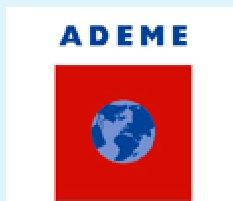


GT « Plans d'échantillonnage et reconstitution de données »

Elaboration d'un plan d'échantillonnage

Laure Malherbe, LCSQA / INERIS



Position du problème

Pour la mesure fixe comme pour la mesure indicative, les directives définissent des objectifs de qualité en ce qui concerne :

- la période minimale à prendre en compte (% de l'année),
- la répartition dans le temps des mesures,
- l'incertitude maximale du résultat final.

Exemple : HAP

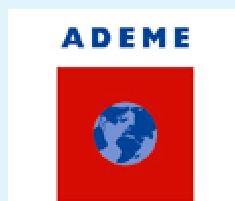
	Période minimale	Incertitude*	Répartition des mesures
Mesure fixe	14 %	50%	« égale » sur l'année
Mesure indicative	6%	50%	

* incertitude de mesure + incertitude liée à la couverture incomplète

Problème:

On souhaite limiter la surveillance à une partie de l'année: combien de mesures sont nécessaires ? A quelles périodes doivent-elles être réalisées...
... pour répondre à la fois aux exigences de qualité et aux contraintes de ressources (disponibilité du personnel, du matériel...) ?

Ces questions relèvent de la **planification de l'échantillonnage**.



Quelques termes d'échantillonnage

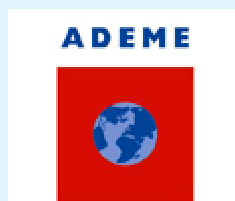
Échantillonnage : tirage d'un certain nombre d'individus dans une population de données (*ex: d'un certain nombre d'heures ou de jours dans une année*)

- à intervalles réguliers : échantillonnage **systématique**
- au hasard : échantillonnage **aléatoire**.

Echantillonnage stratifié : la population est divisée en sous-ensembles, les **strates**, plus homogènes que la population dans son entier. Un échantillon est prélevé dans chaque strate.

Stratification temporelle : année découpée en grandes périodes (ex: quatre trimestres)

Echantillonnage par grappes : Au lieu de prélever des individus isolés (*ex: des heures ou des jours disséminés dans l'année*), on tire des groupes d'individus (*ex: des séries horaires ou des prélèvements journaliers sur plusieurs jours*).



Caractéristiques d'un plan d'échantillonnage

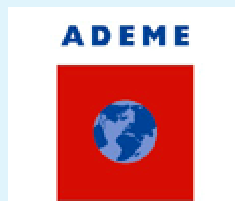
- établi pour une certaine **période**, le plus souvent l'année;
- à partir de séries complètes de concentrations d'années antérieures.

Mode de tirage: systématique ou aléatoire

Paramètres de dimensionnement

- ✓ couverture temporelle (%),
- ✓ nombre de strates temporelles (ns),
- ✓ pour des mesures par grappes :
 - la taille des grappes (en nombre d'heures ou de jours),
- ✓ nombre total de mesures individuelles ou de grappes sur la période considérée,
- ✓ si $ns \geq 2$, nombre de mesures individuelles ou de grappes dans chaque strate.

Eventuellement, **évaluation a priori de la qualité** de la donnée reconstituée, fonction du mode de tirage, de la stratification et du dimensionnement.



Elaboration d'un plan d'échantillonnage aléatoire stratifié

Choix des données pour établir le plan

On souhaite établir un plan d'échantillonnage en un site S, pour l'année N et le polluant P.

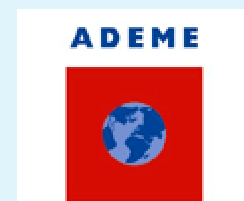
Sélection d'un site permanent S_0 , jugé similaire à S et qui mesure P. Choix d'une série annuelle complète de données de concentration de P.

Contrainte de départ

Qualité (précision) de l'estimation, i.e. on fixe une incertitude maximale $I_{\max}$ en $\mu\text{g}/\text{m}^3$ ou en % (relativement à la moyenne)

Si Z^* est la donnée reconstituée et V la variance de l'estimateur, on cherche à avoir :

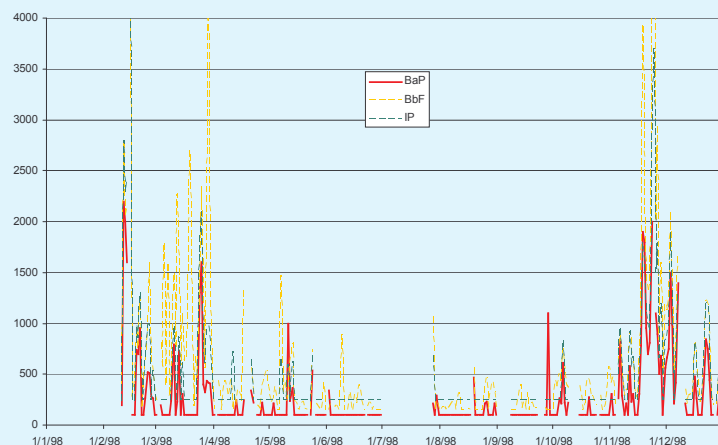
$$2 \cdot \sqrt{V} \leq I_{\max} \left[\mu\text{g} / \text{m}^3 \right] \quad \text{ou} \quad 100 \cdot \frac{2 \cdot \sqrt{V}}{\bar{y}} \leq I_{\max} [\%]$$



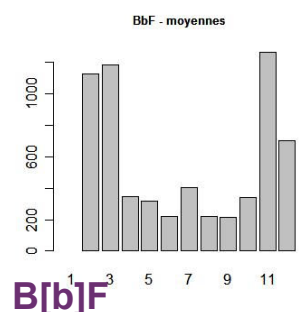
Elaboration d 'un plan d 'échantillonnage stratifié

Application à la mesure de HAP en centre urbain (sous influence du trafic)

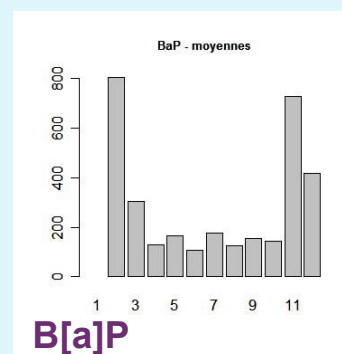
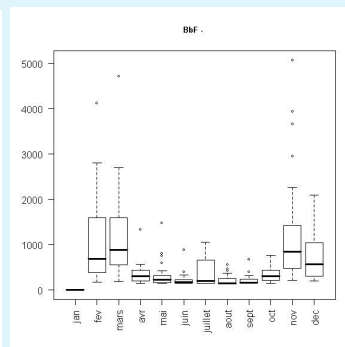
Etape 1a - Etude de la variabilité temporelle des concentrations et stratification temporelle



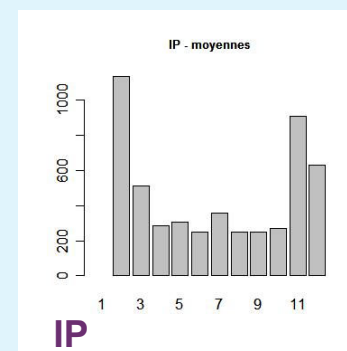
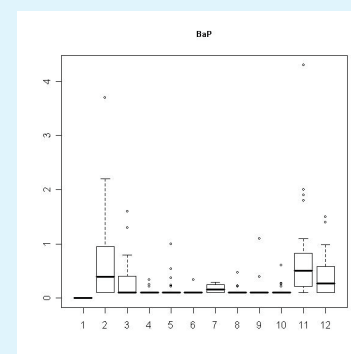
Données AIRPARIF, 1998



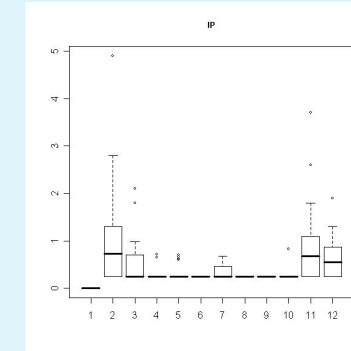
B[b]F



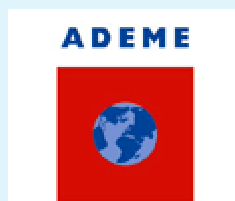
B[a]P



IP



Ex: proposition d 'un découpage en 4 strates
jan-mars / avr.-mai+sept.oct. / juin-août / nov.déc.



