Société de Calcul Mathématique, S. A.
Outils d'aide à la décision
depuis 1995

Méthodes probabilistes pour l'Environnement

Article rédigé par Charline Carlier, Ingénieur de Recherche SCM

Résumé

La rédaction de cet article fait suite à une étude que nous a confiée l'Agence Européenne pour l'Environnement sous le thème « *Improving the results and their attached uncertainty in stratified assesments of drivers – water composition relationships*. ».

Le suivi des données de concentrations de polluants ou de particules dans l'eau, l'air ou le sol est nécessaire afin d'évaluer l'efficacité des politiques environnementales. Ces politiques ont-elles été bien appliquées ? Quels secteurs ont été les plus vertueux ? Etc.

Deux méthodes se rencontrent actuellement : les méthodes statistiques et les méthodes probabilistes.

Les méthodes actuelles d'évaluation utilisent des techniques statistiques simples à mettre en œuvre : calcul de moyenne de concentration sur une région donnée ou un pays donné, mais dont le résultat est pauvre en information et est difficilement exploitable.

Les méthodes probabilistes, mises en place dans le cadre d'un contrat avec l'Agence Européenne de l'Environnement, peuvent sembler plus complexes mais ne sont en fait qu'une mise en forme des données dont on dispose au départ. Le résultat est clair et compréhensible par tous : par exemple, le pourcentage de stations dont la concentration est inférieure à une quantité donnée.

En statistique, si l'on obtient une valeur de concentration élevée pour une région, c'est toute la région qui sera pénalisée et considérée comme polluée. Avec une méthode probabiliste, il possible de déterminer si l'ensemble d'une région est polluée ou si seulement quelques zones le sont, et ainsi de ne pas tirer des conclusions alarmantes pour l'ensemble de cette région.

Les méthodes probabilistes permettent donc de mieux évaluer la qualité environnementale d'une région ou d'un pays. Elles peuvent également permettre de reconstituer des données manquantes dans le temps ou de prévoir des concentrations futures via la méthode de Hypersurface Probabiliste Expérimentale.

Juin 2007

Introduction

L'analyse des données pour l'environnement n'est jamais simple : c'est un domaine récent ; les données sont encore peu nombreuses et généralement éparses. Pourtant les attentes sont considérables, tant auprès du public que des politiques et des industriels.

Des politiques environnementales au niveau national ou européen sont mises en place afin de limiter les pollutions de l'air, du sol et de l'eau. Il faut pouvoir évaluer leur efficacité. Plusieurs questions se posent alors :

- Y a-t-il eu une diminution des polluants concernés ?
- Est-ce que l'évolution est la même pour tous les polluants ?
- Quelles ont été les villes, régions, bassins versants ou pays les plus vertueux ? Et quels sont ceux qui l'ont été le moins ?
- Quels sont les secteurs qui ont le mieux suivi ces politiques (agriculture, industrie…)?

Pour répondre à ces questions, deux méthodes sont disponibles : la méthode statistique et la méthode probabiliste. La première, simple mais critiquable, est celle employée actuellement. La deuxième, plus robuste, est celle que la SCM a mise en place dans le cadre d'un contrat avec l'Agence Européenne pour l'Environnement.

La méthode probabiliste présentée en seconde partie de cet article utilise des données de concentration de polluants dans les eaux de rivières en France. Les mêmes techniques s'appliquent aussi bien à la concentration de polluants dans l'eau de nappes phréatiques, dans l'air ou le sol.

I. La méthode statistique

L'utilisation des statistiques, dans le domaine de l'environnement ou dans tout autre domaine, est devenue courante car elle est simple d'utilisation et est comprise de tous. Il est vrai que cette méthode est applicable à quasiment tous les domaines et à toutes les données mais cela ne signifie par pour autant que les résultats soient exploitables. De plus, l'utilisation des statistiques soulève des problèmes méthodologiques.

L'approche statistique est simple : on souhaite connaître la qualité des eaux dans un bassin versant pour une année donnée. Le plus simple est alors de prendre l'ensemble des données du bassin versant pour l'année en question et de faire la moyenne :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Où :

– n est le nombre de stations pour lesquelles les données sont disponibles, dans la région considéré et pour l'année en question ;

– x_i est la concentration du polluant pour la station *i*.

Le résultat type sera alors : la concentration moyenne en ammonium dans le bassin versant de la Loire en 2005 est de 1.5 mg/l. C'est un résultat que n'importe qui peut comprendre et qui est donc facilement accepté.

Le principal problème des méthodes statistiques est que toute l'information dont on disposait au départ (concentration pour les n stations) a disparu. On dispose au départ de n informations de concentrations et on se retrouve à la fin avec une seule et unique information.

