.N} \cdot \frac{N}{N-1}\text{var}(y)$$

où var(*y*) est la variance totale des concentrations de NO_2 sur l'année,

soit $\text{var}(y) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$.

Remarque : comme nous l'avons évoqué plus haut, la variance de l'estimateur dépend des données sur toute l'année car il fait intervenir ici leur variance.

Ceci était prévisible car la dispersion de l'estimateur est due à la différence entre les échantillons. Si toutes les données sont très proches, tous les échantillons se ressembleront et donc l'estimateur ne variera pas beaucoup selon l'échantillon tiré. Inversement, si les données varient énormément, les échantillons issus du sondage précédent pourront être très différents, et l'estimateur variera beaucoup. Ainsi une variance faible des données impliquera une variance faible de l'estimateur.

Ce sera le cas pour les plans de sondage que nous proposerons, la variance de l'estimateur sera fortement liée à la variance des données. Puisque nous ne pouvons pas influer sur la variance des données (celle-ci est fixée), il conviendra de jouer sur la variabilité entre les échantillons. Tout l'intérêt du plan de sondage est justement de proposer une façon de tirer des échantillons pour qu'ils soient les plus « ressemblants » possibles entre eux du point de vue de l'estimateur, même si les données de base sont très variables, afin de minimiser la variance de l'estimateur. Choisir un plan de sondage pour lequel la variance de l'estimateur est faible reviendra à choisir un plan de sondage pour lequel les échantillons potentiels se ressemblent.

La variance de l'estimateur de la moyenne s'écrit souvent :

$$\text{var}(\hat{\bar{y}}) = \frac{N-n}{n.N}.S^2$$

où S^2 s'appelle la variance corrigée de *y,*

$S^2 = \frac{N}{N-1}\text{var}(y)$, c'est une quantité que nous utiliserons par la suite couramment.

Pour ce plan de sondage, il n'y a pas grand-chose à choisir. La taille de la population *N* ainsi que var(*y*) sont fixés. Le seul paramètre à choisir est donc *n,* la taille de l'échantillon que l'on va tirer. Il est clair que la variance de l'estimateur va décroître avec *n.* La situation est simple : plus *n* sera grand, plus la variance de l'estimateur sera faible et meilleur sera notre plan de sondage.

Théoriquement, *n* doit être pris le plus grand possible, mais dans la pratique sa taille est limitée par des contraintes techniques (idéalement, $n = N$ rendrait nulle la variance de l'estimateur, mais ce n'est alors plus un sondage, c'est un recensement exhaustif).

À un *n* fixé seront associées une variance et donc une erreur d'estimation. On peut tenter de prévoir cette erreur et choisir le *n* en conséquence, c'est une démarche que nous expliquerons dans la troisième partie. Dans cet exemple nous prendrons $n = 1\,000$ arbitrairement. Le plus important est d'estimer cette erreur *via* la variance une fois le sondage effectué. Il faut pour cela estimer S^2 :

$$\hat{S}^2(y) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{\bar{y}})^2$$

et l'intervalle de confiance estimé est donc :

$$IC_{95\,\%} = \left[\hat{\bar{y}} - 1{,}96\sqrt{\frac{N-n}{n.N}.\hat{S}^2}\,;\ \hat{\bar{y}} + 1{,}96\sqrt{\frac{N-n}{n.N}.\hat{S}^2}\right]$$

Applications numériques

Nous allons effectuer ce plan de sondage sur le site « Verdun » de La Rochelle pour y estimer la moyenne de NO_2 en l'an 2000.

En fait nous connaissons toutes les mesures sur ce site, nous allons procéder comme si ce n'était pas le cas puis nous vérifierons les résultats. En 2000, il y avait 32 358 mesures valides sur ce site, ainsi afin d'effectuer la comparaison on considérera que $N = 32\,358$ (au lieu de 35 136 quarts d'heure que contient l'année 2000).

On choisit $n = 1\,000$ quarts d'heure à tirer.

La vraie moyenne annuelle de NO_2 vaut 27,63 $\mu g.m^{-3}$. Sachant que la variance des données sur l'année vaut 374,92, on peut calculer la variance théorique de l'estimateur :

$$\text{var}(\hat{\bar{y}}) = \frac{N-n}{n.N}\cdot\frac{N}{N-1}\text{var}(y) = 0{,}36$$

d'où l'erreur à 95 % théorique : $1{,}96 * \sqrt{\text{var}(\hat{\bar{y}})} = 1{,}18$

Ces valeurs ne sont pas disponibles normalement car elles nécessitent toutes les mesures sur l'année pour les calculer. Nous allons effectuer le sondage et comparer.

Une fois les 1 000 quarts d'heure tirés aléatoirement de façon indépendante, il vient :

$$\hat{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n} y_i = 28{,}24,\ \hat{S}^2(y) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{\bar{y}})^2 = 370{,}95$$

d'où l'erreur à 95 % estimée : $1{,}96\sqrt{\frac{N-n}{n.N}.\hat{S}^2} = 1{,}17$

On peut donc calculer l'intervalle de confiance à 95 % estimé, et prétendre qu'il y a 95 % de chances pour que la vraie moyenne soit dans l'intervalle.

$[28{,}24 - 1{,}17\ ;\ 28{,}24 + 1{,}17] = [27{,}07\ \mu g.m^{-3}\ ;\ 29{,}41\ \mu g.m^{-3}]$.

On vérifie que l'intervalle proposé contient bien la vraie moyenne : 27,63. L'intervalle de confiance repose sur l'estimation de l'erreur qui est très bonne puisque qu'elle est estimée à 1,17 au lieu de 1,18 théoriquement.

Pour montrer la validité des résultats avancés, nous allons procéder à une simulation en effectuant 10 000 sondages identiques au précédent. Chaque sondage nous fournira un échantillon de 1 000 valeurs à partir desquelles nous estimerons la moyenne et calculerons un intervalle de confiance théorique (donc d'erreur 1,18), et un autre estimé. Nous disposerons donc de 10 000 moyennes et de 10 000 intervalles de confiance théoriques et estimés. Nous pourrons alors vérifier d'une part la dispersion de l'estimateur de la moyenne (en calculant la moyenne des estimations et leur variance), d'autre part la validité des intervalles de confiance à 95 % (95 % d'entre eux doivent contenir la vraie moyenne).

Nous aboutissons aux résultats suivants :

En moyenne les estimations valent 27,63 $\mu g.m^{-3}$, et leur variance vaut 0,36. On retrouve bien le résultat théorique qui annonçait qu'en moyenne l'estimateur vaut la vraie valeur, et que sa variance vaut :

$$\text{var}(\hat{\bar{y}}) = \frac{N-n}{n.N}.\frac{N}{N-1}\text{var}(y) = 0{,}36.$$

Les intervalles de confiance calculés à partir de l'erreur théorique contiennent, pour 95,2 % d'entre eux la vraie moyenne ; la proportion est de 95,8 % pour les intervalles de confiance estimés. On confirme bien le résultat théorique affirmant qu'il y a 95 % de chances pour que l'intervalle de confiance contienne la vraie valeur (avec évidemment une moindre exactitude dans le cas de l'intervalle estimé, le taux théorique choisi étant de 95 %).

La démarche adoptée dans cet exemple simple de sondage sera toujours plus ou moins la même. On fixe d'abord les paramètres du sondage (ici *n*) puis on estime le caractère d'intérêt ainsi qu'un intervalle de confiance autour de ce dernier. Pour choisir les paramètres du plan de sondage, on pourra s'inspirer d'un autre site connu, ce que nous exposerons dans la troisième partie. Ici, nous avons pris $n = 1\,000$ arbitrairement, mais quoi qu'il en soit, ce choix n'influe pas sur l'exactitude des résultats finaux. Un meilleur choix des paramètres initiaux influera sur la taille de l'intervalle de confiance, mais même s'ils sont mauvais, l'intervalle de confiance sera juste, plus large mais juste.

Choix d'un plan de sondage

Il existe de nombreuses façons d'effectuer un sondage. La plus élémentaire est de tirer de façon indépendante et équiprobable les individus (pour nous les instants de mesure) comme dans l'exemple précédent. Les contraintes techniques nous poussent à en définir d'autres, il n'est en effet pas envisageable d'aller effectuer 1 000 mesures indépendamment dans l'année.

Notre choix d'une stratégie de sondage s'appuie sur plusieurs exigences : respecter les contraintes techniques liées aux mesures de la qualité de l'air, minimiser autant que possible la variabilité entre les échantillons potentiels, mais aussi permettre une utilisation adaptable à tout site et tout polluant en offrant une relative simplicité de mise en œuvre.

Ces exigences nous ont conduit à retenir les plans de sondage par grappes stratifiés.

Les grappes sont des ensembles de mesures, pour nous des périodes. Une fois une grappe tirée dans l'échantillon, on relève toutes les mesures dans cette grappe. Une grappe peut par exemple être une semaine, et dans ce cas on tirerait des semaines de mesures dans l'année. Ce principe de tirage est adapté aux mesures par camion-laboratoire car on préfère laisser le camion-laboratoire pendant une période sur un site, plutôt que de le déplacer pour chaque mesure. Utiliser des grappes est donc une première nécessité technique, reste à choisir la longueur de ces grappes et leur nombre à tirer durant l'année.

