

Société de Calcul Mathématique SA

*Outils d'aide à la décision*

*depuis 1995*



Cartographie des pollutions :

Outils mathématiques pour la

représentation spatiale et l'évolution temporelle

par :

- Stéphane Belbèze (BRGM) ;
- Rima Abdenbi, Lucas Busson, Cherif Seddik, Bernard Beauzamy (SCMSA)

***Version de travail, 21/12/2024***

En application du Bon de Commande BRGM no 268904, du 14 octobre 2024

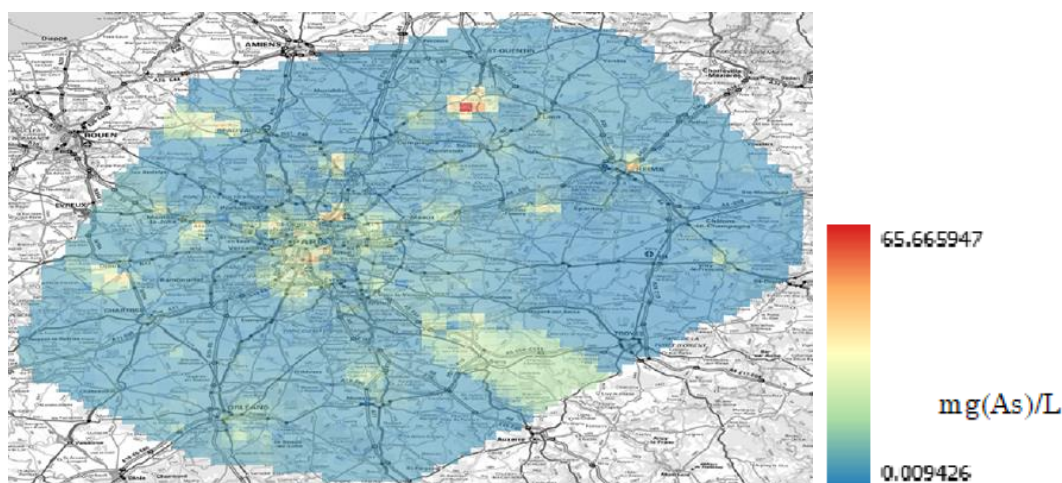
## Résumé Opérationnel

La cartographie des pollutions doit être aussi précise que possible. Cela répond à un besoin de la société civile, pour l'aménagement des territoires ; c'est très légitimement que le Maire d'une commune se demandera "puis-je construire une école à tel endroit, compte-tenu de la présence passée d'une usine ?". Cela répond aussi à une nécessité juridique : si une pollution accidentelle se produit, quelle en est l'ampleur et quelle doit-être l'indemnisation ? Enfin, cela répond à un souhait exprimé par la Gendarmerie Nationale : si on trouve des traces d'arsenic sous les semelles d'un suspect, pouvons-nous savoir quels lieux il a fréquentés ?

Certes, cette cartographie doit être précise, mais elle doit aussi être neutre, ce qui est beaucoup plus difficile ! Il est très facile d'être extrêmement précis si on fait des hypothèses de modèle : là, la propagation sera linéaire, là elle a doublé, etc. On constate malheureusement que, à partir des mêmes données, de multiples institutions réalisent des cartes absolument divergentes, ce qui fait que la société civile n'y comprend plus rien et que le juridique s'enlise dans d'interminables querelles d'experts.

Il y a donc un besoin, clairement exprimé par le BRGM, de réaliser des cartes à la fois précises et neutres, c'est-à-dire dépourvues d'hypothèses factives. C'est ce que réalise la méthode EEPH (Extended Experimental Probabilistic Hypersurface) que nous présentons ici. Elle représente l'adaptation aux questions de pollution des sols de la méthode EPH, introduite en 2004 par la SCM (dans le cadre de contrats avec Framatome), destinée à propager l'information de la manière la plus neutre possible.

Voici la carte obtenue pour l'arsenic en région parisienne ; les couleurs vont du bleu-vert (faible concentration) à l'orange-rouge (forte concentration) :



## I. Introduction

La prise de décision, en ce qui concerne l'aménagement du territoire, requiert une cartographie des risques aussi précise que possible. Parmi ceux-ci, on dénombre les risques naturels (par exemple les inondations) : la cartographie se fait par référence à un historique, aussi long que possible. En ce qui concerne les pollutions, les choses sont beaucoup plus difficiles : en général, on ne dispose que de quelques relevés, pas nécessairement aux bons endroits, et on voudrait en déduire l'étendue spatiale de la pollution ainsi que son évolution dans le temps, au fil des jours et des mois. La question est complètement légitime : le maire d'une commune se demandera s'il peut construire une école à tel endroit, compte-tenu des relevés qu'on lui soumet. La liste des polluants possibles est extrêmement vaste ; le document [Belbèze & al.] mentionne que le BRGM, pour la France, a déjà établi 47 fonds géochimiques pour les sols urbains.

La question est difficile, surtout pour l'étendue spatiale : étant donné un certain nombre de relevés, que peut-on trouver en un point où aucun relevé n'a encore été fait ? La réponse évidente est : n'importe quoi, depuis une pollution nulle jusqu'à une valeur extrême. Pour réaliser une cartographie, il est donc nécessaire de faire des hypothèses de modèle. Malheureusement, la réponse, c'est-à-dire la carte, va dépendre de manière essentielle du type de modèle retenu, et ceci va influencer la décision des Autorités. Il faudra donc :

- Faire aussi peu d'hypothèses que nécessaire ;
- Prendre soin de bien les énumérer, par souci de clarté ;
- Les faire varier, pour voir en quoi la réponse dépend du choix qui a été fait.

En mathématiques, cette dernière préoccupation s'appelle "analyse de sensibilité" et elle est considérée comme essentielle : une présentation qui en est dépourvue doit par principe être rejetée.

Une analyse critique de la littérature disponible sur ces questions a été faite par S. Belbèze [Belbèze] ; elle montre que, dans la plupart des articles, on a recours à des modèles mathématiques adoptés pour la circonstance, sans justification physique, et qu'aucune analyse de sensibilité n'est faite : il y a donc place pour un progrès méthodologique.

Une méthode mathématique permettant de diffuser l'information avec un minimum d'hypothèses a été conçue par la SCM, initialement dans le cadre de contrats avec Framatome (2004). Elle est présentée dans le livre [PIT], 2013, et elle a depuis été largement utilisée, y compris par le BRGM. Elle s'appelle "EPH", pour "Experimental Probabilistic Hypersurface".

Elle présente cependant une insuffisance majeure : elle ne rend pas compte des spécificités du terrain, évidemment essentielles si on s'intéresse à l'état des sols. Des propriétés comme l'inhomogénéité, l'anisotropie, doivent pouvoir être prises en compte.

C'est à quoi cherche à remédier la nouvelle version, appelée EEPH, ou Extended EPH, présentée ici. Elle est capable de prendre en compte les anisotropies, les discontinuités de terrain, etc. Elle se comporte mieux que ne le faisait l'EPH en présence de "clusters" (points d'accumulation : beaucoup de mesures proches les unes des autres).

Parmi les applications déjà réalisées pour l'EPH, en ce qui concerne les pollutions, nous mentionnons l'exemple traité plus loin : reconstitution de la pollution en hydrocarbures dans un port, suite au naufrage d'un pétrolier.

Parmi les applications déjà réalisées pour l'EEPH, ou bien anticipées, citons :

- des cartes relatives à la teneur en divers polluants, par exemple une carte de France de la teneur en arsenic (de telles cartes existent déjà, mais on cherche à améliorer la précision) ;
- l'amélioration de l'analyse d'indices, qui peut intéresser la Gendarmerie : on trouve tel échantillon de sol sous les semelles d'un suspect ; d'où cet échantillon peut-il provenir ? Evidemment, plus la réponse est précise et plus l'aide à la décision est significative.

## II. Présentation du besoin

On dispose généralement, dans une zone donnée, d'un certain nombre de mesures relatives à divers polluants. Ces mesures ont été prises à des dates variables, sont de qualité inégale, et la répartition n'est pas régulière. A partir de cette information, on souhaite :

1. Reconstituer une carte de la pollution, pour chaque polluant, pour l'ensemble de la zone ;
2. Définir un éventuel "plan d'inspection" de la zone, si l'on estime que les mesures sont en nombre insuffisant : où faudra-t-il faire les mesures destinées à compléter l'existant ?
3. Corréler éventuellement la carte de pollution avec des éléments externes : la présence d'une route, d'une usine, etc. ;
4. Etudier l'évolution de la pollution présente, par exemple à échéance d'un an, cinq ans, etc. ;
5. Rechercher le passé de la pollution, par exemple un an avant la mesure, cinq ans avant, etc., ceci dans le but de savoir d'où provient la pollution et comment elle a évolué.

Mathématiquement parlant, on se trouve donc en présence de problèmes d'extrapolation, soit dans l'espace (reconstituer toute une zone), soit dans le temps (vers l'avenir comme vers le passé).

Les méthodes à employer seront nécessairement probabilistes, pour deux raisons :

- Les données de départ sont toujours empreintes d'incertitudes ;
- On ne connaît pas réellement les lois de propagation ; elles dépendent de la nature du terrain, de l'humidité, de la saison, etc. Il serait donc absurde de donner en sortie un résultat

précis, du type : à tel endroit, la concentration en tel polluant est de tant et sera de tant dans deux ans.

Pour cette même raison, les modèles à employer devront être aussi "neutres" que possible et éviter toute hypothèse factice. Dire par exemple "la décroissance est exponentielle", c'est faire une hypothèse de modèle que rien ne justifie. Une telle démarche, donnant un résultat précis là où il n'a pas lieu d'être, est susceptible d'influencer le résultat, conduisant à des décisions erronées.

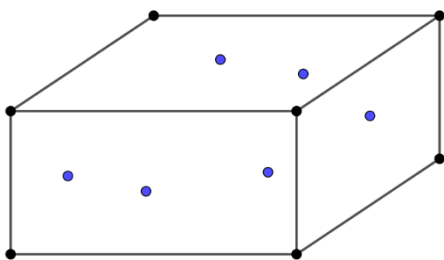
Notre approche méthodologique fournira donc nécessairement, en sortie, une loi de probabilité, du type : compte-tenu des données disponibles, à tel endroit, pour tel polluant, à telle époque (passée, présente ou future), voici la loi de probabilité relative à la concentration propre à ce polluant. On peut retenir l'espérance de la loi pour valeur moyenne, les quantiles 5% et 95% pour intervalle de confiance, etc. Il est en outre intéressant de savoir si la loi est très concentrée (ce qui signifie qu'on est à peu près certain du résultat) ou bien très diffuse (ce qui signifie que notre connaissance est très pauvre).

Si la loi est trop diffuse, cela signifie qu'il faut faire davantage de mesures pour préciser la connaissance : on a ainsi un moyen objectif de juger de la qualité et de la pertinence des données recueillies.

Pour nous, les lois de probabilité seront toujours discrètes, c'est-à-dire consistant en une collection de  $N$  nombres  $p_1, \dots, p_N$ , positifs de somme 1. Les lois continues sont des idéalizations, et ne correspondent pas aux situations réelles.

### III. Représentation spatiale

La question se présente sous une forme simple à définir : des relevés ont été faits en des points  $M_k$  (il s'agit de relevés 3d : la profondeur est connue). Ils ont montré des concentrations  $c_k$  en chaque relevé ; peu importe de quel polluant il s'agit ; la concentration est mesurée en unité appropriée, par exemple en mg de polluant par m<sup>3</sup> de sol. La question est de reconstituer la pollution possible dans un terrain avoisinant, par exemple dans un rayon d'un km autour de la zone de mesure. La reconstitution doit se faire en 3d : on s'intéresse à la pollution possible en profondeur.



La zone à considérer a généralement la forme d'un parallélépipède ; les points de mesure peuvent être absolument quelconques : ils ne sont pas nécessairement espacés de manière régulière. En outre, selon la profondeur, les relevés sont plus ou moins nombreux.

L'outil de base que nous utiliserons pour ce travail est l'hypersurface probabiliste (EPH : Experimental Probabilistic Hypersurface), introduite par la SCM en 2004 dans le cadre de contrats avec Framatome et déjà utilisée par le BRGM à plusieurs reprises, ainsi que par d'autres organismes. L'EPH possède fondamentalement une qualité et un défaut :

## 1. Qualité

Elle est d'information minimale : un Martien débarque, on lui remet une colonne de chiffres : les  $c_k$  relevés aux points  $M_k$  ; le Martien ne sait pas qu'il s'agit de pollutions ni de sols. L'EPH va lui permettre de déterminer la pollution attendue  $c$  en un point quelconque  $M$  ; mieux même, l'EPH va fournir une loi de probabilité de la pollution attendue en ce point (et pas seulement une valeur précise).

## 2. Défaut

Par définition, l'EPH ne sait rien sur rien ; elle ne sait pas qu'il s'agit de sols, et il va falloir la paramétrer en fonction des caractéristiques du sol. Pire, si le sol n'est pas homogène et présente des discontinuités, il va falloir trouver un moyen d'en informer l'EPH, en lui disant que la propagation ne sera pas la même dans toutes les directions et ne suivra pas toujours la même loi dans une direction donnée, en fonction des obstacles ou discontinuités qui seront rencontrés.