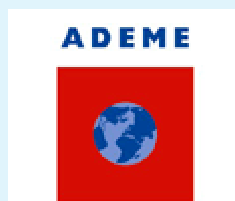
Elaboration d 'un plan d 'échantillonnage stratifié

Etape 2 – Dimensionnement de l'échantillonnage

- 1) Choix de mesures individuelles (ex: prélèvements journaliers non consécutifs) ou de grappes de mesures.
- 2) Dans ce second cas, choix de la **taille des grappes**.
Ex.1 : mesures isolées de 1 jour
Ex.2 : grappes de 3 jours
Ex.3 : grappes de 7 jours
- 3) Détermination du nombre de grappes

Tirage aléatoire : la méthode des plans de sondage permet de calculer directement le **nombre total n de mesures individuelles ou de grappes** que doit inclure l'échantillon pour atteindre la précision recherchée, et le **nombre de mesures individuelles ou de grappes à prélever dans chaque strate temporelle**.

Tirage systématique : pas de formule de dimensionnement, possibilité de procéder par itération : on fixe le nombre et la répartition des mesures et l'on quantifie la précision associée. Si celle-ci n'est pas satisfaisante, on modifie le dimensionnement.



Elaboration d 'un plan d 'échantillonnage aléatoire stratifié

Etape 2 – Dimensionnement de l'échantillonnage (suite)

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

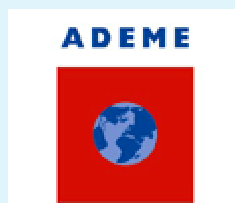
Dimensionnement compatible avec les ressources ?

→ oui...

→ non : on redimensionne l'échantillon et on contrôle la précision correspondante

Exemple : 7 ; 4 ; 3 ; 7 grappes de 3 jours, (total = 21 fois 3 jours) → précision attendue: 36% pour le BaP, 22% pour le B[b]F

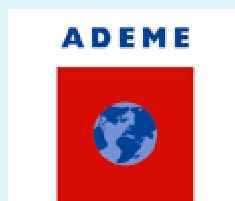
! Calculs réalisés à partir d'une série comprenant plusieurs semaines de données manquantes



Elaboration d 'un plan d 'échantillonnage aléatoire stratifié

Etape 3 – Détermination des dates de mesure

- Dans chaque strate, tirage aléatoire :
 - des dates de début de grappe, pour un échantillonnage aléatoire,
 - de la date de la première mesure, pour un échantillonnage systématique.
- Vérification de la compatibilité des dates avec la disponibilité des ressources.
- Nouveau tirage si nécessaire.



Caractéristiques d'un plan d'échantillonnage

Exemples

Exemple 1

- Échantillonnage **systématique non stratifié**
- **16,4 %** de l'année couverte par des mesures
- **pas de stratification**
- **60 prélèvements** journaliers à raison de **1 prélèvement tous les 6 jours**

-Précision estimée a priori: BaP: 30%; B[a]A: 15%; B[b]F: 35%; B[k]F: 15%; IP: 19% ; NO: 30%; NO₂: 8,4%

Exemple 2

- Échantillonnage **aléatoire stratifié**
- **16,4 %** de l'année couverte par des mesures
- **4 strates**
- **60 prélèvements** dans l'année
- **29, 9, 3, 19** prélèvements en jan-mars, avr.-mai + sept.-oct., juin-août, nov.-déc.

-Précision estimée a priori: BaP: 26%; B[a]A: 17%; B[b]F: 20%; B[k]F: 17%; IP: 18% ; NO: 26%; NO₂: 10,1%

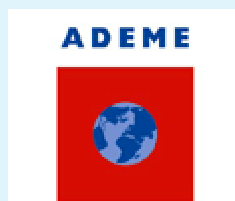
Période d'étude : année

Exemple 3

- Échantillonnage **aléatoire stratifié**
- **17,2 %** de l'année couverte par des mesures
- **4 strates**
- Choix d'un échantillonnage par grappes
 - **grappes de 7 jours** (7 prélèvements consécutifs)
 - **9 grappes** au total dans l'année
 - **5, 1, 1, 3** prélèvements en jan-mars, avr.-mai + sept.-oct., juin-août, nov.-déc.

- Précision estimée a priori : BaP: 36%; B[a]A: 21%; B[b]F: 23%; B[k]F: 28%; IP: 26% ; NO: 37%; NO₂: 12%

* Précision évaluée pour une reconstitution sans redressement.



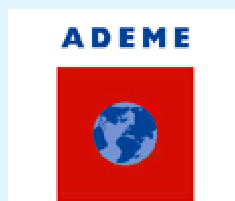
Caractéristiques d'un plan d'échantillonnage

Echantillonnage aléatoire stratifié

- ✓ Etude préalable à la stratification → meilleure connaissance de la variabilité temporelle des concentrations.
- ✓ Possibilité de dimensionner l'échantillonnage pour un objectif de qualité fixé, en adaptant ce dimensionnement à la variabilité temporelle dans chaque strate.
- ✓ Tirage aléatoire dans les strates →
 - permet d'éviter d'éventuels biais d'échantillonnage
 - hypothèse requise par la méthode de reconstitution des « plans de sondage »

Echantillonnage systématique

- ✓ En pratique, souvent utilisé sans stratification car il est simple à définir et permet une bonne répartition des mesures sur l'année.
- ✓ Attention ! La fréquence d'échantillonnage ne doit pas coïncider avec un cycle de variation des concentrations (risque de biaiser l'estimation) ⇒ importance de l'étude préalable à la stratification.
- ✓ Dimensionnement pour une précision souhaitée : pas de calcul direct, méthode itérative.