Voyons maintenant plus en détail pourquoi la moyenne n'est pas un bon indicateur.

Nous disposons des données de deux bassins versants X et Y . Chacun d'entre eux contient 10 stations qui ont enregistré des concentrations en ammonium. Les moyennes annuelles de concentration de ces stations sont présentées dans le tableau ci-dessous.

Stations	**Bassin Versant X**	**Bassin Versant Y**
1	0.05	1
2	0.1	1.3
3	0.2	1.4
4	0.2	1
5	0.1	1
6	0.05	1
7	0.1	1
8	0.2	1.1
9	5	1.2
10	5	1

La moyenne pour ces deux bassins versants est la même, 1.10 mg/l et pourtant les états de pollution des deux bassins sont complètement différents. Le bassin versant X est peu pollué, mis à part les deux dernières stations. En revanche, toutes les stations du bassin versant Y enregistrent des concentrations assez élevées, supérieures à 1 mg/l.

La moyenne indique une pollution égale pour les deux bassins. L'étude des données élémentaires indique que le bassin X est moins pollué que le bassin Y.

Bien sûr, on peut calculer l'écart type qui représente la dispersion des valeurs autour de la moyenne. Plus l'écart type est élevé et plus la dispersion des valeurs autour de la moyenne est grande. L'écart type se calcule de la manière suivante :

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}{n}}$$

L'écart type pour le bassin versant X est de 2.06 et celui du bassin versant Y est 0.15. On peut donc conclure que même si les moyennes sont identiques, les valeurs sont plus dispersées dans le bassin versant X que dans le bassin versant Y.

Nous savons donc que les deux bassins versants ne sont pas pollués de la même manière mais nous n'en savons pas plus.

Les statistiques ne peuvent normalement être employées que si les deux conditions suivantes sont satisfaites simultanément :

1) la loi du phénomène est connue ;

2) l'échantillon est de taille suffisante.

Dans les questions qui nous occupent, aucune de ces deux conditions n'est réellement satisfaite : on ne connaît pas la loi de distribution du phénomène, et l'échantillon est en général très maigre.

Il y a une difficulté supplémentaire, due à la non-représentativité des stations. Souvent, comme il est naturel, elles ont été mises dans des zones supposées polluées (pour jouer un rôle de surveillance) et aussi par facilité d'accès. On ne peut donc pas, sans précaution, considérer qu'une moyenne indiquée par les stations est un bon indicateur global pour la région concernée.

L'utilisation des statistiques dans ce cas là n'est donc pas pertinente ; il est nécessaire de mettre en place une méthode plus robuste qui permet de conserver l'ensemble de l'information tout en restant autant compréhensible pour les autorités et le grand public.

II. La méthode probabiliste

Les probabilités sont souvent montrées du doigt, car trop compliquées à comprendre et à appliquer. Nous allons voir qu'il n'en est rien.

Avec une approche probabiliste, les informations apportées par chaque station sont conservées dans le résultat. On peut voir les probabilités comme un arrangement des données sous une forme qui permet une meilleure compréhension.

La méthode se déroule en plusieurs étapes très simples. Les exemples que nous allons présenter concernent les mesures du phosphore total dans le Golfe de Gascogne.

1ère étape : Stratification

Les sources de pollutions sont souvent multiples, que ce soit la pollution du sol, de l'air ou de l'eau. Par exemple pour l'eau, les sources peuvent être : les zones urbanisées, l'agriculture, le bétail,…

Chacune de ces sources va rejeter des polluants différents, en quantités différentes. Il est donc nécessaire de stratifier les données disponibles, afin qu'elles soient le plus représentatives possibles de la source de pollution. On pourra alors comparer les différentes sources de pollution afin de voir, par exemple, quel domaine a fait le plus de progrès et, à l'opposé, appliquer des politiques environnementales plus strictes aux secteurs les plus pollueurs.

Le nombre de strates et leur contenu dépend bien entendu du type de polluant et du type de pollution que l'on souhaite étudier. La condition finale est que chaque strate soit plus homogène que l'ensemble.

Afin de vérifier l'homogénéité de chaque strate nous définissons un indicateur d'homogénéité : l'écart type relatif (ETR). Il est égal à l'écart type divisé par la moyenne

Prenons un exemple afin de mieux comprendre comment est calculé cet indicateur. Le tableau ci-dessous présente les concentrations moyennes en phosphore total de l'année 1985, pour les quatre stations de la strate Agriculture du bassin du Golfe de Gascogne.