Nous commençons par une présentation rapide de l'EPH ; les détails peuvent être trouvés dans le livre [PIT].

### A. Présentation de l'EPH

Le concept de base, pour l'EPH, est la notion d'entropie :

L'entropie d'une loi de probabilité  $p_1, \dots, p_N$  est :

$$H = - \sum_{n=1}^N p_n \log(p_n)$$

où  $\log$  désigne le logarithme népérien. Il est incorrect de parler d'entropie d'une variable aléatoire (comme on le voit souvent), puisque les valeurs de la variable n'interviennent pas : seules les probabilités interviennent.

Il est évident que l'entropie est positive. Elle est nulle si et seulement si un seul des  $p_n$  vaut 1, tous les autres étant nuls. Elle est maximale si tous les  $p_n = \frac{1}{N}$  ; auquel cas elle vaut  $H = \log(N)$ .

On a donc toujours  $0 \leq H \leq \log(N)$  ; l'entropie est d'autant plus basse que la loi est plus concentrée, d'autant plus haute qu'elle est plus diffuse.

L'entropie est une notion très ancienne ; elle apparaît principalement dans les trois domaines suivants :

- Entropie thermodynamique (Clausius, 1864) ;

- Théorie de l'information (Shannon, 1948) ;
- Entropie topologique et entropie métrique de Kolmogorov-Sinaï (1950, dans le cadre de la théorie des systèmes dynamiques en mathématiques). Voir : [https://fr.wikipedia.org/wiki/Entropie\\_métrique](https://fr.wikipedia.org/wiki/Entropie_métrique)

Si un émetteur est susceptible d'émettre  $N$  lettres, mais qu'il émet toujours la même, on considère que la quantité d'information est maximale ; l'entropie est nulle. Si par contre il émet toutes les lettres avec même probabilité, la quantité d'information est minimale et l'entropie vaut  $\log(N)$ .

Une bonne présentation est faite dans l'article "Energy and Information", Myron Tribus and Edward C. McIrvine, Scientific American , Vol. 225, No. 3 (September 1971), pp. 179-190 :

*"La connaissance sur une question particulière peut être représentée par l'affectation d'une certaine probabilité (notée  $p$ ) aux différentes réponses envisageables à la question. Une connaissance complète d'une question est la capacité d'attribuer une probabilité nulle ( $p = 0$ ) à toutes les réponses imaginables, sauf une. Une personne qui attribue (correctement) la probabilité unitaire ( $p = 1$ ) à une réponse n'a évidemment plus rien à apprendre sur cette question. En observant que la connaissance peut être ainsi codée dans une distribution de probabilité (un ensemble de probabilités affecté à l'ensemble des possibilités), nous pouvons définir l'information comme ce qui provoque un ajustement dans une affectation de probabilités."*

Nous avons besoin de définir convenablement la notion d'information associée à une loi de probabilité : ce besoin est fondamental dans le cadre d'une aide à la décision. La notion d'entropie, comme mesure de l'information, répond bien à ce besoin, pour deux raisons fondamentales :

- La définition est simple ;
- Elle est utilisée depuis plus de 150 ans par diverses communautés scientifiques.

Notre approche requiert de ne jamais ajouter une information factice, c'est-à-dire qui soit issue du modèle et non présente dans les données. Mais ce n'est pas si simple ; commençons par le cas d'une seule mesure. Supposons qu'une mesure réalisée en un unique point  $A$  ait donné comme résultat  $C = c$  pour la concentration en un produit quelconque (peu importe de quoi il s'agit). Deux écoles philosophiques vont s'affronter ; l'une est pessimiste et l'autre optimiste :

- On a mesuré  $C = c$  en un point, donc il y a  $C = c$  partout ;
- Certes, on a mesuré  $C = c$  en un point, mais ailleurs on n'en sait rien, et nous allons donc continuer à penser que  $C = 0$  ailleurs.

Si on a réalisé deux mesures,  $C = c_1$  en  $A_1$  et  $C = c_2$  en  $A_2$ , alors les divergences philosophiques deviennent effrayantes et irréconciliables. Si par exemple  $c_2 = 2c_1$  et que  $A_1, A_2$  sont distants d'un km, certains vont conclure que la pollution double tous les km, et ils tireront immédiatement la sonnette d'alarme.

Même dans le cas où le nombre de mesures est faible, il est nécessaire de disposer d'un outil conceptuel d'information minimale, c'est-à-dire qui se contente d'exploiter les données sans nécessairement tomber dans l'alarmisme.

Pour nous, l'information minimale va être guidée par l'entropie. Nous supposons qu'une mesure a été faite en un point  $A$  ; lorsque nous nous éloignons de ce point, l'information devient de moins en moins précise et l'entropie augmente. La relation qualitative est donnée par :

**Lemme d'information minimale** – Soient  $Z_1, \dots, Z_{K-1}$  des variables aléatoires satisfaisant, pour un  $z$  fixé :

$$0 \leq Z_1 \leq \dots \leq Z_{K-1} \leq z.$$

Nous supposons que chaque variable  $Z_k$  suit une loi uniforme sur son intervalle de définition. Alors l'information minimale est obtenue en attribuant la valeur :

$$z_k = \frac{k z}{K}, \quad k = 1, \dots, K-1$$

à chacune d'elles. Cette valeur est l'espérance de chaque variable, sous les contraintes données plus haut.

La démonstration de ce lemme peut être trouvée dans [PIT].

Le principe général est le suivant : on dispose d'une information  $C_{\min}, C_{\max}$  qui nous dit que la concentration attendue devra nécessairement être comprise entre ces deux bornes. Ce n'est pas une information "à dire d'expert", mais plutôt la compilation de toutes les informations recueillies un peu partout à la surface du globe. Pour nous, l'information  $C_{\min} \leq C \leq C_{\max}$  correspond à l'information minimale, donc à l'entropie maximale.

## B. Outils conceptuels pour la construction de l'EPH

Donnons les grandes lignes ; les détails sont donnés dans [PIT].

### 1. Notations

On travaille nécessairement sur un domaine d'étendue bornée, sur lequel des mesures ont été réalisées. On note  $d_{\max}$  la distance maximale entre les points de mesure et les bornes du domaine. Ceci est important, parce que l'on s'astreint à dire que, aux bornes du domaine, on ne sait rien.

On dispose d'une mesure, représentée par une variable aléatoire  $X$  ; nous dirons qu'il s'agit de la concentration en un certain polluant et les valeurs prises seront notées  $c$ . Nous fixons par référence à un historique des valeurs maximales et minimales pour  $c$ , notées  $c_{\min}$  et  $c_{\max}$  ; habituellement  $c_{\min} = 0$  (aucune pollution). Pour des raisons techniques, on discrétise l'intervalle

$[c_{\min}, c_{\max}]$  en sous-intervalles ; habituellement, cela résulte de la précision de la mesure. Par exemple, si l'intervalle total est  $[0,1]$  grammes, on le subdivisera en 1000 intervalles de taille 1 mg. On note  $\nu$  le nombre de sous-intervalles et  $w$  (width) la taille de chaque sous-intervalle. On note  $b_j$  les bornes des sous-intervalles ; elles sont au nombre de  $\nu + 1$ .

Nous avons besoin de définir un "coefficient de propagation", qui caractérise la manière dont l'information se propage. Par définition, il vaut :

$$\lambda = \frac{\text{Log}(\nu + 1)}{d_{\max}}.$$

Ce choix résulte d'un lemme d'information minimale (voir [PIT]) : au point le plus éloigné, l'information doit être minimale, et vaudra donc  $\text{Log}(\nu + 1)$ .

## 2. Propagation de l'information

Nous sommes en un point quelconque  $M$  du domaine d'investigation et nous voulons estimer l'information propagée par une mesure faite en un point  $M_1$  ; cette mesure a donné la valeur  $c_1$ .

Nous introduisons l'écart-type  $\sigma_1 = \frac{w}{\sqrt{2\pi}} \exp(\lambda d(M, M_1))$  ; l'information générée par la mesure en  $M_1$ , évaluée au point  $M$ , est donnée par la loi de probabilité :

$$p_{M_1 \rightarrow M}(j) = u_1 \exp\left(-\frac{(b_j - c_1)^2}{2\sigma_1^2}\right),$$

où  $c_1$  est la mesure, les  $b_j$  sont les points de discrétisation introduits plus haut et  $u_1$  est une constante de normalisation, pour que  $\sum_j p_j = 1$ .

La forme gaussienne de cette loi de probabilité ne résulte nullement d'un choix empirique. Bien au contraire, on démontre (Sobolev) que, pour une variance fixée, la loi de probabilité d'entropie maximale (donc d'information minimale) est une gaussienne.

L'information propagée en tout point a donc la forme d'une courbe en cloche, dont la valeur maximale est prise en  $c_1$  ; la courbe devient de plus en plus évasée à mesure que l'on s'éloigne du point de mesure, puisque la variance augmente.

## 3. Cas de plusieurs mesures

Supposons que des mesures aient été faites en des points  $M_1, \dots, M_N$  ; elles ont donné les concentrations  $c_1, \dots, c_N$ . Quelle concentration peut-on attendre en un point  $M$  quelconque ?

On note  $d_n = d(M, M_n)$ ,  $n = 1, \dots, N$ .

- Cas de la dimension 1

On pose  $\gamma_n = \frac{d_n^{-1}}{\sum_{k=1}^N d_k^{-1}}$ ; c'est l'inverse de la distance entre  $M$  et  $M_n$ , normalisé :  $\sum_{n=1}^N \gamma_n = 1$ .

La loi de probabilité en  $M$ , reçue du fait des mesures en  $M_1, \dots, M_N$ , sera :

$$P_{(M_1, \dots, M_N) \rightarrow M}(j) = \sum_{n=1}^N \gamma_n p_n(j), \text{ où l'on note simplement } p_n = p_{M_n \rightarrow M}.$$

- Cas des dimensions 2 et 3

Habituellement, les mesures de pollution se font en dimension 2 (problème plan) ou en dimension 3 (problème spatial). La construction précédente doit être modifiée comme suit, en notant  $K$  la dimension de l'espace :

$$\gamma_n = \frac{d_n^{-K}}{\sum_{k=1}^N d_k^{-K}}$$

Ces choix sur la forme du coefficient ne sont pas arbitraires, mais résultent de contraintes imposées par la forme du problème ; voir [PIT].

L'EPH se prête bien aux situations rencontrées par le BRGM : typiquement des situations non-homogènes. Le paramètre de propagation  $\lambda$ , au lieu d'être fixé une fois pour toutes, va dépendre du milieu : il sera différent d'une portion à l'autre du domaine considéré.

Nous donnons plus loin un exemple d'application : il s'agit de reconstituer la présence possible d'hydrocarbures dans un port, à partir d'un certain nombre de relevés. Ces relevés sont 3d (à des profondeurs diverses), mais nous décomposons le port en trois couches : 0 – 20 cm de profondeur, 20 à 30 cm, au-delà de 30 cm, et employons l'EPH séparément sur chacune des trois couches. Nous avons donc trois constructions d'EPH, chacune en 2d. Ceci est dû au fait qu'il a été estimé, à l'époque, que les couches étaient en quelque sorte isolées les unes des autres : la pollution pouvait se propager à l'intérieur d'une même couche, mais non d'une couche à l'autre ; nous ne sommes pas responsables de ce choix, fait par des spécialistes sur des bases physiques. En outre, le besoin était une estimation immédiate de la pollution et non son évolution future.

Pour remédier aux insuffisances de l'EPH, une version améliorée a été introduite par S. Belbèze [Belbèze & al] ; elle s'appelle EEPH : Enhanced EPH. Nous la présentons maintenant, en insistant sur les différences avec l'EPH. Il s'agit essentiellement de prendre en compte les possibles anisotropies, fréquentes dans l'analyse des sols.

### C. Présentation de l'EEPH

L'EPH, comme vu précédemment, repose sur le concept d'entropie. Nous allons voir comment modifier ce concept.

#### 1. La correction de l'entropie en prenant en compte la portée des phénomènes

La portée d'un phénomène (exemple : pollution des sols) désigne la distance limite, notée  $RoI$  pour rayon d'influence, au-delà de laquelle des mesures reflètent uniquement des variations locales. En d'autres termes, des paires de points expérimentaux séparées par une distance  $h \geq RoI$  ont des mesures qui sont de moins en moins similaires.

En géostatistique, la fonction utilisée pour déterminer la similarité entre des points de mesures s'appelle « variogramme expérimental » notée  $\Gamma(h)$ , elle est définie par :

$$\Gamma(h) = \frac{\text{Espérance}(|Z(x_i) - Z(x_i + h)|^2)}{2} \approx \frac{1}{2N(h)} \sum_{i=1}^{N(h)} |Z(x_i) - Z(x_i + h)|^2$$

Où  $Z(x_i)$  et  $Z(x_i + h)$  sont des paires de mesures de coordonnées respectives  $x_i$  et  $x_i + h$ ,  $h$  désigne la distance qui les sépare et  $N(h)$  le nombre de paires de mesures séparées par  $h$ .

