



Mesures en nombre insuffisant

par Bernard Beauzamy
PDG, SCM SA

16 janvier 2008

Résumé opérationnel

Cette Note est un complément méthodologique à plusieurs contrats que nous avons traités au cours de ces dernières années, en particulier :

- Veolia Environnement, Région Ouest, 2005 à 2007 ;
- Direction Générale de l'Energie et des Matières Premières, 2006 et 2007 ;
- Agence Européenne de l'Environnement, 2007.

Dans chaque cas, en effet, une décision doit être prise, sur la base d'une information incomplète : les mesures dont on dispose sont en nombre insuffisant pour permettre une connaissance parfaite du phénomène considéré. Quelles sont les erreurs à éviter ?

La question n'est pas, insistons-y bien, de la simple reconstitution de données manquantes, à partir de données existantes (comme par exemple le débit d'un fleuve connaissant celui des fleuves voisins) : nous savons faire cela, au moins grossièrement. La question se pose lors d'une ignorance complète : une zone tout entière, une époque tout entière, sont complètement dépourvus de données.

Nos conclusions sont les suivantes :

- Il ne faut jamais essayer de compléter la connaissance sur une base uniquement statistique (ou probabiliste, comme on voudra), à partir des données elles-mêmes. Par exemple, il ne faut pas procéder à des régressions linéaires, à des extrapolations, etc., parce que de telles reconstitutions sont absolument factices. Il faut toujours essayer de faire intervenir, au moins de manière simplifiée, la physique du problème : sachant que le processus évolue de telle manière, et que l'on a trouvé telle valeur à telle date ou à tel endroit, voici ce qu'on peut s'attendre à trouver ici.

Par conséquent, toute méthode qui ne fait pas intervenir, au moins de manière grossière, la physique du problème doit systématiquement être bannie.

Mais il ne faut pas non plus tomber dans l'excès inverse, en ayant recours à des modèles physiques trop fins, dont le coût serait excessif, et qui requerraient d'autres données, tout aussi indisponibles. Prenons un exemple précis pour illustrer ceci (il est détaillé plus bas). Supposons que l'on dispose des débits journaliers d'un fleuve pendant une année, et on veut les reconstruire l'année suivante.

Un simple prolongement linéaire n'a aucun sens. On essaiera donc de se servir des données de pluviométrie, généralement disponibles. Mais construire un modèle « pluie-débit », reliant le débit d'un fleuve aux données de pluviométrie aux alentours, est très difficile et requiert des informations sur les coefficients de ruissellement, la géologie, toutes données généralement indisponibles.

On se contentera donc de rechercher les endroits où, pendant l'année disponible, la pluviométrie est la mieux corrélée au débit (avec éventuellement un décalage temporel) et se servir de cette corrélation pour reconstruire l'année de débits manquante à partir de la pluviométrie connue. On a donc bien les ingrédients annoncés : un modèle physique, mais simple et robuste.

- Un réseau de mesure est souvent construit pour répondre à une utilité sociale (connaître le besoin en eau d'une ville, le degré de pollution d'une rivière, le besoin en énergie, etc.). Cette utilité sociale est locale (limitée dans le temps et dans l'espace) et souvent grossière, en ce sens que l'on n'a pas besoin de mesures d'une extrême précision. Il ne faut pas prétendre que cette connaissance sociale est suffisante pour une connaissance scientifique fine du processus.

Par exemple, la connaissance de la température à la surface du globe est dictée par des préoccupations de production énergétique, de dimensionnement de bâtiments et d'infrastructures, et de voyages ; le réseau existant suffit grossièrement à cet usage. Il ne faut pas vouloir en déduire que nous connaissons la température du globe en tout point, en tout instant. La définition d'un réseau de capteurs à usage scientifique est un problème entièrement différent. Les questions : combien de capteurs faut-il ? où les mettre ? avec quelle fréquence récupérer l'information ? doivent alors être abordées avec l'objectif de la connaissance globale. Le coût sera évidemment très élevé, et l'utilité sociale sera faible.

Dans toutes les situations que nous avons rencontrées, le réseau « social » suffit à peine au besoin local et grossier qu'il est supposé couvrir : les capteurs cessent de fonctionner, ou sont mal entretenus, etc. Dans ces conditions, l'utiliser pour un objectif scientifique global et précis est totalement illusoire, et souvent malhonnête.

I. Défauts dans les mesures

Lorsqu'on mesure un processus physique, deux défauts peuvent être rencontrés, séparément ou, plus souvent, tous deux ensemble :

- Il y a des erreurs dans les mesures ;
- Les mesures ne sont pas en nombre suffisant.