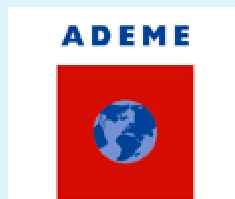
Reconstitution par régression

■ Objectif :

- Reconstituer des indicateurs annuels (moyenne, nb de dépassements de seuil)

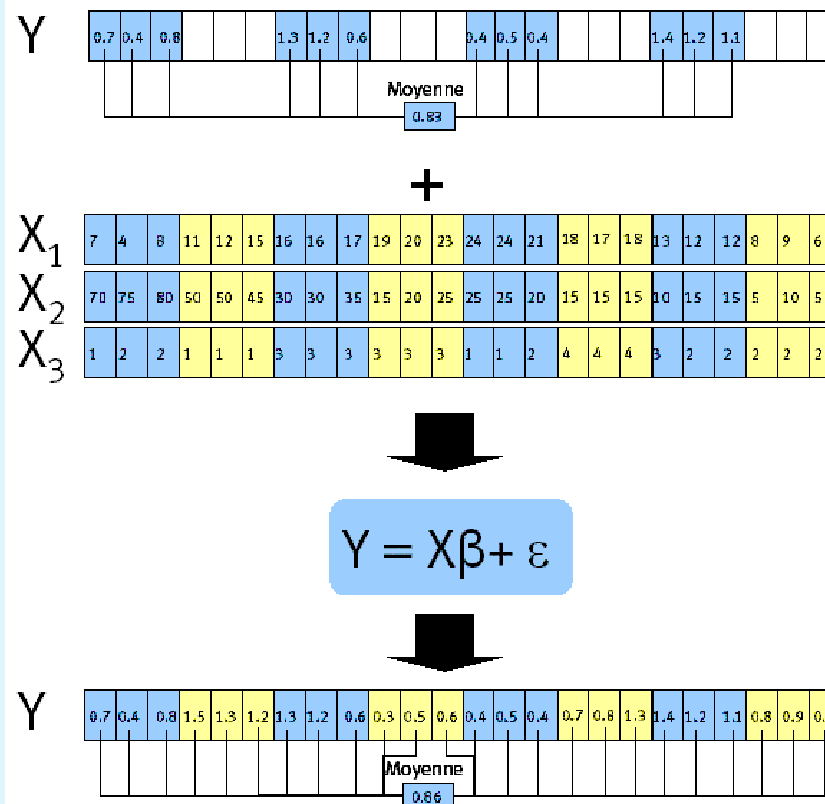
■ Moyens :

- Variable à expliquer : Série chronologique de mesures incomplète
- Variables explicatives : Séries chronologiques de mesures complètes



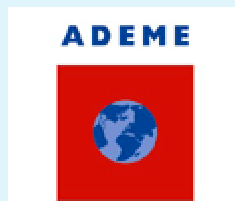
Reconstitution par régression

- Principe : combler les périodes d'absence de mesure par estimation à partir des valeurs connues des variables explicatives.
- Estimation : un modèle de régression détermine, **si elle existe**, une relation entre une variable à expliquer (Y) et des variables explicatives (Xi).



exemple : une relation linéaire

$$Y = a_0 + a_1.X_1 + a_2.X_2 + \dots + a_n.X_n + \varepsilon$$

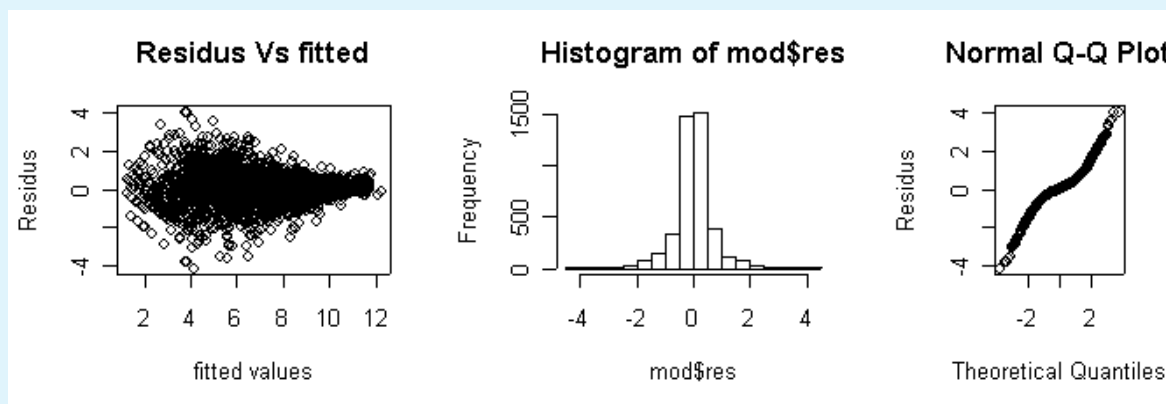
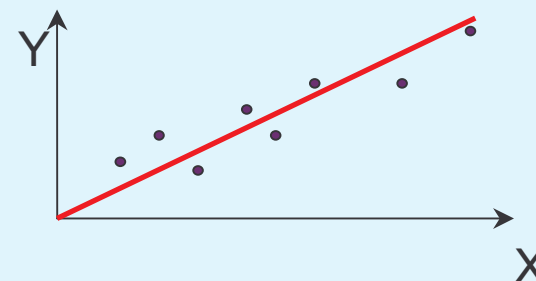
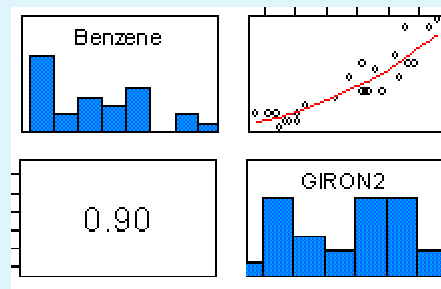


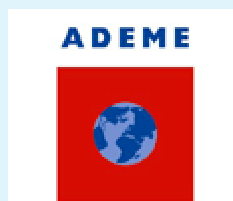
Reconstitution par régression



Etapes :

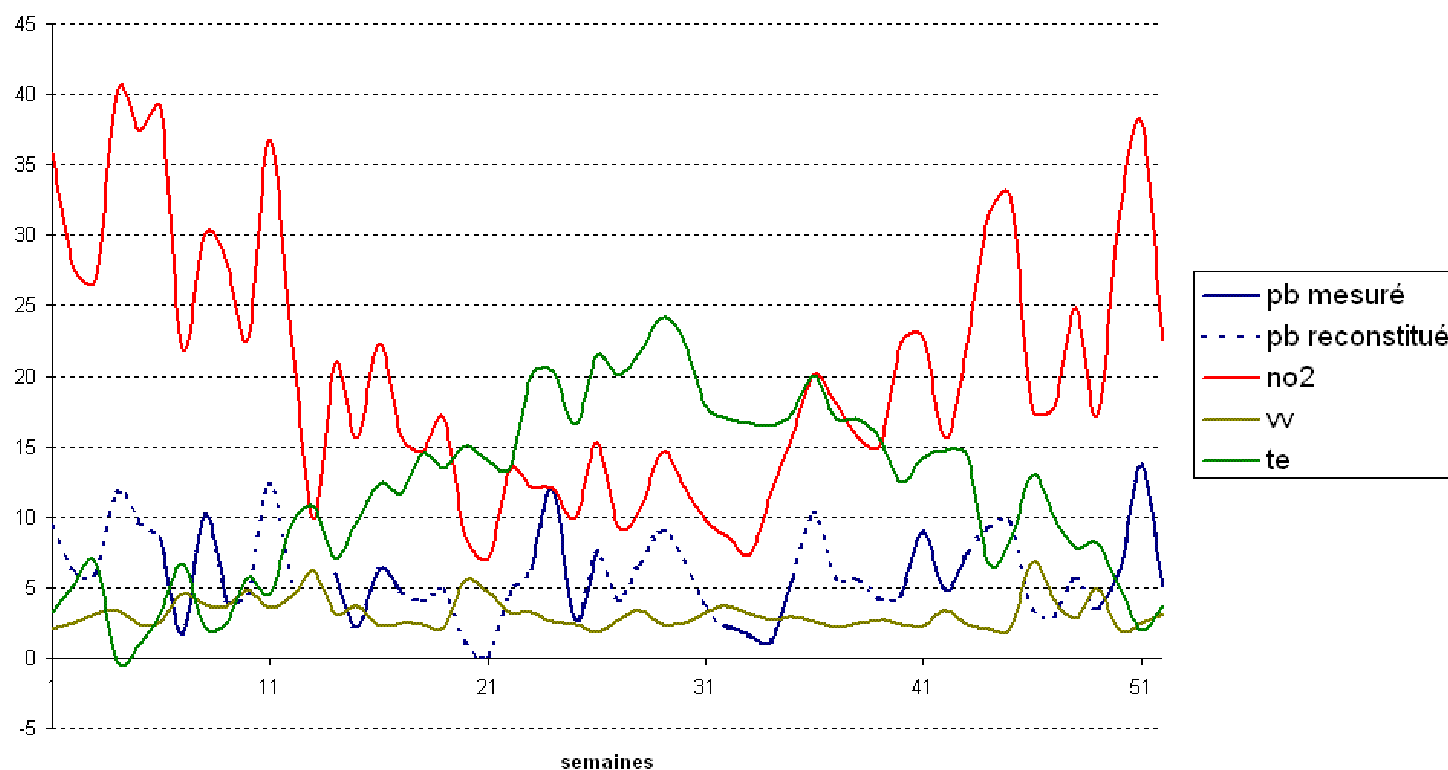
- Sélection des Variables explicatives
- Calcul des coefficients
- Vérification des hypothèses

