Société de Calcul Mathématique, S. A.
Algorithmes et Optimisation

Evolution du nombre de cancers en France

Document présenté au CEA

Direction de la Protection et de la Sûreté Nucléaire

(à l'attention de M. Bertrand Mercier)

par la

Société de Calcul Mathématique S. A.

en application du Marché no 40003007131, notifié le 13 octobre 2007

rédaction Bernard Beauzamy, Francis Jolly et Olga Zeydina

janvier 2008

Siège social et bureaux : 111, Faubourg Saint Honoré, 75008 Paris. Tel 01 42 89 10 89. Fax 01 42 89 10 69. www.scmsa.eu
Société Anonyme au capital de 56 200 Euros. RCS : Paris B 399 991 041. SIRET : 399 991 041 00035. APE : 7219Z

Résumé opérationnel

Nous montrons dans ce rapport comment appliquer des méthodes probabilistes robustes à des situations simples, naturelles en épidémiologie. A partir du nombre de décès pour une maladie donnée (ici les cancers), par région, nous voulons savoir si une région est plus affectée qu'une autre, ou bien si l'évolution se fait ou non dans un sens favorable : de telles questions reviennent fréquemment.

Nous utilisons uniquement l'information de base, sous forme de taux de décès enregistrés (taux pour 100 000 ha), par région et par année. Nous ne prenons pas en compte ici les doses d'éventuelles radiations que les populations auraient pu recevoir ; cette information ne figure pas dans les données dont nous disposons. Nous montrerons comment utiliser une telle information au cours de la phase 3 du présent contrat.

L'utilisation des méthodes robustes que nous présentons ici permet également de mettre en évidence un certain nombre de biais méthodologiques, que l'on rencontre souvent. Une conclusion peut être fausse parce que l'on a oublié de prendre en considération certaines caractéristiques des données. Nous en verrons des exemples plus loin. C'est un aspect particulièrement intéressant dans le cadre du présent contrat, parce que les données, les phénomènes, les conclusions, dont nous parlons sont destinés à être communiqués au public.

1. Les méthodes utilisées

Le présent rapport est divisé en quatre étapes, par ordre de complexité croissante. Pour chacune, nous présentons des résultats-types qu'elle permet d'obtenir et nous faisons une critique méthodologique des hypothèses qui sont faites.

A. Construction d'histogrammes

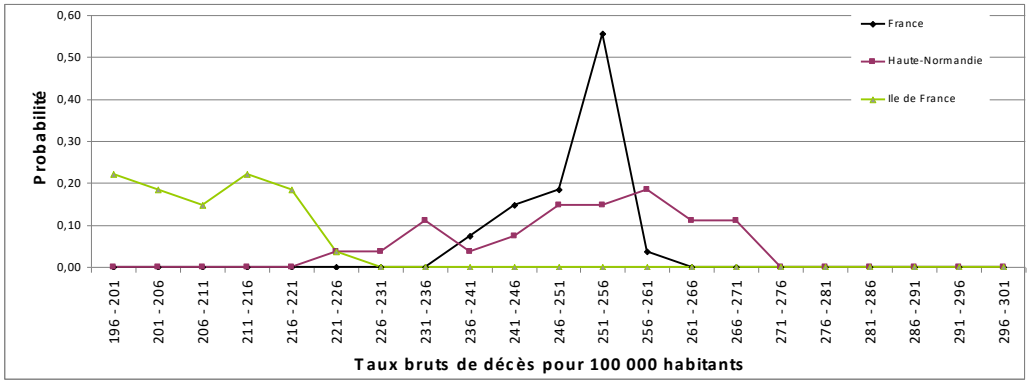
C'est la méthode probabiliste de base, la plus simple. Elle permet d'obtenir les lois de probabilité des taux de décès pour chacune des régions, c'est-à-dire un résultat du type : pour la Haute-Normandie, la probabilité que le taux de décès soit ≥ 256 (par an, pour 100 000 ha) est 0.41. Le sens d'un tel énoncé est : prenons cent mille années, en supposant que rien ne varie dans le cadre de vie ; on peut s'attendre à en avoir 4 100 au cours desquelles, en Haute Normandie, on recensera au moins 256 cancers.

On peut ensuite comparer ces probabilités entre elles, pour les différentes régions. Les résultats sont basés sur l'ensemble des données ; aucune hypothèse artificielle n'est donc faite. Mais dans cette méthode, tout se passe comme si toutes les années étaient équivalentes. Nous l'avons dit plus haut : « en supposant que rien ne varie dans le cadre de vie ». Comme, en pratique, bien des choses varient, cette méthode est discutable sur le plan méthodologique. Comme nous le verrons, c'est avant tout l'augmentation de la durée de vie et le vieillissement de la population qui affectent les résultats.

Voici des exemples de résultats :

Comparaison par régions

La comparaison des régions est faite à partir des lois de probabilité ci-dessous :



Graphique 1 : Lois de probabilité pour deux régions et pour la France

La probabilité d'avoir au moins 256 décès pour 100 000 habitants est de :

- 0 pour l'Ile de France ;
- 0,41 pour la Haute Normandie ;
- 0,04 pour la France.

B. *Analyse des tendances*

Nous analysons les variations des taux de décès au cours du temps. Nous le faisons pour les différentes tranches d'âge, et nous les comparons à la population dans son ensemble. Nous comparons les taux de décès pour la France entière et les taux par région. Nous établissons également des prévisions de taux de décès pour les années à venir.

Cette analyse est fondée sur des méthodes d'ajustement linéaire simples. Elle prend le temps en considération, mais elle présente deux défauts majeurs :

- Elle donne une valeur précise pour la prévision, liée à l'ajustement linéaire, et non pas une probabilité ;

- Elle utilise de la même manière les données récentes et les données anciennes : il n'y a pas de pondération en fonction de l'ancienneté.

Voici des exemples de résultats :

Résultats globaux

Le taux de cancer en France augmente d'une année sur l'autre, pour l'ensemble de la population. Pour 100 000 habitants, la proportion de décès par cancer sera plus élevée en 2010 qu'elle ne l'était en 1990. Conclusion : on meurt davantage de cancer que par le passé.

Ce résultat paraît inquiétant ; en réalité il est uniquement dû à l'accroissement de la durée de vie. En effet, sur la tranche d'âge 0 – 85 ans, ce taux est stable, et il diminue pour toutes les tranches inférieures.

Explication réelle : effectivement, on meurt davantage de cancer que par le passé, mais on vit plus longtemps et auparavant on mourait d'autre chose, plus jeune.

Pronostics

A partir des données obtenues, les droites de régression linéaire permettent la prédiction du nombre de décès par tumeur. La prédiction peut être réalisée en fonction de la région, du sexe et de la tranche d'âge. Ainsi, à partir des données de la région Ile de France, et pour la population de moins de 85 ans, on obtient une prédiction du taux de décès par cancer en 2008 de 160,2 pour 100 000 habitants.

Evolution par tranches d'âge

L'évolution du taux de décès par cancer ne suit pas la même tendance selon la tranche d'âge considérée. Pour la population française, on observe une diminution du taux de décès par cancer pour la plupart des tranches d'âge:

Tranches d'âge	Evolution du taux de décès par tumeur de 1979 à 2005
<1	Diminution
01-04	Diminution
05-14	Diminution
15-24	Diminution
25-34	Diminution
35-44	Diminution
45-54	Diminution
55-64	Diminution
65-74	Diminution
75-84	Diminution
85-94	Augmentation
>94	Augmentation

Tableau 1 : Tendance du taux de décès par tumeur en fonction de la tranche d'âge pour l'ensemble de la population française

Même si le taux de décès par cancer diminue pour toutes les tranches d'âge de 0 à 85 ans, le taux de décès global sur la tranche 0-85 ans reste stable : cela tient au déplacement de la population vers les tranches d'âge élevées. Les effectifs de chaque tranche ne sont pas constants.