Ceci s'applique évidemment à tout processus réel : pollution dans une ville, pression sanguine, température en un lieu, consommation énergétique, pour ne citer que quelques exemples évidents.

Culturellement parlant, les deux sont traités de manière très différente.

Les responsables de chaque domaine savent parfaitement que toute mesure est affectée d'une certaine incertitude, qui peut généralement être évaluée, au moins de manière grossière. Les outils mathématiques permettant d'estimer les conséquences sur le résultat existent, et si on s'en donnait la peine, on pourrait presque toujours donner une incertitude sur le résultat. Lorsque ce n'est pas le cas, on peut donc généralement incriminer l'incompétence, ou la négligence, des responsables. Un décideur a le devoir d'écarter toute solution qui lui est présentée sans être assortie d'une analyse sur les incertitudes (que ce soit sous forme de loi de probabilité, d'intervalle de confiance, ou ce que l'on voudra).

L'insuffisance du nombre de mesures est beaucoup plus difficile à estimer : quand peut-on dire que les mesures sont en nombre suffisant ? Quelles sont les conséquences, lorsque ce n'est pas le cas ? C'est de ce problème que nous allons traiter ici.

L'insuffisance des mesures est souvent traitée par les responsables avec plus de légèreté encore que les incertitudes : on complète, on interpole, on extrapole, tout ceci sans justification ni précaution, souvent sans même le dire. Les conséquences sur l'objectif recherché, qui est une aide à la décision, sont souvent funestes.

Nous allons commencer par traiter les aspects mathématiques du problème.

II. Deux modes de mesures : mesures régulières et mesures aléatoires

En pratique, on rencontre deux modes de mesures :

- Mesures régulières : on effectue des mesures à intervalles de temps constants, ou bien à distance constante. C'est ce que l'on appelle un « échantillonnage régulier » (le mot « échantillon », utilisé en statistiques, signifie « ensemble de mesures »).

C'est le cas pour la température à Paris, par exemple, qui est mesurée toutes les dix minutes (elle est mesurée en continu, mais l'indication est remise à jour toutes les dix minutes). Ce sera aussi le cas pour un réseau de capteurs de pollution, qui, une fois par jour, transmettra des relevés à un site central.

- Mesures irrégulières : les mesures n'ont pas pu être prises de manière régulière, soit parce qu'on ne l'a pas voulu, soit parce qu'on ne l'a pas pu.

Il arrive très souvent, en effet, surtout pour les informations du passé, que peu de détails soient disponibles : pour un séisme ancien, pour une épidémie, pour une inondation, on ne dispose que de données très partielles, du type « à telle date, la hauteur d'eau du fleuve dépassait tel niveau ». Cela n'a rien à voir avec un relevé régulier des hauteurs.

Il arrive aussi que l'on ne veuille pas recueillir toutes les mesures, généralement pour réaliser des économies : c'est typiquement le cas pour les sondages. On se contente d'interroger un « échantillon représentatif de la population » et de recueillir les informations relatives à cet échantillon. On en déduit, tant bien que mal, les informations relatives à la population entière.

Ceci s'applique en particulier au contrôle de la qualité, qui est une forme de sondage : on ne vérifie pas toutes les pièces en sortie d'une chaîne de fabrication, mais un échantillon pris au hasard.

Sur un territoire, la mise en place de stations permettant de suivre la pollution (par exemple dans des fleuves) se fait de manière irrégulière : on ne met ces stations que dans les zones que l'on estime concernées, ce qui pose un problème d'interprétation statistique (voir notre travail [1] pour l'Agence Européenne de l'Environnement). La plupart des réseaux de surveillance fonctionnent sur le même principe : on ne met des capteurs que dans les zones à surveiller, ce qui est complètement légitime à la fois sur le plan physique et sur le plan financier, mais il ne faut pas interpréter les résultats comme un sondage.

A. Cas d'un processus périodique

La plupart des phénomènes naturels ont une composante périodique : météo en fonction des saisons, cycle diurne pour le corps humain, etc. Il est donc légitime de commencer par regarder ce qui se passe pour un phénomène complètement périodique.

Considérons un phénomène très simple, mais d'oscillation rapide : un signal, donné par une équation du type :

$$y = \sin(2000\pi t)$$

où t est donné en secondes. Au bout du premier millième de seconde, le signal se reproduit à l'identique ; il y a 1000 répétitions à l'identique en une seconde.