N° station	Concentration (mg/l)
05131000	0.0828
05081000	0.5042
05073000	0.7243
05129000	0.0830

La moyenne de ces quatre stations est :

$$\overline{X_{GA,A}} = \frac{C_{05131000} + C_{05081000} + C_{05073000} + C_{05129000}}{4} = 0.3486$$

Et son écart type est égal à :

$$\sigma_{GA,A} = \sqrt{\frac{\left(C_{05131000} - \overline{X_{GA,A}}\right)^2 + \left(C_{05081000} - \overline{X_{GA,A}}\right)^2 + \left(C_{05073000} - \overline{X_{GA,A}}\right)^2 + \left(C_{05129000} - \overline{X_{GA,A}}\right)^2}{4}}$$

$$= 0.320$$

Notre indicateur d'homogénéité est alors égal à :

$$ETR_{GA,A} = \frac{\sigma_{GA,A}}{\overline{X_{GA,A}}} = 91.7\%$$

L'ETR de la strate A du bassin Golfe de Gascogne pour l'année 1985 est élevé, le jeu de données n'est donc pas homogène.

Cet indicateur permet également la comparaison entre plusieurs jeux de données de tailles et d'origines différentes.

Prenons un exemple afin de mieux comprendre comment utiliser cet indicateur. On dispose de deux jeux de données et l'on souhaite déterminer lequel est le plus homogène. On calcule dans un premier temps la moyenne et l'écart type de chacun des jeux de données :

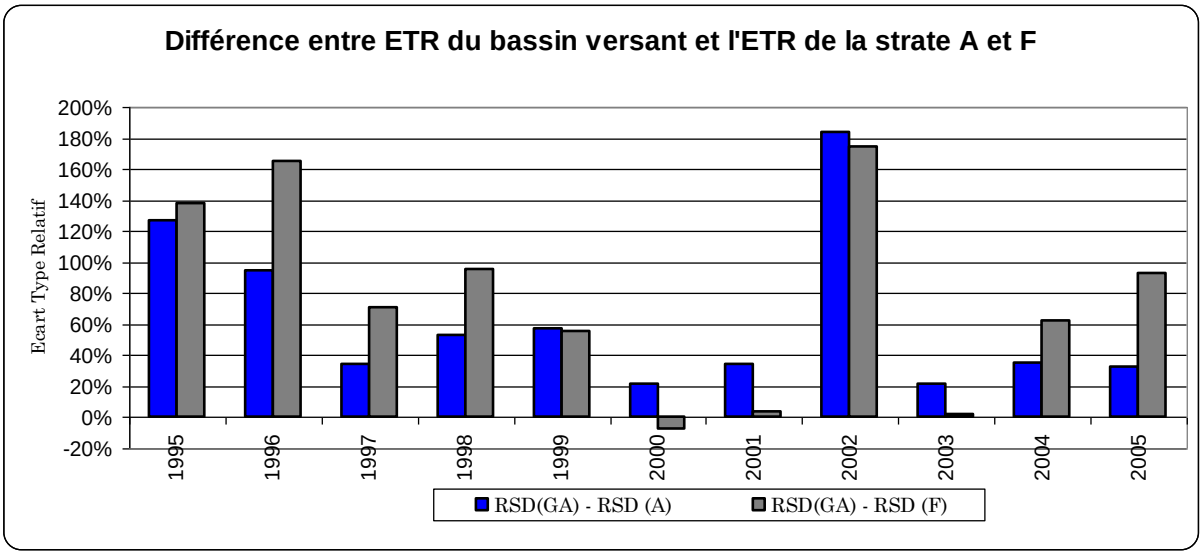
Jeu de donnés n°1: $Moyenne = 0.1$ mg/l ; Ecart type $= 0.2$ mg/l ; $ETR = 200$ %.

Jeu de donnés n°2: $Moyenne = 2$ mg/l ; Ecart type $= 0.2$ mg/l ; $ETR = 10$ %.

Si on regardait uniquement l'écart-type, on pourrait conclure que la dispersion est la même et donc que les deux jeux de données sont aussi homogènes l'un que l'autre. L'étude de l'ETR montre que les deux jeux de données sont complètement différents et que le jeu de données n°2 est beaucoup plus homogène que le premier.

Si les strates sont plus homogènes que l'ensemble, alors on peut conclure que les strates ont été bien construites et que la stratification est utile.

La figure ci-dessous présente la différence entre le bassin du Golfe de Gascogne et deux de ces strates (A : Agriculture et F : Forêt) entre 1995 et 2005. Lorsque la différence entre l'ETR du bassin et celle de la strate est positive, cela signifie que la strate est plus homogène que le bassin. Dans l'exemple présenté ci-dessous, seules les données de la strate F, durant l'année 2000, sont plus hétérogènes que les données du bassin.



