



Analyse statistique d'une base de données

"Tritium dans l'eau de pluie"

Rapport adressé à

l'IRSN

par la

Société de Calcul Mathématique SA

en application de votre commande DG00/12010287 du 11.12.2012

rédaction Bernard Beauzamy et Olga Zeydina

## Résumé Opérationnel

A la demande de l'IRSN, nous avons procédé à l'analyse critique de 992 données, concernant la présence de tritium dans l'eau de pluie, recueillies entre le 28/12/2008 et le 22/10/2012, sur les sites de FONTENAY-AUX-ROSES, GINASSERVIS, GRENOBLE, SACLAY, SAINT-AUBIN, SAINT-PAUL-LEZ-DURANCE, soit six emplacements au total.

La question posée tient à la présentation de ces données (qui sont publiques), du fait de "censures" : pour 765 de ces données, ce n'est pas la valeur exacte qui est publiée, mais une valeur "inférieure à un seuil". Une telle présentation s'appelle "censure" en statistiques ; la situation la plus fréquente est celle de patients subissant un traitement médical et ayant quitté ce traitement avant la fin : on ne connaît leur comportement que jusqu'à une certaine date. C'est aussi le cas lorsque des essais, portant sur la rupture d'une pièce, sont interrompus avant que la pièce ne rompe : on sait alors simplement que "durée de vie de la pièce > durée de l'essai".

La première question posée par l'IRSN concerne la nature de ces "censures" : sont-elles raisonnables, sont-elles homogènes ?

Une seconde question porte sur les données non censurées ; elles sont assorties d'une incertitude. Ces incertitudes sont-elles convenablement présentées, sont-elles utiles au public ?

A ces deux questions, nous apportons une réponse simple et claire :

D'un point de vue méthodologique, cette base de données est de très mauvaise qualité ; elle ne peut que susciter la méfiance, voire l'exaspération, même auprès du lecteur le plus indulgent.

Nous commençons par résumer notre analyse et nous faisons ensuite quelques suggestions.

### *A. Résumé de l'analyse*

#### **1. Les censures**

Les censures sont réalisées de manière totalement arbitraire, et diffèrent considérablement d'un site à l'autre, et même, pour le même site, d'une date à une autre. Par exemple, pour Saint Paul lez Durance, elles vont de 1.75 à 4.21 Becquerel par litre (l'unité sera constamment celle-ci dans tout ce rapport, et elle sera désormais omise).

Sur le site de Grenoble, les mesures ont été faites successivement par deux organismes différents :

- du 04/01/2009 au 23/11/2009, par un organisme appelé GRE/EAU dépendant du CEA ;
- du 04/01/2010 au 22/10/2012, par une "Station extérieure de surveillance de l'environnement".

Bien entendu, les deux organismes ne trouvent pas la même chose :

- La proportion de fois où SE annonce une valeur supérieure à 3 (censurée ou non) est 61%, et 88% pour GRE.
- La proportion de données censurées pour SE est 34 % et 53% pour GRE.

On peut donc conclure clairement que les censures sont faites de manière totalement incohérente.

## 2. Les incertitudes

Les données non censurées sont assorties d'une incertitude. Intéressons-nous à l'incertitude relative, définie comme incertitude / valeur de la mesure. Cette incertitude relative s'étale sur une très large plage : de 10% à 98%, ce qui est déjà surprenant.

Si on se limite à une plage de mesure plus réduite, par exemple de 4 à 5, les marges d'incertitude relative vont encore de 40% à 96%.

Mais le comble est atteint avec les incohérences de ces incertitudes. On trouve 4075 incohérences, c'est-à-dire des situations où la mesure satisfait  $m_1 < m_2$  tandis que l'incertitude est dans l'autre sens :  $i_1 > i_2$ . Parmi ces 4075 incohérences, 1547 correspondent à des relevés faits par le même organisme au même endroit : 1475 par "ill" à Grenoble et 122 par le CEA.

### *B. Nos suggestions*

#### 1. Métrologie du tritium

Le tritium est un isotope de l'hydrogène qui existe à l'état naturel, produit par les rayons cosmiques. Il est également produit par l'industrie nucléaire, et le réseau de surveillance est là pour rassurer la population quant aux quantités de produits radioactifs rejetés dans l'atmosphère.

Nous ne connaissons pas le fonctionnement des appareils qui sont employés, mais, très vraisemblablement, il s'agit d'appareils assez grossiers. L'objectif n'est pas ici de parve-

nir à des mesures fines et précises, ce qui requerrait du temps et des moyens considérables.

A partir des années 2005, nous avons aidé le laboratoire de l'IRSN "DSMR/SATE" (M. Pierre Funk) à améliorer les "déclarations d'écart", en ce qui concerne l'uranium et le plutonium, et cela a nécessité la constitution d'un plan d'expériences et l'acquisition par ce laboratoire de masses certifiées. Le travail a pris plusieurs années, et M. Funk nous a dit récemment que l'IRSN était maintenant leader mondial dans ce domaine. Nous ne pensons pas qu'un travail similaire soit envisageable ici.

La question est donc simplement de la présentation mathématique qui est faite des données, et non de leur amélioration.

## **2. Suggestions pour les données**

Nous considérons qu'il est plus simple, plus commode et plus légitime de publier les données brutes, quelles que soient les valeurs de celles-ci. Il est très probable qu'elles seront soumises à une forte variabilité, tout comme le sont, en un point donné, la température et la pression. L'opinion publique s'habitue ainsi à cette variabilité, plutôt que de réclamer "un seuil en dessous duquel il ne se passe rien".

La variabilité dépend certainement des conditions de l'expérience (que nous ne connaissons pas) et aussi de la capacité des instruments à différencier le tritium d'autres matières radioactives.

On peut aussi s'interroger sur le sens qu'il faut donner au résultat : le tritium a une période de plus de 12 ans ; autrement dit, celui que l'on détecte peut avoir été émis plusieurs années auparavant.