Si l'on sait que le signal est périodique, il est inutile de faire des mesures tous les millièmes de seconde : le second intervalle est identique au premier. Par contre, la question se pose de savoir combien de mesures faire entre $t = 0$ et $t = 1/1000$; la réponse n'a rien d'évident : elle dépend de la forme du signal et de sa rapidité de variation.

Il peut parfaitement arriver qu'une période se décompose en une moitié de calme plat, suivie d'une moitié de variation rapide :

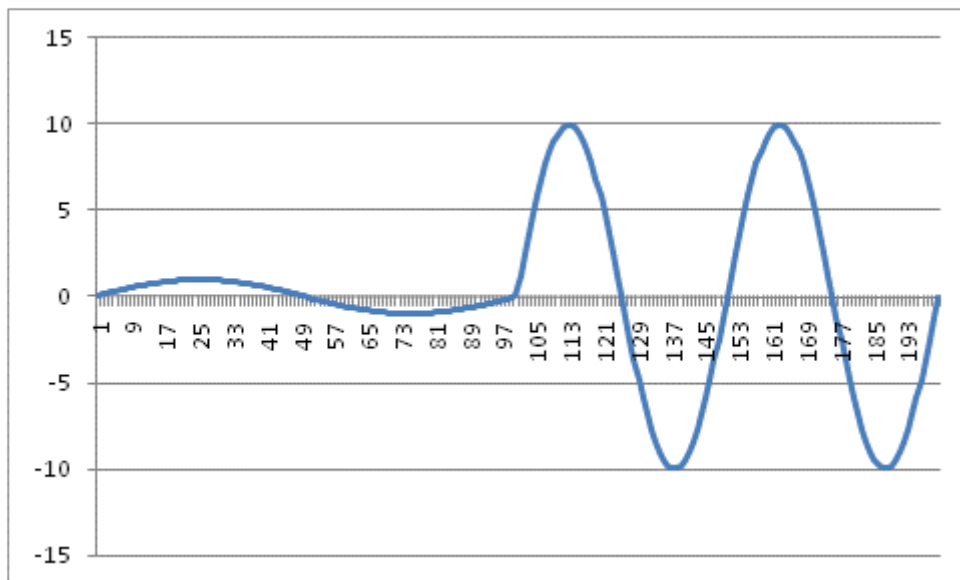


Figure 1 : variation du signal pendant une période

En ce cas, une mesure faite pendant la première moitié de la période ne révèle rien sur la seconde.

Une règle fondamentale apparaît donc très simplement : l'intervalle entre mesures est fonction de la vitesse de variation du processus. La question n'est donc pas de la fréquence du processus (ici toutes les millisecondes) mais de la rapidité de ses variations pendant chaque période. Dans le cas présent, nous voyons clairement que trois mesures, prises à intervalles réguliers pendant chaque milliseconde, ne donneront pas de bons résultats : on ne pourra pas reconstituer les pics de la seconde moitié de l'intervalle. Il faudra au moins dix mesures pour y parvenir.

On voit ici un danger de la régularité de la mesure : si le processus est régulier et la mesure aussi, avec une fréquence inappropriée, on ne s'apercevra jamais des défauts. La mesure donne toujours quelque chose de cohérent, de périodique, mais sans refléter correctement le processus physique. On a une reconstitution fautive et on ne s'en aperçoit jamais.

Echantillonnage

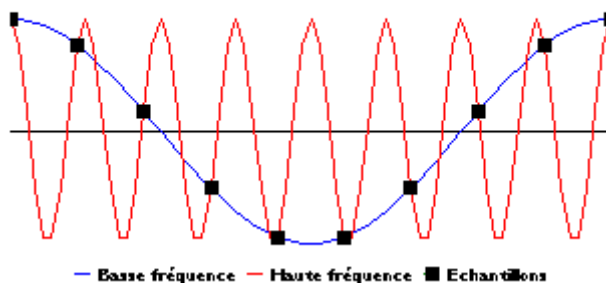


Figure 2 : un exemple caractéristique de sous-échantillonnage

La figure ci-dessus est extraite de Wikipedia « Théorème d'échantillonnage de Nyquist-Shannon » ; elle montre un exemple caractéristique de sous-échantillonnage : le vrai signal est en rouge et le signal reconstitué en bleu. On ne se rend jamais compte de la vitesse de variation réelle du vrai signal.

Ici, la mise en place d'une méthode d'échantillonnage aléatoire (c'est-à-dire à pas irrégulier) permettra au moins la mise en évidence de pics que l'on ne soupçonnait pas.