Pour calculer le rayon d'influence  $RoI$ , nous traçons  $\Gamma(h)$  en fonction de  $h$  (c'est une fonction croissante) et déterminons  $h$  tel que la fonction atteint un palier, c'est-à-dire atteint un maximum et devient constante.

Pour intégrer cette information dans l'EPH, nous modifions le paramètre  $d_{max}$  qui intervient dans le calcul de l'entropie en tout point de mesure  $K$  :

$$\lambda = \frac{\text{Log}(v + 1)}{d_{max}}$$

La propagation ne se fait plus jusqu'aux limites  $[x_{min}, x_{max}]$  du domaine, mais est limitée aux bornes  $[x_i - RoI, x_i + RoI]$ , en fonction du rayon d'influence, comme montré par la figure ci-dessous :

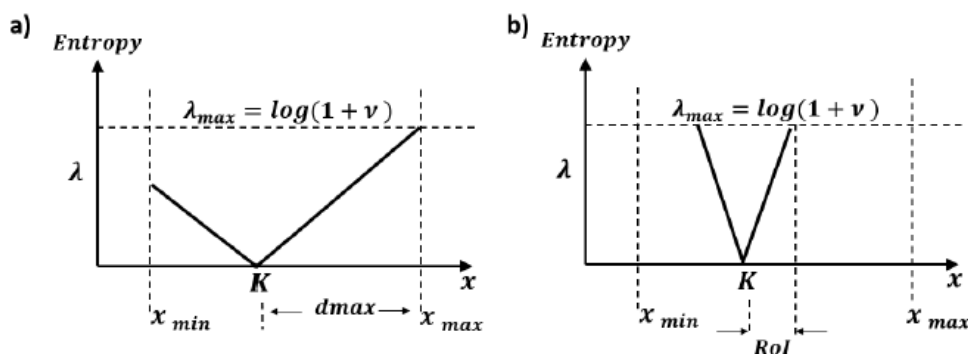


Fig. Prise en compte de la portée des phénomènes : modification de l'entropie  
Source : ISLANDR\_1.3

Ainsi, le paramètre  $\lambda$  lié à l'entropie de la mesure au point  $K$  est modifié de la manière suivante :

$$\lambda = \frac{\text{Log}(v + 1)}{RoI}$$

## 2. La prise en compte de l'anisotropie géométrique

L'anisotropie géométrique désigne une propriété physique où la portée d'un phénomène (expliquée au point précédent) dépend de la direction d'étude. En d'autres termes, les variations des mesures ne sont pas uniformes dans toutes les directions, car la portée est plus ou moins grande selon la direction. Par exemple, la dispersion de polluants dans l'eau est privilégiée dans la direction du courant.

S'il est établi à l'aide d'un variogramme expérimental, comme vu au point précédent, que la portée d'un phénomène varie en fonction de la direction, cette connaissance est intégrée dans le calcul de l'EEPH en la transformant en isotropie : variation uniforme dans toutes les directions.

### a. Notations :

Soit un point de mesure de coordonnées  $(x_1, y_1)$  dans un repère d'étude d'axes  $X$  et  $Y$ .

Les portées maximales  $a_{max}$  et minimales  $a_{min}$  (exprimées en mètres) s'observent selon deux directions orthogonales : respectivement l'axe principal de l'anisotropie et l'axe secondaire de l'anisotropie.

### b. Rotation des axes :

On commence par aligner l'axe principal de l'anisotropie sur l'axe d'étude  $Y$  et l'axe secondaire de l'anisotropie sur l'axe d'étude  $X$ . Cette rotation est définie par la matrice de rotation  $R$  suivante, où  $\alpha$  est l'angle entre l'axe d'étude  $Y$  et l'axe principal de l'anisotropie :

$$R = \begin{bmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{bmatrix}$$

### c. Mise à l'échelle des distances :

Une fois les axes alignés, on réduit les distances dans la direction principale d'un facteur  $\frac{1}{a_{max}}$ , et les distances dans la direction secondaire d'un facteur  $\frac{1}{a_{min}}$ .

Ainsi, les nouvelles coordonnées  $(x'_1, y'_1)$  de  $(x_1, y_1)$  s'obtiennent par la transformation :

$$\begin{bmatrix} x'_1 \\ y'_1 \end{bmatrix} = \begin{bmatrix} 1/a_{max} & 0 \\ 0 & 1/a_{min} \end{bmatrix} R \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$

### 3. Prise en compte des incertitudes

Soient  $A_1, A_2, \dots, A_n$   $n$  points de mesure et  $X$  le point cible où nous souhaitons reconstruire l'information. Nous rappelons que l'EPH de base s'obtient par la formule :

$$P_{x_j}(X) = \gamma_1 P_{1,j}(X) + \dots + \gamma_n P_{n,j}(X) = \sum_{i=1}^n \gamma_i P_{A_i,j}$$

Il s'agit de la somme des contributions des  $n$  points  $A_i$  au point  $X$ , pondérées en fonction de la distance entre le point cible et chaque point de mesure.

Les mesures, ainsi que les paramètres des mesures (tels que la position), sont entachés d'incertitudes. Sans correction, ces valeurs peuvent biaiser les estimations de densité de probabilité obtenues par l'EPH.

#### a. Mesures imprécises :

Nous distinguons deux types de mesures dont le calcul de l'incertitude peut simplement être intégré dans l'EPH :

- Des mesures dites « censurées » : valeurs inférieures à la limite de détection fixée (LQ).
- Des mesures dont l'incertitude est exprimée comme un pourcentage relatif. Par exemple une teneur en polluant de  $50g \pm 10\%$ .

Pour intégrer dans l'EPH la prise en compte des incertitudes sur les mesures, nous pondérons le calcul par un poids  $q_k$  associé aux  $m$  mesures imprécises d'indice  $k$  :

Nous tirons  $m_k$  valeurs suivant une loi uniforme sur :

- $[0, LQ]$  s'il s'agit d'une mesure censurée ;
- $[-E, E]$  s'il s'agit d'une mesure avec une incertitude exprimée comme un pourcentage relatif de  $\pm E$ , où  $E$  est le taux d'incertitude sur la mesure, exprimé en unité de mesure.

puis, calculons l'EPH en toute valeur pour construire une probabilité d'occurrence  $q_k$ .

Nous tirons selon une loi uniforme pour ne pas donner de forme particulière à notre incertitude. L'algorithme peut néanmoins être paramétré sur n'importe quelle loi de probabilité « à dire d'expert » ; il n'y a pas de raison particulière à préférer une loi plutôt qu'une autre.

#### b. Paramètres imprécis :

L'EPH de base sait parfaitement incorporer le cas d'incertitudes sur les positions des mesures (paramètres imprécis) en définissant une déviation aléatoire de la position observée.

Pour intégrer dans l'EPH la prise en compte des incertitudes sur les paramètres, nous pondérons le calcul par un poids  $r_{k'}$  associé aux  $l$  paramètres imprécis d'indice  $k'$ .

c. Formule de l'EPH prenant en compte les incertitudes :

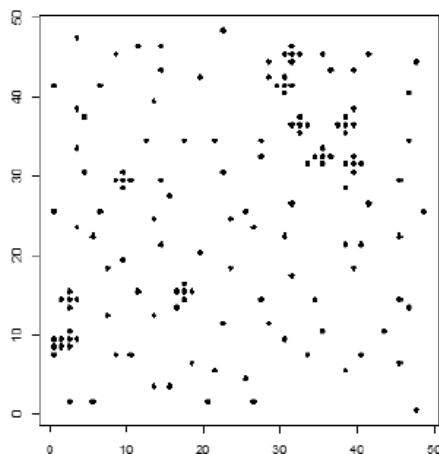
On aura donc, si  $q_k$  est la probabilité de la mesure imprécise  $m_k$ , et si  $r_k$  est la probabilité du paramètre imprécis  $l_{k'}$ , une pondération pour le calcul de l'EPH en un point  $X$  donnée par :

$$P_{x_j}(X) = \sum_{i=1}^n \gamma_i \sum_{k=1}^m q_k \sum_{k'=1}^l r_{k'} P_{A_i, j, k, k'}$$

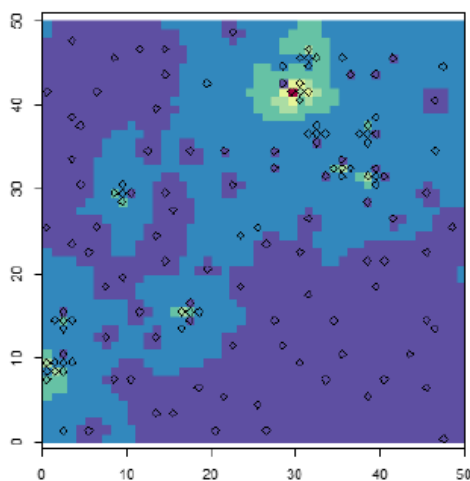
4. La correction des cartes issues de l'EPH

Nous illustrons cette correction par un exemple [ISLANDR] :

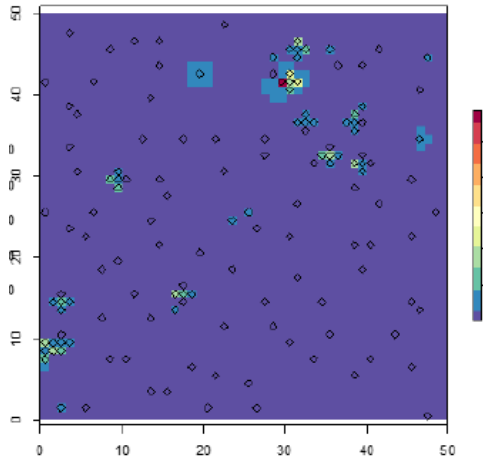
La figure suivante représente une carte de 140 mesures de pollution des sols (points noirs) dans un système de coordonnées à 2 dimensions [0 à 50 km] :



L'EPH neutre permet de générer la carte de propagation de mesures suivante ; la légende de couleur représente l'intensité de la pollution sur une échelle de faible (violet) à forte (rouge) :



Pour un géo-statisticien, cette propagation est trop groupée autour des points de mesure, on parle alors de « clusters ». Pour remédier à cela, un algorithme de dé-clustering (dégroupage) du nom de « Hclust 3.6.2 de la librairie R » a été utilisé pour corriger la carte :



Le principe de l'algorithme est de rompre la continuité entre les clusters, comme le montre la figure ci-dessus.

## IV. Exemple de mise en œuvre de la méthode

Cet exemple utilise l'EPH d'origine.

### A. Cartographie probabiliste de la pollution dans un port

Nous décomposons le port en trois couches : 0 – 20 cm de profondeur, 20 à 30 cm, au-delà de 30 cm, parce que la surface des trois couches n'est pas la même, parce que les mesures faites ne sont pas les mêmes, et parce que le comportement vis-à-vis de la pollution n'est pas le même (sédimentation). Cette précaution méthodologique nous paraît essentielle.

Pour la première couche (0-20 cm), l'EPH conduit à une estimation moyenne (espérance de la loi) qui est de 3,4 tonnes de fioul, et pour la seconde couche, de 19,4 tonnes.

Ces valeurs pour la seconde couche sont élevées : ceci traduit simplement le fait que les mesures sont en nombre très faible. En un point situé à grande distance des points de mesure, par exemple sur le bord nord ou nord ouest du port, on ne sait à peu près rien, puisque tous les points de mesure sont au sud.

Dans ces conditions, la stricte application des principes probabilistes conduit, pour une maille de la zone nord, à une estimation relativement élevée (de l'ordre de 1 700 mg/kg), et comme cette zone est très vaste, la quantité totale est importante. Le moyen de réduire cette estimation serait de faire davantage de mesures sur cette couche : nous indiquons à quel endroit elles doivent être faites.

Sur une très vaste étendue, si l'on ne dispose que de très peu de mesures, même si une petite proportion seulement a révélé des résultats élevés, l'intégrale globale sera significative.

Pour la troisième couche (au-delà de 30 cm), comme aucune mesure n'a été faite, aucune extrapolation ne peut être réalisée, qu'elle soit probabiliste ou déterministe.

Pour mettre en place de futurs plans d'expérience, nous définissons quatre zones de risque de présence de fioul sur la carte de la seconde couche (20-30 cm). A partir de ces zones, nous simulons un plan d'expérience constitués de vingt mesures supplémentaires dans cette couche. Si l'absence de fioul est avérée pour toutes les mesures, l'estimation dans cette couche ne serait plus que de 3,6 t. Au total, la quantité de fioul présent dans le port serait de 7 t. Avec davantage de mesures négatives, la quantité totale serait encore réduite, bien évidemment.

Rien ne permet d'affirmer que la zone non mesurée contient du fioul. La dispersion du fioul dans le port n'étant pas homogène, il est impossible de généraliser l'estimation de cette manière.