La variance en fonction des taux de décès augmente d'année en année et l'indépendance des régions est de plus en plus marquée. Ce qui signifie que les taux de décès par cancers entre les régions sont de plus en plus dispersés d'année en année. Cette remarque va dans le sens d'un récent rapport des Académies, qui dit qu'il n'y a pas de « cause com-

mune » aux cancers (par exemple, pas ou peu de causes de nature environnementale) ; tout se passe comme si le cancer était dû au hasard (ou, plus exactement, à des phénomènes inexpliqués, qui n'agissent pas de manière identique selon les régions).

C. *Comparaisons probabilistes robustes*

Les comparaisons d'une zone à l'autre sont l'objet de polémiques permanentes : encore aujourd'hui, les experts se battent pour savoir s'il y a ou non un excès de cancers au voisinage de La Hague. Cela tient aux tests statistiques factices qui sont utilisés (voir notre rapport 1).

Nous utilisons ici une méthode nouvelle dans le présent contexte ; elle permet d'obtenir des résultats robustes, sous forme probabiliste, de type suivant : sachant qu'une région 1 a tant de décès, et une région 2 tant de décès, quelle est la probabilité que la région 1 soit plus touchée que la région 2 ?

Cette méthode d'évaluation du risque à partir d'une donnée du présent a été mise en œuvre par nous dans le cadre d'un contrat européen. La théorie est développée dans le livre de B. Beauzamy « Méthodes probabilistes pour l'étude des phénomènes réels » [1] ; elle est ancienne. La méthode a, en particulier, été mise en œuvre par la Direction de la Sûreté Nucléaire de Défense, pour évaluer les risques liés aux manipulations des têtes nucléaires et par le CEA/Saclay pour évaluer les risques liés au survol des avions.

Voici un exemple de résultat :

Pour la région Champagne, le nombre de décès par cancer en 2005 était de 3 564 pour 1 338 340 habitants (soit 266 décès pour 100 000 habitants). A la même période, le nombre de décès pour la France était de 155 407 pour 60 634 800 habitants (soit 256 décès pour 100 000 habitants).

La méthode probabiliste développée dans le logiciel EvalRisk permet d'obtenir la conclusion suivante : la probabilité que la région Champagne soit plus affectée que le pays est de 0,9881.

L'inconvénient de cette méthode est qu'elle n'exploite qu'une seule année, et non l'ensemble de l'historique.

D. *Construction de l'EPH (Experimental Probabilistic Hypersurface).*

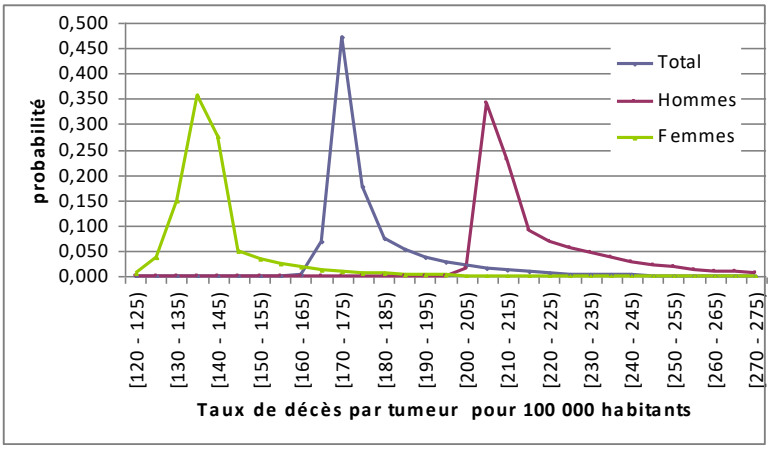
La méthode de l'Hypersurface Probabiliste, introduite par nous dans le cadre d'un contrat avec Framatome (2004), puis développée dans le cadre de contrats avec l'IRSN et le CEA (2006, 2007), permet une prédiction probabiliste à partir de l'ensemble de l'historique : nombre de décès attendus, dans telle région, pour telle tranche d'âge. Elle se distingue des méthodes usuelles (prolongements linéaires) par le fait que le résultat est donné sous forme probabiliste (donc robuste) et par le fait que le passé proche influe plus que le passé lointain.

Elle remédie donc à toutes les critiques que nous avons faites au cours des paragraphes précédents.

L'Hypersurface Probabiliste donne un poids plus important aux données les plus récentes. La prédiction par l'EPH pour 2008 est donc complètement différente des lois de probabilité obtenues à partir d'histogrammes, vues plus haut, qui considèrent les données avec le même poids qu'elles soient anciennes ou récentes.

L'Hypersurface Probabiliste fait partie de la thèse de doctorat de Olga Zeydina (« Méthodes probabilistes pour la Sûreté Nucléaire »), thèse cofinancée par le CEA et l'IRSN. Les bases de la construction de l'EPH peuvent être trouvées dans [2]. L'application de la construction au cas de l'épidémiologie est donnée dans un document séparé annexé au présent rapport. Ici, nous nous contenterons donc de présenter les résultats.

Les lois de probabilité suivantes sont obtenues par application de l'EPH :



Graphique 2 : Prédiction pour 2008 des lois de probabilité obtenue par l'EPH pour la région Ile de France

E. Les biais méthodologiques

Ici, nous travaillons par région, par tranche d'âge, sur des taux bruts de cancer pour 100 000 habitants : c'est ainsi que les données sont présentées et nous n'y pouvons rien. Mais il est intéressant de constater que cette présentation, qui paraît naturelle, amène naturellement à des erreurs !

- Lorsque les taux de décès de plusieurs groupes de population diminuent, il se peut cependant que le taux d'ensemble augmente !

- Les taux de décès ne s'ajoutent pas : il ne suffit pas d'ajouter les taux pour les différentes régions pour avoir le taux pour la France entière.

Le phénomène sous-jacent qui doit être pris en compte est l'allongement de la durée de vie, déplacement des effectifs vers les classe d'âge les plus élevées, comme nous allons le voir.

Table des matières

I. Les données utilisées

Les données utilisées dans ce rapport proviennent du site Internet du CépiDC (Centre d'épidémiologie sur les causes médicales de décès) et de l'INSERM (Institut National de la Santé et de la Recherche Médicale).

Elles concernent le nombre de décès causés par une maladie donnée, calculé pour 100 000 habitants (taux bruts de décès pour 100 000 habitants). La période d'observation est comprise entre 1979 et 2005. L'étude réalisée prend en compte toute cette période, soit 27 années consécutives. Pour ces 27 années, l'information est détaillée par tranche d'âge et par maladie.

Les tranches d'âge sont données par intervalle de 10 ans de 5 à 95 ans. Le tableau ci-dessous indique les 12 tranches d'âge utilisées :

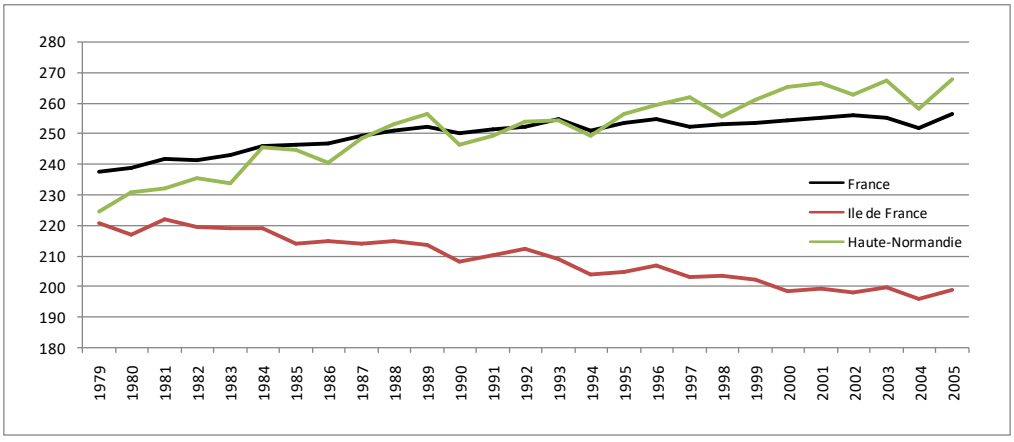
Tranches d'âge											
<1	01-04	05-14	15-24	25-34	35-44	45-54	55-64	65-74	75-84	85-94	95+

Tableau 2 : Tranches d'âge considérées dans l'étude

Le taux de décès ainsi que le nombre de décès sont donnés pour chacune des tranches d'âge ainsi que pour la population totale.