Une fois les strates construites, on peut passer à la mise en œuvre de la méthode probabiliste.

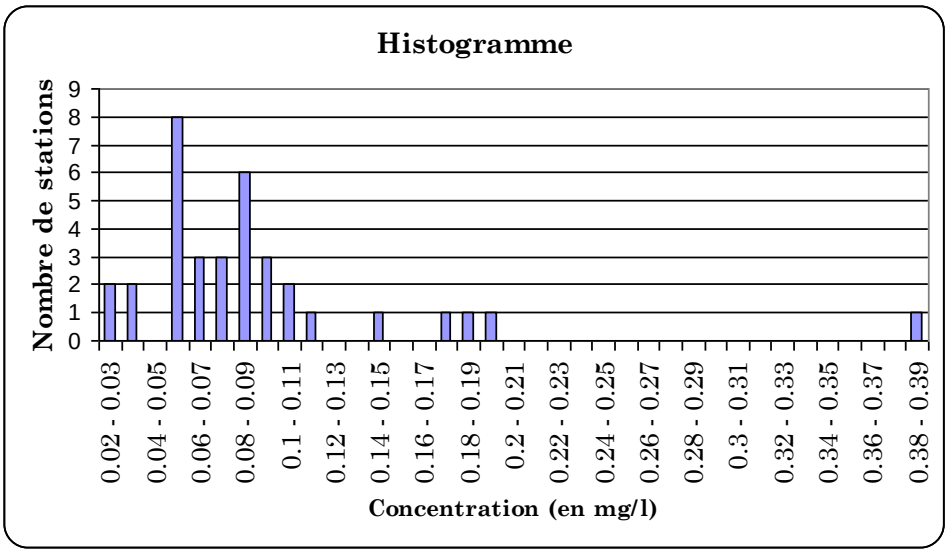
2ème étape : Construction d'un histogramme

Pour un bassin versant (ou une strate) donné et une année donnée, nous disposons de n valeurs de concentrations correspondant aux concentrations enregistrées par les n stations du bassin. Ces stations donnent des valeurs comprises dans un intervalle de concentration, $[0;1]$ par exemple. On divise cet intervalle en petits intervalles de 0.01 mg/l par exemple. La taille des intervalles dépend de la précision que l'on souhaite sur les données mais également de la taille de l'intervalle de concentration.

Pour notre exemple nous prendrons des intervalles de 0.01 mg/l. On obtient donc 100 intervalles de concentration.

Pour construire l'histogramme, on comptabilise le nombre de stations qui tombent dans chacun des intervalles. On va donc compter le nombre de stations dont la concentration est comprise entre $[0;0.01[$, puis le nombre de stations dont la concentration est comprise entre $[0.01;0.02[$ et ainsi de suite, jusqu'au dernier intervalle. On obtient alors une table des occurrences à partir de laquelle on va tracer un histogramme.

La figure ci-dessous présente un histogramme pour les mesures du phosphore total dans la strate Agriculture du Golfe de Gascogne en 2005.



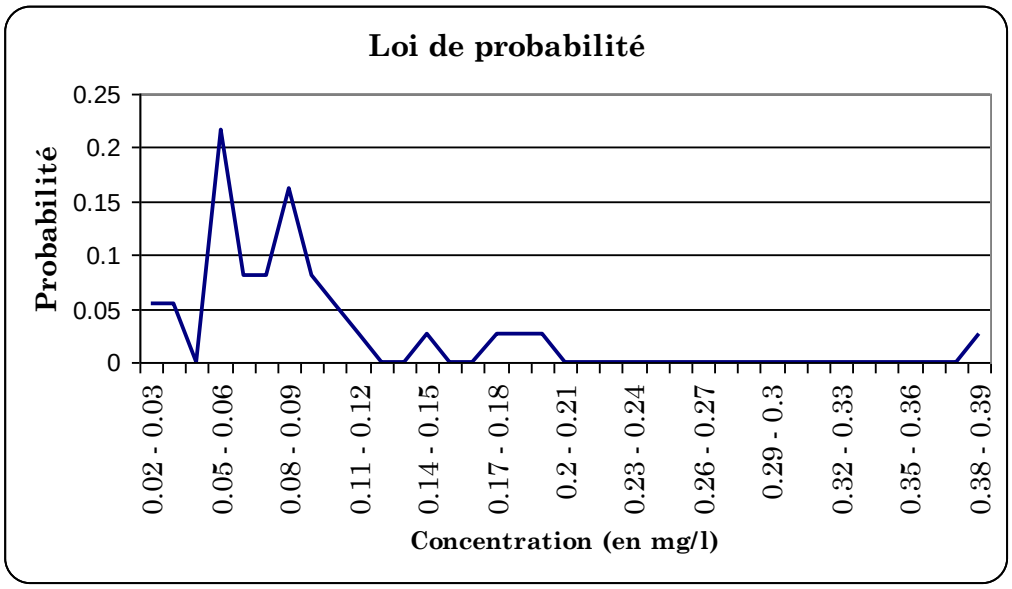
En 2005, 8 stations ont enregistré une concentration de phosphore total comprise dans l'intervalle 0.05 mg/l – 0.06 mg/l