B. Défauts de l'échantillonnage aléatoire

Le principal défaut de l'échantillonnage aléatoire est qu'il est généralement insuffisant. Prenons un exemple concret.

Voici le même signal sinusoïdal que précédemment (le graphe n'est tracé que jusqu'à 0.01) :

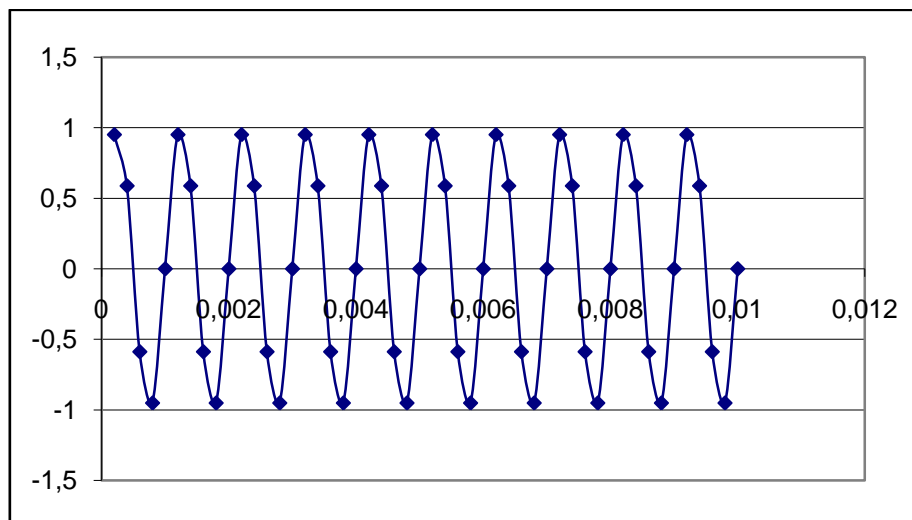


Figure 3 : un signal aléatoire

et voici deux échantillons, de cent mesures chacun :