Une autre idée naturelle consiste à répartir les mesures sur l'année. Nous savons que les concentrations des polluants sont fortement liées aux périodes de l'année, et notre échantillon doit donc brasser toutes les saisons. Ce choix devrait aider à minimiser la variabilité entre les échantillons. Nous allons donc découper l'année en différentes périodes appelées strates (typiquement les saisons), et nous imposerons un nombre de grappes à tirer dans chaque strate. Par exemple les strates peuvent être les saisons, il y en a donc quatre, et l'on impose de tirer un nombre fixe de semaines (les grappes) par saison. Il nous reste à choisir la longueur des strates et leur nombre.

Ainsi nous proposons d'effectuer des sondages par grappes stratifiés dont les paramètres à choisir seront la taille et le nombre des strates composant l'année, la taille et le nombre de grappes à tirer dans chaque strate. Ces paramètres sont à choisir *a priori* car ils déterminent les périodes de mesure, nous allons discuter de leur choix. Aussi, une fois les mesures effectuées, nous pouvons estimer à partir de ces dernières nos caractères d'intérêt (la moyenne et le taux de dépassement de seuil) ; nous allons voir comment optimiser cette étape.

Choix des paramètres

Nous verrons dans la partie suivante les formules donnant la variance de l'estimateur de la moyenne en fonction de tous les paramètres du plan de sondage retenu. Il apparaît que la variance est d'autant plus petite que les concentrations dans les strates sont homogènes. Mais comment savoir *a priori* quelles seront les périodes de l'année pour lesquelles les concentrations d'un polluant seront assez homogènes ? Nous savons, par l'analyse de sites connus, que l'on peut systématiquement découper l'année en périodes assez homogènes mais ce découpage n'est jamais le même d'une année sur l'autre. Il est possible de comprendre cette homogénéité *a posteriori* par l'analyse des phénomènes influant sur la qualité de l'air, mais comment prédire cela pour une année future ? Nous pouvons simplement dire que le découpage respecte plus ou moins les saisons. C'est pourquoi il ne nous semble pas nécessaire à ce stade, alors qu'on est censé prévoir un découpage d'une année future, d'affiner inutilement des choix quelque peu arbitraires, et nous préférons imposer une taille identique pour toutes les strates par souci de simplicité. Les formules que nous présentons sont proposées dans le cadre général, mais les simulations que nous avons effectuées se placent sous l'hypothèse d'un découpage en strates égales.

Par ailleurs, nous choisirons des grappes de taille identique. Cette hypothèse simplifie l'implémentation du plan de sondage et allège les formules. De plus, elle n'est pas restrictive car, nous le verrons dans les simulations, il est plus important de jouer sur le nombre de grappes à tirer que sur leur taille pour améliorer la variance de l'estimateur. Les tailles possibles des grappes sont fixées par des contraintes techniques, et le choix d'une taille sera à intégrer aux choix des autres paramètres.

À ce stade, il nous faut donc encore choisir le nombre de strates, la taille des grappes à tirer et surtout leur nombre. Pour ne pas choisir tout cela de façon complètement hasardeuse, nous pouvons nous appuyer sur un site proche connu d'une année antérieure. Nous pourrons voir selon les paramètres choisis quelles sont sur ce site la variance de l'estimateur et donc la précision attendue. Ceci nous donnera une idée de la précision que l'on peut espérer sur le site à explorer pour l'année d'étude avec le choix de ces paramètres. Évidemment cette étape, qui sera détaillée dans la partie suivante, ne sert qu'à proposer des paramètres pour le plan de sondage. La justesse des résultats finaux ne dépend pas de ce choix.

Finalement, le choix *a priori* des paramètres, même s'il est « soufflé » par une simulation sur un site supposé proche, ne repose pas sur des connaissances pointues du site et du polluant à explorer, connaissances dont nous voulons nous affranchir dans l'optique d'une prospection de sites inconnus.

Estimation des caractères d'intérêt

Une fois le sondage effectué, la situation est quelque peu différente puisque nous avons à disposition les mesures, et nous pouvons estimer directement notre caractère d'intérêt à partir de celles-ci. Mais nous disposons également des données d'autres sites fixes connus à partir desquelles nous pouvons tirer de l'information améliorant nos estimations. Des phénomènes de biais ont en effet pu survenir lors d'une période de mesure, dus par exemple à des conditions météorologiques inhabituelles pour la saison, qui rendent alors les relevés dans cette période non représentatifs de la strate. Or ces phénomènes sont intégrés dans les sites proches, et nous pouvons *a posteriori* les repérer afin de « redresser » l'estimation.

Cette idée est le principe des estimateurs redressés que nous présenterons dans la prochaine partie, qui s'appuient sur un site auxiliaire supposé proche. Cette notion de proximité n'est pas arbitraire car on peut la quantifier sur les périodes de mesure et elle est prise en compte dans les calculs. Si les concentrations du site auxiliaire ne s'avèrent pas fluctuer comme celles du site d'étude, les estimateurs redressés utiliseront d'autant moins les informations relatives à ce site auxiliaire, comme nous le verrons par la suite.

L'intérêt de ce genre de redressement est qu'il n'est pas utile de connaître les vraies raisons du biais (est-ce une hausse des températures, des conditions de pression atmosphérique atypiques, des directions de vent rares ?) mais simplement d'en constater l'incidence sur les concentrations d'un site proche. Cela nous permet d'améliorer nos estimations sans utiliser de connaissances pointues des phénomènes agissant sur le site et le polluant étudiés, autrement dit sans faire appel à un effort de modélisation.

Mise en œuvre du plan par grappes stratifié

Nous rentrons à présent dans les implémentations du plan de sondage. Nous présentons tout d'abord le plan par grappes stratifié avec les formules utiles à son exploitation. Nous proposons ensuite un protocole de choix des paramètres qui s'appuie sur des simulations sur un site connu supposé proche d'une année antérieure. Nous terminons enfin par la présentation de quelques estimateurs redressés utilisant un site auxiliaire, dont nous discuterons l'utilisation.

Les plans par grappes stratifiés

Nous allons donc, d'une part, stratifier l'année pour utiliser le fait que la concentration du polluant (le NO_2 ici) est fortement liée à la période de l'année, et d'autre part, effectuer un tirage par grappes pour tenir compte des contraintes techniques de mesure.

Les notations seront les suivantes : notre plan fait intervenir H strates dans l'année, nous tirons m grappes dans l'échantillon sur M que contient la population. Chaque strate h contiendra en tout M_h grappes, dont m_h seront tirées. De façon générale on indiquera par l'indice g les quantités relatives aux grappes et par l'indice h celles relatives aux strates.

L'estimateur de la moyenne standard est la moyenne sur l'échantillon, nous avons besoin de l'expression de sa variance pour proposer des intervalles de confiance. Elle vaut dans ce cas :

$$\mathrm{var}(\hat{\bar{y}}) = \frac{1}{M^2}\sum_{h=1}^{H} M_h^2 . \mathrm{var}(\hat{\bar{y}}_{gh})$$

où $\hat{\bar{y}}_{gh}$ est l'estimateur de la moyenne dans la strate h, le g rappelant que l'échantillon dans cette strate est composé de grappes.

Cette écriture n'est pas exploitable comme telle, nous donnons la formule développée ci-dessous. Mais elle met en évidence une propriété fondamentale des plans stratifiés. La variance de l'estimateur de la moyenne est une moyenne des variances des estimateurs dans chaque strate. C'est le gain d'un plan par strates par rapport à un plan simple : la variance de l'estimateur du plan stratifié ne prend en compte que ce qui se passe à l'intérieur de chaque strate, et non entre les strates. Ainsi, on a tout intérêt à trouver des strates homogènes pour minimiser leur propre variance, et ainsi minimiser la variance de l'estimateur. Malheureusement, comme nous l'avons déjà évoqué, prédire des périodes de mesures homogènes une année auparavant n'est pas faisable si ce n'est en utilisant le parallèle grossier avec les saisons. Ceci explique notre choix de stratifier pour tenir compte de ces tendances classiques, mais sans affinements arbitraires.

La formule générale de la variance de l'estimateur de la moyenne dans le cas du plan stratifié s'écrit comme ci-dessus avec :

$$\mathrm{var}(\hat{\bar{y}}_{gh}) = \frac{1}{N_h^2} . \frac{M_h - m_h}{M_h - 1} . \frac{M_h}{m_h} . \sum_{j=1}^{M_h} (N_{gj} . \bar{y}_{gj} - \frac{N_h . \bar{y}_h)^2}{M_h}$$

où N_h est la taille totale de la strate h, M_h le nombre total de grappes qu'elle contient ; m_h le nombre de grappes de la strate h dans l'échantillon ; N_{gj} la taille de la grappe j, $\bar{y}_{gj}$ sa moyenne ; $\bar{y}_h$ la moyenne dans la strate h.

Cette formule est valable dans le cadre le plus général possible, c'est-à-dire lorsque les strates peuvent être de taille respective différente N_h, de même pour les grappes de taille respective N_{gj}.