Lorsque des données sont censurées, même si elles le sont correctement (ce qui n'est pas le cas ici), il devient impossible de faire des statistiques, par exemple sur leur évolution. Si on fait la moyenne de toutes les données du tableau, en prenant comme valeur la valeur seuil lorsqu'il existe, on trouve 3.58, par contre si on prend comme estimation la moitié en cas de censure (ce qui est tout aussi légitime), on ne trouve plus que 2.53.

## **3. Suggestions pour les incertitudes**

Les incertitudes qui sont présentées ici sont totalement incohérentes. Elles reposent probablement sur des hypothèses "lois d'erreurs gaussiennes", mais mal appliquées, ou appliquées différemment selon les endroits.

En vérité, il n'y a aucune raison que les lois d'erreur soient gaussiennes, et il est très probable que personne ne connaît les véritables lois d'erreur de ces appareils. Les déterminer requerrait le même travail que celui que nous avons fait pour le laboratoire de M.

Funk : comparer l'indication des appareils avec une concentration en tritium connue par d'autres moyens (mettre une masse certifiée de tritium dans un volume d'eau connu).

La connaissance vraie des lois d'erreur est donc, très vraisemblablement, une tâche ir-réaliste. Nous suggérons donc de donner simplement une évaluation de l'incertitude, faite à dire d'expert, et englobant toutes les incertitudes sur tous les appareils : par exemple 20% dans telle plage de valeurs, 40% dans telle autre plage de valeurs (habituellement, la précision d'un instrument n'est pas la même tout au long de la plage de mesure).

Dire "l'incertitude est de 2.24" (première ligne du tableau remis par l'IRSN), pour une valeur de mesure de 3.03, est une mystification scientifique inacceptable : personne ne sait comment elle est calculée ; donner deux chiffres après la virgule est ridicule, quand on voit que des valeurs différentes et contradictoires sont données ailleurs. Il vaudrait mieux dire : "lorsque la mesure est voisine de 3, l'incertitude est de l'ordre de 60%".

## Sommaire

Résumé Opérationnel .....	2
A. Résumé de l'analyse .....	2
1. Les censures .....	2
2. Les incertitudes .....	3
B. Nos suggestions .....	3
1. Métrologie du tritium .....	3
2. Suggestions pour les données.....	4
3. Suggestions pour les incertitudes .....	4
I. Introduction .....	7
II. Les censures .....	7
A. Histogramme général.....	7
B. Histogrammes par site .....	8
C. Evolution au cours du temps.....	11
D. Le cas du site de Grenoble .....	11
III. Etude des données non censurées.....	12
A. Histogramme général.....	12
B. Répartition selon les sites .....	13
IV. Analyse des incertitudes .....	14
A. Histogramme général.....	14
B. Plage réduite .....	15
V. Détection des incohérences dans les incertitudes.....	15
VI. Fichier Excel remis.....	16

## I. Introduction

Le présent document est adressé à l'IRSN en réponse à la commande DG00/12010287 du 11.12.2012.

## II. Les censures

Les censures paraissent réalisées de manière totalement arbitraire. Elles diffèrent considérablement d'un site à l'autre. Les valeurs de seuil s'étendent entre 1,75 et 5,8.

### A. Histogramme général

Voici l'histogramme de leur répartition :

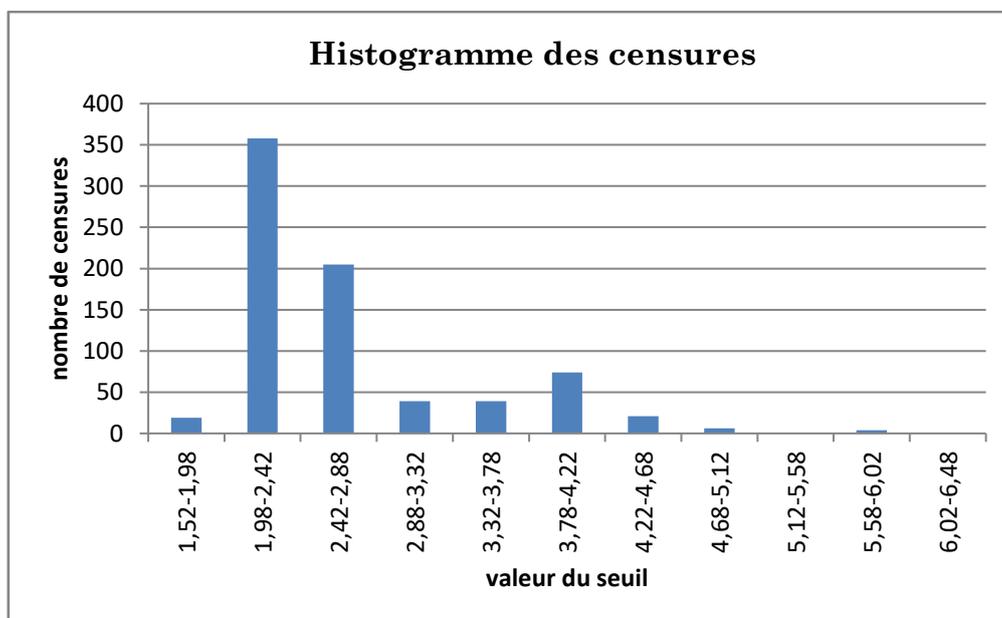


Figure 1 : Histogramme des censures en général

Il n'y a pas d'homogénéité dans la décision de tronquer à un certain seuil, ce qui ne peut qu'exciter la suspicion : ne décide-t-on pas de tronquer "quand cela arrange quelqu'un" ?

Les choses ne sont pas plus satisfaisantes si on se limite à un site en particulier.