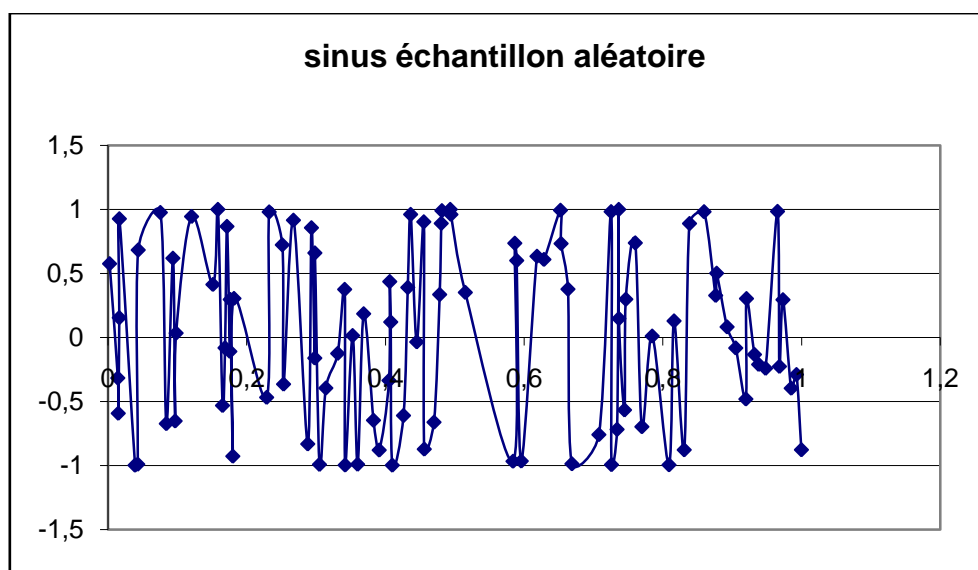


Figure 4 : échantillonnage 1

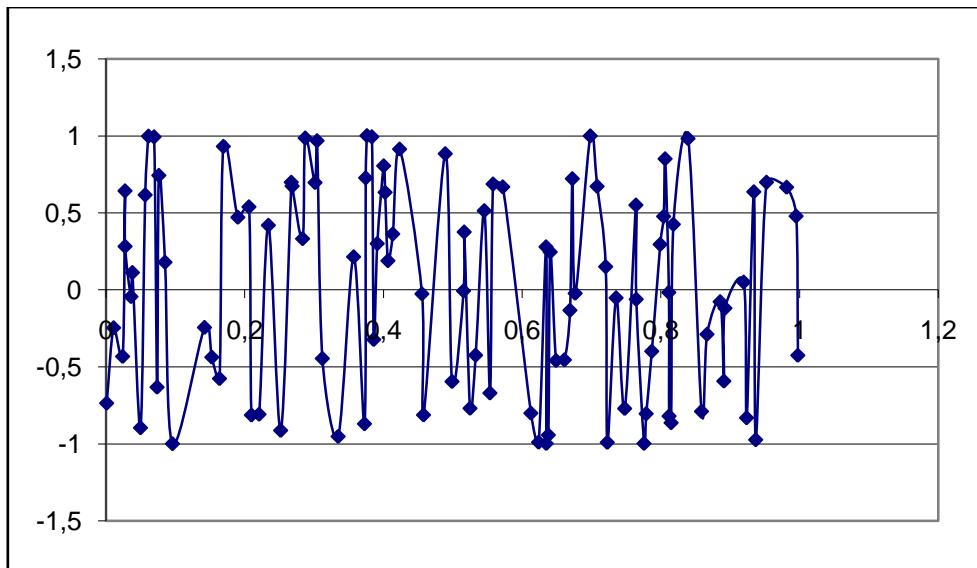


Figure 5 : échantillonnage 2

Les deux échantillons ne semblent pas avoir grand' chose en commun.

Si maintenant on s'intéresse à une valeur caractéristique, comme la moyenne, on trouve ceci :

La vraie moyenne, sur l'ensemble du phénomène, est nulle. Le calcul avec dix-mille points donne une approximation correcte, à savoir -3×10^{-15} . Par contre, le premier échantillon de 100 mesures donne une moyenne de 0.04 et le second de 0.01.

III. Cas d'un processus réel

Un processus réel n'est jamais véritablement périodique, mais il peut avoir une composante périodique, comme la température en un point donné. Pratiquement toutes les activités humaines ont une composante périodique (journalière, annuelle, etc.).

Nous prenons ici le cas des températures journalières moyennes relevées à Paris, du 1^{er} janvier 1900 au 31 décembre 2000 (soit 36 890 données). La moyenne de l'ensemble est 11,5°C.

Voici le début du graphe :

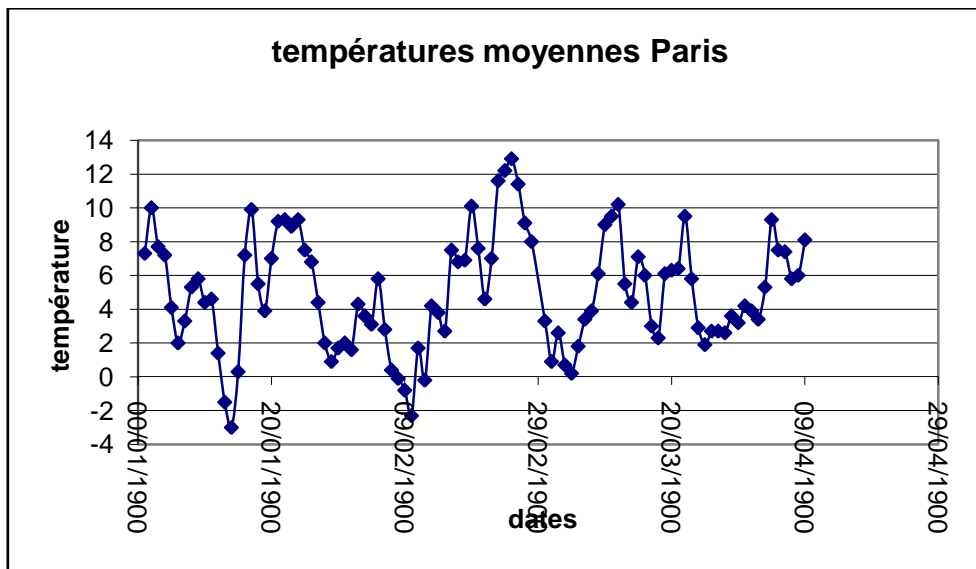


Figure 6 : températures journalières moyennes à Paris

Supposons maintenant que, parmi toutes ces dates, nous en ayons gardé 100 prises au hasard. Voici le graphe obtenu :