Un estimateur de la variance de l'estimateur de la moyenne est :

$$\hat{\mathrm{var}}(\hat{\bar{y}}) = \frac{1}{M^2}\sum_{h=1}^{H} M_h^2 . \hat{\mathrm{var}}(\hat{\bar{y}}_{gh})$$

$$\text{avec } \hat{\mathrm{var}}(\hat{\bar{y}}_{gh}) = \frac{1}{N_h^2} . \frac{M_h - m_h}{m_h - 1} . \frac{M_h}{m_h} . \sum_{j=1}^{m_h} (N_{gj} . \bar{y}_{gj} - \frac{N_h . \hat{\bar{y}}_h)^2}{M_h} .$$

Dans la pratique, il est préférable de choisir une taille unique pour toutes les grappes, ce qui simplifie l'implémentation et n'a pas d'incidence forte sur la

qualité du plan. Ceci dit, des raisons techniques peuvent contraindre à varier la taille des grappes au cours de l'année sondée, il faudra alors utiliser la formule générale ci-dessus pour calculer la variance de l'estimateur.

Dans le cas où les grappes ont toutes la même taille, le plan par grappes stratifié peut être vu comme un plan stratifié classique pour lequel les individus sont les grappes (résumées par leur moyenne). L'expression de la variance se simplifie (dans ce cas les N_{gj} sont constants égaux à N_g et $N_h = N_g.M_h$), et on obtient la formule de variance suivante :

$$\text{var}(\hat{\bar{y}}) = \frac{1}{M^2}\sum_{h=1}^{H} M_h . \frac{M_h - m_h}{m_h} . S_{gh}^2 \qquad (1)$$

$$\text{avec } S_{gh}^2 = \frac{1}{M_h - 1}\sum_{j=1}^{M_h} (\bar{y}_{gj} - \bar{y}_h)^2.$$

S_{gh}^2 fait intervenir une somme sur toutes les grappes de la strate *h,* cette quantité représente la variance corrigée inter-grappes de la strate *h.*

Un estimateur de la variance est alors :

$$\hat{\text{var}}(\hat{\bar{y}}) = \frac{1}{M^2}\sum_{h=1}^{H} M_h \cdot \frac{M_h - m_h}{m_h} \cdot \hat{S}_{gh}^2 \qquad (2)$$

$$\text{avec } \hat{S}_{gh}^2 = \frac{1}{m_h - 1}\sum_{j=1}^{m_h} (\bar{y}_{gj} - \hat{\bar{y}}_h)^2.$$

Cette dernière expression est la formule à utiliser dans le cas standard d'un plan par grappes stratifié où toutes les grappes ont même taille (la formule est valable dans le cas où les strates n'ont pas même taille). À partir de cette variance estimée, nous pouvons proposer un intervalle de confiance estimé comme dans la première partie :

$$IC_{95\,\%} = [\hat{\bar{y}} - 1{,}96\sqrt{\hat{\text{var}}(\hat{\bar{y}})}\ ;\ \hat{\bar{y}} + 1{,}96\sqrt{\hat{\text{var}}(\hat{\bar{y}})}]. \qquad (3)$$

Cet intervalle de confiance reste valable même si nous ne sommes plus dans le cadre simple de tirages indépendants de quarts d'heure. Cette hypothèse se retrouve ici dans le tirage des grappes, fait de façon indépendante.

Remarque : nous constatons que pour le calcul de la variance de l'estimateur de la moyenne, nous n'avons besoin que de la moyenne des grappes et non de toutes les mesures qu'elle contient. Concrètement, cela signifie qu'il suffit de disposer de la moyenne des mesures sur la grappe pour mettre en œuvre l'estimation de la moyenne et de l'erreur associée. Ainsi, un simple tube à diffusion résumant une grappe par sa moyenne donnera une estimation aussi bonne qu'un analyseur qui fournirait chaque concentration quart-horaire.

Choix des paramètres du plan de sondage

Avant toute chose, il est important de noter que quels que soient les paramètres du plan de sondage choisis, les estimations proposées ci-dessus seront toujours justes, que ce soit pour la moyenne ou pour l'intervalle de confiance.

Un choix judicieux des paramètres diminuera simplement la variance de l'estimateur par rapport à un choix plus arbitraire, pour un coût global du sondage identique (même temps de mobilisation des appareils ou des personnes par exemple). Il est évidemment intéressant de choisir, à précision fixée, le plan qui nécessite le moins de moyens.

L'hypothèse de base est que l'on dispose d'une série de mesures annuelle d'un site de même type, d'une année antérieure ou moyenne (moyennée quart d'heure par quart d'heure sur plusieurs années par exemple). Ce choix peut s'avérer difficile, nous le discuterons dans la partie consacrée aux simulations à La Rochelle, mais encore une fois il n'est pas décisif puisqu'il ne sert qu'à se donner une idée des paramètres à choisir ; dans le pire des cas, le choix est maladroit et les paramètres seront mal optimisés, cependant les estimations finales seront justes.

Nous supposons donc que nous sommes en possession d'une telle série auxiliaire. L'avantage majeur est que nous connaissons parfaitement cette série et qu'il n'y a pas d'estimations à faire. Ainsi, on peut calculer la variance exacte de l'estimateur sur cette dernière pour un plan fixé. Une première méthode est de proposer des paramètres pour un plan de sondage, et de calculer la variance de l'estimateur afin de choisir le plan pour lequel la variance de l'estimateur est minimale pour les contraintes techniques fixées.

Un autre point de vue, et c'est celui que nous adopterons, est de se fixer la précision d'estimation que l'on recherche, autrement dit la taille maximale de l'intervalle de confiance à 95 %. Cela revient à se fixer la variance de l'estimateur. Nous disposons alors de résultats pour nous aider à choisir les paramètres.

Pour un plan par strates composé de H strates de taille respective M_h, la taille minimale m^* de l'échantillon qu'il faut tirer pour obtenir la variance V de l'estimateur que l'on s'est fixée est :

$$m^* = \frac{\left(\sum_{h=1}^{H} M_h . S_h\right)^2}{M^2 . V + \sum_{h=1}^{H} M_h . S_h^2}$$

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

les $y_{h,j}$ sont les individus dans la strate h, et $\bar{y}_h$ est leur moyenne.

Il reste alors à savoir comment on doit répartir cet échantillon selon les strates (combien tirer d'individus par strate ?). La variance recherchée sera atteinte pour la taille d'échantillon m^* seulement si la répartition dans les strates est optimale. De façon générale la répartition optimale dans les strates pour un échantillon de taille m est la suivante (en notant m_h la taille de l'échantillon à tirer dans la strate h) :

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

Ces résultats étant à disposition, nous pouvons proposer un protocole de choix optimal des paramètres. Nous imposons un plan dont les grappes sont toutes de même taille. Dans ce cas, notre plan par grappes stratifié s'apparente à un plan stratifié classique dont les individus sont les grappes, les formules ci-dessus sont donc utilisables en utilisant comme valeur des individus la moyenne dans les grappes. M est alors le nombre de grappes totales présentes dans l'année, m le nombre de grappes dans l'échantillon, M_h le nombre de grappes présentes dans la strate h et m_h le nombre de grappes tirées dans la strate h. S_h^2 est remplacée par S_{gh}^2 la variance corrigée intergrappes dans la strate h, rappelons que cette variance se calcule à ce stade sur la série auxiliaire.

Le protocole est le suivant :

1. on se fixe une précision ε attendue à 95 %, donc une variance d'estimateur V puisque $\varepsilon = 1{,}96.\sqrt{V}$ (on peut remplacer le 95 % par $(1-\alpha)*100$ % avec α entre 0 et 1, et remplacer 1,96 par $z_{1-\alpha/2}$, le quantile d'ordre $1-\alpha/2$ d'une loi normale centrée réduite) ;

2. on propose un découpage en strates de l'année, donc on se fixe H et les N_h (tailles des strates) ;

3. on propose une taille des grappes N_g. Ceci fixe donc les M_h ;

4. on calcule m^*, le nombre minimal de grappes qu'il faudra tirer pour atteindre la précision fixée avec les paramètres proposés du plan.

$$m^* = \frac{\left(\sum_{h=1}^{H} M_h . S_{gh}\right)^2}{M^2 . \grave{V} + \sum_{h=1}^{H} M_h . S_{gh}^2} \qquad (4)$$

avec $S_{gh}^2 = \frac{1}{M_h - 1} \sum_{j=1}^{M_h} (\bar{y}_{gh,j} - \bar{y}_h)^2$ et $S_{gh} = \sqrt{S_{gh}^2}$

$\bar{y}_{gh,j}$ est la moyenne de la grappe j dans la strate h, et $\bar{y}_h$ la moyenne dans la strate h ;

5. on calcule la répartition optimale dans les strates des m^* grappes pour obtenir la précision voulue, autrement dit le nombre de grappes qu'il faudra tirer dans chaque strate h :

$$m_h = \frac{m^* . M_h . S_{gh}}{\sum_{l=1}^{H} M_l . S_{gl}},\ h = 1..H \qquad (5)$$

6. les m_h calculés en 5. n'étant pas forcément des entiers, on les approche par l'entier le plus proche. On sait donc combien de grappes il faudra tirer par strate pour appliquer ce sondage ;

7. il convient de vérifier la précision de ce plan en calculant la variance théorique de l'estimateur donnée en (1). Elle ne sera pas égale à celle attendue en 1. car l'approximation des m_h en 6. fait perdre de l'optimalité ;

8. on réitère le processus depuis l'étape 2. en proposant une autre stratification et/ou une autre taille des grappes pour la même précision attendue.

Ce protocole permet d'avoir plusieurs propositions de plan pour une même précision. Il suffit alors de choisir le plus pratique à mettre en œuvre ou le moins coûteux.