## B. Histogrammes par site

Pour Fontenay-aux-Roses, les seuils varient entre 2,5 et 3,16. Voici l'histogramme :

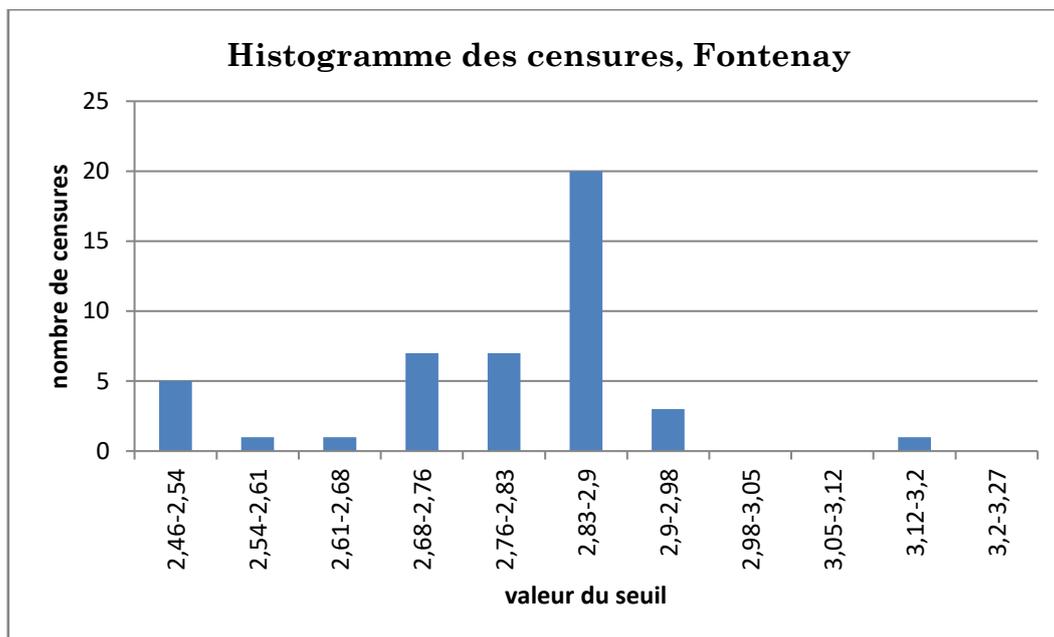


Figure 2 : Histogramme des censures, Fontenay

Pour Ginasservis, elles vont de 1,79 à 2,72 avec l'histogramme suivant :

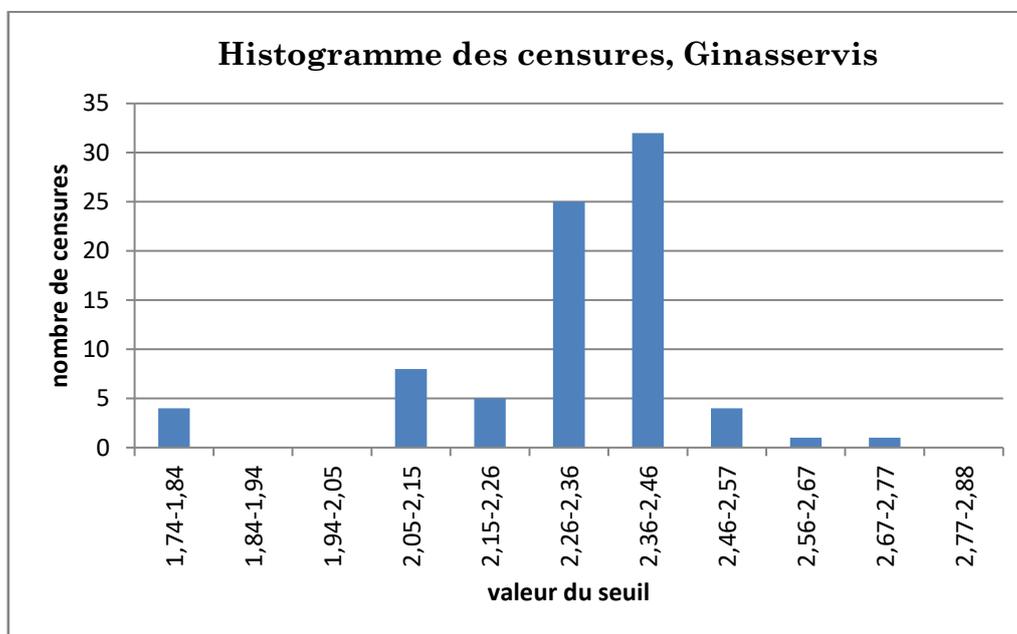


Figure 3 : Histogramme des censures, Ginasservis

Pour Grenoble, elles vont de 2.38 à 3.59, selon l'histogramme suivant :

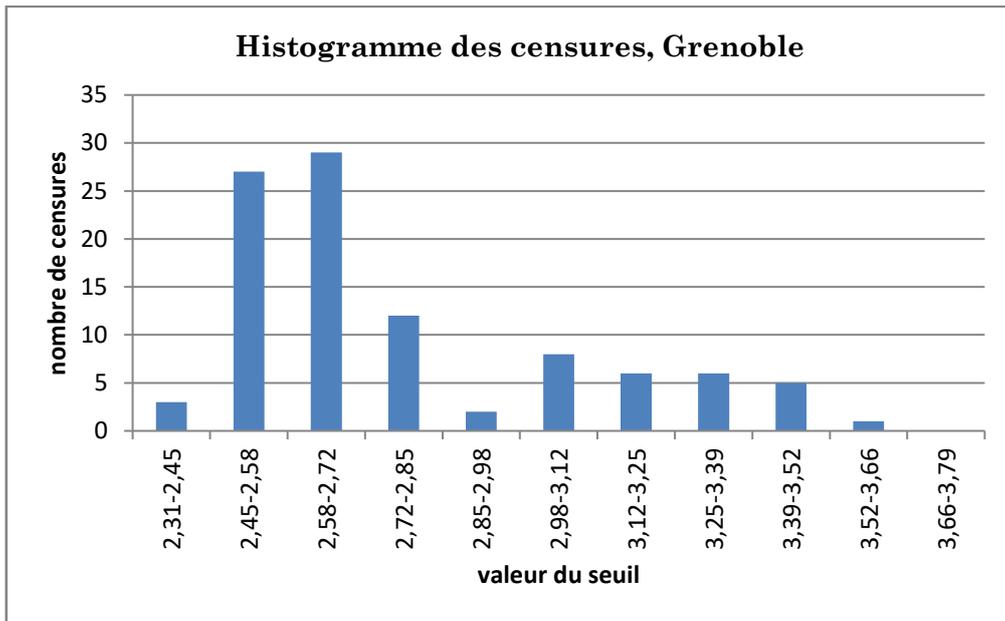


Figure 4 : Histogramme des censures, Grenoble

Pour Saclay, de 3.3 à 5.8, avec l'historgramme suivant :