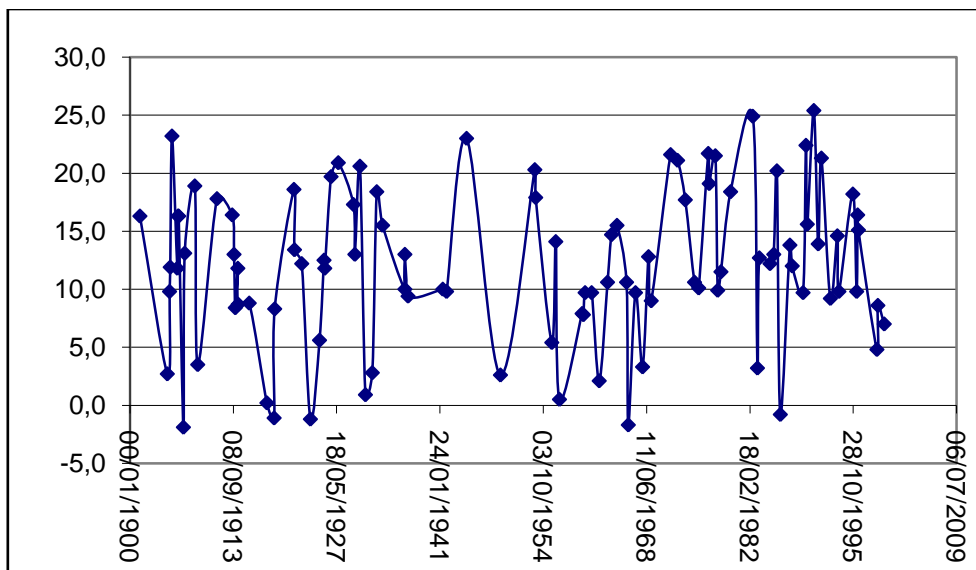


Figure 7 : échantillon de 100 dates prises au hasard

La moyenne en ce cas est 12,1°C.

Si on fait une nouvelle extraction, indépendamment de la précédente, on obtient des résultats différents :

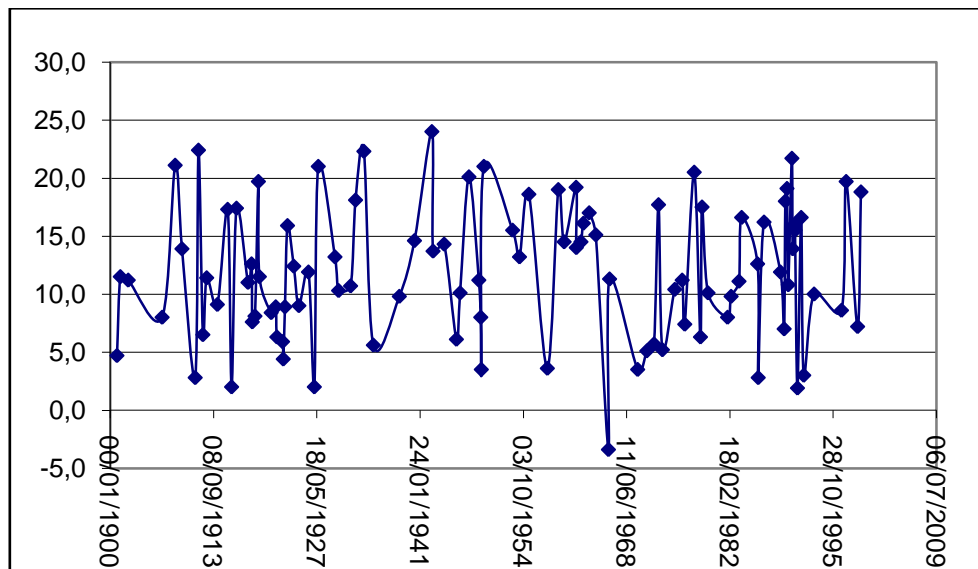


Figure 8 : autre échantillonnage

et la moyenne est $11,9^{\circ}\text{C}$.

Bien entendu, d'autres extractions donneraient d'autres résultats, avec des moyennes différentes.

Le processus est suffisamment irrégulier pour que l'extraction de 100 dates seulement, sur près de 14 000, ne soit pas un indicateur très précis. A partir de ces 100 dates, on ne pourra pas dire, par exemple, s'il y a eu réchauffement ou non.

Remarquons pourtant que nous travaillons sur des moyennes journalières. Si les données étaient des températures instantanées, l'effet du sous-échantillonnage serait encore plus visible. De même, évidemment, si nos données étaient prises au hasard entre de nombreux lieux : ici, toutes sont en un seul lieu (Paris).