2ème étape : Construction des lois de probabilité

Une fois l'histogramme construit, il est facile d'en déduire une loi de probabilité. Pour cela, il suffit de diviser les valeurs obtenues pour chaque année par la somme.

Pour l'année 2005, dans la strate A du Golfe de Gascogne le nombre de stations est de 37. On divise alors le nombre de stations obtenu dans chaque intervalle par 37 pour ob-

tenir la loi de probabilité pour cette année. Le résultat est présenté dans la figure ci-dessous.



En 2005, la probabilité qu'une station ait enregistré une concentration de phosphore total comprise dans l'intervalle 0.05 mg/l – 0.06 mg/l est de 0.216. On peut également dire cela de la manière suivante : en 2005, 21,6 % des stations de la strate Agriculture du golfe de Gascogne, ont enregistré une concentration de phosphore total comprise dans l'intervalle 0.05 mg/l – 0.06 mg/l.

On obtient alors pour chaque année une loi de probabilité.

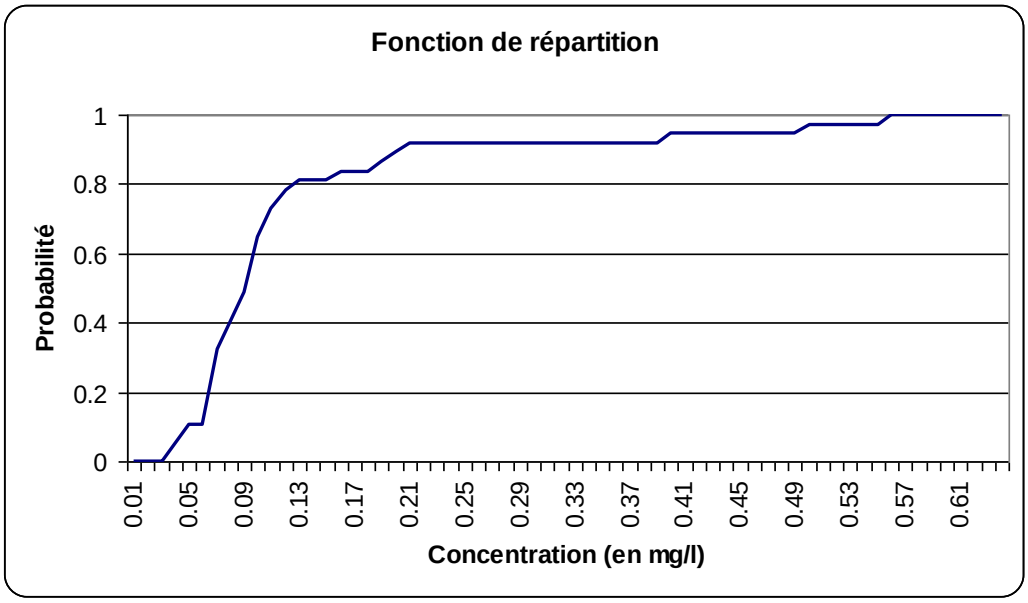
Etape 2bis : Construction de la fonction de répartition

L'inconvénient de l'histogramme ou de la loi de probabilité est qu'une partie de l'information est perdue, puisqu'on représente uniquement des intervalles. On ne sait plus où étaient les valeurs au sein des intervalles, mais seulement qu'on est tombé tant de fois entre telle et telle borne. L'information dont on dispose généralement est déjà très pauvre (souvent des moyennes annuelles) ; on préfèrera donc utiliser la fonction de répartition.