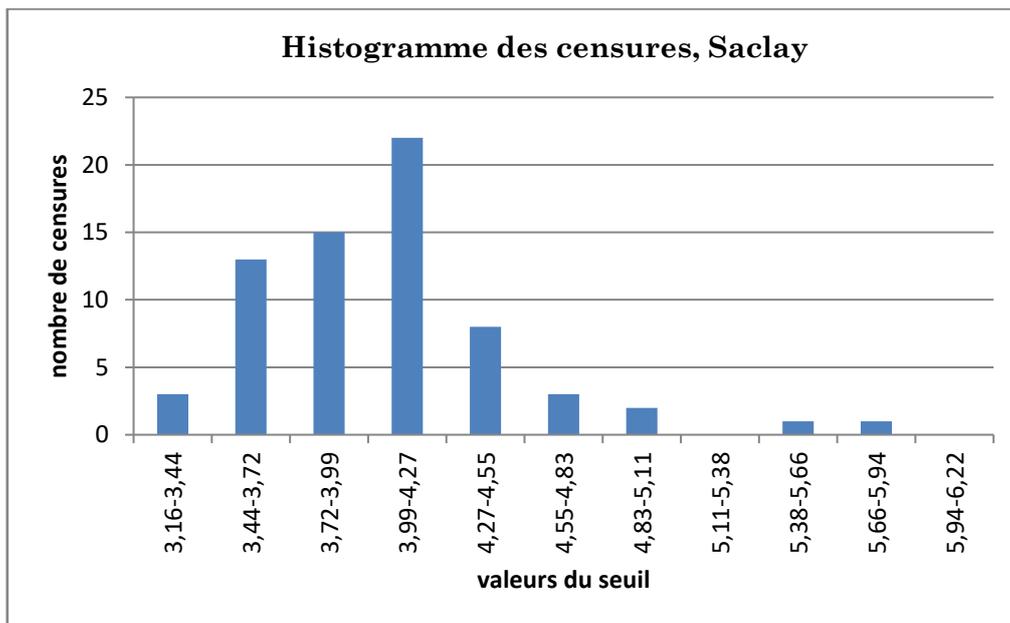


Figure 5 : Histogramme des censures, Saclay

Pour Saint-Aubin, de 3 à 5.8, avec l'histogramme :

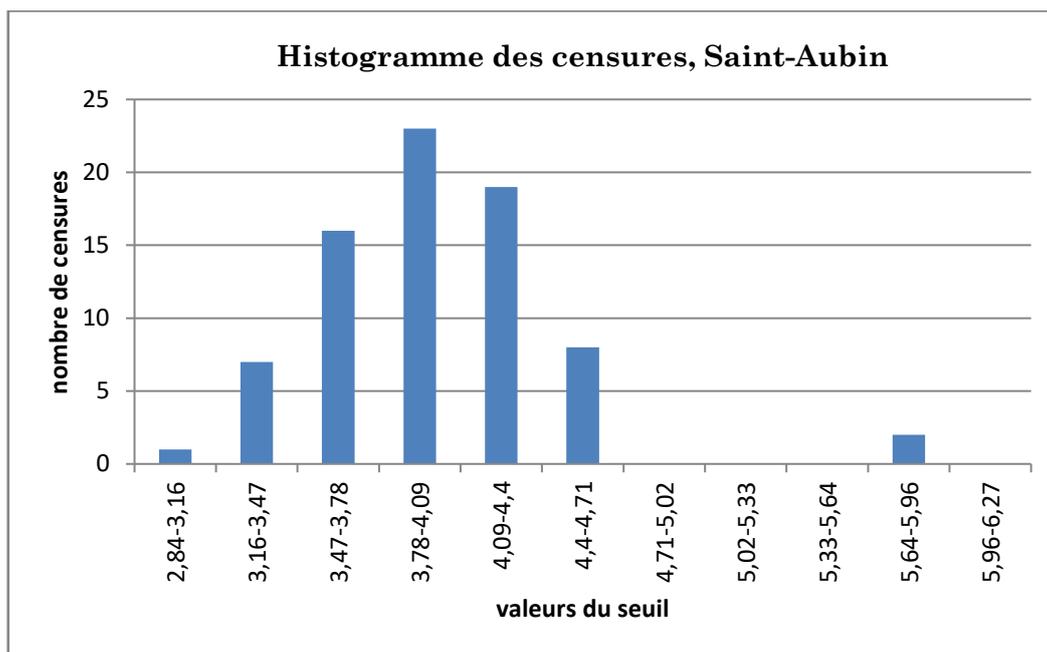


Figure 6 : Histogramme des censures, Saint-Aubin

Pour Saint Paul lez Durance, elles vont de 1.75 à 4.21, selon l'histogramme :