La base de données est importante ; pour présenter la méthodologie employée, nous nous limitons aux tumeurs (de tous types).

Le graphique suivant présente l'évolution du nombre de tumeurs au cours du temps, pour deux régions et pour la France entière.



Graphique 1 : Taux de décès par tumeur en fonction de l'année

L'axe des abscisses représente les années (de 1979 jusqu'à 2005), l'axe des ordonnées représente le nombre de décès par tumeur, pour une année donnée, calculé pour 100 000 habitants pour toutes les classes d'âge confondues.

Le taux augmente d'année en année pour la France et la Haute Normandie. Comme nous le verrons, ceci sera expliqué par l'allongement de la durée de la vie.

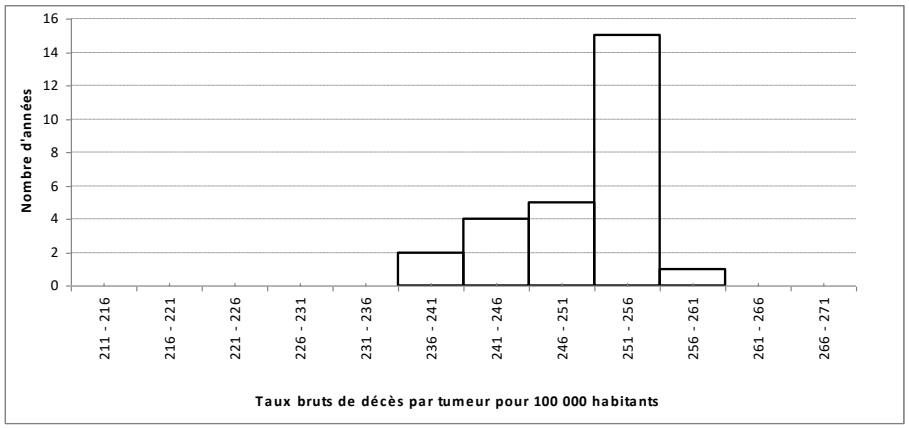
Le graphique montre clairement que la région Ile de France est moins atteinte que la région Haute Normandie et que le pays entier.

II. Histogramme

A. Construction de l'histogramme

Pour construire un histogramme, pour chaque région, on considère le nombre de cancers pour chaque année : on dispose donc d'un échantillon de 27 résultats expérimentaux (un par année, de 1979 à 2005). Les intervalles de variation sont différents d'une région à l'autre : le minimum de décès enregistrés est 196 et 356 est le maximum. Cet intervalle est divisé en petits sous-intervalles de largeur 5, et on compte, parmi les 27 années, le nombre de fois où l'on est tombé dans chaque petit sous-intervalle.

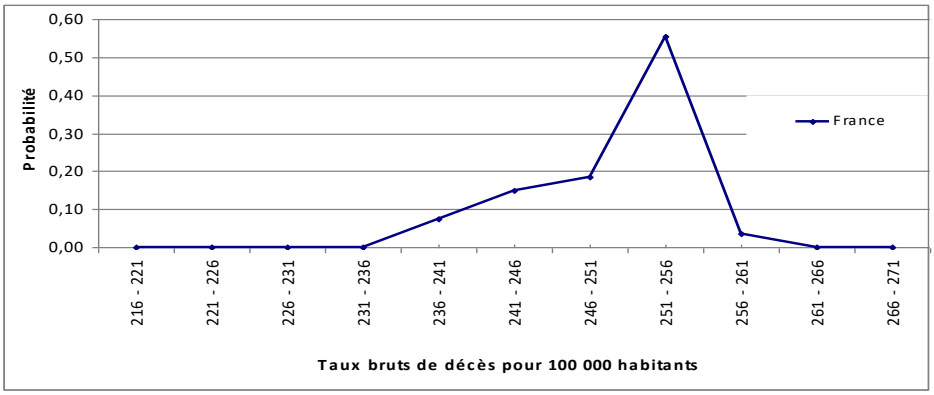
Voici la construction, pour la France entière ; les graphes par région sont différents.



Graphique 2 : Histogramme du taux de décès par tumeur pour la France

Par exemple, un taux de décès compris entre 246 et 251 pour 100 000 habitants a été enregistré 5 fois entre 1979 et 2005 en France.

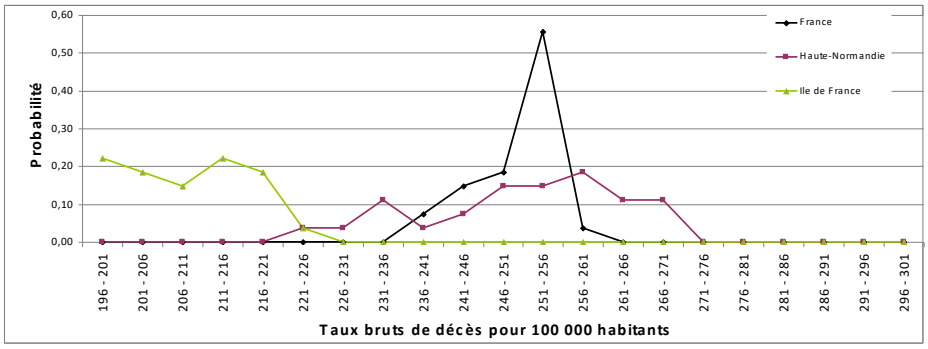
A partir de cet histogramme, il est facile d'obtenir la loi de probabilité pour la région ou pour le pays entier. Pour cela, il suffit de diviser le nombre de fois où le taux de décès est compris dans l'intervalle par le nombre d'années total, soit 27. On obtient le résultat suivant :



Graphique 3 : loi de probabilité du taux de décès par tumeur pour la France

Pour le pays entier, la probabilité que le taux de décès par tumeur pour 100 000 habitants soit compris entre 241 et 246 est de 0,15.

Nous appliquons la même méthode pour les régions Haute Normandie et Ile de France.



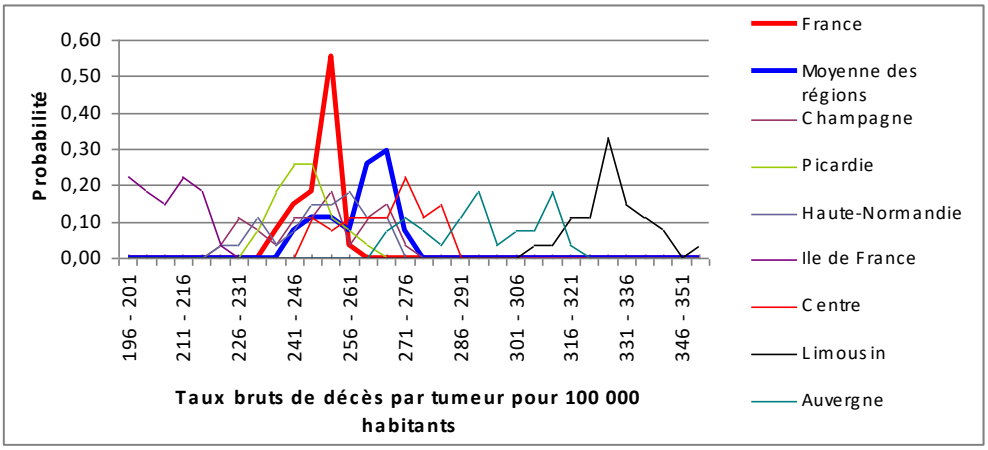
Graphique 4 : Loi de probabilité du taux de décès par tumeur pour deux régions et la France

Il est intéressant de remarquer que ces lois de probabilité ne sont pas gaussiennes, ni même symétriques.