Pour prendre en compte toutes les particularités liées aux données récoltées lors des trois campagnes de mesures, nous procédons de la manière suivante :

- Au lieu d'affecter les mesures à des mailles de 25 m de côté ou de moyenniser l'ensemble des mesures pour la totalité du port, nous utilisons la méthode de l'hypersurface probabiliste (EPH) qui permet de tenir compte de la proximité des zones avec et sans fioul.
- La zone étudiée est discrétisée en maille d'1 m de côté, ce qui permettra de tenir compte de la distance entre les points de mesures.
- La quantité de fioul est reconstituée par couche de vase. Avec les données dont nous disposons, nous allons étudier les couches de 0 à 20 et de 20 à 30 cm. Au-delà de 30 cm aucune donnée n'a été collectée.

Dans la mesure où aucune donnée n'existe au-delà de 30 cm, il est évident que :

- Aucune extrapolation d'une couche à une autre ne saurait être justifiée, car chaque couche correspond à une sédimentation différente, à une époque différente ;
- Aucune méthode mathématique ne permet de reconstituer des données quand on n'en a pas du tout.

### *B. Estimation de la couche de 0 à 20 cm*

La surface de la couche de 0 à 20 cm de vase est de 76 875 m<sup>2</sup> ; elle est représentée en clair sur la figure 1 :

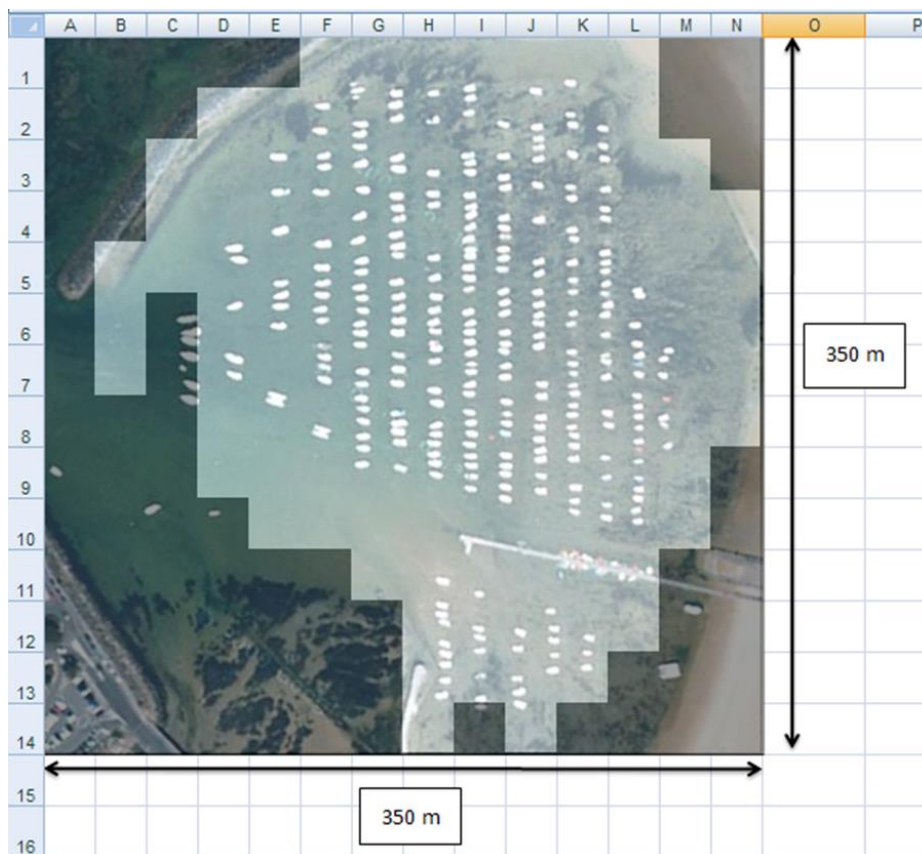


Figure 1 : Surface de la couche de 0 à 20 cm de vase

Nous utilisons les 51 mesures comprises entre 0 et 20 cm issues des campagnes de mesure. Parmi ces mesures, une seule est attribuée à la pollution, avec comme valeur 7 610 mg d'HCT / kg de vase. Pour les autres mesures, la densité en pétrole est nulle.

L'hypersurface probabiliste permet alors de reconstruire la densité en pétrole des 76 824 cellules vides à partir des 51 données disponibles. On obtient la carte du port (figure 2) en fonction de la densité de fioul dans la couche entre 0 et 20 cm. La valeur maximale est indiquée en rouge, les zones où la présence de fioul est nulle sont en bleu.

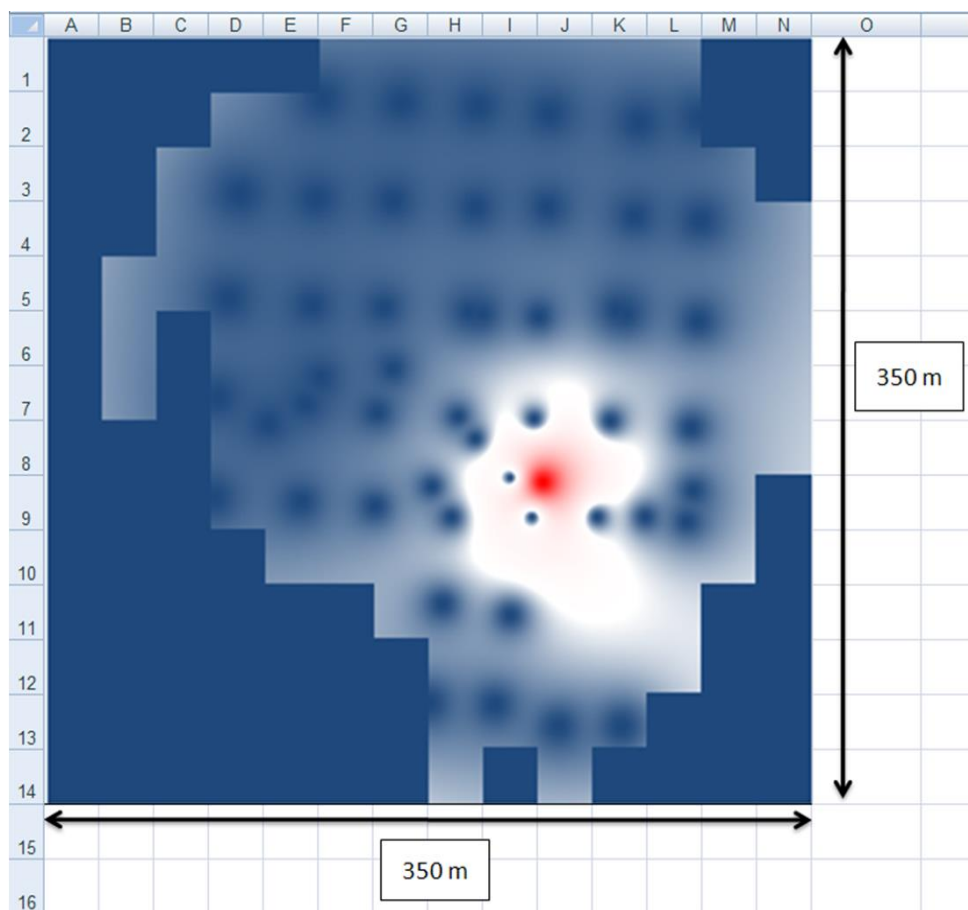


Figure 2 : Carte de la densité de fioul dans la couche de 0 à 20 cm

Connaissant la masse de vase dans chaque cellule, l'estimation de la quantité de fioul présent dans la couche entre 0 et 20 cm de vase dans le port est de 3,4 t.

Dans cette première couche, la quantité de mesures est suffisante pour restreindre la zone de forte densité. La zone où la densité en pétrole est supérieure à 1 000 mg d'HCT / kg de vase ne représente que 2,6% de la surface de la couche, soit environ 2 000 m<sup>2</sup>.

### C. Estimation de la couche de 20 à 30 cm

La surface de la couche de 20 à 30 cm de vase est de 63 750 m<sup>2</sup>; elle est représentée en clair sur la figure 3 :

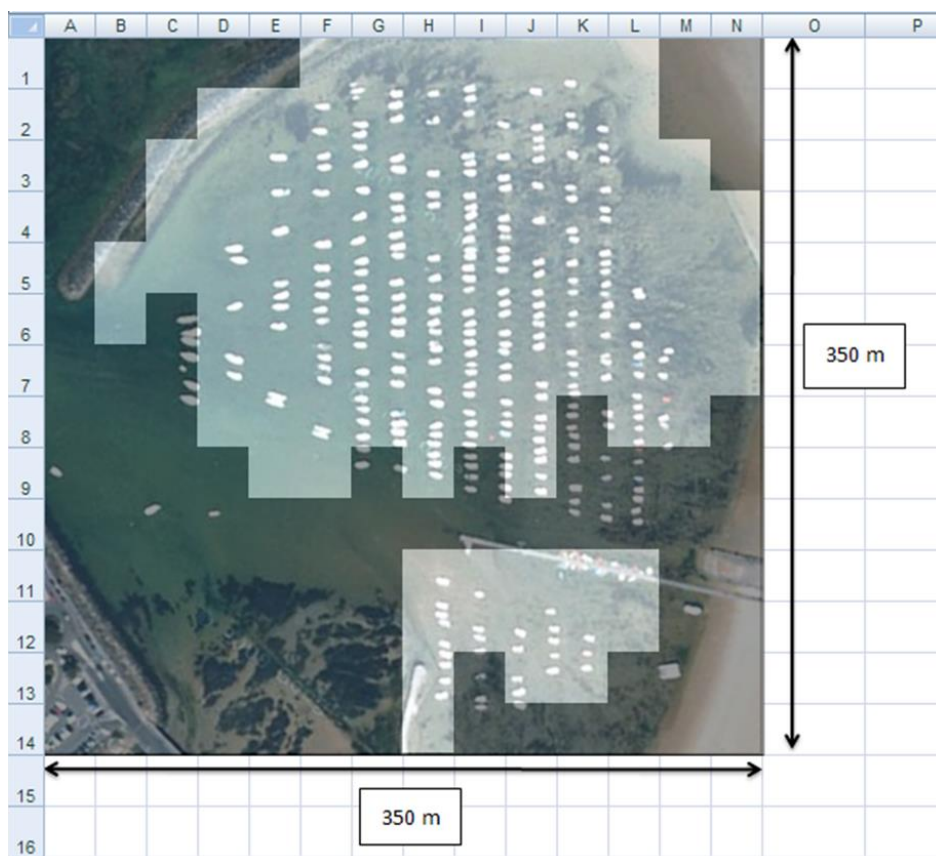


Figure 3 : Surface de la couche de 20 à 30 cm de vase

Parmi les six mesures disponibles, deux sont attribuées à la pollution, avec comme valeur 3 160 et 9 290  $mg$  d'HCT /  $kg$  de vase. Les autres mesures ne correspondant pas à du pétrole, la densité en pétrole est nulle.

L'hypersurface probabiliste permet alors de reconstruire la densité en pétrole des 63 744 cellules vides à partir des six données disponibles. On obtient la carte du port (figure 4) en fonction de la densité de fioul dans la couche entre 20 et 30 cm. La valeur maximale est indiquée en rouge, les zones où la présence de fioul est nulle sont en bleu.

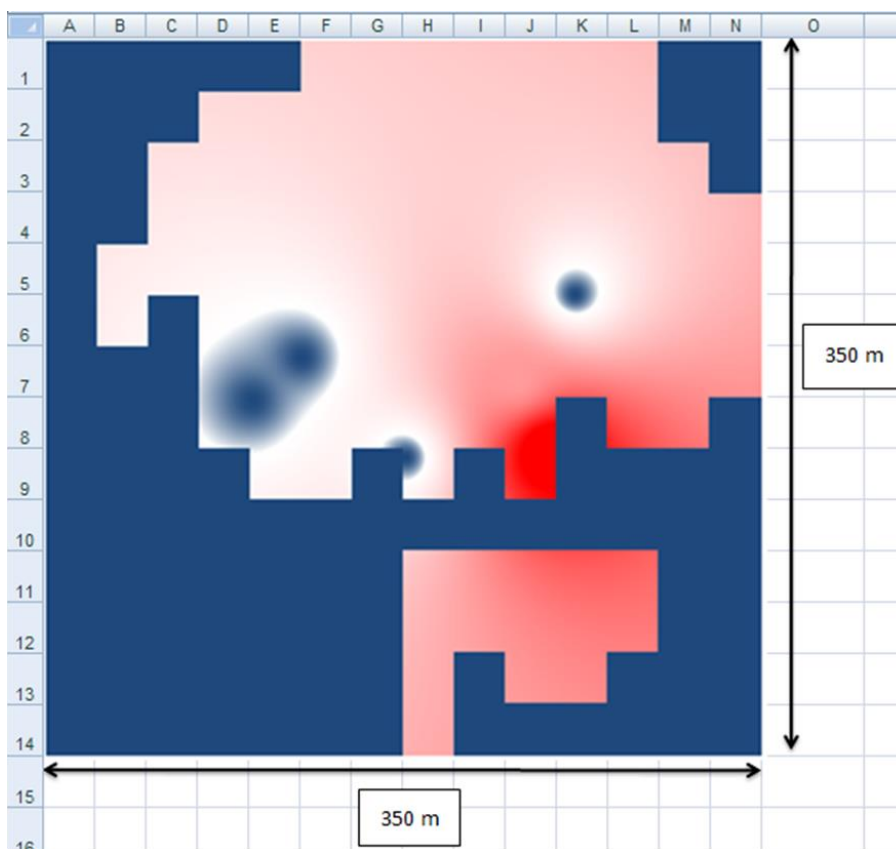


Figure 4 : Carte de la densité de fioul dans la couche de 20 à 30 cm

Connaissant la masse de vase dans chaque cellule, l'estimation de la quantité de fioul présent dans la couche entre 20 et 30 cm de vase dans le port est de 19,4 t.