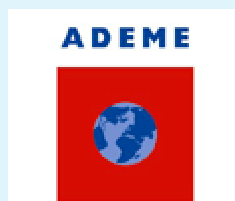




Reconstitution par régression

■ Exemple : données de Plomb à Laval en 2006





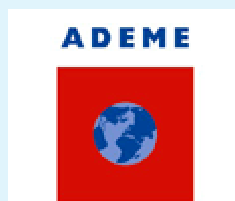
Reconstitution par régression

■ Exemple : données de Plomb à Laval en 2006

variable à expliquer (Y)

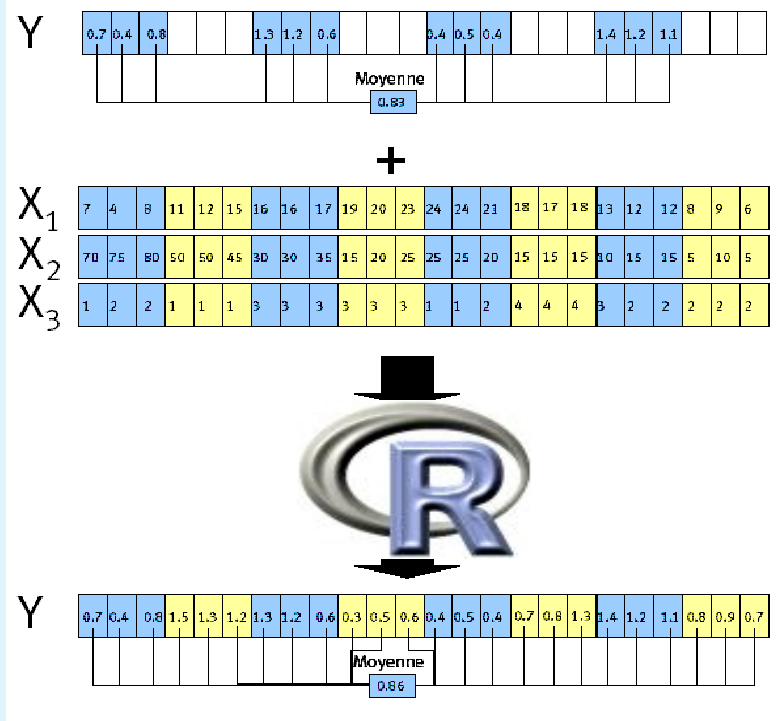
variables explicatives (Xi)

semaine	Pb mesuré	Pb reconstitué	no2	o3	vv	te	hr
s1		9.44	35.94	11.13	2.05	3.19	92.62
s2		6.01	27.65	23.14	2.50	5.19	94.19
s3		5.94	26.80	33.33	3.18	6.82	94.30
s4		11.68	40.33	19.10	3.36	-0.34	79.01
s5		9.69	37.43	20.26	2.40	1.01	89.05
s6	8.31	8.31	39.01	29.20	2.59	3.22	84.74
s7	1.69	1.69	22.02	57.02	4.54	6.73	91.13
s8	10.09	10.09	30.12	34.05	3.80	2.11	85.12
s9	4.15	4.15	28.07	54.64	3.66	2.44	86.51
...	...	...					
moyenne annuelle	5.67	5.93 [5.85;6.01]	...	...	...	...	...

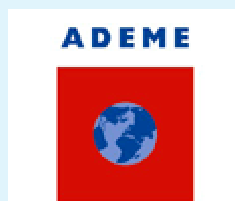


Reconstitution par régression

■ Pratique :



- ## ■ Bibliographie :
- MOILIME D., PENVEN F., 2005.
Reconstitution de données. Mise en œuvre de modèle de régression dans le cadre de la mesure cyclique. Rapport AIRPL.

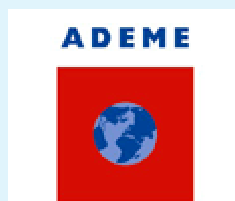


Reconstitution par application de la théorie des plans de sondage

■ Objectifs

- Estimer une concentration moyenne sur une période à partir de
 - Mesures partielles
 - Connaissances complémentaires sur le comportement du polluant
- Évaluer la précision de cette estimation
 - Intervalle de confiance

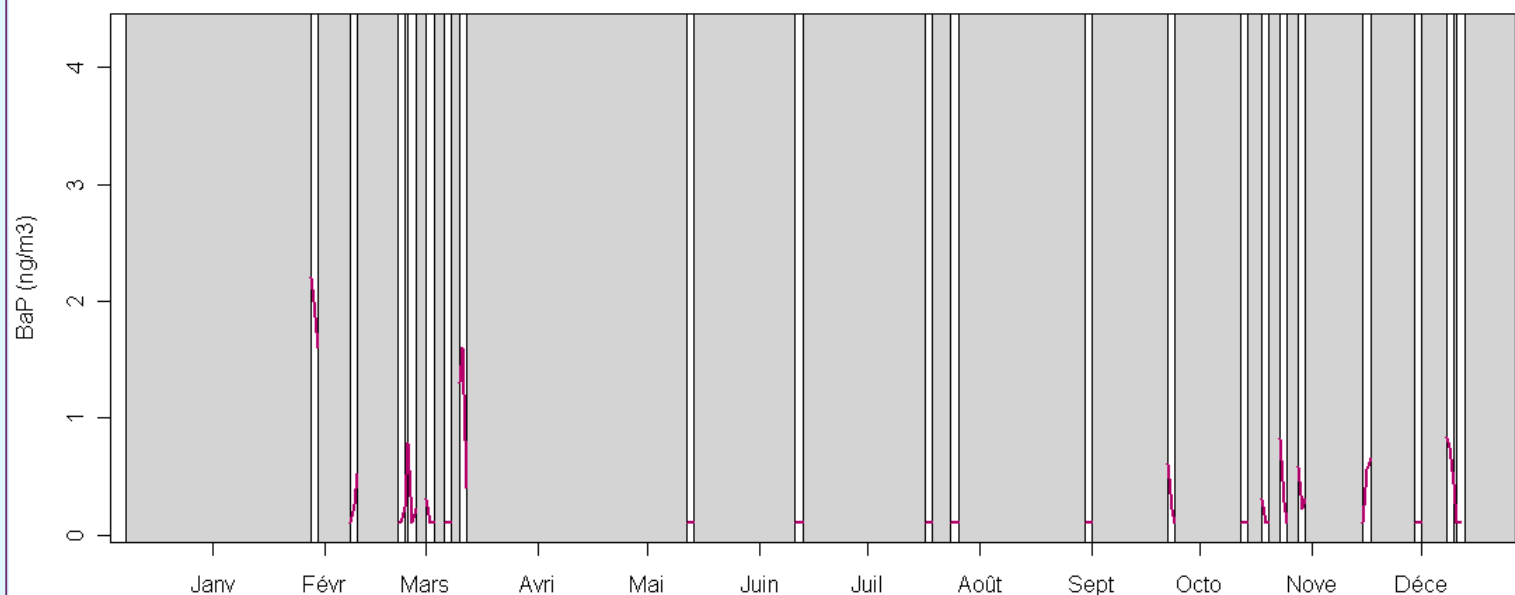
$$\text{moyenne} = \frac{1}{H} \sum_{h=1}^H \frac{1}{n_h} \sum_{i=1}^{n_h} x_i$$