L'histogramme pour toute la France est plus concentré que celui pour chaque région. Cela semble indiquer l'indépendance des régions entre elles. Nous allons détailler ceci.

B. Indépendance des régions entre elles

La loi de probabilité de la moyenne des 22 régions ainsi que de quelques-unes de ces régions françaises sont représentées sur le graphique ci-dessous :



Graphique 5 : Taux de décès par tumeur pour 100 000 habitants

Notons d'abord que la loi pour la France n'est pas la même que pour la moyenne des régions, car toutes n'ont pas la même population : nous travaillons sur des taux pour 100 000 ha. Ici, pour parler d'indépendance, c'est la moyenne qui nous intéresse et non le taux pour la France entière.

La loi de probabilité des moyennes des régions est plus concentrée que les lois des régions prise séparément. Ceci est un indice d'indépendance. Pour l'expliquer, supposons à l'inverse que, pendant 5 ans, un événement se soit produit (exemple : nuage radioactif) qui conduise partout (ou au moins dans de nombreuses régions) à une augmentation du nombre de cancers : la moyenne serait alors élevée. Or elle est faible, ce qui signifie des compensations entre régions : les années où certaines régions ont beaucoup de cancers, les autres en ont peu.

Prenons par exemple l'intervalle $[261-266[$: s'il y avait un nombre d'années important et un nombre de régions important où l'on trouve ce nombre de cancers, le taux total pour la France ne serait pas nul. Or il est très faible : cela indique que, lorsque certaines régions ont des taux de cancers importants pendant plusieurs années, d'autres ont des taux faibles.

C. Histogramme cumulé

La probabilité cumulée est un outil souvent utilisé, car elle permet des visualisations faciles. Il s'agit de la probabilité que le Taux Brut (TB) soit au dessus d'un certain seuil, par exemple :

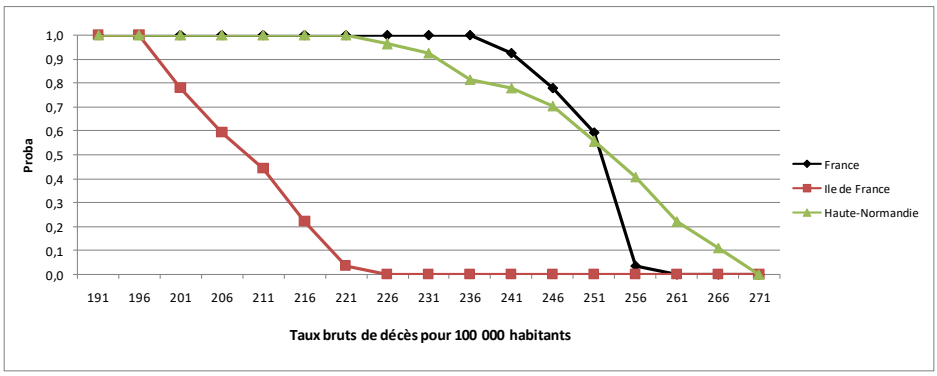
$$\mathrm{P}_{Ile\ de\ France}\ \ TB \geq 256\ \ = 0$$
$$\mathrm{P}_{Haute-Normandie}\ \ TB \geq 256\ \ = 0.41$$
$$\mathrm{P}_{France}\ \ TB \geq 256\ \ = 0.04$$

La probabilité d'atteindre un taux au moins égal à 256 décès pour 100 000 habitants est de :

- 0 pour l'Ile de France ;
- 0,41 pour la Normandie ;
- 0,04 pour la France.

L'interprétation de ces résultats se fait de la manière suivante : pour un million d'années, si rien ne change, on peut s'attendre à en trouver 410 000 pour lesquelles, en Ile de France, il y aura plus de 256 cancers.

Les histogrammes cumulés sont :

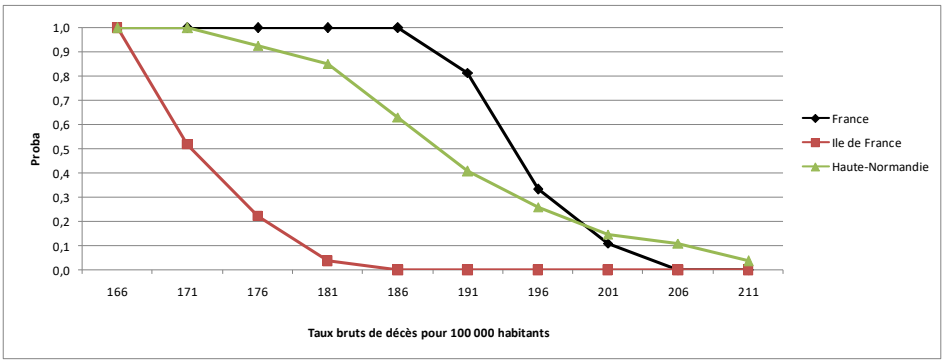


Graphique 6 : Loi de probabilité cumulée du taux de décès par tumeur

Pour la région Ile de France, la probabilité d'avoir un taux brut de décès supérieur à 266 pour 100 000 habitants est de 0.

Pour la région Haute Normandie, la probabilité d'avoir un taux brut de décès supérieur à 266 pour 100 000 habitants est de 0,11.

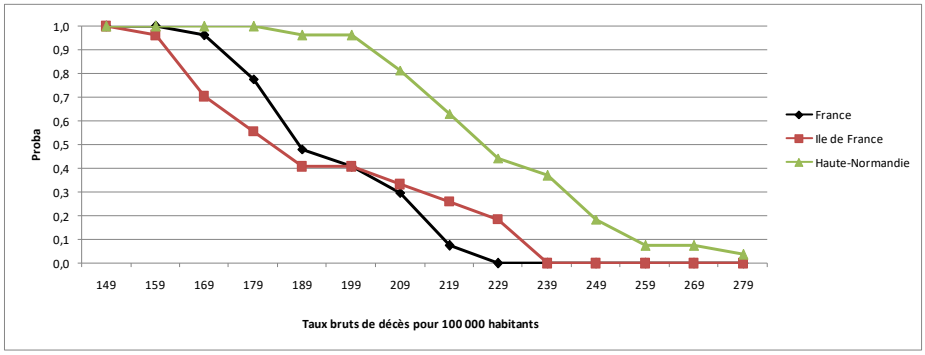
Cette méthode peut être appliquée pour une seule catégorie de la population: si l'on considère uniquement la population féminine, on obtient le graphique suivant :



Graphique 7 : Loi de probabilité cumulée du taux de décès par tumeur pour la population féminine

L'information obtenue ici n'est plus la même que précédemment (pour la totalité de la population). Les deux régions sont moins affectées que la France.

Sur le graphique suivant, nous ne considérons qu'une tranche d'âge, par exemple les personnes appartenant à la tranche d'âge 45-54 ans.



Graphique 8 : Loi de probabilité cumulée du taux de décès par tumeur pour la tranche d'âge 45-54 ans

Les lois de probabilité cumulée rendent la comparaison entre les régions plus évidente.