Cette estimation est élevée, tout simplement parce que nous ne disposons que de six mesures. Grossièrement, et ceci est facile à comprendre, à grande distance des six points, par exemple dans toute la région Nord-Ouest, l'estimation due à l'EPH n'est pas nulle, parce que l'influence des points où la pollution est forte se fait sentir. Certes, en chaque point de la zone, l'estimation de pollution est faible, mais ceci vaut pour une zone très large. En effet, la zone où la densité en pétrole est supérieure à 1 000 mg d'HCT / kg de vase représente 76,2% de la surface de la couche, soit environ 48 500 m<sup>2</sup>.

La méthode de l'hypersurface probabiliste ne permet pas uniquement d'estimer une valeur déterministe, mais une loi de probabilité (chaque valeur déterministe est obtenue en déterminant l'espérance mathématique de chaque loi).

La loi de probabilité de chaque cellule n'est pas la même partout : près des valeurs observées, elle est fortement concentrée alors qu'elle devient plus diffuse dans les zones éloignées des mesures.

Par exemple, prenons le point de coordonnées (1 ; 126) situé à l'extrémité Nord-Ouest du port dans lequel aucune mesure n'a été effectuée. En utilisant, l'hypersurface probabiliste, l'espérance de la loi de probabilité obtenue à partir de la propagation des six mesures est de 1 708 mg d'HCT / kg de vase.

La loi de probabilité en ce point est représentée sur la figure 5 :

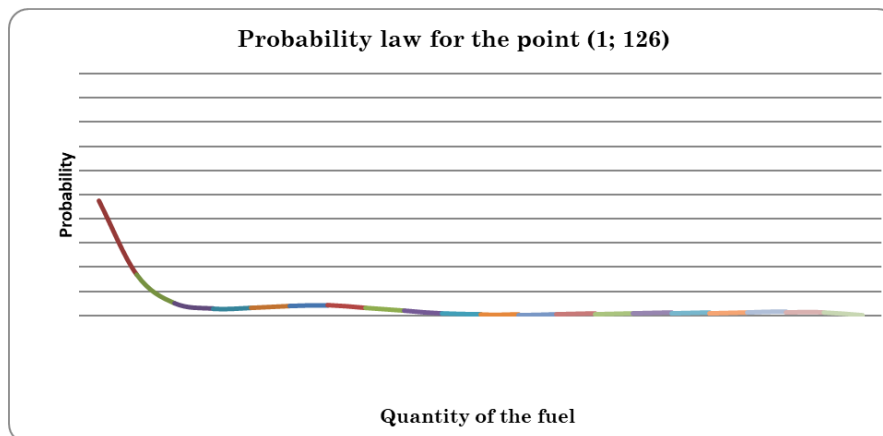


Figure 5 : Loi de probabilité de la densité de pétrole au point de coordonnées (1 ; 126)

La probabilité que la densité en pétrole soit très faible est importante. Cependant, la loi n'étant pas assez concentrée vers les faibles valeurs du fait de l'éloignement par rapport aux mesures, la probabilité que cette zone soit fortement dense en fioul est non négligeable.

Par conséquent, la seule manière de connaître avec plus de précision la quantité de fioul présent dans la couche de 20 à 30 cm est d'augmenter le nombre de mesures afin de vérifier l'ensemble de la zone et de limiter la distance entre les points de mesures et les points estimés.

#### D. Plan d'expérience pour la couche de 20 à 30 cm

##### 1. Détermination de zones à risque

Pour choisir les zones du port où il est nécessaire d'effectuer de nouvelles mesures, la couche de 20 à 30 cm est décomposée en quatre zones en fonction du risque de présence de pétrole.

Pour cela, nous fixons le seuil de 1 000 mg d'HCT / kg de vase. Les quatre zones (figure 6) ont les caractéristiques suivantes :

- La zone n°1 (en bleu) correspond à la surface de la couche pour laquelle la probabilité de dépasser le seuil est inférieure à 10%. Aucune mesure supplémentaire n'est nécessaire dans cette zone.
- La zone n°2 (en blanc) correspond à la surface de la couche pour laquelle la probabilité de dépasser le seuil est comprise entre 10% et 50%. Cette zone où l'information est minime nécessite de réaliser des mesures de vérification. On pourra procéder à des mesures à un intervalle de 50 m.
- La zone n°3 (en jaune) correspond à la surface de la couche pour laquelle la probabilité de dépasser le seuil est comprise entre 50% et 90%. Dans cette zone, on pourra procéder à des mesures à un intervalle de 25 m.

- La zone n°4 (en rouge) correspond à la surface de la couche pour laquelle la probabilité de dépasser le seuil est supérieure à 90%. Des mesures doivent être effectuées dans cette zone afin de limiter la surface de forte densité.

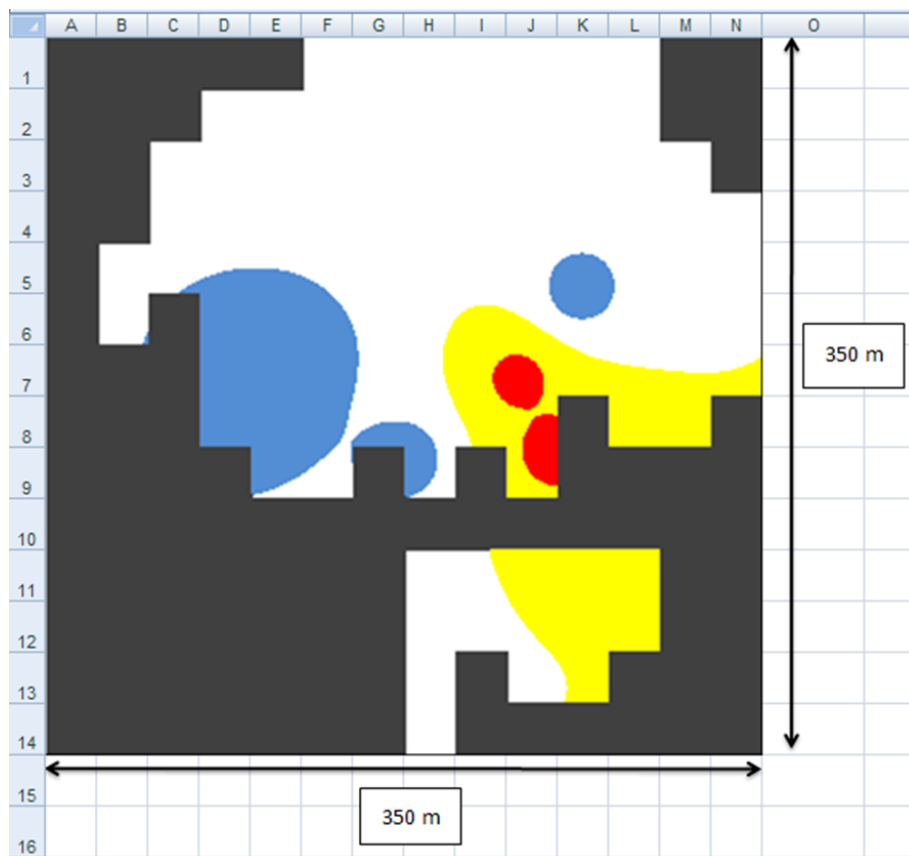


Figure 6 : Carte des 4 zones de risque de la couche de 20 à 30 cm

## 2. Simulation d'un plan d'expérience

Supposons que nous disposons d'un budget permettant de réaliser 20 mesures supplémentaires dans la couche de 20 à 30 cm du port.

Nous choisissons de simuler le plan d'expérience suivant :

- 10 mesures (en bleu) sont réalisées dans la zone n°2 située au Nord-Ouest du port avec un intervalle de 50 m ;
- 10 mesures (en jaune) sont réalisées dans la zone n°3 située autour de la zone où la densité est probablement la plus forte avec un intervalle de 25 m.

Les mesures sont réalisées au centre des mailles de 25 m de côté suivantes :

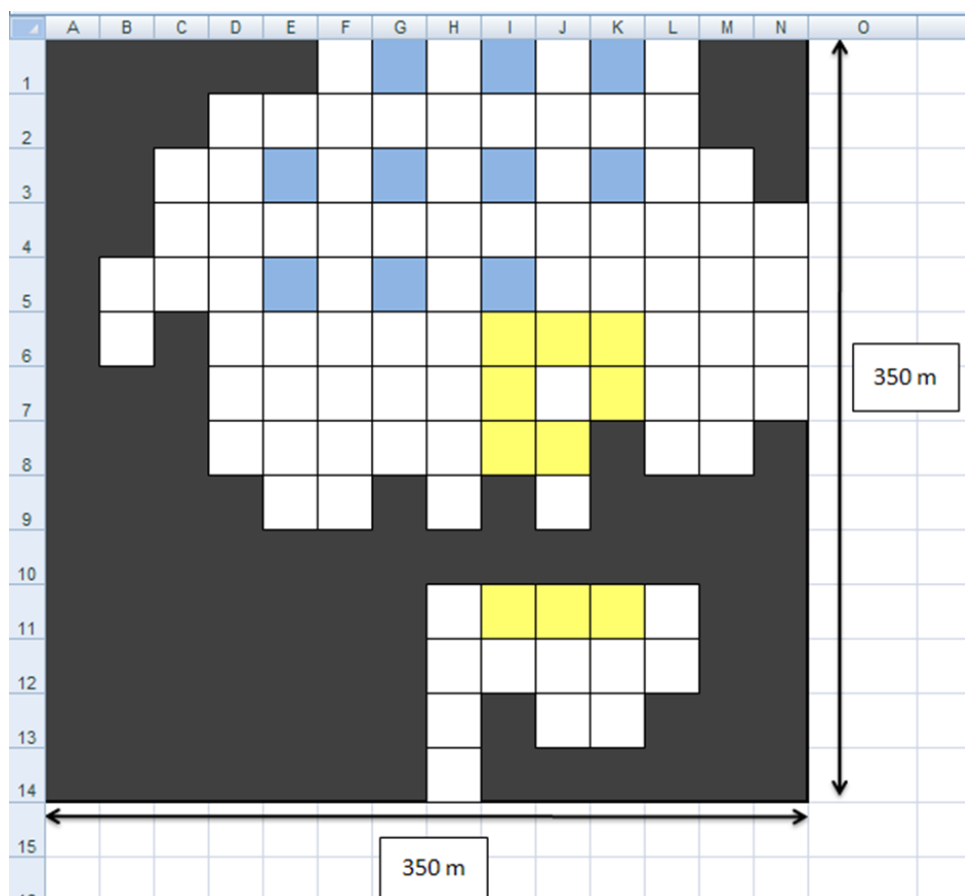


Figure 7 : Position des 20 points de mesures supplémentaires

En supposant que l'absence de pétrole soit avérée pour les 20 points de mesure, nous utilisons la méthode de l'hypersurface probabiliste en affectant des densités nulles à tous les points supplémentaires.

On obtient une nouvelle carte du port (figure 8) en fonction de la densité de fioul dans la couche entre 20 et 30 cm. La valeur maximale est indiquée en rouge, les zones où la présence de fioul est nulle sont en bleu.

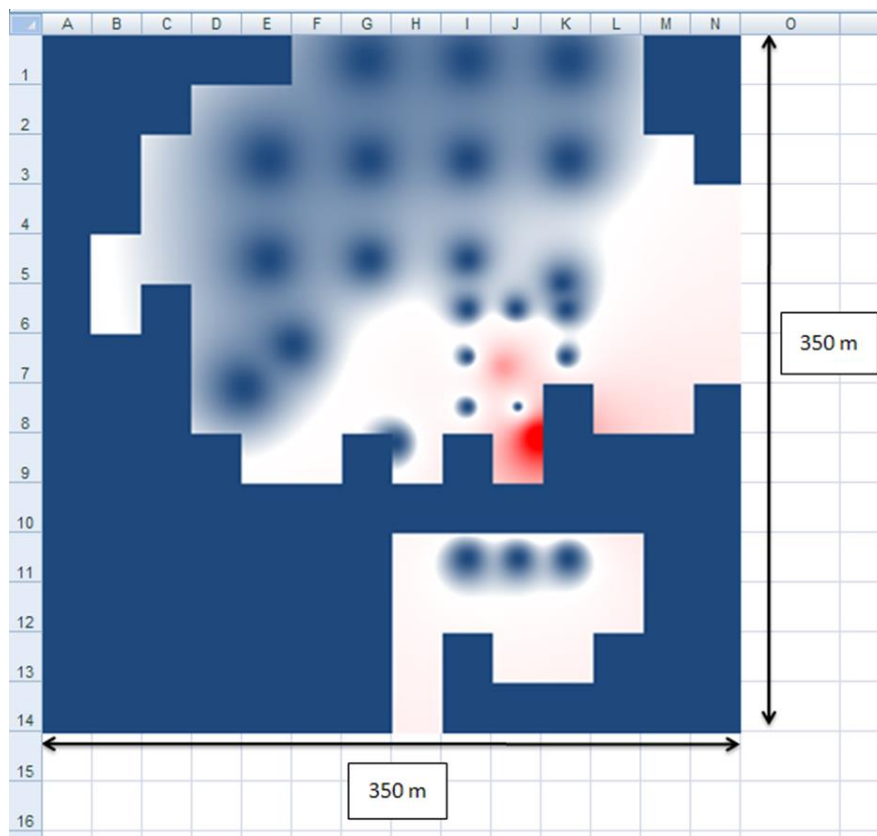


Figure 8 : Nouvelle carte de la densité de fioul dans la couche de 20 à 30 cm suite à la simulation d'un plan d'expérience

Après la simulation du plan d'expérience, la nouvelle estimation de la quantité de fioul présent dans la couche entre 20 et 30 cm de vase est de 3,6 t.