Encore une fois, tous les calculs effectués dans ce protocole utilisent les données sur la série auxiliaire supposée proche, la subtilité des choix ne doit donc pas être exagérée par rapport à la proximité supposée de la série et de l'année.

Exploitation des données avec information auxiliaire

Dans cette partie, nous allons proposer des méthodes d'estimation utilisant de l'information auxiliaire. Contrairement à la partie précédente, cette information n'est disponible qu'une fois le plan de sondage effectué. Elle est utile pour redresser une estimation classique ne s'appuyant que sur les données tirées par sondage.

On suppose encore une fois que l'on dispose d'une série proche. Le degré de proximité de cette série auxiliaire avec celle sondée sera quantifié et interviendra dans le redressement de l'estimation. Il s'agit de la covariance corrigée entre les deux séries calculée comme suit :

$S_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x}).(y_i - \bar{y})$ où x représente la série auxiliaire, y notre série d'intérêt et N la taille de la population.

Plus cette quantité est proche du produit des écarts types de x et de y, plus le comportement des séries est proche (il y a égalité lorsque $x = y$, elle vaut 0 lorsque x et y sont des variables indépendantes). Notons que cette quantité nécessite toutes les valeurs des deux séries, il conviendra donc de l'estimer lorsque nous l'utiliserons.

$\hat{S}_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\bar{x}}).(y_i - \hat{\bar{y}})$ avec n la taille de l'échantillon, $\hat{\bar{y}}$ la moyenne de l'échantillon (l'estimateur de la moyenne) et $\hat{\bar{x}}$ la moyenne de la série auxiliaire sur la période de sondage.

Le principe est simple : puisque l'on connaît parfaitement la série auxiliaire, il nous est possible de comparer la vraie moyenne annuelle de cette dernière avec la moyenne estimée sur la période de sondage. On utilise alors l'erreur observée pour redresser plus ou moins l'estimation de la série d'intérêt.

La pertinence du redressement dépend de la proximité des deux séries. C'est pourquoi leur covariance interviendra dans le calcul de l'intervalle de confiance, induisant une erreur d'estimation après redressement plus ou moins grande.

Nous proposons trois estimateurs de la moyenne redressés : l'estimateur par la différence, l'estimateur par le quotient et l'estimateur par la régression. Nous donnerons leur forme et leur variance (utile pour le calcul de l'intervalle de confiance) puis nous discuterons duquel choisir selon les cas.

L'estimateur par la différence

Nous le notons $\hat{\bar{y}}_D$, il est donné par :

$$\hat{\bar{y}}_D = \hat{\bar{y}} + \bar{x} - \hat{\bar{x}}.$$

C'est l'estimateur redressé le plus évident : à l'estimateur de la moyenne classique $\hat{\bar{y}}$ on retranche l'erreur observée sur la série auxiliaire $(\hat{\bar{x}} - \bar{x})$.

Comme nous travaillons avec des plans par grappes stratifiés, il est naturel de redresser l'estimateur strate par strate (il n'y a aucune raison pour que l'estimation soit de même qualité dans chaque strate).

Ainsi la forme de l'estimateur redressé par la différence s'écrira plutôt :

$\hat{\bar{y}}_D = \frac{1}{H}\sum_{h=1}^{H}(\hat{\bar{y}}_h + \bar{x}_h - \hat{\bar{x}}_h)$ où $\hat{\bar{y}}_h$ est l'estimateur de la moyenne classique dans la strate *h*, $\hat{\bar{x}}_h$ la moyenne sur l'échantillon de la série auxiliaire et $\bar{x}_h$ la vraie moyenne de la série auxiliaire dans la strate *h*.

Sa variance dans le cas où les grappes sont de même taille est la suivante (dans ce cas les grappes jouent le rôle des individus pour un plan par strate classique) :

$$\mathrm{var}(\hat{\bar{y}}_D) = \frac{1}{M^2}\sum_{h=1}^{H} M_h.\frac{M_h - m_h}{m_h}(S^2_{y,h} + S^2_{x,h} - 2.S_{xy,h})$$

avec $S^2_{y,h}$ la variance corrigée dans la strate *h* pour la série d'intérêt *y*, $S^2_{x,h}$ la variance corrigée dans la strate *h* pour la série auxiliaire *x*, $S_{xy,h}$ la covariance corrigée entre les deux séries, *M* le nombre total de grappes dans la population, M_h le nombre total de grappes dans la strate *h*, m_h le nombre de grappes tirées dans la strate *h*.

Les variances et covariances corrigées sont bien sûr à calculer entre les grappes. Par exemple

$$S_{xy,h} = \sqrt{S^2_{xy,h}} = \sqrt{\frac{1}{M_h - 1}\sum_{j=1}^{M_h}(\bar{y}_{gh,j} - \bar{y}_h).(\bar{x}_{gh,j} - \bar{x}_h)}$$

où $\bar{y}_{gh,j}$ est la moyenne de la grappe *j* dans la strate *h*, $\bar{y}_h$ la moyenne dans la strate *h*, et de même pour le caractère auxiliaire *x*, la somme se fait sur toutes les grappes de la strate *h*.

L'estimateur de la variance est alors :

$$\hat{\mathrm{var}}(\hat{\bar{y}}_D) = \frac{1}{M^2}\sum_{h=1}^{H} M_h.\frac{M_h - m_h}{m_h}(\hat{S}^2_{y,h} + \hat{S}^2_{y,h} - 2.\hat{S}_{xy,h})$$

avec $\hat{S}^2_{y,h} = \frac{1}{m_h - 1}\sum_{j=1}^{m_h}(\bar{y}_{gh,j} - \hat{\bar{y}}_h)^2$, la somme se faisant sur les grappes de la strate *h* tirées dans l'échantillon, et de même :

$$\hat{S}^2_{x,h} = \frac{1}{m_h - 1}\sum_{j=1}^{m_h}(\bar{x}_{gh,j} - \hat{\bar{x}}_h)^2,$$

$$\hat{S}_{xy,h} = \sqrt{\hat{S}^2_{xy,h}} = \sqrt{\frac{1}{m_h - 1}\sum_{j=1}^{m_h}(\bar{y}_{gh,j} - \hat{\bar{y}}_h).(\bar{x}_{gh,j} - \hat{\bar{x}}_h)}.$$

L'intervalle de confiance à 95 % estimé pour l'estimateur par la différence se calcule grâce aux formules précédentes :

$$IC_{95\,\%D} = [\hat{\bar{y}}_D - 1{,}96\sqrt{\hat{\mathrm{var}}(\hat{\bar{y}}_D)}\ ;\ \hat{\bar{y}}_D + 1{,}96\sqrt{\hat{\mathrm{var}}(\hat{\bar{y}}_D)}] \quad (6)$$

L'estimateur par le quotient

L'idée est sensiblement la même que pour l'estimateur par la différence. Simplement, au lieu de corriger de façon additive, on corrige de façon multiplicative :

$$\hat{\bar{y}}_Q = \hat{\bar{y}} \cdot \frac{\bar{x}}{\hat{\bar{x}}}.$$

Dans notre plan par grappes stratifié, on considère ce redressement strate par strate :

$\hat{\bar{y}}_Q = \frac{1}{H}\sum_{h=1}^{H}\hat{\bar{y}}_h \cdot \frac{\bar{x}_h}{\hat{\bar{x}}_h}$ avec les mêmes notations que pour l'estimateur par la différence.

Cet estimateur est plus difficile à étudier que ceux que nous avons rencontrés jusqu'alors. Son espérance n'est pas égale à la vraie moyenne, ce qui signifie que si on tire beaucoup d'échantillons, la moyenne des estimations ne donnera pas la bonne valeur. Ceci dit, plus l'échantillon est grand et plus la différence devient négligeable. Cependant, on peut quand même proposer un intervalle de confiance en calculant son erreur quadratique moyenne (EQM). Elle représente l'erreur moyenne au carré que commet l'estimateur. Lorsque l'espérance de l'estimateur est égale à la bonne valeur, ce qui était le cas pour tous les estimateurs rencontrés jusqu'alors, l'erreur quadratique moyenne est égale à la variance.

Ainsi l'EQM joue ici exactement le même rôle que la variance des estimateurs, elle servira à calculer de la même manière l'erreur d'estimation permettant de fournir un intervalle de confiance.

L'EQM de l'estimateur par le quotient dans le cas de grappes de même taille peut être approchée de la manière suivante :

$$EQM(\hat{\bar{y}}_Q) \approx \frac{1}{M^2}\sum_{h=1}^{H} M_h.\frac{M_h - m_h}{m_h}(S^2_{y,h} + r^2 S^2_{x,h} - 2.r.S_{xy,h})$$

avec $r = \frac{\bar{y}}{\bar{x}}$, le rapport des vraies moyennes des deux séries, les autres notations sont les mêmes que précédemment.

Un estimateur de l'EQM est dans ce cas :

$$\hat{EQM}(\hat{\bar{y}}_Q) = \frac{1}{M^2}\sum_{h=1}^{H} M_h.\frac{M_h - m_h}{m_h}(\hat{S}^2_{y,h} + \hat{r}^2\hat{S}^2_{x,h} - 2.\hat{r}.\hat{S}_{xy,h})$$

avec $\hat{r} = \frac{\hat{\bar{y}}}{\hat{\bar{x}}}$, les autres notations étant inchangées.