Remarquons également que l'extraction de 100 données sur 36 000 (une sur 360) est un bien meilleur échantillonnage que le sondage de 1000 personnes sur soixante millions (une sur soixante mille). Mais lorsqu'on réalise un sondage à des fins électorales, on a par avance confectionné un panel représentatif : on sait quelles catégories de gens il faut interroger, et combien sont nécessaires dans chacune des catégories ; en d'autres termes, on connaît déjà la loi de probabilité du phénomène. Dans le cas de la température, on ne connaît pas cette loi.

IV. Reconstitution de données manquantes

Dans une situation d'absence de données, sur une zone ou un intervalle de temps, beaucoup d'organismes adoptent l'une des deux attitudes suivantes :

- ou bien ils font comme si les données disponibles étaient représentatives de l'ensemble et nous venons de voir que cette attitude est erronée ;
- ou bien ils « refabriquent » les données manquantes.

Vouloir reconstituer les données manquantes est tout à fait légitime, pour une aide à la décision (décider d'irriguer, de produire de l'énergie, etc.), mais complètement illégitime pour la connaissance scientifique d'ensemble, car les données reconstituées, par définition, ne comportent pas d'information scientifique propre.

Il faut en outre être vigilant quant aux méthodes de reconstruction :

Une méthode purement statistique, par exemple d'ajustement linéaire, est totalement à proscrire, car elle fait une hypothèse factice (ici de linéarité).

Pour faire comprendre ceci, prenons l'exemple d'une rivière, dont les débits journaliers sont connus sur deux ans, Y_1 et Y_3 et on cherche à reconstituer l'année intermédiaire manquante Y_2 : ce n'est pas du tout la situation de notre livre [3] où nous disposons de 19 fleuves ; quand l'un manquait, nous nous appuyions sur les autres. Ici, il n'y en a qu'un.

La tentation de procéder à un ajustement linéaire en utilisant les deux années connues est évidemment absurde, parce qu'elle donnera une variation linéaire pendant l'année manquante.

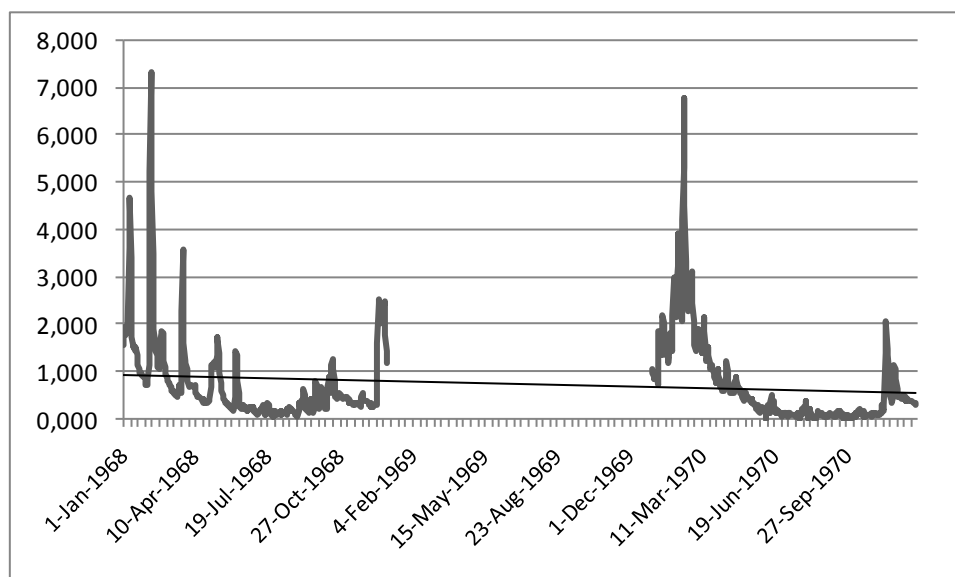


Figure 9 : reconstitution linéaire de la Sèvre Nantaise en 1969 à partir des données de 1968 et 1970

Pour reconstituer correctement, il faut utiliser une information physique, en l'occurrence la pluie, dont les mesures sont connues. On peut pour cela travailler de manière fine, en construisant un « modèle pluie-débit », qui exploitera la géographie du terrain, sa porosité, les coefficients de ruissellement, etc. De tels modèles sont coûteux à construire et à alimenter. Ou bien, on travaillera de manière grossière, et c'est ce que nous recommandons : en se servant des deux années pour lesquelles les données sont disponibles, on cherchera, parmi les lieux alentours, ceux pour lesquels la pluie en ce lieu est la mieux corrélée au débit du fleuve (avec éventuellement un décalage temporel), et on se servira de cette corrélation pour reconstituer les données de débit manquantes.