Au lieu de comptabiliser le nombre de stations dans des intervalles, on comptabilise le nombre de stations qui ont enregistré une concentration C inférieure à α :

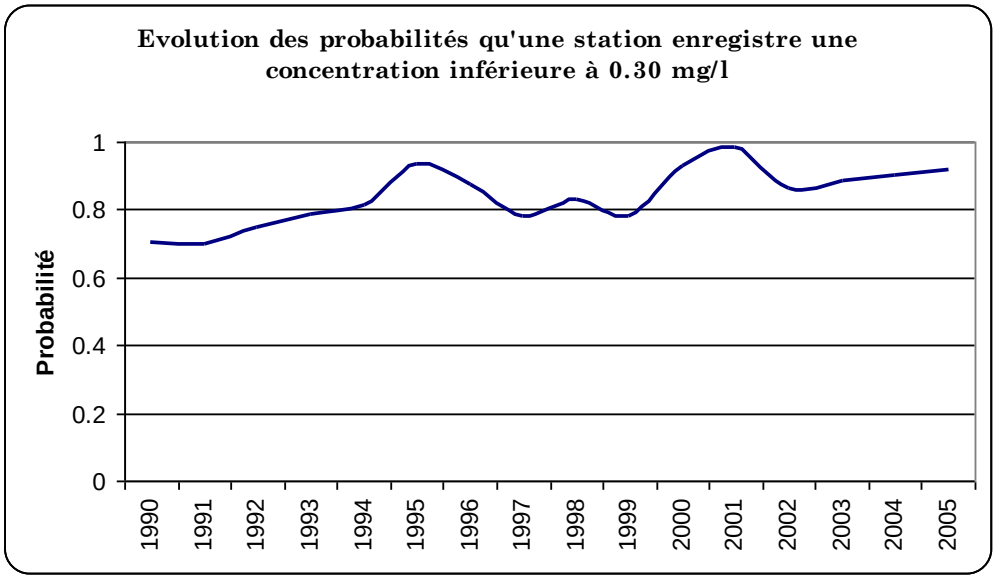
$$F(\alpha) = P\ \ C \leq \alpha\ \ = \int_{-\infty}^{\alpha} f(x)dx$$

On obtient alors le graphique suivant :



La probabilité qu'une station enregistre une concentration inférieure à 0.30 mg/l est de 0.92. On peut dire également cela de la manière suivante : 92 % des stations de la strate A du Golfe de Gascogne ont enregistré une concentration inférieure à 0.30 mg/l en 2005.

On obtient ainsi une fonction de répartition pour chaque année. On peut par la suite regarder l'évolution de la probabilité d'être en dessous d'un seuil de concentration dans le temps. La figure ci-dessous présente l'évolution des probabilités qu'une station enregistre une concentration inférieure 0.30 mg/l entre 1990 et 2005.



La figure ci-dessus montre que la plupart des stations deviennent de plus en plus vertueuses. Dans les années à venir on peut prédire que toutes les stations enregistreront des concentrations inférieures à 0.30 mg/l.

A partir de simples moyennes annuelles de concentration pour des stations on obtient une grande quantité d'informations sur leur comportement passé, présent et futur.

III. Méthode de l'Hypersurface Probabiliste

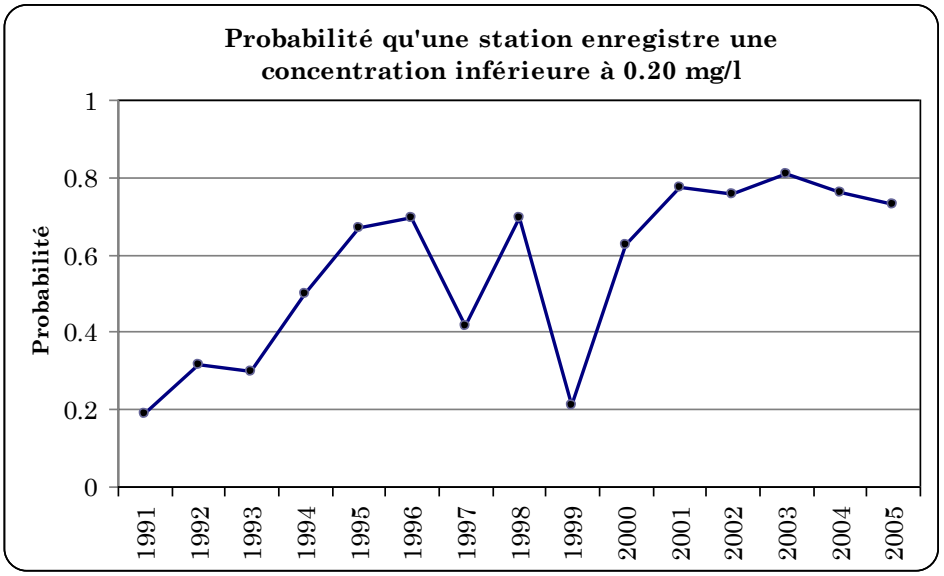
L'une des préoccupations principales, lorsque l'on parle d'environnement, est bien sûr l'état actuel de cet environnement, mais également et surtout son état futur. La méthode la plus employée pour prévoir les valeurs dans un futur plus ou moins proche est la régression linéaire. Elle consiste à calculer une droite qui s'ajuste au plus près des données en utilisant la méthode des moindres carrés. Cette méthode est très simple d'utilisation et très facile à mettre en œuvre avec Excel, ce qui la rend très attractive. Malheureusement simplicité de la méthode ne rime pas toujours avec exactitude des résultats.