L'intervalle de confiance à 95 % estimé pour l'estimateur par le quotient se calcule grâce aux formules précédentes :

$$IC_{95\,\%,\,Q} = [\hat{\bar{y}}_Q - 1{,}96\sqrt{\hat{EQM}(\hat{\bar{y}}_Q)}\ ;\ \hat{\bar{y}}_Q + 1{,}96\sqrt{\hat{EQM}(\hat{\bar{y}}_Q)}]. \quad (7)$$

L'estimateur par la régression

L'estimateur par la régression est donné par :

$$\hat{\bar{y}}_R = \hat{\bar{y}} + \hat{b}.(\bar{x} - \hat{\bar{x}}) \text{ avec } \hat{b} = \frac{\hat{S}_{xy}}{\hat{S}^2_x}.$$

Dans le cas d'un plan par grappes stratifié, on redresse strate par strate :

$\hat{\bar{y}}_R = \frac{1}{H}\sum_{h=1}^{H} \hat{\bar{y}}_h + \hat{b}_h.(\bar{x}_h - \hat{\bar{x}}_h)$ où $\hat{b}_h = \frac{\hat{S}_{xy,h}}{\hat{S}^2_{x,h}}$, les notations étant les mêmes qu'avant.

De même que l'estimateur par le quotient, cet estimateur a des propriétés difficiles à obtenir. Nous devons considérer l'EQM au lieu de la variance pour le calcul de l'intervalle de confiance, et dans le cas de grappes de même taille, elle est approchée par :

$$EQM(\hat{\bar{y}}_R) \approx \frac{1}{M^2}\sum_{h=1}^{H} M_h.\frac{M_h - m_h}{m_h} S^2_{y,h}.(1-\rho^2_h)$$

où $\rho^2_h = \frac{S^2_{xy,h}}{S^2_{x,h}.S^2_{y,h}}$.

Un estimateur de l'EQM dans le cas de grappes de même taille est :

$$\hat{EQM}(\hat{\bar{y}}_R) = \frac{1}{M^2}\sum_{h=1}^{H} M_h.\frac{M_h - m_h}{m_h} \hat{S}^2_{y,h}.(1-\hat{\rho}^2_h)$$

où $\hat{\rho}^2_h = \frac{\hat{S}^2_{xy,h}}{\hat{S}^2_{x,h}.\hat{S}^2_{y,h}}$, les notations sont inchangées.

L'intervalle de confiance à 95 % estimé pour l'estimateur par la régression se calcule grâce aux formules précédentes :

$$IC_{95\,\%,\,R} = [\hat{\bar{y}}_R - 1{,}96\sqrt{\hat{EQM}(\hat{\bar{y}}_R)}\ ;\ \hat{\bar{y}}_R + 1{,}96\sqrt{\hat{EQM}(\hat{\bar{y}}_R)}]. \quad (8)$$

Utilisation des estimateurs redressés

Le paragraphe « Choix des paramètres du plan de sondage » p. 558, propose un protocole du plan par grappes stratifié. Nous allons maintenant nous intéresser à la démarche à adopter lorsque les données ont été tirées selon le plan retenu.

Plusieurs estimateurs de la moyenne sont à notre disposition : l'estimateur classique non redressé, et ceux redressés par la différence, le quotient et la régression. Chacun d'entre eux proposera une estimation différente, et surtout un intervalle de confiance plus ou moins grand selon leur variance estimée (ou EQM). Quel résultat choisir ?

Théoriquement, le choix n'est pas difficile, il suffit de retenir l'estimateur qui propose l'intervalle de confiance le plus petit. Nous savons en effet que chacun d'entre eux propose un intervalle de confiance à 95 %, le risque de se tromper est donc le même pour chaque estimation (5 %), autant prendre alors l'intervalle le plus précis.

Malheureusement, il se pose un problème majeur. Les intervalles de confiance prédits à 95 % sont calculés à partir de la variance des estimateurs, or nous sommes obligés d'estimer cette variance. Cette estimation, si elle est médiocre, induira un intervalle de confiance faux, c'est-à-dire qu'il risque d'être trop petit et le risque que la vraie moyenne ne soit pas à l'intérieur sera supérieur aux 5 % attendus.

La qualité de l'estimation des variances des estimateurs repose sur le nombre total de grappes présentes dans l'échantillon m_h. S'il est inférieur à 15, comme nous le verrons dans les simulations p. 562, l'estimation de la variance risque d'être médiocre. Les contraintes techniques ne permettant pas toujours de tirer un nombre confortable de grappes, il faut être vigilant sur la qualité des intervalles de confiance estimés.

En l'occurrence, lorsque le nombre total de grappes tirées est faible (moins de 15), il est bon de faire une simulation sur une série auxiliaire connue pour se rendre compte de la qualité des estimations. Cela consiste à tirer un très grand nombre d'échantillons selon le plan de sondage à tester, d'estimer à chaque fois les intervalles de confiance pour chaque estimateur, et de compter finalement combien d'intervalles de confiance contiennent la vraie moyenne annuelle. La proportion d'intervalles estimés contenant la vraie moyenne devrait être de 95 % pour chaque estimateur. L'imprécision due à l'estimation de la variance induira une autre proportion pour chaque estimateur. Cette dernière, si elle est trop faible (moins de 85 % par exemple), invalide l'estimation des intervalles de confiance pour l'estimateur concerné et le plan de sondage testé.

Enfin, on peut se poser la question de la pertinence de la proximité supposée entre le site étudié et le site auxiliaire. Nous rappelons que cette proximité est quantifiée par la covariance entre les deux séries, ainsi si le comportement des deux séries n'est pas proche, la covariance sera quasi nulle. Dans ce cas, on peut le vérifier sur les formules des variances ou des EQM des estimateurs redressés, l'incertitude proposée dans l'intervalle de confiance sera grande. Ainsi une visualisation des intervalles de confiance nous renseignera tout de suite sur la pertinence du choix du site auxiliaire, sans qu'il soit nécessaire de s'en convaincre par des raisons annexes. Il est de toute manière toujours possible de ne pas retenir les estimations proposées par redressement faute de pertinence, et de ne considérer que l'estimation directe sans site auxiliaire.

Estimation d'un dépassement de seuil

Principe

Jusqu'ici, nous nous sommes intéressés exclusivement à l'estimation de la moyenne annuelle d'un polluant. Nous allons proposer dans cette partie une adaptation des méthodes précédentes pour estimer le taux moyen de dépassement d'un seuil de pollution fixé.

Supposons que l'on se soit donné un seuil à ne pas dépasser. Pour calculer le taux de dépassement de ce seuil, il suffit de recoder la série avec des 1 lorsque la valeur est supérieure au seuil, et 0 sinon. On calcule alors la moyenne de cette nouvelle série, et on obtient le taux de dépassement.

Ainsi le problème d'estimation de ce taux revient à l'estimation d'une moyenne. Il suffit donc d'appliquer les méthodes d'estimation présentées précédemment, et on obtiendra un taux moyen estimé avec un intervalle de confiance associé.

Discussion

Une autre démarche aurait pu être de rechercher le percentile 98 (par exemple) de la série parce que ce dernier intervient dans des normes. Malheureusement, l'estimation d'un percentile est beaucoup plus délicate que l'estimation d'une moyenne, et pour l'entreprendre de façon acceptable, cela nécessite un grand nombre de valeurs dans l'échantillon, ce qui n'est pas notre cas (nous rappelons que dans le cas d'un plan par grappes stratifié, une grappe est résumée par une valeur).

Le fait de fixer le seuil, et de calculer le taux de dépassement permet de se ramener à une moyenne, et cela facilite l'estimation.

Pourtant, ce seuil ne doit pas être non plus fixé n'importe comment. Si on le choisit trop élevé (par exemple juste au-dessus de la plus grande valeur observée dans l'échantillon), il n'y aura aucune valeur dans l'échantillon supérieure à lui, et le taux estimé vaudra 0. De plus, la variation des dépassements de seuil sera nulle puisqu'on n'en observe aucune, et donc l'intervalle de confiance du taux sera [0 ; 0], ce qui n'est probablement pas réaliste (on peut concevoir sans problème qu'en dehors de l'échantillon, une valeur puisse le dépasser).

Ainsi, il convient de se fixer un seuil adapté à l'échantillon à disposition. Cela permet une estimation plus fiable, avec des bornes de l'intervalle de confiance qui, elles, tiennent compte de la fluctuation probable des mesures non observées.

Enfin, il convient d'être conscient que si, en principe, l'estimation du dépassement d'un seuil est équivalente à l'estimation d'une moyenne, techniquement la réalisation du sondage est plus contraignante. Pour l'estimation d'une moyenne, nous n'avions besoin que de la moyenne observée dans chaque grappe ; ainsi un simple tube à diffusion résumant la moyenne des concentrations d'un polluant pendant la période de la grappe était suffisant. Pour l'estimation du seuil, il faut la moyenne de dépassement du seuil dans chaque grappe, ce qui n'est plus mesurable par un simple tube à diffusion. Cela nécessite des relevés classiques type quart-horaire dans chaque grappe (typiquement à l'aide d'un camion-laboratoire) afin de pouvoir recoder chaque mesure en 0 et 1 et calculer le dépassement de seuil moyen dans la grappe. Ce type de relevés permet bien sûr une estimation de la moyenne classique, il permet aussi d'estimer le taux moyen de dépassement de seuil pour différents seuils fixés après le sondage.