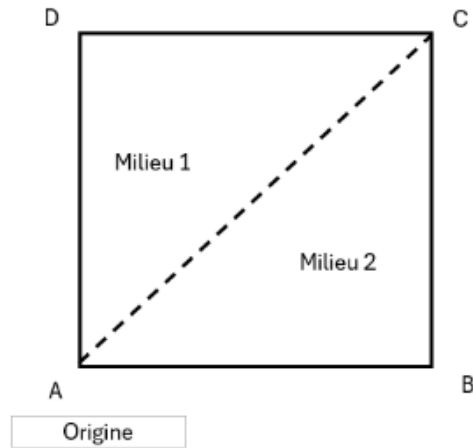
Dans ce cas, la quantité de fioul présent dans le port serait de 7 t.

## V. Utilisation de l'EPH dans des milieux non homogènes

Dans cette partie, nous présentons la méthode de l'EPH appliquée à des milieux non homogènes c'est-à-dire à des milieux de nature différente séparés par une délimitation. La propagation de l'information dépendra de la nature du milieu. Il est important de noter que les formules de l'EPH restent identiques quelle que soit la situation (homogène ou non homogène, 2D ou 3D). La seule différence réside dans les calculs des distances entre les points et dans le choix du coefficient de propagation.

### A. Milieu non-homogène en deux dimensions.

Comme indiqué précédemment, les calculs de distances dépendent de la configuration de la situation pour laquelle l'EPH doit s'appliquer. Les différences entre les situations dépendent de la position de la ligne de délimitation entre les milieux. Cette délimitation peut être simple (une droite parallèle à l'un des axes) ou complexe (une délimitation non linéaire). Par souci de simplicité, nous traiterons un cas où la limite est une droite coupant le plan en diagonale. Notre milieu d'étude est présenté dans la figure suivante :



Soit un repère orthonormé d'origine A. L'axe des abscisses est confondu avec la droite (AB) et l'axe des ordonnées est confondu avec la droite (AD). La droite (AC) sépare le plan en deux milieux distincts et a pour équation  $x = y$ . Tous les points de mesures sont compris dans le carré ABCD de côté 5. Les coordonnées des 4 points du carré sont donc A(0, 0), B(5, 0), C(5, 5) et D(0, 5). Une fois la situation définie, nous pouvons procéder à la mise en place de l'EPH.

Nous disposons de 100 points de mesures  $M_i$ . Leurs coordonnées sont comprises entre 0 et 5. Les valeurs  $c_i$  des mesures sont comprises entre 0 et 150. La première étape de l'EPH est de déterminer le coefficient de propagation  $\lambda$  de chacun des milieux. Par définition, il est proportionnel à la caractéristique C du milieu. Celle-ci peut être le coefficient de diffusion du milieu ou sa densité. C'est un paramètre qui doit être défini à l'avance.  $\lambda$  et C sont proportionnels entre eux via un coefficient  $\delta$ . Il est calculé grâce à la formule :  $\delta = \frac{\ln(1+v)}{c_1d_1 + c_2d_2}$ . Il correspond à la plus forte atténuation, ce qui signifie que la composante  $c_1d_1 + c_2d_2$  doit être la plus élevée possible. Cela revient à calculer, pour chacun des points de mesure, la distance qui les sépare des quatre sommets du carré et de sélectionner la valeur la plus élevée.

Dans un milieu homogène, ce calcul est simple : il suffit d'appliquer la formule de la distance 4 fois pour chacun des points de mesures. Pour un milieu non-homogène, la difficulté réside dans le franchissement de la limite entre les deux milieux. Pour les distances aux points A et C, les calculs sont aisés, car cette ligne n'est jamais franchie. Pour les deux autres sommets, les choses sont plus difficiles. Si un point M se situe dans le milieu 1, sa distance au point D est facile à calculer. Mais la distance au point B est plus compliquée à déterminer. En effet, la distance MB traverse les deux milieux.

Selon la composante  $c_1d_1 + c_2d_2$ , il faut calculer les distances en fonction du milieu qu'elles traversent. Une distance unique doit traverser un seul milieu. En cas de franchissement de la droite de démarcation, la distance doit être coupée en deux. Il faut dans un premier temps déterminer la distance entre le point M et la ligne de démarcation, puis la distance entre la ligne de démarcation et le sommet B. Dans les faits, cela revient à calculer le point d'intersection entre les droites (AC) et (BM). Pour cela, il faut déterminer l'équation des deux droites, puis résoudre un système de deux équations à deux inconnues pour trouver le point  $X_r$ , point d'intersection des deux droites. Une fois ce point connu, la composante  $c_1d_1 + c_2d_2$  devient facile à calculer en veillant bien à définir les distances  $d_1$  et  $d_2$ .

Les étapes suivantes sont similaires à celles pour un milieu homogène. Il faut calculer le paramètre  $\sigma$  qui est la racine carrée de la variance pour chacun des points de mesures. Sa formule est légèrement différente lorsqu'on étudie un milieu non-homogène. Si le point de mesure  $M_i$  et le point manquant  $M$  (celui dont nous voulons déterminer la valeur via l'EPH) sont dans le même milieu, la forme de  $\sigma$  reste identique. Mais si ces deux points sont dans des milieux différents, alors la composante  $\lambda d$  dans la formule de  $\sigma$  est remplacée par  $S = \lambda_1 d_1 + \lambda_2 d_2$ . Il faut alors déterminer le point d'intersection entre les droites (AC) et ( $M_iM$ ). Une fois  $\sigma$  connu, le dernier paramètre  $\gamma$  peut être calculé.

La dernière étape est le calcul des lois de probabilités  $p(j, M_i, M)$  qui est la propagation de l'information de  $M_i$  sur  $M$  dans la classe  $j$ . Ces probabilités doivent ensuite être normalisées pour que, quel que soit  $M_i$ , la somme des probabilités  $p(j, M_i, M)$  soit égale à 1. Nous combinons ensuite les lois de probabilités pour obtenir  $p(j, M)$  qui est la propagation de l'information issue de toutes les mesures  $M_i$  sur  $M$ . Une fois cette loi obtenue, la valeur recherchée associée à  $M$  est liée au maximum de la loi  $p(j, M)$ . La valeur la plus élevée de la probabilité donne la valeur de la mesure en  $M$ .

Comme nous venons de le montrer, l'application de l'EPH en milieu non-homogène ne se différencie pas d'une application en milieu homogène dans sa structure de base. La méthode générale ne change pas. La seule différence réside dans le calcul des distances où le passage d'un milieu à l'autre doit être pris en compte.

### **Exemple numérique :**

Soient 100 points de mesures  $M_i$  dont les coordonnées  $x$  et  $y$  sont comprises entre 0 et 5. Les valeurs de concentrations  $c$  en polluants des points  $M$  sont comprises entre 0 et 150 mg/L. Les attributs  $C_1$  et  $C_2$  des milieux 1 et 2 sont respectivement 1.5 et 2. Les concentrations de polluants sont discrétisées en 30 classes de taille 5 mg/L. Soit le point  $M$  de coordonnées (2, 2.5) dont nous voulons connaître la concentration en polluants. Le point  $M$  appartient au milieu 1 et l'application de l'EPH donne la valeur de 97.5 mg/L comme concentration  $c$ .

### *B. Milieu non-homogène en trois dimensions*

Le passage en trois dimensions complique le calcul des distances entre les points. Dans un espace en trois dimensions, le plan de séparation entre les milieux peut être placé de manière très différente. Pour notre étude, dans un souci de simplicité, le plan de séparation entre les deux milieux a une équation simple.

L'espace de travail est un cube avec un plan qui le sépare en deux selon l'axe des abscisses. Le plan a pour équation  $x = 3$ . L'origine du repère est placée sur le point A. Les coordonnées des points sont comprises entre 0 et 5 pour les trois axes du repère. L'axe (Ox) est confondu avec la droite (AE), l'axe (Oy) avec la droite (AB) et l'axe (Oz) avec la droite (AD). Les 8 points du cube ont pour coordonnées : A(0, 0, 0), B(0, 5, 0), C(0, 5, 5), D(0, 0, 5), E(5, 0, 0), F(5, 5, 0), G(5, 5, 5) et H(5, 0, 5).

Comme pour un milieu en 2D, la difficulté réside dans le franchissement du plan de séparation lors du calcul d'une distance. Lorsque deux points sont dans le même milieu, le calcul est simple.

Mais lorsque les deux points sont dans un milieu différent, il est obligatoire de déterminer le point d'intersection entre le plan et la droite portée par les deux points. En effet, les deux milieux ayant un attribut différent, il faut séparer la distance en deux et calculer la distance parcourue dans chacun des milieux.

Soit un point  $P(X, Y, Z)$  situé dans le milieu 2,  $A$  le sommet du cube de coordonnées  $(0, 0, 0)$  et  $R(x_R, y_R, z_R)$  le point d'intersection entre le plan de séparation et la droite  $(AP)$ . Nous voulons calculer la distance  $AP$ . Dans l'EPH, la distance est associée au coefficient de propagation du milieu  $\lambda$  proportionnel à l'attribut du milieu  $c$ . Il faut donc séparer la distance en deux selon le milieu traversé. La distance  $AP$  n'est pas exploitable ; il faut calculer  $AR$  puis  $RP$  et donc déterminer  $R$ .

Pour calculer le point d'intersection entre le plan de séparation et la droite, il faut connaître leur équation respective. L'équation cartésienne du plan est simple :  $x = 4$ . Dans l'espace, l'équation paramétrique d'une droite est un système de trois équation de la forme  $x=a+bt$ .

La droite  $(AP)$  passe par les points  $A(0, 0, 0)$  et  $P(X, Y, Z)$ .

Elle a pour vecteur directeur  $\overrightarrow{PA} = (-X, -Y, -Z)$ .

La droite a pour équations  $x = X - Xt$ ,  $y = Y - Yt$  et  $z = Z - Zt$ .

Comme le plan a pour équation  $x = 4$ , alors  $x_R = 4$ . Cela permet de calculer le  $t$  correspondant et donc de calculer  $y_R$  et  $z_R$ . Ainsi les coordonnées de  $R$  sont déterminées.

En suivant cette méthode, nous pouvons calculer toutes les distances entre les points de mesures et les 8 sommets du cube. Le coefficient  $\delta$  reliant l'attribut du milieu et le coefficient de propagation peut être calculé en sélectionnant la composante  $C_1d_1 + C_2d_2$  la plus élevée possible. Une fois les coefficients  $\lambda$  connus, nous pouvons calculer les paramètres  $\sigma$  et  $\gamma$ . Le calcul de la distance entre un point de mesure et le point manquant  $M$  est similaire à celui vu précédemment. Les dernières étapes de l'EPH sont semblables à celles en milieu homogène.

En trois dimensions, l'information se propage dans toutes les directions de la même manière. La seule différence dans cette propagation vient de la nature du milieu : le coefficient de propagation est directement lié aux paramètres du milieu. Le passage entre les milieux est considéré comme totalement poreux et ne freine pas la diffusion de l'information. La difficulté des calculs lors de l'étude d'un milieu en trois dimensions est proportionnelle à la complexité du plan de séparation entre les milieux. Plus le plan est complexe, plus les calculs sont compliqués. Il peut y avoir aussi plusieurs plans de séparations qui se croisent entre eux. Lors d'une étude, il est possible d'implémenter des règles supplémentaires dans l'EPH pour traduire les propriétés du milieu. Par exemple, dans le cadre d'une étude sur la pollution, celle-ci peut se propager d'un milieu à l'autre dans un seul sens et pas dans l'autre. Ainsi, le plan de séparation pourra être poreux et dans un sens et imperméable dans l'autre. Il peut aussi freiner la diffusion de la pollution et donc le poids des mesures sur le point manquant. Tout dépend de la nature du milieu étudié.