Cette approche consiste à insérer un modèle mécaniste grossier, souvent linéaire ou linéaire par morceaux. Mais cette linéarité n'a rien à voir avec la reconstruction statistique linéaire ! dire que le débit est proportionnel à la pluie est grosso-modo correct, au moins dans certaines limites.

Nous avons utilisé cette approche mécaniste grossière dans un travail en collaboration avec l'INERIS (écotoxicologie), 1997-2000 [5] : le modèle mécaniste consistait à dire que le poison pénétrait dans l'organisme selon certaines lois simples, était éliminé selon d'autres lois simples, et que son action était proportionnelle à certaines concentrations.

Bien sûr, il faut ensuite chercher à valider le modèle ; dans bien des cas on s'aperçoit qu'il n'est pas correct, ou pas suffisant. C'est une première approche, beaucoup plus satisfaisante que les deux extrêmes que l'on rencontre : uniquement des ajustements statistiques et pas de modèle du tout, ou bien des modèles physiques extrêmement fins, mais inexploitable en pratique, faute de données pertinentes pour les caler.

V. Utilité sociale et connaissance scientifique

Beaucoup de mesures sont faites avec pour objectif une utilité sociale : définir les besoins en eau, en énergie, contrôler la qualité, prévoir un résultat, etc. Il suffit alors d'une information relativement grossière, compatible avec les incertitudes sur les mesures.

Par exemple, la mesure des débits des fleuves en Vendée [2] répond à l'objectif de veiller à l'apport en eau pour la population. Cette mesure est grossièrement suffisante pour l'objectif visé (mais nous avons été amenés à reconstruire de nombreuses données manquantes, voir [3]).

De la même façon, la mesure des températures en France sert à EdF pour prévoir la production d'électricité. Mais cette mesure n'a pas à être précise, locale, dans la mesure où EdF utilise seulement une moyenne nationale, et non pas une moyenne différenciée par région.

Les réseaux de mesures qui sont établis fonctionnent en général avec pour objectif l'utilité sociale qui les définit. Ils s'efforcent d'être suffisants pour ce niveau de besoin et, dans la plupart des cas que nous connaissons, ils n'y parviennent qu'avec difficulté (mesures manquantes, erreurs nombreuses, défauts de calibration, etc.).

Pour savoir si les habitants de Vendée ont assez d'eau ou non, la connaissance du débit journalier moyen d'un fleuve est largement suffisante. Mais la connaissance scientifique du phénomène est quelque chose de tout à fait différent, de beaucoup plus précis : débit en chaque point (près des berges, au centre, en surface, au fond), variation avec le temps, etc.

Il en va de même de la température : la connaissance de la température journalière moyenne en un nombre suffisant de stations est suffisante pour EdF. Mais la connaissance scientifique du phénomène requerrait un nombre de capteurs beaucoup plus important (voir l'annexe de notre Note au SGDN [4] à ce propos) : densité plus grande, répartition sur toute la surface de la Terre, dans les courants marins, dans la haute atmosphère, mesures plus fréquentes.

Cette distinction entre utilité sociale et connaissance scientifique est tout à fait essentielle, et nous la résumerons de la manière suivante :

Les réseaux de capteurs installés ont, quel que soit le domaine, les plus grandes difficultés à répondre correctement à l'utilité sociale qui leur est impartie. Il est complètement illusoire de vouloir les utiliser de manière déterminante dans la connaissance scientifique des phénomènes auxquels ils se rattachent. Les exemples ci-dessus montrent que les mesures sont en nombre très insuffisant.

Références

- [1] Charline Carlier : Méthodes probabilistes pour l'Environnement. Rapport adressé par la SCM à l'Agence Européenne pour l'Environnement, 2007.
- [2] Contrats SCM pour Veolia Environnement, Région Ouest, 2005-2007.
- [3] Bernard Beauzamy et Olga Zeydina : Méthodes probabilistes pour la reconstruction de données manquantes. Editions de la SCM, 2007.
- [4] Bernard Beauzamy : Le réchauffement climatique : mystifications et falsifications. Note adressée au Secrétariat Général de la Défense Nationale en 2001 ; réactualisée en 2006.
- [5] Thèse de Vincent Bonnomet à l'INERIS (codirigée par Bernard Beauzamy et Eric Vindimian) : Modélisation mathématique des effets toxiques sur les espèces vivantes. Université de Lyon 1, 2002.