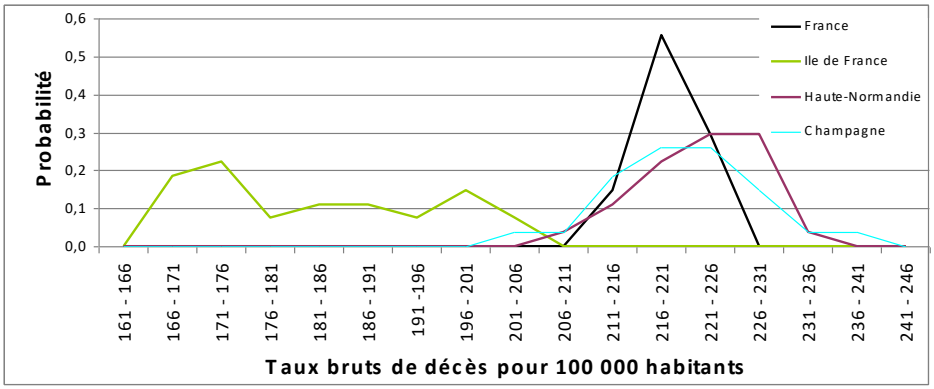
L'histogramme, ou sa variante « probabilité cumulée », permet une réponse simple à la question : quelle est la probabilité que le nombre de cancers dépasse telle valeur ? Mais il fait explicitement l'hypothèse que toutes les années se valent. Or ce n'est pas exact : entre 1979 et 2005, beaucoup de choses ont changé, notamment l'accroissement de la durée de vie.

Notons que l'augmentation du nombre d'habitants est prise en compte puisqu'on considère des taux pour 100 000 ha, et non un nombre total de cancers.

D. Tranches d'âge

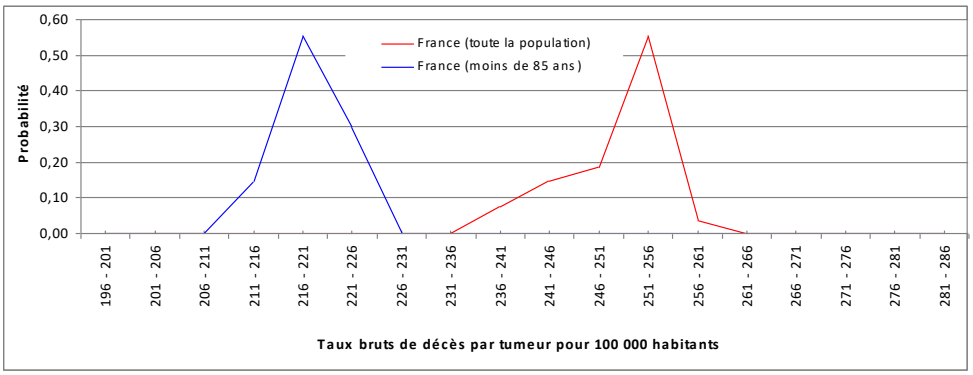
Sur un plan purement méthodologique, l'histogramme pour la population entière n'est pas satisfaisant, du fait du biais lié à l'accroissement de la durée de vie. Pour éliminer ce biais, il faut considérer une tranche d'âge fixe, par exemple 0 – 85 ans

Les histogrammes précédents sont réalisés en prenant en compte toutes les classes d'âge et toutes les années de 1979 à 2005. La durée de vie de la population a changé entre 1979 et 2005, ce qui crée un biais méthodologique, comme nous l'avons déjà dit. Les gens vivent plus vieux ; ils meurent de cancer à un âge avancé, ce qui ne se voyait pas auparavant, car ils mouraient d'autre chose plus jeune. Pour mettre ce fait en évidence, nous nous limitons à la tranche d'âge 0 – 85 ans.



Graphique 9 : Taux de décès par tumeur pour la population de moins de 85 ans

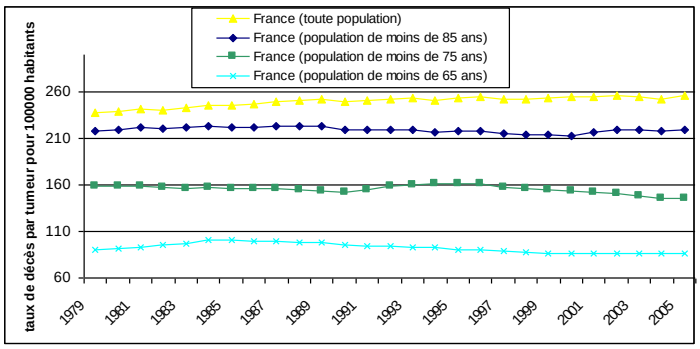
Les lois de probabilité obtenues n'ont plus de biais lié au vieillissement de la population. La population âgée de plus de 85 ans n'étant pas prise en compte, les taux bruts sont donc modifiés et sont inférieurs à ceux obtenus pour l'ensemble de la population. Cette différence est visible sur le graphique ci-dessous :



Graphique 10 : Taux bruts de décès par tumeur pour 100 000 habitants

Le taux de décès par tumeur est plus important pour la population la plus âgée. Cette importance du taux de décès conduit donc à une diminution du taux lorsque l'on ne prend plus en compte les deux tranches d'âge supérieures à 85 ans.

L'évolution du taux de décès n'est pas identique en fonction de la tranche d'âge considérée. Le graphique suivant représente l'évolution du taux de décès par tumeur pour la population française :



Graphique 11 : Evolution du taux de décès par tumeur en fonction de la tranche d'âge pour la population française

Les courbes sur le graphique ci-dessus montrent :

- une augmentation du taux de décès pour l'ensemble de la population française ;
- une stabilité pour la population de moins de 85 ans ;
- une diminution pour les populations de moins de 75 ans et en deçà.

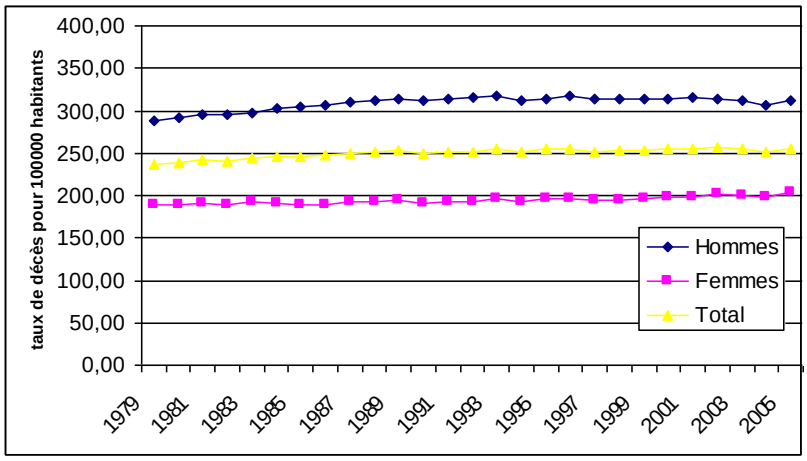
L'augmentation de la durée de vie est la principale cause de la variation des taux de décès en fonction de l'âge. Nous analysons ces variations en fonction de la tranche d'âge dans le paragraphe suivant.

III. Analyse des tendances

A. Observation de la tendance au cours du temps

Si l'on représente le taux de décès par tumeur pour 100 000 habitants, on observe une augmentation au cours du temps ; ceci est dû à l'accroissement de l'espérance de vie, comme expliqué plus haut.

Le graphique suivant représente le taux de décès par tumeur pour 100 000 habitants.



Graphique 12 : Evolution du taux de décès par tumeur pour la population française

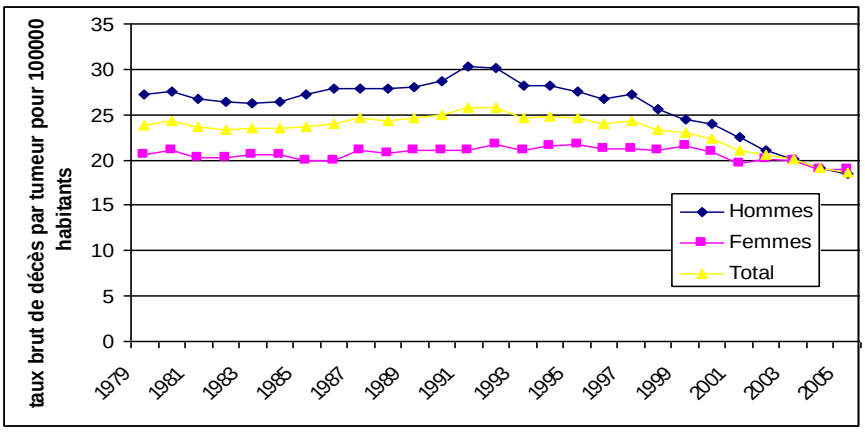
Par contre, si le taux de décès pour 100 000 habitants est représenté par tranche d'âge, le résultat obtenu est différent.