## Exemple numérique :

Soient 100 points de mesures  $M_i$  dont les coordonnées  $x$ ,  $y$  et  $z$  sont comprises entre 0 et 5. Les valeurs de concentrations  $c$  en polluants des points  $M$  sont comprises entre 0 et 150 mg/L. Les attributs  $C_1$  et  $C_2$  des milieux 1 et 2 sont respectivement 1.5 et 2. Les concentrations de polluants sont discrétisées en 30 classes de taille 5 mg/L. Soit le point  $M$  de coordonnées (2, 2.5, 3) dont nous voulons connaître la concentration en polluants. Le point  $M$  appartient au milieu 1 et l'application de l'EPH donne la valeur de 47.5 mg/L comme concentration  $c$ .

## VI. Evolution temporelle

La structure même de l'EPH, par sa flexibilité, se prête bien à l'étude des évolutions temporelles. L'axe est maintenant celui du temps (problème en 1d), avec par exemple pour borne un an. On mesure la pollution, à un endroit, à un instant  $t=0$  et on voudrait savoir ce qu'elle peut être, en ce même endroit, à un instant ultérieur.

Nous distinguerons trois situations :

### A. Absence d'information

On ne sait absolument rien sur la physique de la pollution. On a mesuré  $c$  en un point  $M_0$  à l'instant  $t=0$ . On se demande quelle sera la pollution en ce même point, en un instant ultérieur. Pour fixer les idées, disons que la pollution est entre 0 et 1g, mesurée au mg près : donc  $\nu=1000$  et  $w=10^{-3}$ .

Le coefficient de propagation sera  $\lambda = \frac{\text{Log}(\nu+1)}{365}$  si l'horizon de temps est d'une année et l'unité de temps le jour.

L'écart-type au temps  $t$  (en jours) sera :

$$\sigma_1 = \frac{w}{\sqrt{2\pi}} \exp(\lambda t).$$

L'information générée par la mesure au temps 0, évaluée au temps  $t$ , est donnée par la loi de probabilité :

$$p_t(j) = u_1 \exp\left(-\frac{(jw-c)^2}{2\sigma_1^2}\right) \quad (1)$$

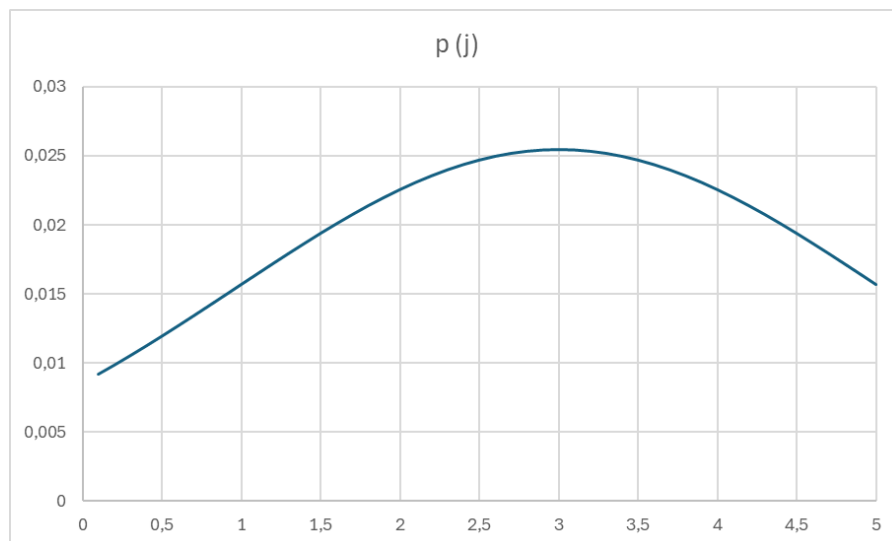
C'est une courbe en cloche, dont le maximum est toujours atteint pour  $x=c$ , mais la cloche va en s'évasant au fur et à mesure que le temps passe : l'information devient de moins en moins précise.

En d'autres termes, on pronostiquera toujours, au point de mesure, que la pollution vaut le  $c$  d'origine (que dire d'autre ?), mais on en est de moins en moins sûr au fur et à mesure que le temps passe.

Si maintenant on recherche la pollution au temps  $t$  en un point quelconque de l'espace, on appliquera l'EPH spatiale, en partant de la valeur donnée par la formule (1) et non en partant de la valeur  $c$ . Chaque valeur  $jw$  de la pollution possible sera affectée de la probabilité correspondante.

### Exemple numérique :

Soit un point dans l'espace de coordonnées  $M(2, 5, 9)$ . La concentration en polluants mesurée est de 3 mg/L. Nous étudions l'évolution de la pollution en ce point sur une année. Nous utilisons les formules de l'EPH décrites dans le paragraphe précédent. L'horizon de temps limite choisi est de 1 an. La pollution est étudiée à  $t = 365$  j. Le graphique montrant les résultats est présenté ci-dessous :



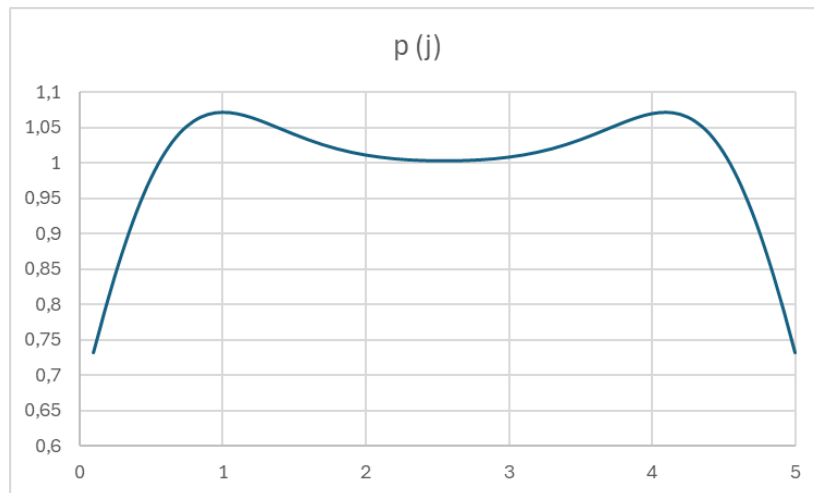
Comme démontré précédemment, plus l'horizon de temps est lointain plus la courbe des probabilités sera évasée, tout en conservant un maximum autour de la valeur mesurée.

Ensuite nous étudions la pollution en point quelconque de l'espace toujours à l'horizon de temps  $t = 365$  j. Nous appliquons l'EPH spatiale après l'EPH temporelle.

Soit un point de l'espace  $X$  de coordonnées  $(8, 8, 3)$ . L'espace défini est un cube d'arête 10. Nous voulons connaître la concentration de polluants au point  $X$  à partir de la concentration au point  $M$ . Les paramètres de l'EPH sont  $d_{\max} = 13$ ,  $\lambda = 0.302$  et  $\sigma = 0.602$ . La distance  $d$  entre  $M$  et  $X$  est égale à 9. Comme il n'y a que deux points à étudier, le paramètre  $\gamma$  est égal à 1. Les concentrations en polluants sont comprises entre 0 et 5 mg/L et sont discrétisées de la même manière que pour l'EPH temporel.

Dans ce cas, il n'y a pas de points de mesures, mais nous utilisons les valeurs  $j$  de l'EPH temporelle que nous propageons sur le point  $X$ . Ainsi nous avons autant de points propagés que de classes  $j$  pour la concentration.

Pour chaque point propagé, la valeur  $\sigma$  est la même, seule la valeur  $j$  change dans le calcul de la probabilité. Comme  $\gamma$  est égal à 1, la combinaison des lois de probabilité revient à additionner les probabilités des points pour chaque classe  $j$ . Le résultat est présenté dans le graphique ci-dessous :



La courbe obtenue est symétrique autour de la valeur 2.5. La valeur de la concentration en polluants pour le point X est égale à 4.05 mg/L.

### B. La pollution décroît naturellement, mais ne se propage pas

Il se peut que, pour des raisons physiques, on sache que la concentration en polluant va diminuer de manière naturelle (par exemple la radioactivité diminue avec le temps).

Admettons par exemple que, au temps  $t$ , la pollution soit  $c\varphi(t)$  où  $\varphi$  est une fonction décroissante, avec  $\varphi(0) = 1$ . Alors la construction précédente se simplifie : on n'a plus besoin de l'EPH temporelle. A l'instant  $t$ , on propage la valeur  $c\varphi(t)$  au moyen de l'EPH spatiale.

### C. La pollution décroît et se propage

Si les règles de décroissance et de propagation sont très bien connues, pour des raisons physiques, on pourra, à partir de ces règles, reconstituer la pollution en tout point et en tout instant. On dispose en général d'un logiciel qui incorpore les lois de la physique, mais la question se pose : comment tenir compte des incertitudes ? Un logiciel retourne une réponse précise si, en entrée, on lui fournit des données précises.

L'espace des configurations peut avoir une dimension très élevée : outre les trois paramètres d'espace, des paramètres liés à l'environnement : hydrographie, température, débit, etc. Une exploration explicite de l'espace des configurations est souvent impossible : on ne peut faire que quelques milliers ou millions de runs du code, et si l'espace des configurations comporte 15 paramètres, chacun pouvant prendre 10 valeurs, il y a  $10^{15}$  situations à explorer.

L'EPH permet alors de propager les résultats fournis par le code, à partir des points où un run a été effectué, à tous les points où aucun run n'a été réalisé, et en particulier d'identifier les zones à risque, c'est-à-dire les situations où une variable pourrait prendre des valeurs que l'on estime dangereuses.

## VII. Exemples de mise en œuvre de la méthode EEPH

Dans cette partie, nous présentons une mise en œuvre de l'EEPH sur des exemples de haute expertise : l'analyse spatio-temporelle des concentrations d'arsenic (As) et de sélénium (Se) dans la nappe de la craie alimentant en eau potable le sud parisien, pour la période d'octobre à décembre 2022.

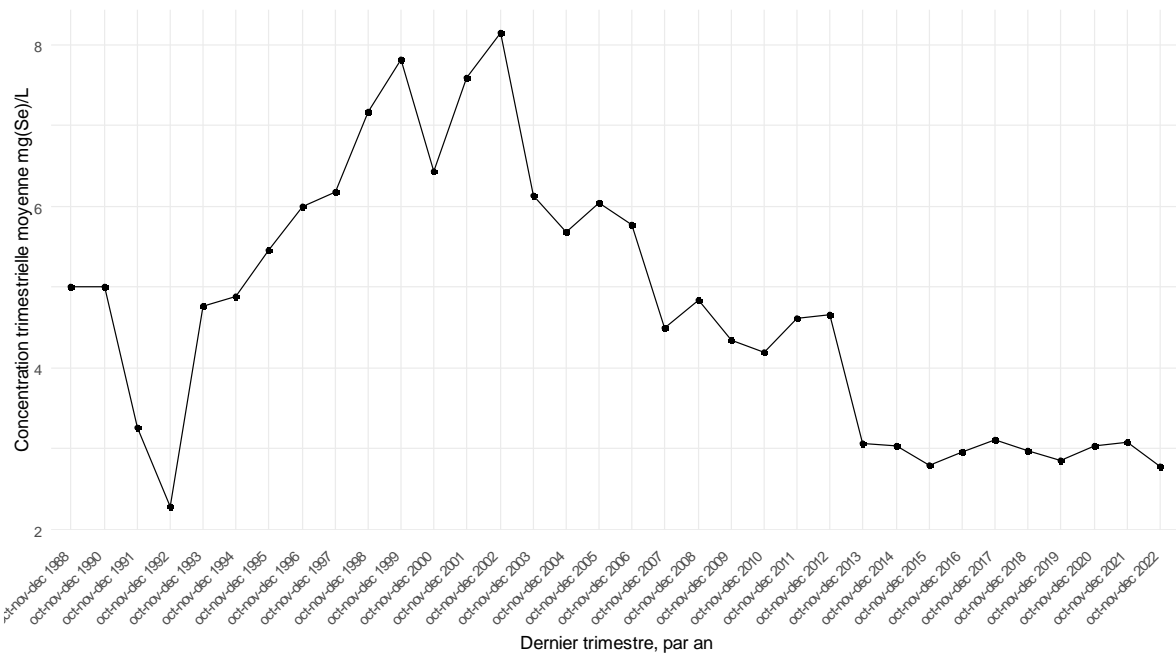
Les données sont issues de la base ADES et sont créditées au projet HOUSES de l'Agence Nationale de la Recherche française (ANR). La base regroupe plusieurs données en basses-eaux en France, pour la période de mars 1987 à décembre 2022 :

- Positions spatiales des relevés expérimentaux ;
- Mesures de concentrations d'arsenic (As) et de sélénium (Se) en chaque position, exprimées en mg/L ;
- Dates des débuts des prélèvements pour chaque position ;
- Commentaires d'experts sur la validité de chaque mesure.

### *a. Pré-traitement des données*

Les valeurs aberrantes sont filtrées en définissant des plages de variations valides des concentrations (mg/L). Pour l'arsenic, la plage de concentrations valides, fixée par des experts, est  $0.01 < mg(As)/L < 500$ . Ainsi, toutes les valeurs  $\leq 0.01$  mg/L ou  $\geq 500$  mg/L sont éliminées. Pour le sélénium, la plage retenue est  $0.5 < mg(Se)/L < 50$ .