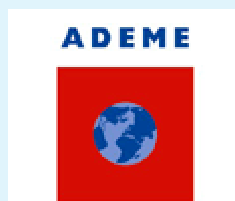


Exemple de BaP en centre urbain, sous influence du trafic (source AirParif)

■ Plan d'échantillonnage

- 21 * 3 jours (d'après la méthode présentée auparavant)

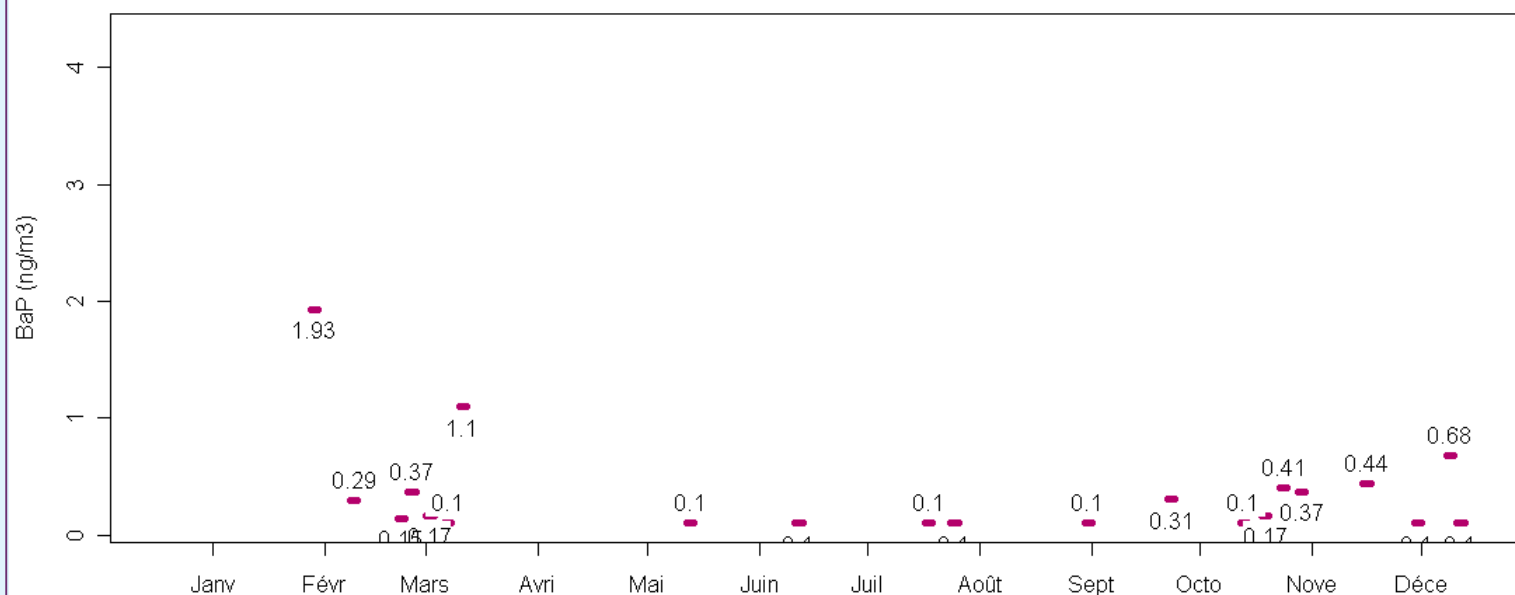




Prise en compte des grappes

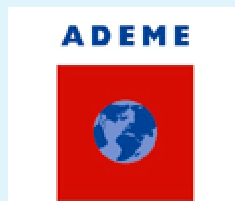
■ Résultats

- Moyenne estimée $0.35 \text{ ng/m}^3 \pm 49\%$
- Moyenne réelle 0.29 ng/m^3



« Plans d'échantillonnage et reconstitution de données », méthode 'plans de sondage'
Vladislav NAVEL ATMO PC