Lorsque les données suivent une tendance claire et linéaire, l'utilisation de la droite de régression peut être une bonne solution. Mais lorsque les données sont chaotiques, il est alors impossible de détecter une tendance passée et donc de déduire une tendance pour les années à venir.

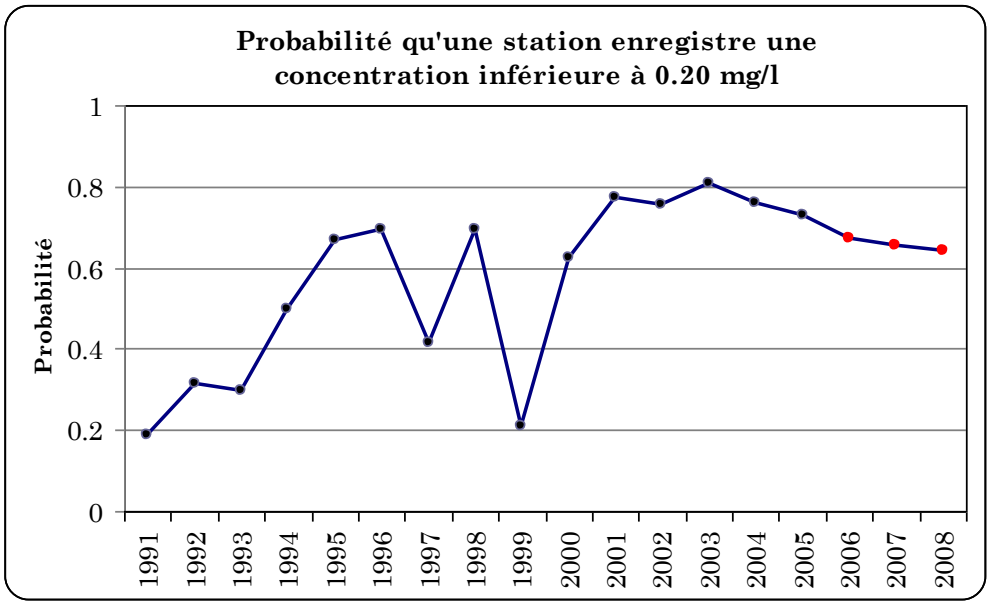
On utilisera donc une méthode que nous avons mise en œuvre il y a quelques années : l'Hypersurface Probabiliste Expérimentale (EPH). Le résultat de l'EPH est, pour chaque point où l'information n'est pas connue, une loi de probabilité qui est une « propagation » de l'information à partir de points existants.

Pour plus de renseignements sur cette méthode voir le rapport n°4 rédigé par Olga Zeydina pour l'Institut de Radioprotection et Sûreté nucléaire disponible sur notre site Internet (http://www.scmsa.com/RMM/IRSN_SCMSA_EPH4.pdf).

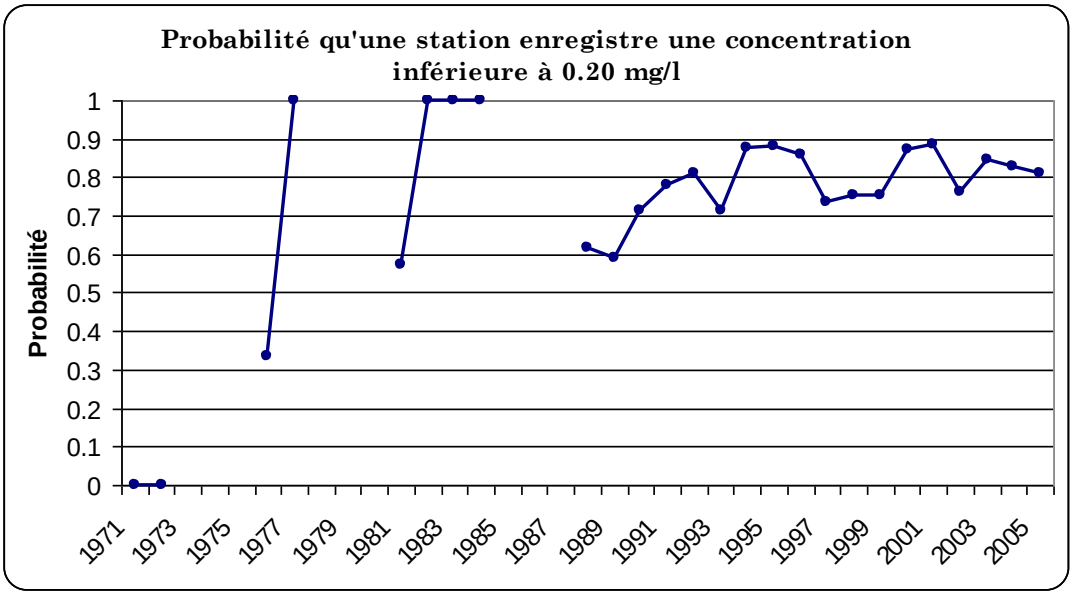
La figure ci-dessous présente l'évolution de la probabilité qu'une station enregistre une concentration inférieure à 0.20 mg/l dans la Strate A du Golfe de Gascogne.