Pour obtenir le taux de décès sur la tranche d'âge de 15 à 45 ans, il ne suffit pas de faire la moyenne des taux de décès des trois tranches d'âge [15 ;25[, [25 ;35[, [35 ;45[car la population n'est pas la même.

Le taux de décès de la tranche d'âge [15 ;45[est obtenu à partir du nombre de décès et du nombre d'habitants :

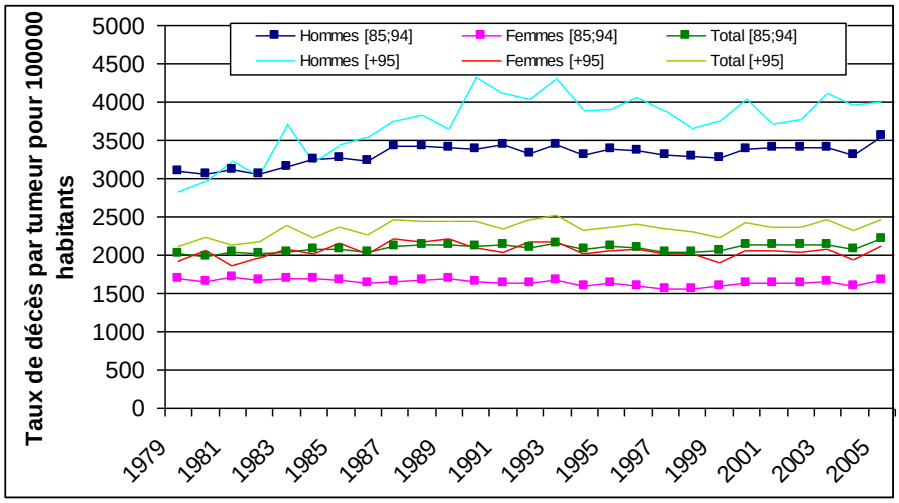
$$\text{Taux de décès}_{15-45ans} = \frac{\text{nombre décès}_{15-45ans}}{\text{nombre d'habitants}_{15-45ans}}$$

Pour les tranches d'âge allant de 15 à 45 ans, on observe une diminution du taux de décès par tumeur :



Graphique 13 : Evolution du taux de décès par tumeur pour la population française âgée de 15 à 45 ans

L'augmentation de l'espérance de vie a conduit à une augmentation du nombre de décès par cancer : parmi les 12 tranches d'âge observées, l'augmentation du taux de décès pour 100 000 habitants n'apparaît que pour les tranches d'âge [85 ; 94] et [95 +], comme on le voit sur le graphique qui suit.

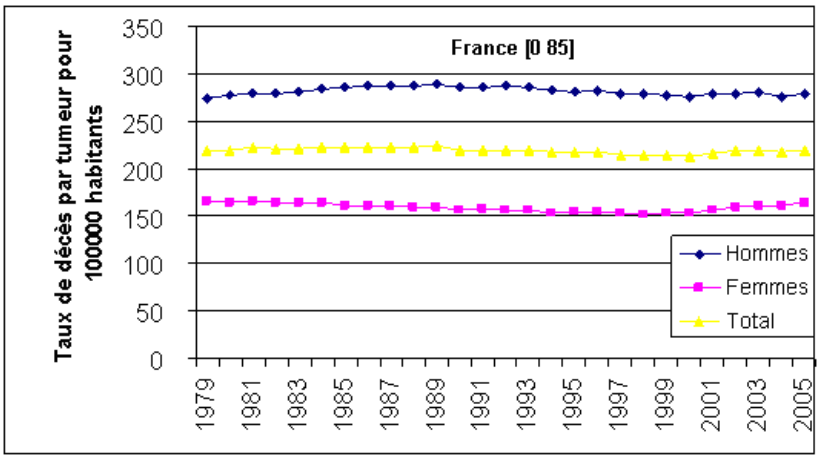


Graphique 14 : Taux de décès en fonction du sexe et de la tranche d'âge

Notons aussi que la détection et l'identification des causes de décès par cancer s'est probablement améliorée avec le temps, ce qui est susceptible d'introduire un biais méthodologique supplémentaire, mais nous n'avons aucun moyen de le savoir ici.

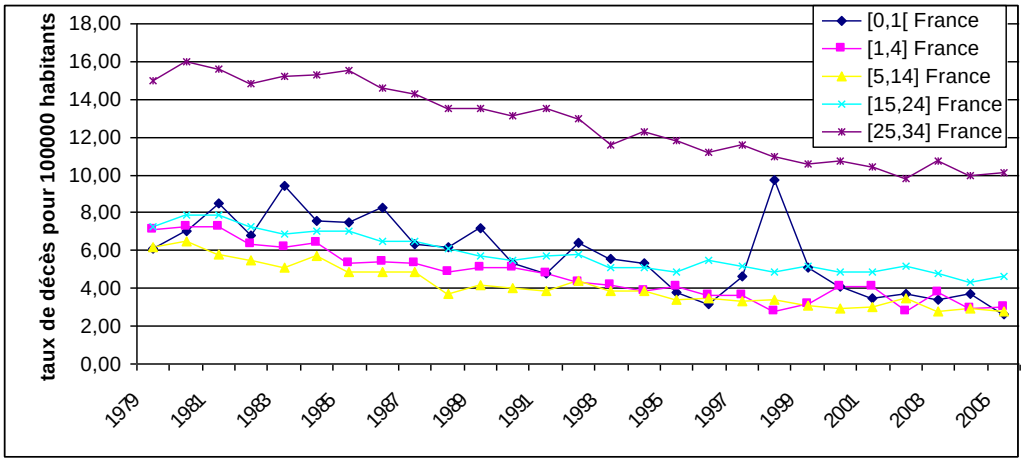
La diminution du taux décès pour chacune des tranches d'âge n'entraîne pas nécessairement une diminution globale du taux de décès :

Le graphique ci-dessous représente le taux de décès pour la tranche d'âge [0 ;85] :

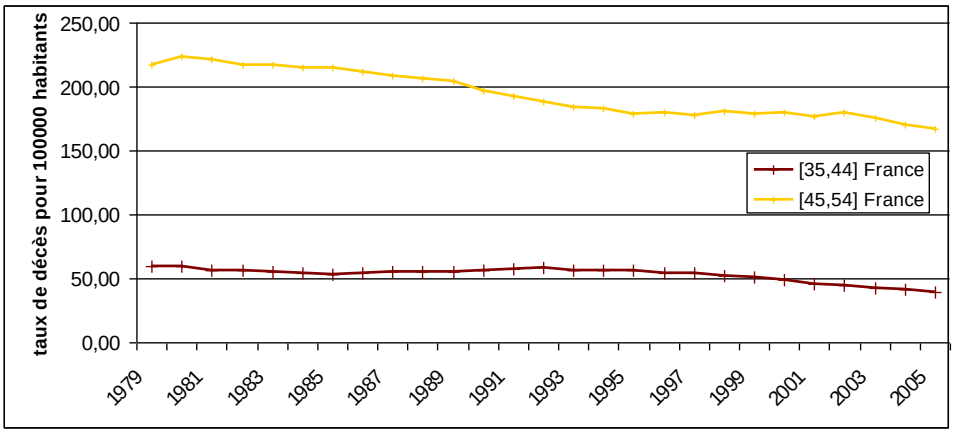


Graphique 15 : Taux de décès pour la population française âgée de moins de 85 ans