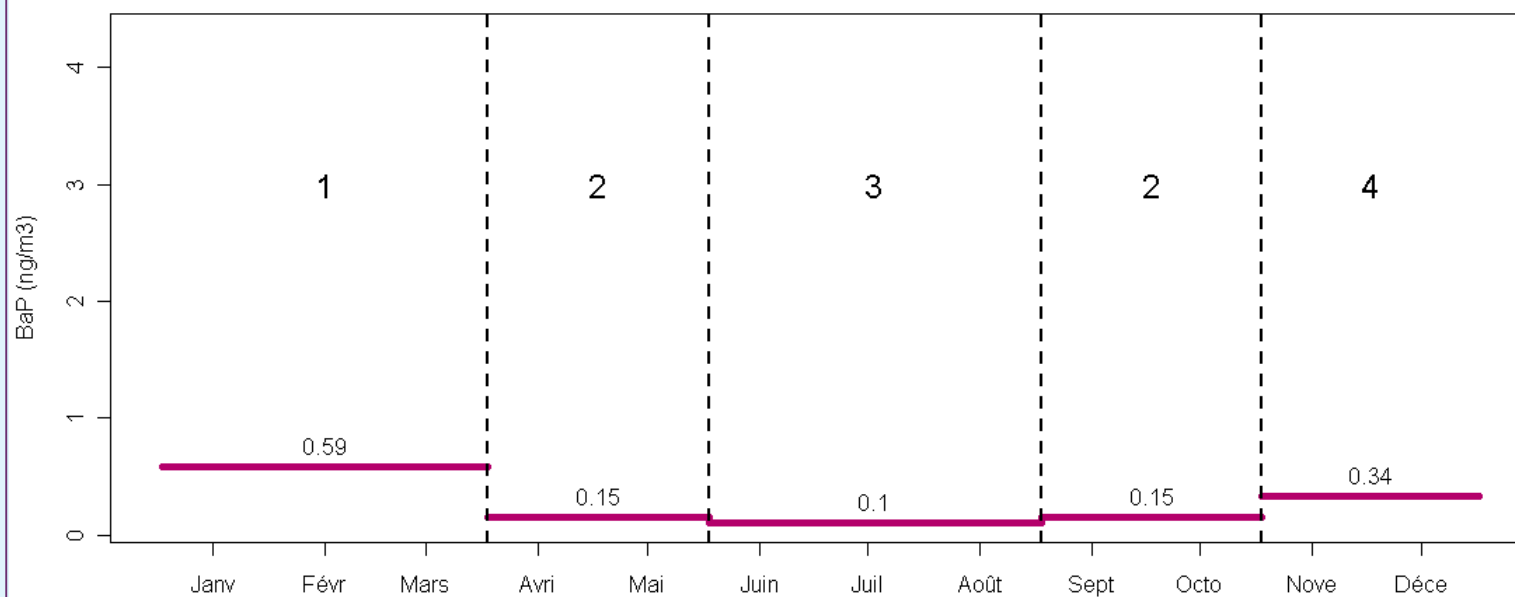
Journée LCSQA
Paris, 19 octobre 07



Prise en compte des strates

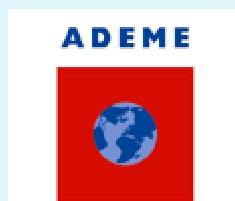
■ Résultats

- Moyenne estimée $0.28 \text{ ng/m}^3 \pm 42\%$
- Moyenne réelle 0.29 ng/m^3



« Plans d'échantillonnage et reconstitution de données », méthode 'plans de sondage'
Vladislav NAVEL ATMO PC

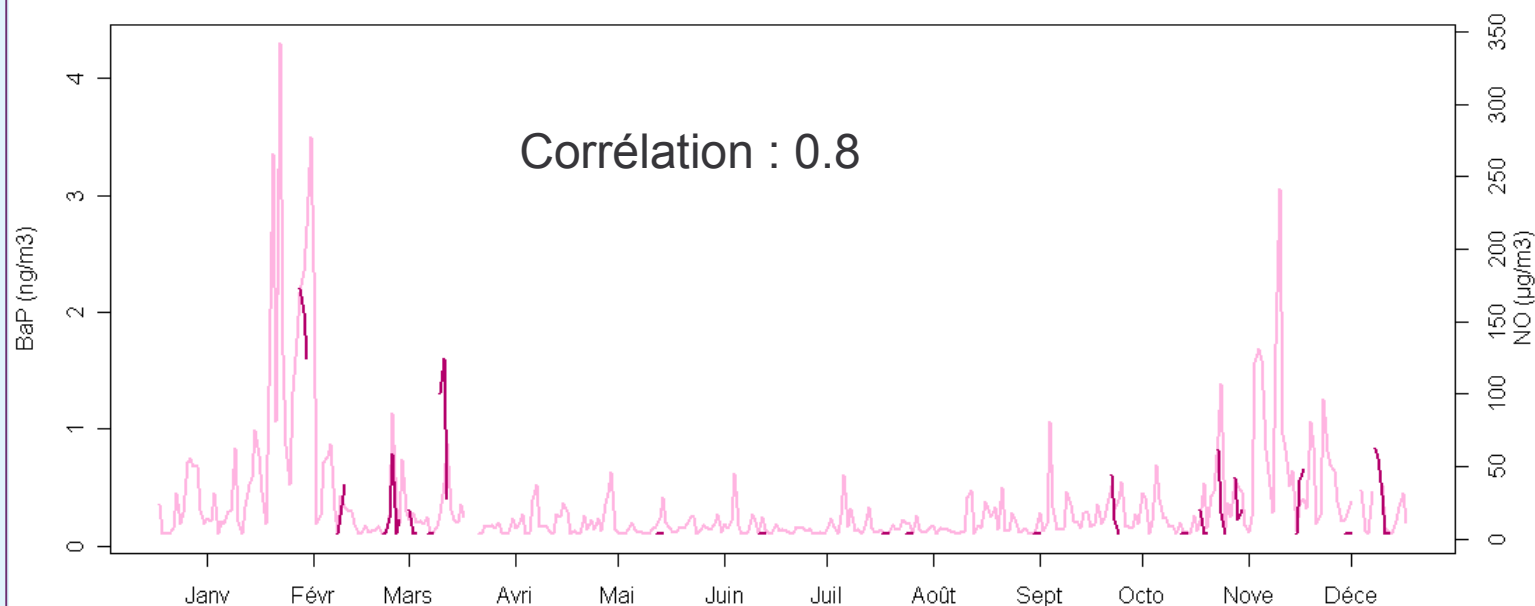
Journée LCSQA
Paris, 19 octobre 07



Prise en compte d'une série auxiliaire

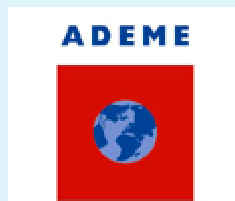
■ Résultats

- Moyenne estimée $0.30 \text{ ng/m}^3 \pm 26\%$
- Moyenne réelle 0.29 ng/m^3



« Plans d'échantillonnage et reconstitution de données », méthode 'plans de sondage'
Vladislav NAVEL ATMO PC