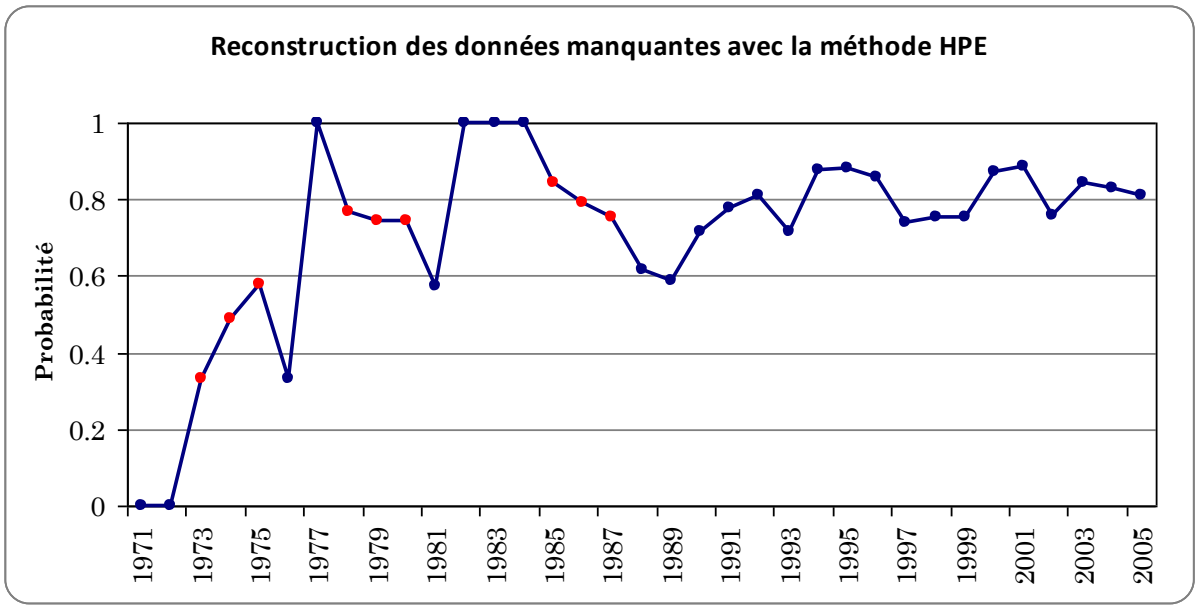
Les données sont chaotiques ; il n'est donc pas correct d'appliquer une droite de régression dans ce cas-là. On préfèrera l'utilisation de l'EPH afin de prédire les concentrations pour les années à venir. Le graphique ci-dessous présente les résultats de la prédiction (en rouge) pour 2006, 2007 et 2008.



Cette méthode est utilisée ici comme une méthode de prédiction, mais elle peut être également utilisée comme une méthode de reconstitution des valeurs manquantes dans le passé. La figure ci-dessous présente un exemple pour un autre bassin, celui de la Bretagne, pour la strate Bétail.



Les valeurs des années 1973 à 1975, 1978 à 1980 et 1985 à 1987 sont manquantes. La méthode EPH permet la reconstruction de ces valeurs à partir des années antérieures et postérieures. Le résultat final est présenté dans la figure ci-dessous.


Reconstruction des données manquantes avec la méthode HPE
Probabilité
1
0.8
0.6
0.4
0.2
0
1971
1973
1975
1977
1979
1981
1983
1985
1987
1989
1991
1993
1995
1997
1999
2001
2003
2005