Simulations

Cas pratique

Nous effectuons dans cette partie une simulation complète d'un sondage sur le site de La Rochelle. Son objectif est d'illustrer par un exemple la méthode exposée plus haut, en montrant les précautions qu'il convient de prendre dans la lecture des résultats.

On décide de sonder le site de Vaugoin en NO_2 afin d'en estimer sa moyenne annuelle en 2000 et la proportion de dépassement du seuil 45 $\mu g.m^{-3}$. Cela se passe en deux étapes. La première constitue le choix du plan de sondage, selon le protocole présenté p. 559. Ce choix s'effectue avant l'année à sonder pour fixer les jours de mesures. On s'appuie, pour ce faire, sur le site d'Aytré (moyenne sur les années précédentes) qui possède une station fixe fournissant des relevés permanents.

On cherche un plan pour lequel $L = 2$ (longueur de l'intervalle de confiance).

Nous appliquons le protocole de choix dont la formule (4) propose différents m (nombre total de grappes à tirer) selon la stratification (H, le nombre de strates) et la taille des grappes (N_g), paramètres que nous faisons varier.

Parmi toutes les configurations proposées (toutes sont censées fournir la même précision), nous retenons $m = 15{,}67$ proposé pour $H = 6$ et $N_g = 6*96$.

On prend donc $m = 16$, dont la répartition optimale du nombre de grappes dans les strates, calculée par la formule (5) du protocole est :

m_h=[4,1671 ; 1,8841 ; 3,2579 ; 1,5628 ; 3,2042 ; 2,9239]

On choisit m_h =[4 ; 2 ; 3 ; 2 ; 3 ; 2] (car il faut plus d'une grappe dans chaque strate pour estimer la variance).

Nous vérifions maintenant la variance de ce plan pour le site d'Aytré grâce à la formule (1) p. 558 : 0,2861, ce qui fait une longueur attendue pour l'intervalle de confiance de $2*1{,}96*\sqrt{0{,}2861} = 2{,}1$.

Ce n'est pas le 2 désiré initialement (à cause des approximations sur m et m_h) mais la précision est satisfaisante. On décide donc de lancer le tirage aléatoire des jours où il faut effectuer les mesures selon le plan de sondage de paramètres $H = 6$, $N_g = 6$ et m_h = [4 ; 2 ; 3 ; 2 ; 3 ; 2].

On effectue les mesures à Vaugoin pour les jours retenus, et on relève en même temps les mesures pour ces mêmes jours sur le site auxiliaire d'Aytré.

Commence alors la seconde étape : le calcul des estimations. Nous sommes rendus à la fin de l'année sondée et nous avons toutes les séries qu'il nous faut : l'échantillon pour Vaugoin, la série complète ainsi que l'échantillon pour le site auxiliaire d'Aytré.

Pour l'estimation de la moyenne, nous disposons de quatre estimateurs : celui classique non redressé, c'est-à-dire ne s'appuyant pas sur le site auxiliaire d'Aytré, dont l'intervalle de confiance se calcule par la formule (3) ; celui redressé par la différence, donné par la formule (6) ; celui redressé par le quotient, donné par la formule (7) ; et enfin celui redressé par la régression donné par la formule (8). Les résultats sont présentés dans le tableau suivant :

	Classique	Régression	Différence	Quotient
Moyenne estimée	15,46	14,07	13,92	13,88
IC	[13,47 ; 17,45]	[13,57 ; 14,57]	[12,71 ; 15,12]	[12,45 ; 15,31]
Longueur IC	3,97	1	2,42	2,86

Pour l'estimation du taux de dépassement du seuil, nous recodons les séries en 0-1 selon que les mesures dépassent ou non 45 $\mu g.m^{-3}$. Nous sommes ramenés à une estimation de moyenne, et nous utilisons les mêmes formules que précédemment pour les séries recodées. Les résultats sont les suivants :

	Classique	Régression	Différence	Quotient
Proportion estimée	4,86	4,2	4,32	NaN
IC	[3,51 ; 6,22]	[NaN ; NaN]	[3,41 ; 5,24]	[NaN ; NaN]
Longueur IC	2,71	NaN	1,83	NaN

Les « NaN » *(Not-a-Number)* indiquent une valeur manquante ; c'est le cas pour l'estimation d'une proportion lorsque le taux est estimé à 0 et que l'on effectue une division comme dans la méthode du quotient.

Une fois ces estimations effectuées, se posent deux questions : sont-elles fiables, à savoir les intervalles de confiance annoncés à 95 % théoriquement méritent-ils réellement cette confiance (on sait que les résultats souffrent de beaucoup d'approximations) ? et finalement, quelle estimation retenir ?

Pour vérifier la qualité de l'estimation, on effectue une simulation pour le même plan sur un site fixe (le site de Verdun en l'occurrence) redressé avec Aytré. La simulation consiste à tirer beaucoup d'échantillons (ici 10 000), à calculer les estimations des IC, et à regarder finalement combien d'entre eux contiennent la vraie valeur, connue sur ces sites annexes fixes. Le site de Verdun n'est pas censé être très proche de celui sondé, mais le cadre étant similaire, cette simulation nous aidera à déceler les éventuelles mauvaises estimations à écarter.

On obtient les résultats suivants pour la moyenne :

	Classique	Régression	Différence	Quotient
Taux de couverture	92,20 %	43,80 %	84,40 %	87 %
Longueur IC moyenne	3,7	0,83	2,56	4,12

Et pour le taux de dépassement de 45 $\mu g.m^{-3}$:

	Classique	Régression	Différence	Quotient
Taux de couverture	91,90 %	36,80 %	84,90 %	57 %
Longueur IC moyenne	5,68	NaN	4,37	NaN

La simulation révèle des taux de couverture trop éloignés du 95 % attendu. Ces mauvais taux de couverture sont dus à la mauvaise estimation de l'erreur d'estimation. C'est un problème qui a déjà été évoqué : il est dû au peu d'information que contient l'échantillon dans certaines strates, représentées uniquement par deux ou trois grappes. Cela ne remet pas en cause l'estimation de la moyenne ou de la proportion, mais celle de leur erreur associée *via* l'intervalle de confiance.

On peut tenter d'appréhender cette erreur d'une autre manière. Une idée consiste à estimer l'erreur pour un autre plan, moins précis que le plan initial, formé à partir de ce dernier en regroupant des strates ; typiquement ici pour un plan à trois strates de répartition respective des grappes m_h =[6 ; 5 ; 5]. L'erreur estimée n'est pas la vraie erreur associée à notre plan de sondage initial, mais on sait qu'elle est plus grande théoriquement, et surtout qu'elle sera mieux estimée. De plus, cette erreur fictive ne devrait pas être éloignée de la vraie car les deux plans sont assez proches. On préfère donc estimer une erreur que l'on sait être supérieure à la vraie, mais bien l'estimer. Nous allons recommencer nos estimations avec ces nouveaux paramètres (comme pour un plan où m_h =[6 ; 5 ; 5]), et effectuer une simulation annexe comme précédemment pour mesurer le gain de cette nouvelle démarche.

Les estimations sur les données du sondage à Vaugoin redressé avec Aytré deviennent alors :

- pour la moyenne :

	Classique	Régression	Différence	Quotient
Moyenne estimée	15,46	14,36	13,91	13,81
IC	[13,17 ; 17,74]	[13,49 ; 15,23]	[12,96 ; 14,85]	[12,82 ; 14,78]
Longueur IC	4,57	1,74	1,89	1,96

- pour le dépassement du seuil 45 $\mu g.m^{-3}$:

	Classique	Régression	Différence	Quotient
Proportion estimée	4,86	4,58	4,61	4,67
IC	[2,89 ; 6,83]	[3,73 ; 5,42]	[3,61 ; 5,60]	[3,82 ; 5,52]
Longueur IC	3,94	1,7	1,99	1,7

On effectue de nouveau une simulation pour l'estimation du site fixe de Verdun redressé grâce à Aytré, mais en estimant l'erreur comme pour un plan où m_h =[6 ; 5 ; 5].

On obtient les taux de couverture suivants :

- pour la moyenne :

	Classique	Régression	Différence	Quotient
Taux de couverture	99,00 %	88,10 %	92,80 %	93 %
Longueur IC moyenne	4,69	2,01	2,63	4,04

- pour le taux de dépassement de 45 $\mu g.m^{-3}$:

	Classique	Régression	Différence	Quotient
Taux de couverture	99,50 %	92,30 %	98,70 %	82 %
Longueur IC moyenne	7,59	4,64	6,32	8,32

Les résultats semblent plus fiables avec cette procédure. Les taux de couverture sont pour la plupart des estimateurs assez proche du 95 % attendu, même si certains estimateurs inspirent encore un peu de méfiance, comme celui par le quotient pour la proportion ou celui par la régression pour la moyenne, il sera peut-être sage de les écarter par la suite.

On décide donc de retenir les estimations obtenues par cette dernière démarche, c'est-à-dire en estimant les intervalles de confiance comme si le plan initial avait été réalisé avec m_h = [6 ; 5 ; 5].