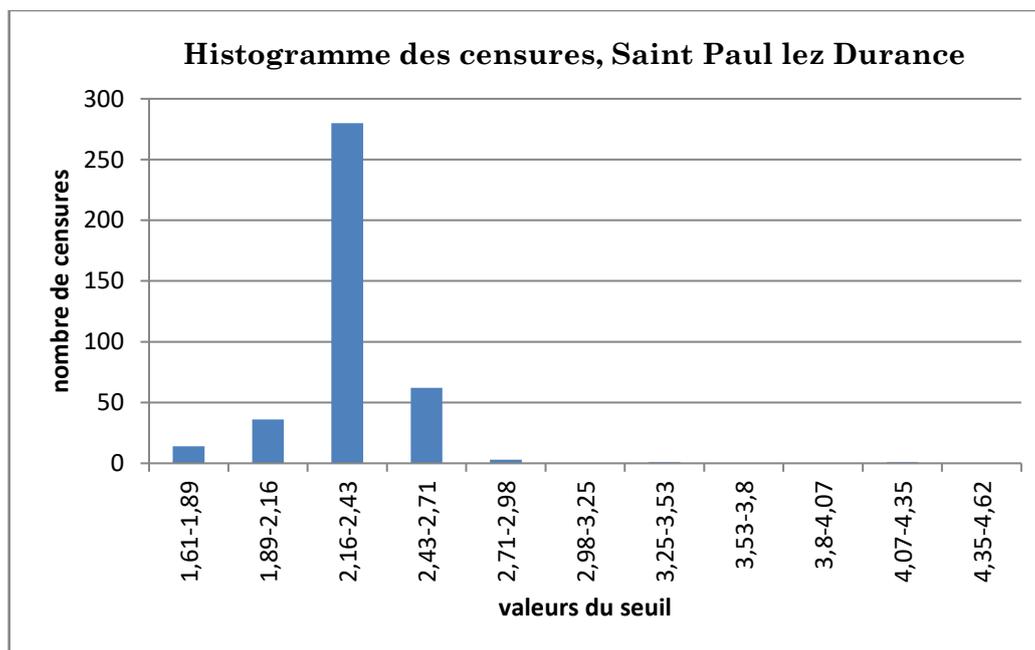


Figure 7 : Histogramme des censures, Saint Paul lez Durance

On peut donc dire que pour tous les sites sans exception, il y a une forte variabilité du choix du seuil de censure.

### C. Evolution au cours du temps

Si on regarde l'évolution des valeurs de seuil, en un lieu donné, on constate qu'elles ne sont pas non plus constantes. Voici le cas de Saint Paul lez Durance :

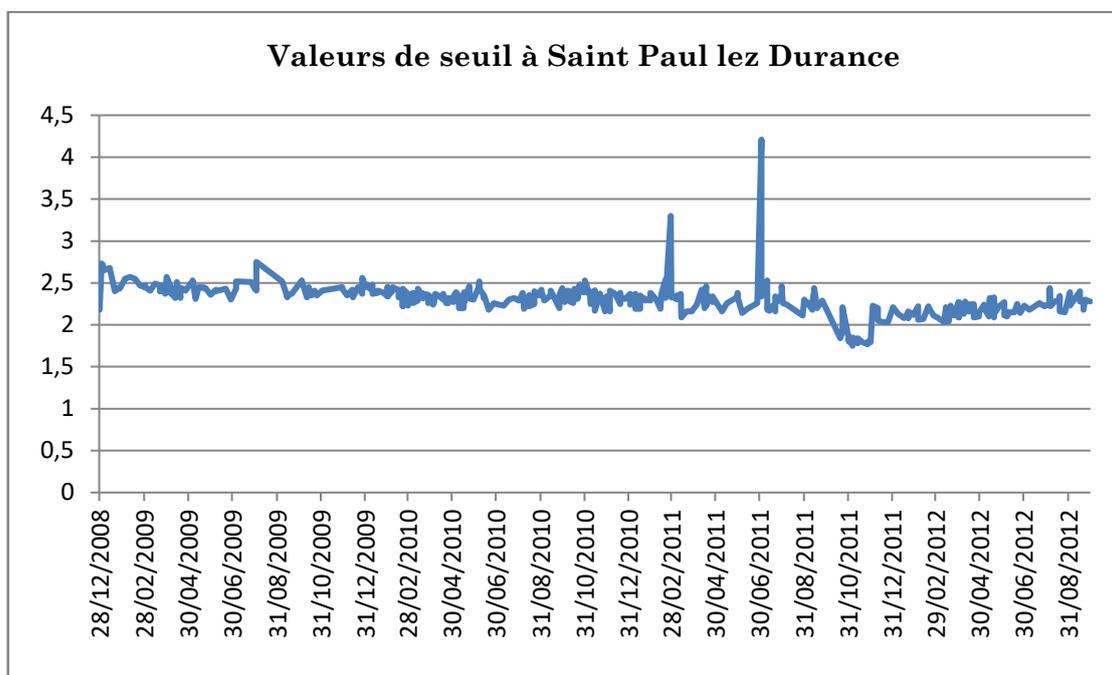


Figure 8 : Valeurs de seuil à Saint Paul lez Durance

### D. Le cas du site de Grenoble

Il est intéressant, parce que les mesures ont été successivement faites par deux organismes différents :

- du 04/01/2009 au 23/11/2009, par un organisme appelé GRE/EAU dépendant du CEA (en fait, il y a deux versions : GRE / EAU / Z113 et GRE / EAU / CHORIE) ; nous le noterons simplement GRE.
- du 04/01/2010 au 22/10/2012, par un organisme appelé "Station extérieure de surveillance de l'environnement", en abrégé SE.

Les deux organismes ne trouvent pas la même chose. Voici la comparaison :