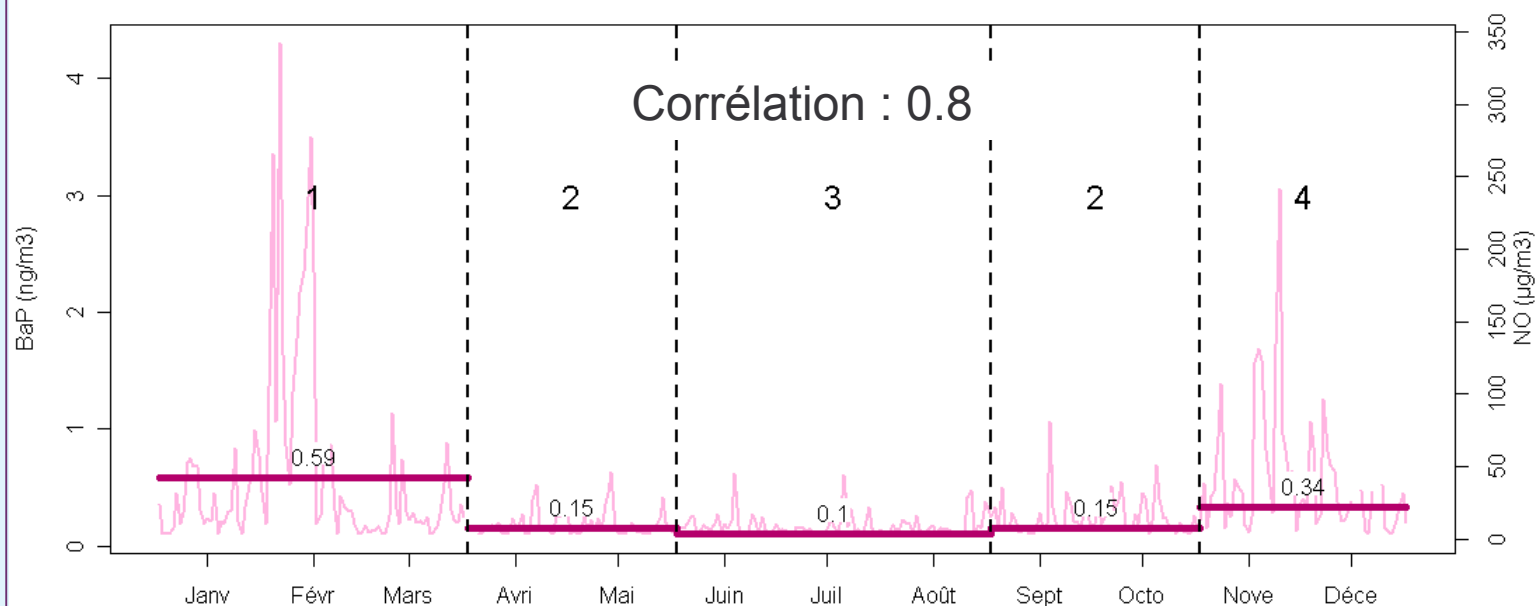
Journée LCSQA
Paris, 19 octobre 07

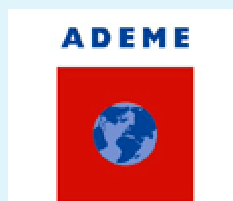


Prise en compte des strates et d'une série auxiliaire

■ Résultats

- Moyenne estimée $0.30 \text{ ng/m}^3 \pm 19\%$
- Moyenne réelle 0.29 ng/m^3



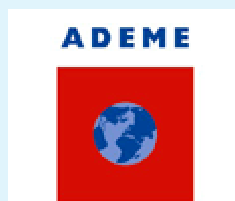


Reconstitution à l'aide d'une stratification paramétrique

Application de la norme
ISO 9359
(Echantillonnage aléatoire stratifié)

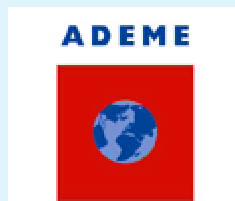
PRINCIPE

REUNIR DES DONNEES PAR GROUPES DE NIVEAUX HOMOGENES,
DANS DES STRATES DETERMINEES PAR UNE COMBINAISON
DE VARIABLES AUXILIAIRES
ASPECT ALEATOIRE + METEO



Etape 1 : Sélection et étude des variables d'influence sur les concentrations

- Exploitation d'une série annuelle
 - de mesures (station de référence) [année (n-x) OU n]
 - et de paramètres associés (« variables explicatives »)
- Pour déterminer les liens entre les concentrations et des
 - paramètres météorologiques :
 - les températures, les vitesses de vent,
 - les directions de vent sélectionnées par rapport à un émetteur (route, zone urbaine...) s'il est l'objet de l'évaluation de la qualité de l'air,
 - paramètres temporels :
 - les heures du jour (photochimie et variations horaires des émissions anthropiques, heures de pointe)
 - les jours ouvrés et non ouvrés
 - les saisons (variations de la température, de l'ensoleillement...)



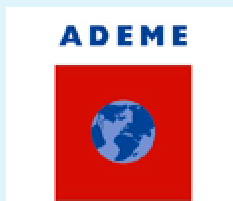
Etape 2 : Classement en strates des séquences horaires en fonction des paramètres sélectionnés

Exemple pour NO [année (n-x) OU n] :

- Identification des valeurs des paramètres météorologiques et temporels qui correspondent aux [NO] les plus élevées :
 - Température : < 8, 10 ou 12°C **selon les régions,**
 - Vitesse du vent : < 1,5 à 2 m/s, **toutes les régions**
 - Horaire de 6 à 9h et de 15 à 21h, **toutes les régions**
 - Jour : du lundi au vendredi. **toutes les régions**

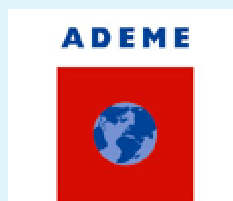
(Etudes déjà réalisées et résultats reproductibles)

[NO]	v < 1,5 m/s	T < 10°C	lundi au vendredi	6 à 9h & 15 à 21h	Σ	N° strate (« i »)
élevée	✓	✓	✓	✓	4	4
	✓		✓	✓	3	4
			✓	✓	2	3
				✓	1	2
faible					0	1

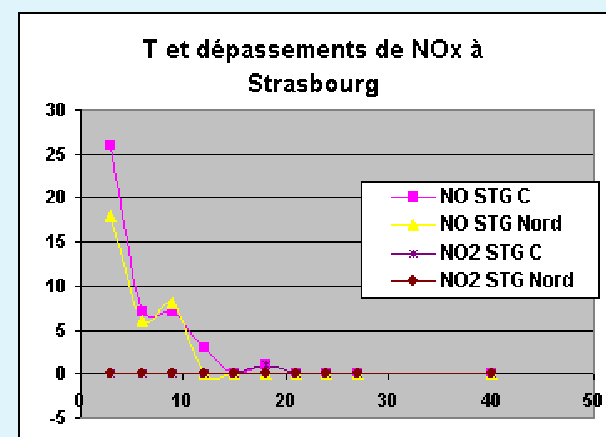
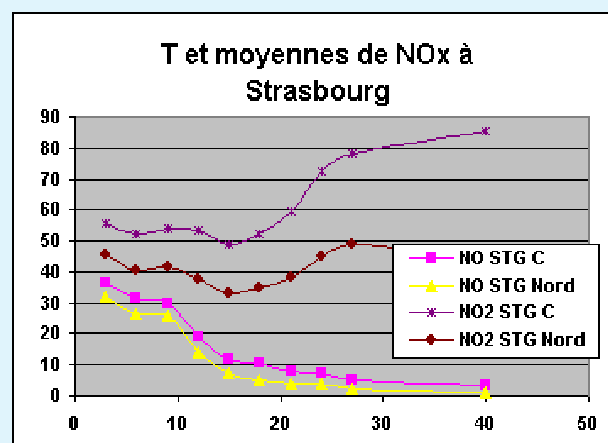
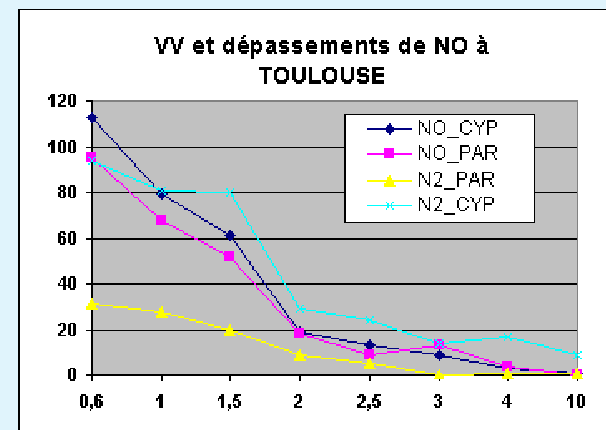
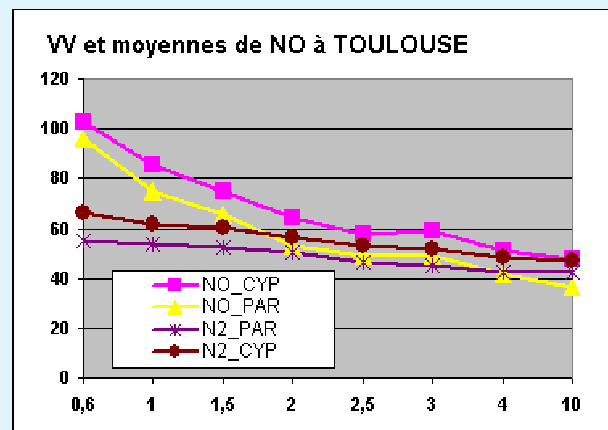