Nous retenons donc les estimations suivantes pour la moyenne :

	Classique	Régression	Différence	Quotient
Moyenne estimée	15,46	14,36	13,91	13,81
IC	[13,17 ; 17,74]	[13,49 ; 15,23]	[12,96 ; 14,85]	[12,82 ; 14,78]
Longueur IC	4,57	1,74	1,89	1,96

Et pour la proportion de dépassement du seuil 45 $\mu g.m^{-3}$:

	Classique	Régression	Différence	Quotient
Proportion estimée	4,86	4,58	4,61	4,67
IC	[2,89 ; 6,83]	[3,73 ; 5,42]	[3,61 ; 5,60]	[3,82 ; 5,52]
Longueur IC	3,94	1,7	1,99	1,7

Il convient à présent de proposer une seule estimation parmi les quatre proposées. Le choix est maintenant simple : toutes ces estimations étant supposées fiables au vu de la simulation annexe effectuée, il suffit de choisir celle qui propose l'intervalle de confiance le plus petit. Pour la moyenne on aurait alors envie de retenir l'estimation par la régression, cependant celle par la différence est à peine plus large et l'estimateur par la différence jouit d'une confiance nettement meilleure d'après la simulation. Il est alors plus prudent de se confier à celui-ci. Quant à la proportion, on pourrait hésiter entre l'estimation par le quotient ou la régression, mais le taux de couverture pour l'estimateur par le quotient dans la simulation est assez faible (82 %), on préfèrera donc se fier à celle par la régression qui semble assez fiable (plus de 92 % de taux de couverture dans la simulation).

La moyenne à Vaugoin est donc estimée à 13,91 $\mu g.m^{-3}$, et estimée appartenir à 95 % à [12,96 $\mu g.m^{-3}$; 14,85], tandis que le taux de dépassement du seuil 45 $\mu g.m^{-3}$ est estimé à 4,58 %, et estimé appartenir à 95 % à [3,73 ; 5,42].

Nous connaissons en réalité la vraie moyenne annuelle et le vrai taux de dépassement de 45 $\mu g.m^{-3}$ sur la station de Vaugoin, ils valent respectivement 14,1 $\mu g.m^{-3}$ et 4,8 %.

Consistance de l'estimation des intervalles de confiance

Nous en avons déjà parlé et le cas pratique ci-dessus l'illustre, l'application d'un plan de sondage avec ses estimations ne se limite pas à l'implémentation aveugle de formules, mais nécessite quelque prudence dans la lecture des résultats. Le problème vient du taux de couverture des intervalles de confiance estimés, qui peut être loin de celui attendu théoriquement et n'est valide en toute rigueur que lorsque le nombre de grappes est élevé.

Dans le cas pratique précédent, le nombre de grappes que l'on tirait était insuffisant pour l'estimation de l'intervalle de confiance et la simulation annexe nous alertait sur la validité des taux de couverture. Nous avons contourné la difficulté en estimant une autre erreur, que nous savions plus grande que celle recherchée : cela nous a permis de mieux l'estimer, et la simulation annexe a confirmé notre démarche.

Nous présentons dans cette partie une série de simulations consistant à calculer le taux de recouvrement selon différents plans de sondage. Le but de ces simulations est de tenter de déterminer un nombre minimal de relevés à partir desquels l'estimation de l'intervalle de confiance semble acceptable.

Les simulations consistent en l'estimation de la moyenne pour le site de Vaugoin redressé avec le site d'Aytré pour l'année 2000. Pour chaque plan de sondage, nous avons effectué 5 000 tirages, donc 5 000 estimations d'intervalles de confiance, ce qui nous a permis de compter combien contenaient la vraie moyenne, ce qui nous fournit le taux de couverture. Les plans de sondage sont tous formés d'une seule strate, le nombre de grappes varie de 2 à 53 et leur taille varie de 1 à 30 jours. Afin de limiter le temps de calcul, les plans de sondage dont le nombre de jours de mesure excède 180 n'ont pas été pris en compte.

Nous avons représenté sur les graphiques suivants le taux de couverture relevé pour chaque plan. Sur l'axe des abscisses se trouve la taille des grappes, sur celui des ordonnées leur nombre. Le taux de couverture est représenté par des niveaux de bleus différents selon qu'il est inférieur ou supérieur à 90 % (rappelons que le taux attendu est de 95 %). Ainsi les graphiques nous donnent une ligne de niveau à partir de laquelle on peut raisonnablement accepter l'estimation de l'intervalle de confiance. Il y a un graphique par estimateur. Notons que pour l'estimateur par la régression, une seconde ligne de niveau à 85 % a été ajoutée pour plus de lisibilité.

Ces simulations montrent que : (1) la taille des grappes a peu d'influence sur la qualité de l'estimation par rapport au nombre de grappes tiré, et ce quel que soit l'estimateur ; (2) pour tous les estimateurs, hormis celui pour la régression, il semble qu'un minimum de 10 ou 15 grappes soit à conseiller pour avoir une estimation fiable. Rappelons que ces simulations ont été effectuées dans le cas d'une seule strate. Si l'on considérait des plans à plusieurs strates, il faudrait appliquer cette règle dans chaque strate, car chacune d'elle est soumise à l'estimation d'une erreur.

Ces choix ne concernent que la fiabilité de l'estimation de l'intervalle de confiance. Il est intéressant par ailleurs de regarder ce qu'il en est de la longueur de l'intervalle de confiance estimé selon les plans. Ce sont les graphiques suivants sur lesquels est représentée la longueur moyenne des intervalles de confiance (exprimés en $\mu g.m^{-3}$).

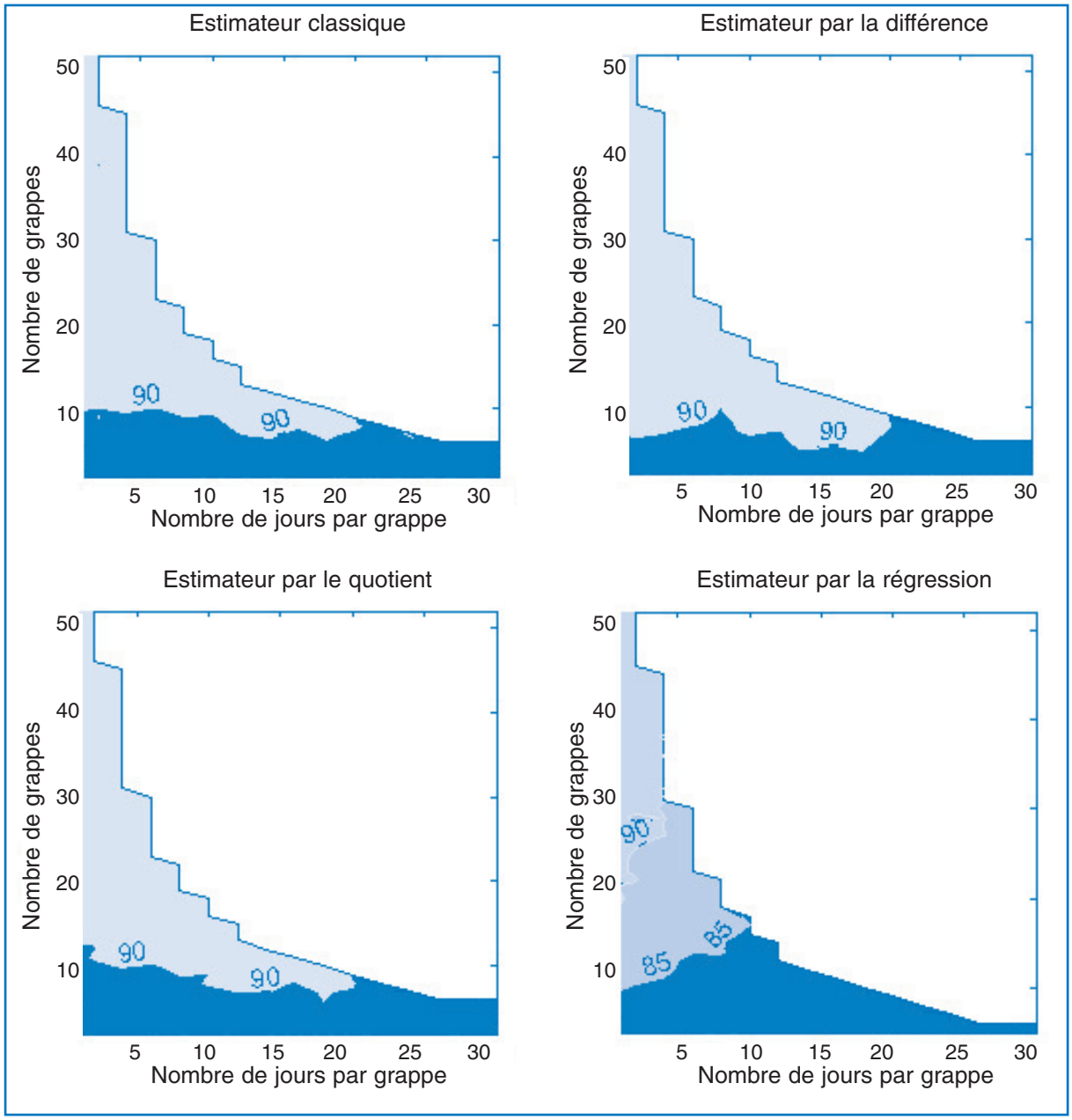

Figure 1.
Taux de couverture des estimateurs selon le plan.
Covery rate of the confidence intervals.