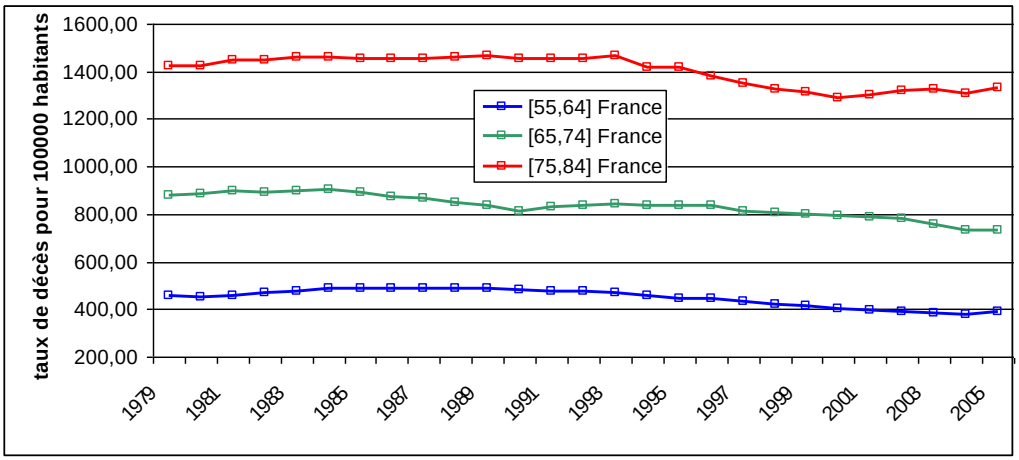
Sur le graphique précédent, nous n'observons aucune diminution du taux de décès. Pourtant comme le montrent les trois graphiques suivants, une diminution du taux de décès est observée sur chacun des intervalles d'âge :



Graphique 16 : Taux de décès pour 5 tranches d'âge de la population française



Graphique 17 : Taux de décès pour 2 tranches d'âge de la population française



Graphique 18 : Taux de décès pour 3 tranches d'âge de la population française

Une diminution du taux de décès pour chacune des tranches d'âge n'entraîne pas nécessairement la diminution du taux de décès de l'ensemble de la population.

Prenons un exemple fictif simple en comparant seulement deux années (1985 et 2005) et trois tranches d'âge ([80 ;82], [83 ;85],[86 ;88]) :

A partir du nombre de décès suivant :

	[80-82]	[83-85]	[86-88]
1985	5	15	4
2005	4	13	18

Tableau 3 : Nombre de décès

En prenant une population totale de :

	[80-82]	[83-85]	[86-88]
1985	1000	1000	200
2005	1000	1000	1000

Tableau 4 : Population pour chaque tranche d'âge

On obtient alors les taux de décès (pour 1000 habitants) suivants :

	[80-82]	[83-85]	[86-88]
1985	5	15	20
2005	4	13	18

Tableau 5 : Taux de décès pour 1000 habitants

Sur cet exemple, le taux de décès a diminué pour toutes les tranches d'âge et pourtant, si l'on calcule les taux de décès pour les deux années en rassemblant les trois tranches d'âge, on s'aperçoit qu'il augmente, comme nous allons le voir.

On note n_i le nombre de décès pour la tranche d'âge i et q_i la population.

$$\text{taux de décès pour 1000 habitants en 1985} = \frac{\sum_{i=1}^{3} n_i}{\sum_{i=1}^{3} q_i} 1000 = \frac{5+15+4}{1000+1000+200} 1000 \simeq 10,9$$

$$\text{taux de décès pour 1000 habitants en 2000} = \frac{\sum_{i=1}^{3} n_i}{\sum_{i=1}^{3} q_i} 1000 = \frac{4+13+18}{1000+1000+1000} 1000 \simeq 11,7$$

Cet exemple nous montre qu'une diminution du taux de décès pour chacun des groupes n'entraîne pas obligatoirement une diminution du taux sur l'ensemble de ces groupes, du fait de la variation de taille de certains groupes (ici, le groupe le plus âgé devient plus important entre 1985 et 2000).

L'augmentation de la population française pour les dernières tranches d'âge (due à l'augmentation de l'espérance de vie) a conduit à une augmentation du taux de décès sur l'ensemble des tranches d'âge.

B. Lien entre la moyenne des taux et le taux de la France

On note t_j le taux pour la j-ème région, n_j le nombre de cancers pour cette région et q_j la population. On a :

$$t_j = n_j \times \frac{100000}{q_j}$$

La moyenne des taux est :

$$M = \frac{1}{n}\sum_{j=1}^{n} n_j \frac{100000}{q_j}$$

Le taux moyen pour la France est :

$$M_F = \frac{\sum_{j=1}^{n} n_j \times 100000}{\sum_{j=1}^{n} q_j}$$

Les deux coïncident si toutes les régions ont la même population ($q_j = q$)

Si on affecte chaque t_j du coefficient barycentrique $\tau_j = \frac{q_j}{\sum q_j}$, on a $\sum \tau_j = 1$ et

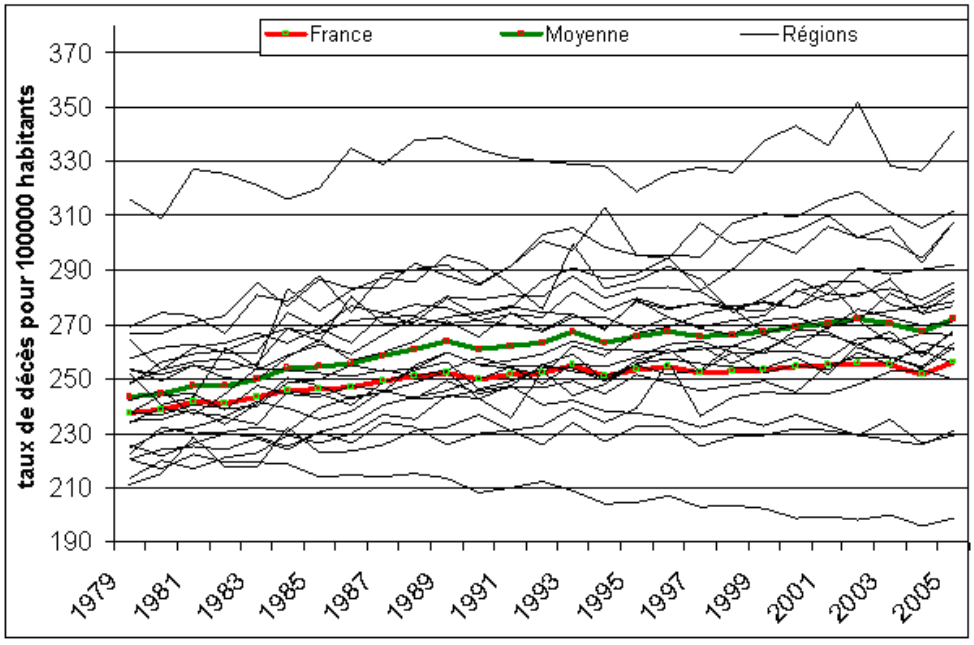
$$\sum \tau_j t_j = \sum \tau_j n_j \times \frac{100000}{q_j} = \sum \frac{10000}{\sum q_j} n_j = 100000 \frac{\sum n_j}{\sum q_j} = 100000 \frac{N}{Q}$$

où N est le nombre total de cancers et Q la population totale.

On constate donc, dans tous les cas, que la moyenne pondérée des taux (pondérée par les coefficients barycentriques τ_j) est égale au taux pour la France.

$$T = \sum t_j \tau_j$$

Le graphique ci dessous présente l'évolution du taux de décès par cancer pour toutes les régions. La moyenne des taux de décès des régions est aussi présentée avec le taux de décès de la France (la moyenne des taux des régions est différente du taux de la France car la population n'est pas la même pour les régions).