Etape 3 : Reconstitution des données annuelles

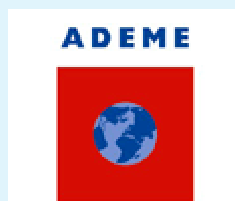
- Ventilation des valeurs horaires de la campagne dans les strates selon les paramètres associés
- Calcul par campagne et par strate « i » des :
 - moyennes et écarts-types associés,
 - nombres de dépassements de seuils prédéterminés
 - centiles.
- Reconstitution sur l'année complète
 - pondération des calculs par les taux de remplissage de ces strates (campagnes / année)
 - calcul des incertitudes associées.
- Redressement à l'aide d'une station de référence
 - Campagne simulée même période que la campagne,
 - Mêmes valeurs des paramètres d'influence



Exemples d'influences de VV et T sur NO et NO₂, moyennes et dépassements C90

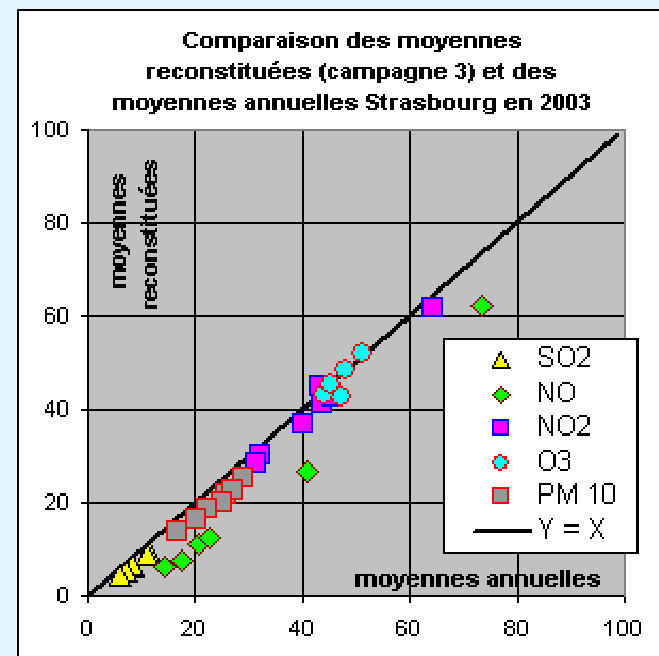
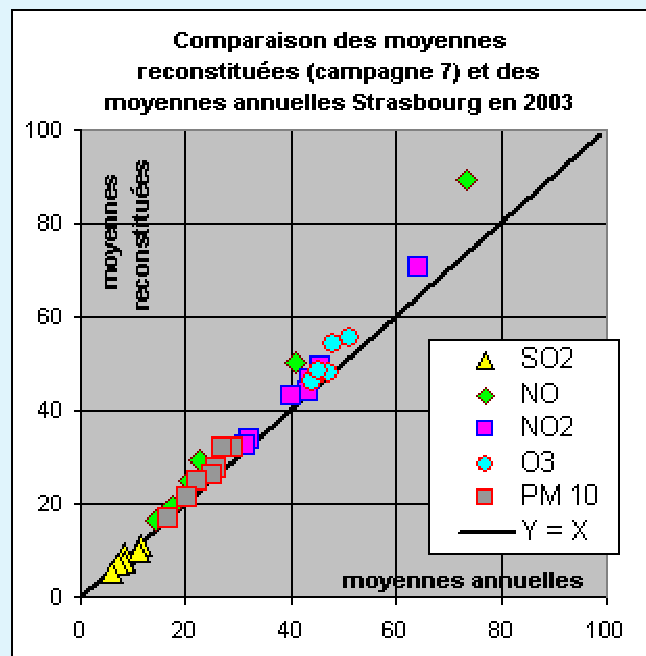


Stratégie différente si l'objectif = moyennes ou dépassements
Campagnes réparties sur l'année ou périodes à risques

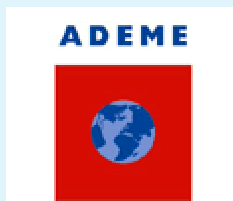


RESULTATS : choix des dates de campagne !

8 campagnes de 4x15 jours à 8 dates différentes

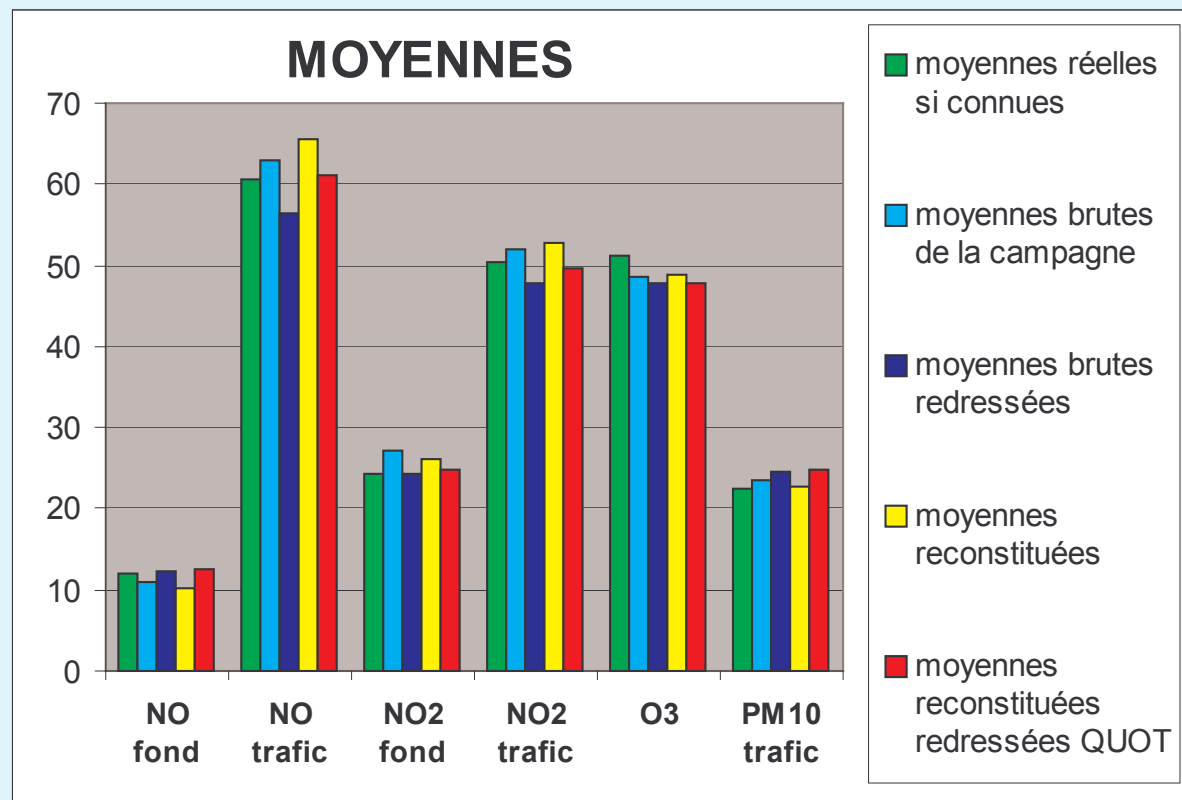


Entre « sur » et « sous » estimation =>
selon les dates des campagnes !



Influence de la reconstitution et du redressement

1 Campagne de 15 jours par trimestre



Nécessité de redresser / station de référence