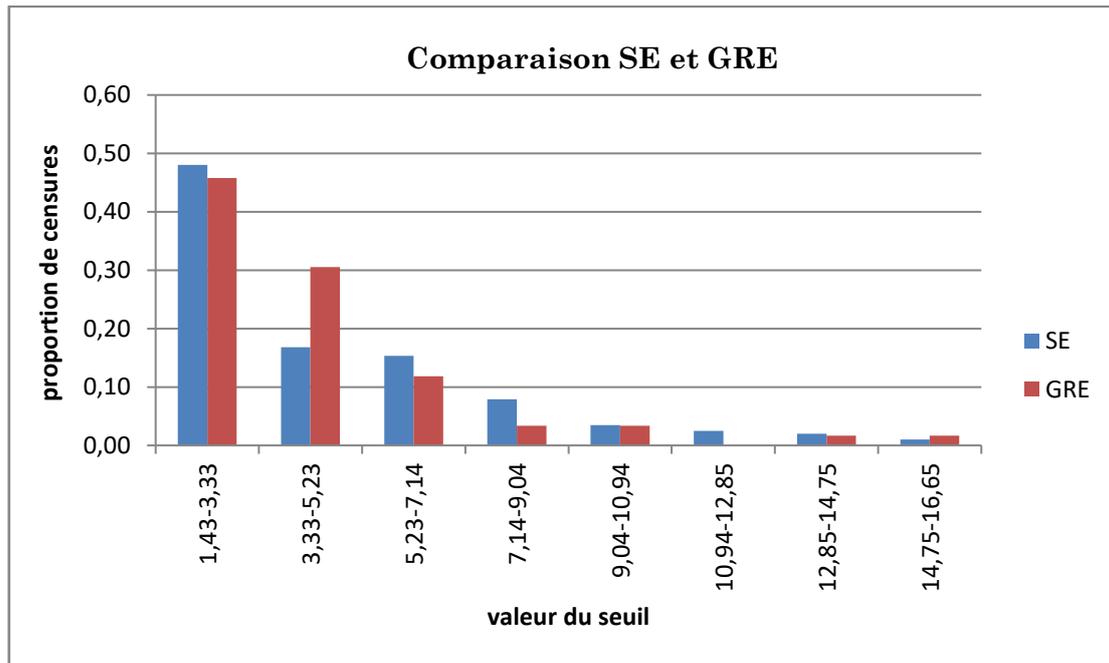


Figure 9 : Comparaison entre SE et GRE

La proportion de fois où SE annonce une valeur supérieure à 3 (censurée ou non) est 61%, et 88% pour GRE.

La proportion de données censurées pour SE est 34 % et 53% pour GRE.

### III. Étude des données non censurées

Il y a au total 227 données non censurées et chacune est assortie de son incertitude. Ces données s'étalent sur une plage très importante : de 2.7 à 82.

#### A. Histogramme général

Voici l'histogramme général des mesures non censurées :

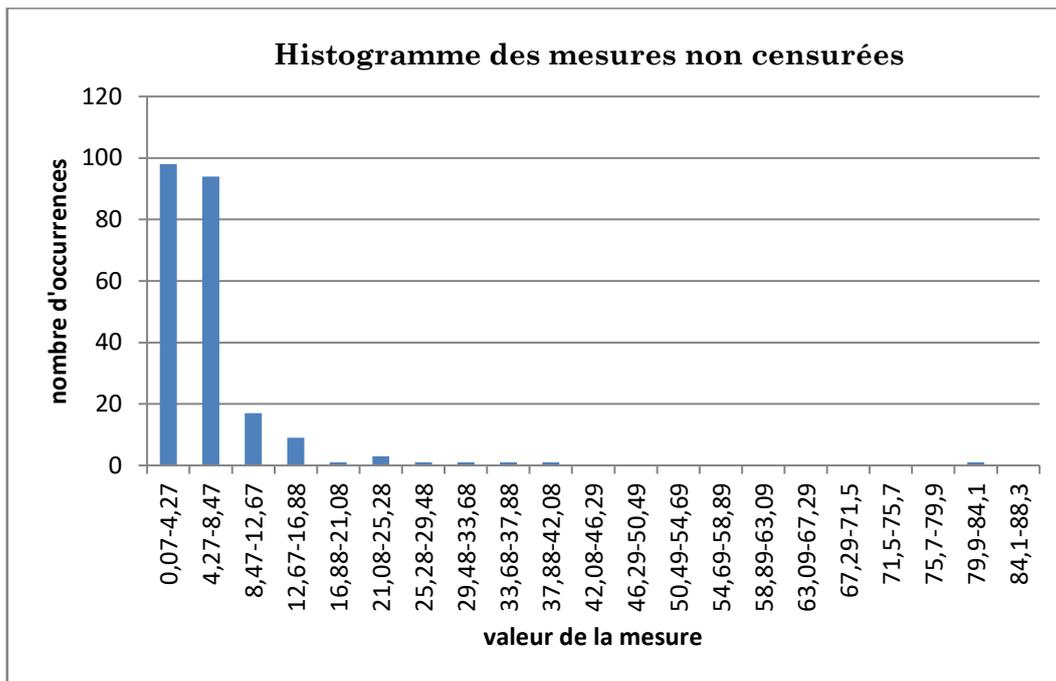


Figure 10 : Histogramme des mesures non censurées

### B. Répartition selon les sites

Voici le tableau de répartition (en pourcentages) des valeurs non censurées selon les sites (la somme de chaque colonne fait 1) :

Intervalles	Fontenay	Ginasservis	Grenoble	Saclay	Saint Aubin	Durance
0,07-4,27	0	1	0,36	0,22	0,2	0,97
4,27-8,47	1	0	0,44	0,61	0,8	0,03
8,47-12,67	0	0	0,10	0,06	0	0
12,67-16,88	0	0	0,06	0,00	0	0
16,88-21,08	0	0	0,01	0,00	0	0
21,08-25,28	0	0	0,02	0,00	0	0
25,28-29,48	0	0	0,01	0,00	0	0
29,48-33,68	0	0	0,01	0,00	0	0
33,68-37,88	0	0	0,00	0,06	0	0
37,88-42,08	0	0	0,01	0	0	0
42,08-46,29	0	0	0	0	0	0
46,29-50,49	0	0	0	0	0	0
50,49-54,69	0	0	0	0	0	0
54,69-58,89	0	0	0	0	0	0
58,89-63,09	0	0	0	0	0	0
63,09-67,29	0	0	0	0	0	0
67,29-71,5	0	0	0	0	0	0
71,5-75,7	0	0	0	0	0	0
75,7-79,9	0	0	0	0	0	0

79,9-84,1	0	0	0	0,06	0	0
84,1-88,3	0	0	0	0	0	0

Tableau 1 : Répartition des valeurs non censurées

Les histogrammes sont très différents selon les sites, mais il n'y a rien de surprenant à cela : la concentration en tritium peut varier d'un endroit à l'autre.