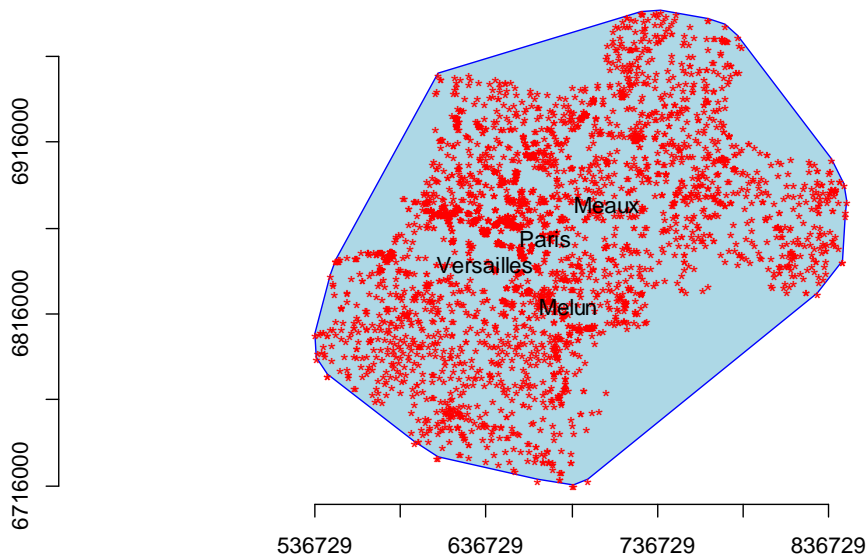
Les données temporelles sont groupées par mois et années et fusionnées avec les données spatiales. Cela permet d'obtenir une vision globale de la concentration moyenne pour chaque site. Pour le sélénium, nous obtenons la courbe suivante qui retrace la moyenne des concentrations pour le trimestre oct-nov-dec de chaque année :



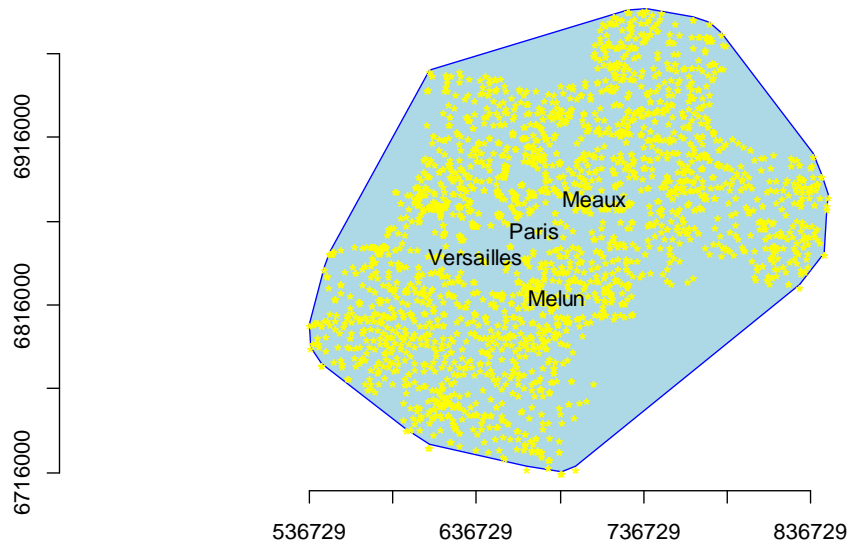
Nous en concluons que les données ont une variabilité significative d'une année à l'autre. Nous faisons de même pour l'arsenic.

Au total, 3648 mesures valides d'arsenic et 2659 mesures valides de sélénium ont été effectuées entre 1987 et 2022 dans la nappe de la craie alimentant en eau potable le sud parisien. Les points de mesures sont marqués en rouge (As) et vert (Se) sur les figures suivantes :

### Carte des mesures de pollution en arsenic autour de Paris



## Carte des mesures de pollution en sélénium autour de Paris



Les axes représentent les coordonnées dans le système EPSG : 2154.

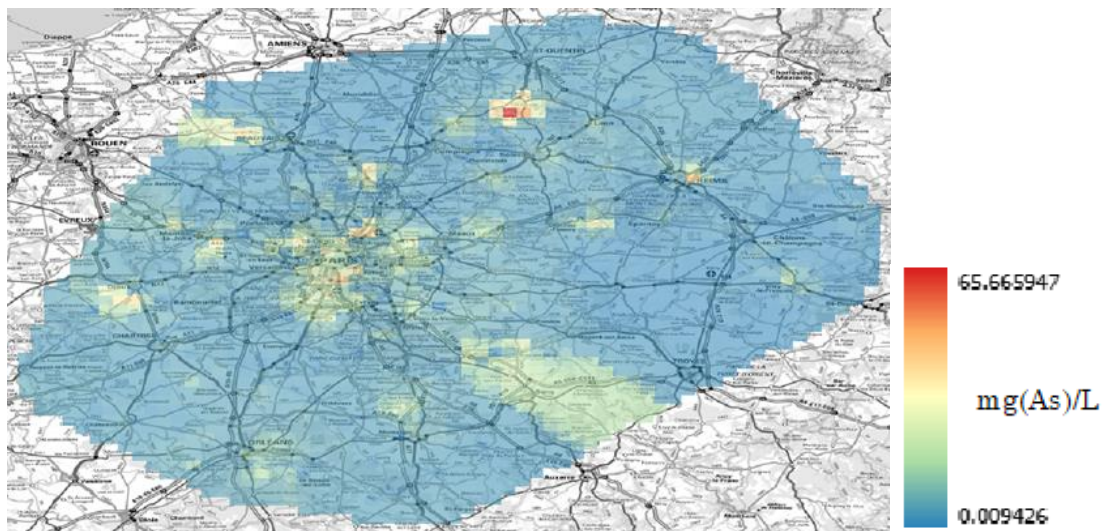
### *b. Pré-tests EEPH :*

Pour cet exemple, ISLANDR utilise la méthode l'EEPH détaillée plus haut comme base d'interpolation. Elle a été codée sous le langage de programmation R sous la forme d'une bibliothèque nommée IISDIA.

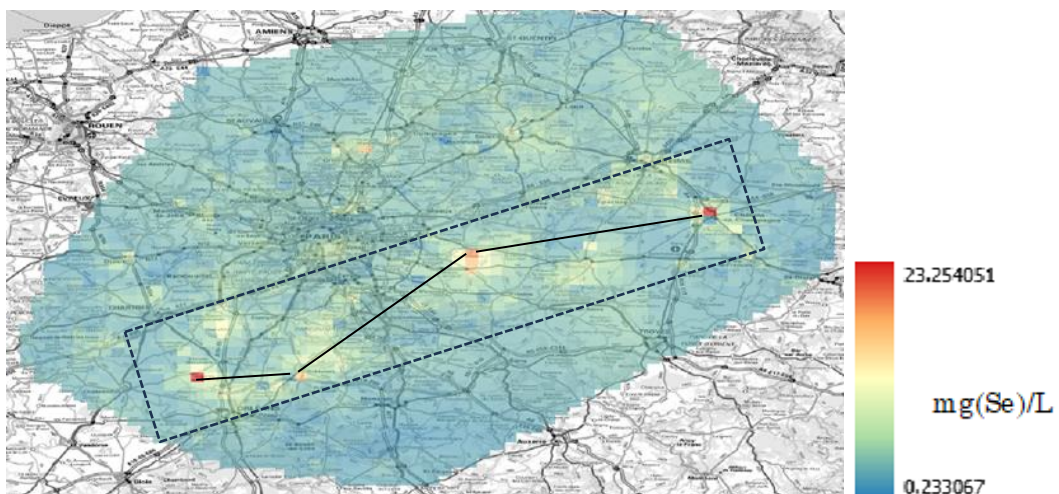
Les paramètres suivants sont utilisés pour l'interpolation :

- Les limites de quantifications des mesures (concentrations limites) sont :  
 $C_{min}(As) = 0.01 \text{ mg/L}$ ,  $C_{max}(As) = 65 \text{ mg/L}$  pour l'arsenic ;  
 $C_{min}(Se) = 0.2 \text{ mg/L}$ ,  $C_{max}(Se) = 23 \text{ mg/L}$  pour le sélénium ;
- La taille des intervalles de discrétisation des mesures est de 5 mg/L pour l'arsenic et 2 mg/L pour le sélénium ;
- La plage temporelle choisie est octobre à décembre 2022.

Les figures suivantes montrent les résultats de l'interpolation en mg/L de chaque polluant, allant du bleu-vert (faible concentration) à l'orange-rouge (forte concentration) :



*Fig. Carte d'interpolation des mesures d'arsenic en eaux basses autour de Paris par EEPH  
Source : ISLANDR*



*Fig. Carte d'interpolation des mesures de sélénium en eaux basses autour de Paris par EEPH  
Source : ISLANDR*

Les estimations des concentrations d'arsenic les plus fortes sont observées près des rivières reliées à la nappe phréatique ; les experts peuvent alors se pencher sur l'hypothèse selon laquelle de l'eau acide peut avoir infiltré la nappe profonde. Quant aux estimations des concentrations de sélénium les plus fortes, elles tracent une ligne en forme de "Z" ; l'hypothèse est qu'il s'agissait d'une rivière fossile.

### *c. Conclusion*

Cette application constitue un intérêt particulier pour les géo-statisticiens car il est difficile de coller des modèles continus à l'Arsenic et au sélénium, et donc d'utiliser des méthodes géostatistiques classiques. De plus, la nature spatio-temporelle des données ajoute une complexité supplémentaire qui dépasse les algorithmes conventionnels.

La méthode EEPH est entièrement probabiliste et se prête bien à l'étude des évolutions spatio-temporelles. Les hypothèses de l'EEPH sont minimales (anisotropie, données éparses, portée des phénomènes) et l'interpolation ne se base pas sur un modèle géostatistique d'expert et ne transforme pas les données.

#### *d. Pistes d'amélioration*

L'exemple des eaux souterraines choisi ici n'intègre pas d'épistèmes géologiques, car ce milieu se caractérise par une continuité qui facilite son interpolation par hypersurface probabiliste. D'autres cas, comme les discontinuités marquées (failles, limites de couches), nécessitent la prise en compte des épistèmes géologiques. Ces dernières regroupent les connaissances propres à la géologie, permettant de modéliser les relations complexes entre propriétés physiques (porosité, densité, granulométrie, minéralogies). La thèse de S. Belbèze en cours explore l'intégration de ces épistèmes dans les interpolateurs afin d'améliorer la représentation des systèmes géologiques discontinus.

## **VIII. Limites de l'EPH**

Dans un contrat récent, nous avons tenté d'utiliser l'EPH pour une prévision temporelle, mais le résultat n'a pas été satisfaisant. La donnée était très saisonnière. Ou bien nous utilisions l'EPH séparément pour tous les mois de janvier, février, etc., et la prévision était trop restrictive, ou bien nous utilisions l'EPH avec l'ensemble des données, et le caractère saisonnier disparaissait. Comme il a été dit tout au début, l'EPH est un modèle d'information minimale, et incorporer une information du type "données saisonnières" n'est pas simple : il faudrait des hypothèses additionnelles, difficiles à justifier.

## **IX. Références**

[Belbèze & al] Defining urban soil geochemical backgrounds: A review for application to the French context. Stéphane Belbèze, Jérémy Rohmer, Philippe Négrel, Dominique Guyonnet  
Journal of Geochemical Exploration 254 (2023) 107298

[Belbèze] ISLANDR\_1.3 Interpolation FR\_v2 : Overview of soil contamination in Europe, Interpolation algorithm, document de travail

[PIT] Probabilistic Information Transfer, by Olga Zeydina and Bernard Beauzamy  
ISBN 978-2-9521458-6-2, ISSN 1767-1175, relié, 208 pages. Avril 2013. Editions de la SCM.

## Table des matières

Résumé Opérationnel.....	2
I. Introduction.....	3
II. Présentation du besoin .....	4
III. Représentation spatiale.....	5
1. Qualité.....	6
2. Défaut.....	6
A. Présentation de l'EPH.....	6
B. Outils conceptuels pour la construction de l'EPH.....	8
1. Notations .....	8
2. Propagation de l'information .....	9
3. Cas de plusieurs mesures .....	9
C. Présentation de l'EEPH .....	11
IV. Exemple de mise en œuvre de la méthode .....	15
A. Cartographie probabiliste de la pollution dans un port.....	15
B. Estimation de la couche de 0 à 20 cm.....	16
C. Estimation de la couche de 20 à 30 cm.....	18
D. Plan d'expérience pour la couche de 20 à 30 cm.....	21
1. Détermination de zones à risque .....	21
2. Simulation d'un plan d'expérience.....	22
V. Utilisation de l'EPH dans des milieux non homogènes .....	24
A. Milieu non-homogène en deux dimensions.....	24
B. Milieu non-homogène en trois dimensions .....	26
VI. Evolution temporelle.....	28
A. Absence d'information.....	28
B. La pollution décroît naturellement, mais ne se propage pas .....	30
C. La pollution décroît et se propage.....	30
VII. Exemples de mise en œuvre de la méthode EEPH .....	31
VIII. Limites de l'EPH .....	35
IX. Références .....	36