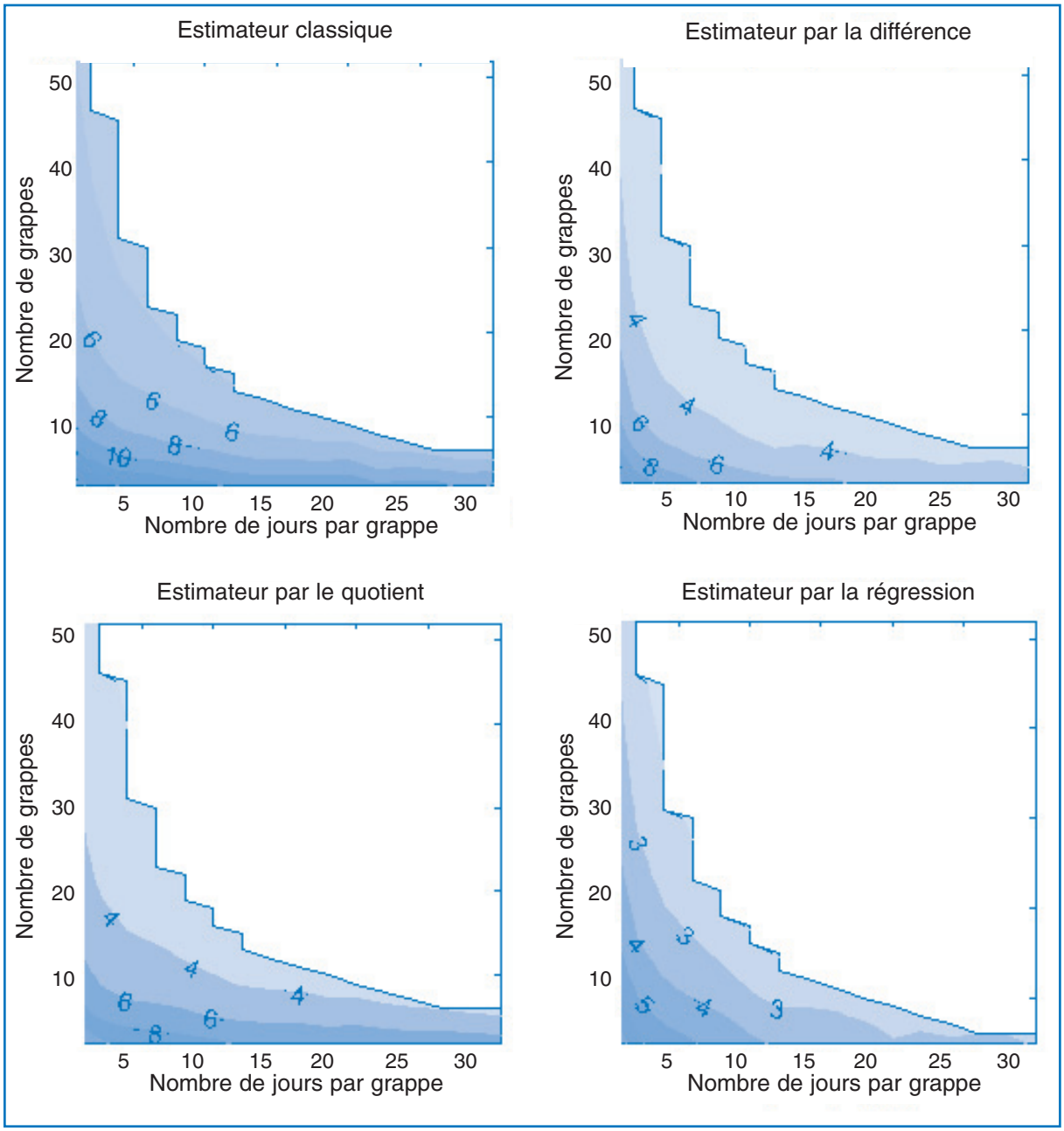

Figure 2.
Longueur des IC selon le plan.
Size of the confidence intervals.

On remarque que la précision des intervalles augmente avec le nombre de grappes et avec leur taille. La différence entre les estimateurs est que pour un même plan, on n'obtient pas la même précision. Il apparaît clairement que le moins précis est l'estimateur classique non redressé, puis viennent ensuite les estimateurs redressés par la différence et le quotient avec des précisions comparables, enfin l'estimateur par la régression propose les intervalles les plus précis.

Ainsi, il est toujours préférable de choisir un estimateur utilisant des informations auxiliaires. Cependant, il faut toujours prendre garde à la fiabilité de l'estimation, qui est moindre par exemple dans le cas de l'estimateur par la régression. On ne décidera de le retenir que dans le cas où le plan fait intervenir beaucoup de grappes.

D'autres simulations du même type ont été réalisées en prenant un autre site auxiliaire (Verdun au lieu d'Aytré). On peut s'attendre à une différence qualitative sur la précision des intervalles pour les estimateurs redressés : il se trouve en effet que le site de Verdun redresse moins bien nos estimations que celui d'Aytré, ainsi pour un même plan et un même estimateur, le redressement par Verdun fournira un intervalle moins précis. Cependant, les conclusions

concernant la fiabilité des estimations sont tout à fait comparables : on obtient des taux de couverture satisfaisants à partir de 10 ou 15 grappes dans l'échantillon. De même, la variation de la longueur des intervalles selon le plan est la même : l'intervalle est d'autant plus précis que l'on augmente la taille des grappes ou leur nombre.

Conclusion

Nous avons voulu proposer un cadre rigoureux aux mesures mobiles de pollution atmosphérique grâce à la théorie des sondages aléatoires. De telles campagnes de mesures sont déjà effectuées dans certains réseaux mais ne s'appuient souvent que sur le bon sens, et ne permettent pas de proposer d'estimations avec calcul d'erreur rigoureuses. Nous avons vu qu'il existe un cadre théorique pouvant légitimer tout cela, permettre de calculer l'erreur et même de l'anticiper dans une certaine mesure afin d'ajuster les paramètres du plan de sondage.

L'approche par mesures mobiles se veut souvent prospective et son exploitation ne doit pas nécessiter trop de connaissances pointues concernant le site d'étude. Contrairement à l'exploitation par modélisation, les sondages permettent cela, car des estimations de moyenne avec leur erreur d'échantillonnage sont possibles dès une campagne par tubes passifs. Par ailleurs, la mise en œuvre d'un plan de sondage tel que nous l'avons présenté est relativement simple (quelques formules sont à implémenter), et leur usage est donc large. Évidemment, les sondages n'apporteront jamais la précision et l'information que peut contenir un bon modèle, mais force est de constater que les modèles n'existent pas pour tous les polluants et tous les sites ; de plus leur coût d'utilisation est sans comparaison avec une campagne par tubes passifs ou camion-laboratoire.

Nous avons vu qu'un sondage comporte deux étapes : (1) le choix des paramètres du plan de sondage, (2) l'estimation de la moyenne du polluant et de la proportion de mesures au-delà d'un certain seuil. Pour la première étape, on peut choisir les paramètres en fonction de la précision de l'intervalle de confiance qu'on souhaite atteindre, en s'appuyant sur un site fixe connu. Cette étape ne sert qu'à « souffler » les paramètres, et n'influe pas sur la qualité des estimations futures. La seconde étape peut s'effectuer uniquement à partir des mesures relevées, *via* l'estimateur classique, mais aussi en utilisant un site auxiliaire en vue d'améliorer les estimations. Elle permet de déterminer des intervalles de confiance pour la moyenne ou la proportion de mesures au-delà d'un certain seuil.

Reste que la fiabilité des estimations est à contrôler, surtout si le plan ne comporte pas assez de périodes de mesure. Il est conseillé d'effectuer une simulation annexe, comme nous l'avons fait dans le cas pratique p. 562, afin de se rendre compte de cette fiabilité. Si les estimations semblent médiocres, on peut écarter certains estimateurs ou recalculer d'autres intervalles de confiance à partir d'un plan moins précis, comme nous le montrons dans le cas pratique. Pour s'affranchir de tout souci, les simulations de la p. 564 nous suggèrent de choisir un minimum de 10 à 15 grappes de mesures par strate, dans le cas du NO_2. Des simulations à plus grande échelle seraient nécessaires pour confirmer ce genre de suggestions et en proposer de nouvelles pour d'autres polluants. Par ailleurs, il serait intéressant de réfléchir à une estimation de l'erreur plus subtile, qui permettrait une meilleure fiabilité des intervalles de confiance lorsqu'il n'y a pas assez de périodes de mesures.

Entre 1998 et 2001, ATMO Poitou-Charentes a réalisé, dans le cadre du Plan régional de la qualité de l'air, une surveillance dans toutes les communes de plus de 10 000 habitants de la région. Dès 2004, ATMO Poitou-Charentes envisage de renouveler cette étude en utilisant le présent cadre de la théorie des sondages, afin de non pas seulement se donner une idée de la qualité de l'air dans ces communes, mais de proposer une estimation fiable de cette dernière à l'aide d'intervalles de confiance.

Mots clés

Sondage. Estimation. Intervalle de confiance. Mesures mobiles. Plan stratifié. Tubes passifs.

Keywords

Sampling. Estimation. Confidence interval. Moving measures. Stratified plan. Passive tubes.

Références

1. Tillé Y. La théorie des sondages. Ed. Dunod 2001.
2. Droesbeke JJ, Fichet B, Tassi P. Les Sondages. *Economica* 1987.