#### IV. Analyse des incertitudes

Toutes les données non censurées sont assorties d'une incertitude, comme nous l'avons dit. Il est normal que les incertitudes soient différentes si les données sont différentes ; aussi, pour procéder à des comparaisons, nous allons travailler sur les incertitudes relatives, définies par :

$$\text{incertitude relative} = \frac{\text{incertitude}}{\text{donnée}}$$

Les incertitudes relatives sont insérées dans l'onglet 4 appelé « non censurées », colonne 4, du fichier Excel joint.

##### A. Histogramme général

Voici l'historgramme des incertitudes relatives :

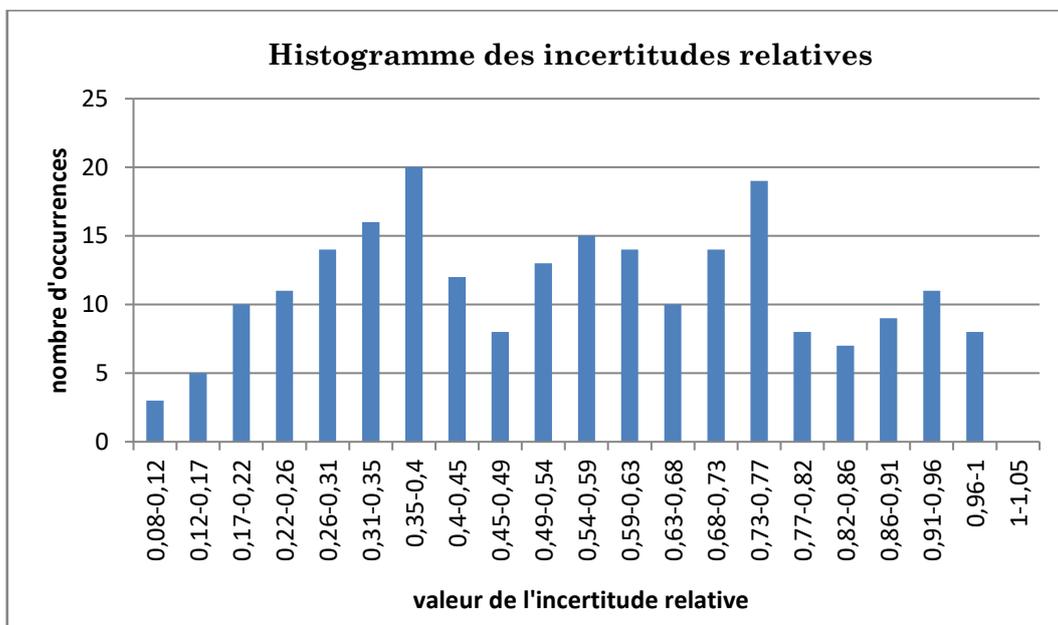


Figure 21 : Histogramme des incertitudes relatives

Ces incertitudes relatives s'étendent sur une large plage : de 10% à 98%, ce qui est déjà surprenant en soi.

## B. Plage réduite

Nous nous intéressons maintenant à une situation plus restrictive, afin de comparer les erreurs relatives : celle où la mesure donne une valeur entre 4 et 5, plage qui est évidemment plus réduite. Même à l'intérieur de cette plage réduite, les erreurs relatives sont sujettes à une grande dispersion ; voici l'histogramme :

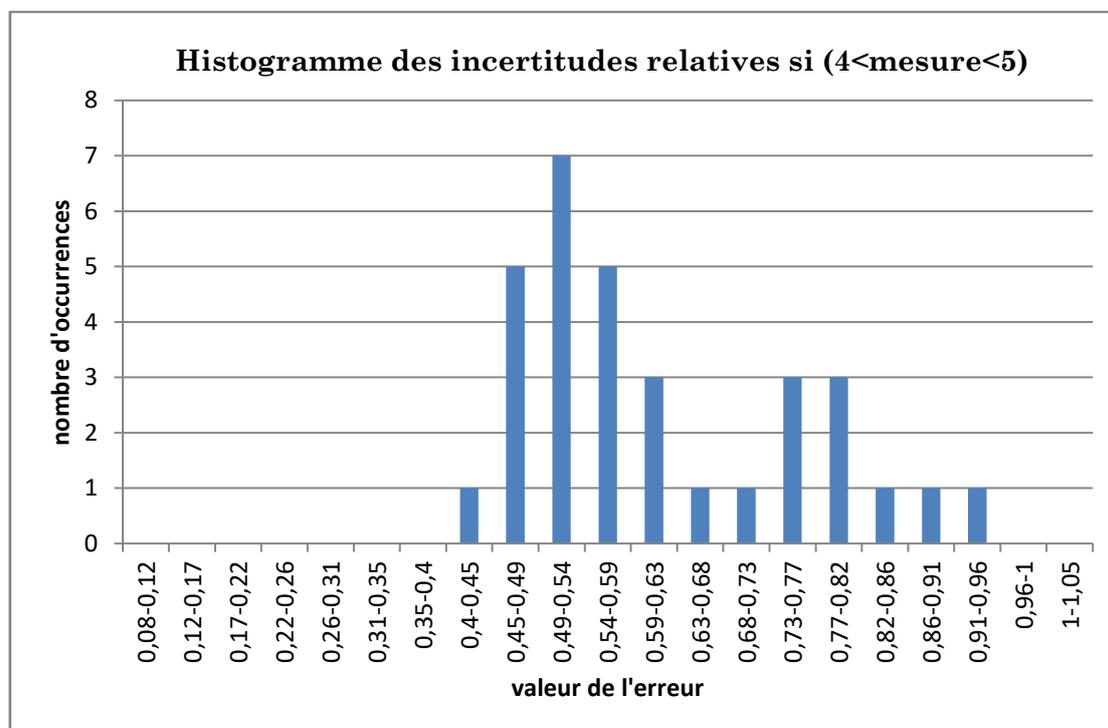


Figure 32 : Histogramme des incertitudes réduites

## V. Détection des incohérences dans les incertitudes

Il y a de très nombreuses situations où les déclarations d'incertitudes sont incohérentes avec les mesures : situations où les deux varient en sens opposé, c'est-à-dire :

$$mesure\_1 > mesure\_2, \text{ mais } incertitude\_1 < incertitude\_2$$

Nous en avons relevé un total de 4075 ; voir onglet 5 du fichier Excel joint appelé « incohérences » pour la liste complète.

Voici les dix premières :

no ligne 1	mesure 1	incertitude 1	no ligne 2	mesure 2	incertitude 2
1	3,03	2,24	46	3,1	2,11
1	3,03	2,24	75	6,19	2,17
1	3,03	2,24	88	4,2	2,22
1	3,03	2,24	179	3,17	2,18

1	3,03	2,24	264	3,51	2,21
1	3,03	2,24	270	5,78	2,2
1	3,03	2,24	278	3,24	2,2
1	3,03	2,24	334	3,37	2,12
1	3,03	2,24	335	3,44	2,13
1	3,03	2,24	338	4,47	2,06

*Tableau 2 : Liste des incohérences*

Sur la première ligne de ce tableau, on lit que si l'on compare les lignes 1 et 46 du tableau de données d'origine, on trouve pour la première ligne une valeur de mesure de 3.03 et une valeur d'incertitude de 2.24, tandis que pour la 46<sup>ème</sup> ligne du tableau, la mesure est supérieure (3.1), tandis que l'incertitude est inférieure (2.11). Les deux ont été relevés par le même organisme, au même endroit.

Sur ces 4075 incohérences, 1547 concernent des mesures faites par le même organisme au même endroit, plus précisément l'organisme appelé "ill", 1475 fois (Grenoble) et le CEA 122 fois. Voir les détails sur le second tableau de l'onglet 5.

## **VI. Fichier Excel remis**

Un fichier Excel est remis en complément du présent rapport ; on y retrouvera tous les tableaux. Les traitements sont faits en VBA sous Excel. Chaque "macro" se trouve associé à un onglet du classeur.

Le premier onglet du fichier remis est identique aux données fournies par l'IRSN, à ceci près : nous avons rajouté une colonne "censure" (en colonne D) et au lieu de <2.61 en colonne B, nous laissons 2.61 en colonne B et mettons le signe "<" en colonne 4. De cette façon, la colonne 2 est entièrement numérique et la colonne 4 désigne clairement les situations de censure.

## VII. Courrier reçu de l'IRSN

Nous répondons par la présente au courrier reçu de l'IRSN (M. Manificat), en date du 17 janvier 2013, faisant suite à la réception de notre rapport du 15 janvier 2013 :

"la censure en mesure radiologique est faite par un test d'hypothèse comparant une mesure à une autre. On compare la mesure d'un échantillon radioactif à une mesure d'un échantillon similaire non radioactif. En caricaturant, on compare l'échantillon peut être radioactif à un échantillon aussi peu radioactif que possible et l'on regarde s'il y a moyen de voir une différence. Le seuil du test d'hypothèse (et donc le niveau de censure) est le résultat lui même d'une mesure. Celle ci varie en fonction de l'instrument, du temps de mesure etc. Il est donc tout à fait normal d'observer cette variation de la censure.

Nous ne sommes pas du tout dans le cas de la censure utilisée dans le monde médical (dite censure "à droite"-censure arbitraire) mais plutôt d'une censure "à gauche" (censure par rapport à un seuil issu d'une mesure)."

## VIII. Notre réponse

### *A. Présentation de données au public*

Le fait que l'IRSN présente des données traitées au public (en l'occurrence censurées) fait l'obligation à cet organisme de préciser comment elles ont été traitées. Il est donc nécessaire d'accompagner les fichiers d'une note explicative, précisant en quoi consiste la censure et comment elle est calculée. Pour prendre une comparaison, les instituts de sondage ont l'obligation légale d'expliquer comment leurs sondages sont faits.

### *B. Les tests d'hypothèse*

Le courrier de l'IRSN explique que la censure résulte d'un "test d'hypothèse". Le principe même d'un test d'hypothèse, en la présente circonstance, représente une erreur méthodologique majeure, parce que, par définition, on ne connaît pas la loi du phénomène considéré (en l'occurrence la concentration en tritium dans l'eau de pluie).

Le test d'hypothèse fait par l'IRSN ne nous a pas été communiqué (non plus qu'au public), mais il est vraisemblablement du type suivant : nous supposons que  $T > S$  ( $T$  étant la concentration en tritium et  $S$  un seuil), et, suivant la valeur de l'échantillon recueilli, nous acceptons ou nous refusons cette hypothèse. Si nous l'acceptons, la donnée est considérée comme fiable (et elle est communiquée au public), si nous la refusons, la donnée est considérée comme censurée, et c'est le seuil qui est communiqué au public.

Cette démarche n'est absolument pas correcte, parce que la loi de  $T$  est inconnue. L'IRSN fait vraisemblablement l'hypothèse (ici encore nous n'en savons rien) que la loi de  $T$  est une gaussienne.

### *C. Censure à gauche ou à droite ?*

Les fichiers envoyés par l'IRSN sont tous du type  $T < S$  ; cela s'appelle en statistiques "censure à gauche". Il n'y a aucune ambiguïté sur ce point.

### *D. La présentation des censures*

Mais, quelle que soit la méthode utilisée pour déterminer la censure – méthode que nous ne connaissons pas et que nous ne pouvons que deviner – la présentation des censures est incohérente et inacceptable pour le public : nous maintenons la totalité de nos conclusions.

### *E. Les incertitudes*

Les données non censurées sont présentées avec une incertitude : le calcul de ces incertitudes est incohérent et inacceptable pour le public : nous maintenons la totalité de nos conclusions.

### *F. En résumé*

Nous maintenons notre recommandation à l'IRSN de n'effectuer aucun travail statistique sur les données (et certainement pas des travaux inacceptables sur le plan de la méthodologie) et de publier les données brutes.