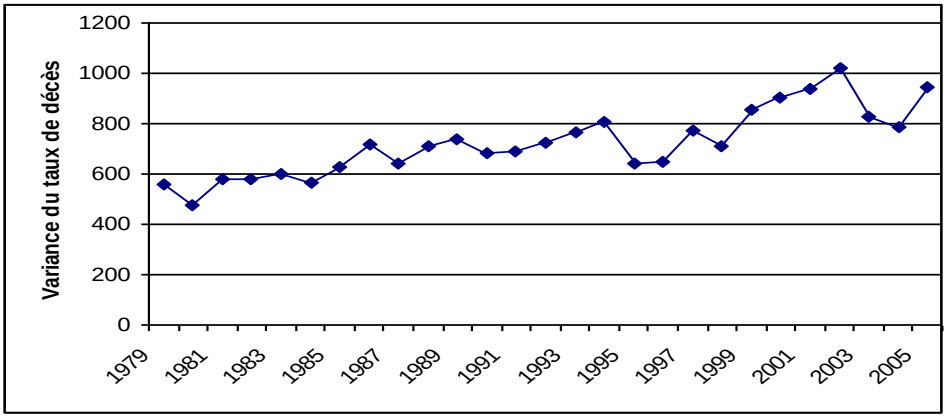


Graphique 19 : Taux de décès par tumeur *pour 100 000 habitants*

L'importance de la population n'est pas prise en compte dans la moyenne des taux des régions. La région Ile de France étant la région la plus peuplée, celle-ci a plus d'influence sur le taux français que sur la moyenne des taux. Comme cette région à le taux de décès le plus faible, le taux de la France est supérieur à la moyenne des taux.

L'augmentation est constante, il n'y a pas d'augmentation brutale du taux de décès parmi ces 27 années.

La dispersion des résultats, d'une région à l'autre, peut être étudiée d'une année sur l'autre afin de vérifier s'il existe une corrélation selon les régions. On calcule donc la variance pour évaluer la dispersion des résultats. On peut remarquer sur le graphique suivant l'augmentation de la variance, qui indique une plus grande dispersion pour les années les plus récentes.

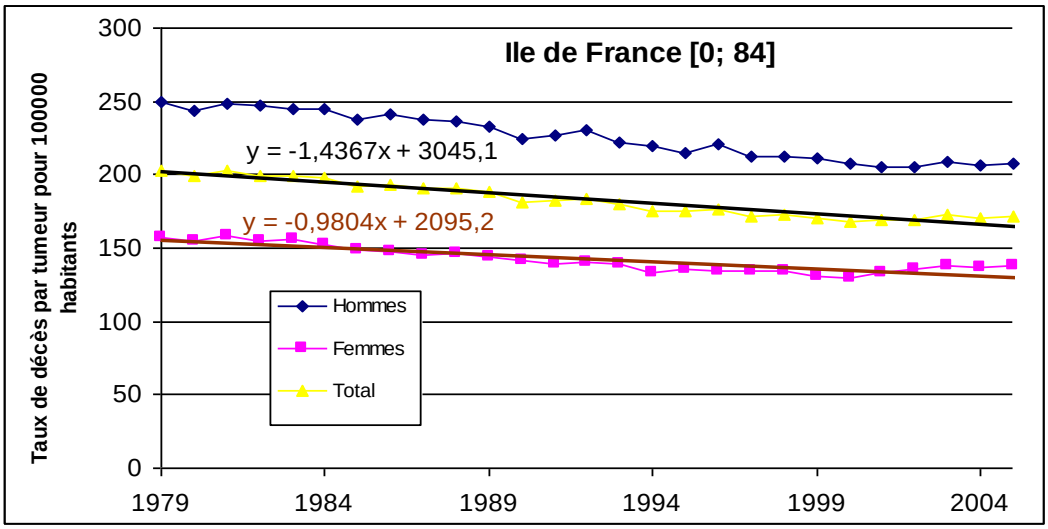


Graphique 20 : Variance du taux de décès par tumeur en fonction de l'année

L'étude de la variance d'une année sur l'autre montre que les données sont moins regroupées en 2005 qu'en 1979 (on peut aussi le remarquer sur le graphique 18). En langage intuitif, cela signifie que les régions deviennent de plus en plus indépendantes, de moins en moins liées entre elles.

C. Prédictions par ajustement linéaire

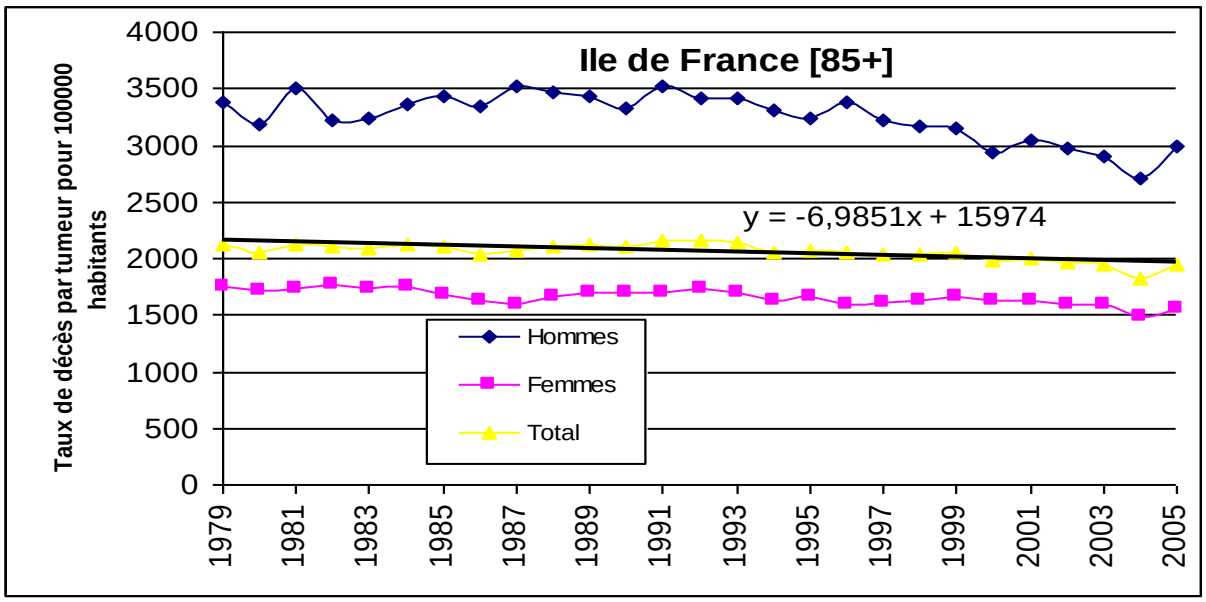
Il est possible de réaliser des prédictions pour 2008, à partir des données historiques. Ces prédictions sont simplement obtenues à partir des courbes de tendance réalisées sous Excel. Le graphique ci-dessous représente l'évolution dans le temps du taux de décès par tumeur pour 100 000 habitants dans la région Ile de France. La droite de régression linéaire ainsi que son équation sont également représentées sur le graphique.



Graphique 21 : Taux de décès par tumeur pour 100 000 habitants en Ile de France pour la population de moins de 85 ans

Pour la prédiction du taux de décès en 2008, l'équation linéaire obtenue est $y = -1,4367x + 3045,1$. La méthode linéaire permet d'obtenir le résultat suivant : en 2008, le taux de décès par cancer prédit est de 160,2 pour 100 000 habitants.

En considérant uniquement la population féminine, on obtient l'équation linéaire $y = -0,9804x + 2095,2$ qui permet d'obtenir une prédiction de 126,6 décès pour 100 000 femmes pour l'année 2008.



Graphique 22 : Taux de décès par cancer pour 100 000 habitants en Ile de France pour la population d'au moins 85 ans

Pour la prédiction du taux de décès total pour la population de plus de 85 ans en Ile de France en 2008, l'équation linéaire obtenue est $y = -6,9851x + 15974$, et le taux de décès prédit est de 1948 pour 100 000 habitants.

La méthode linéaire est simple à mettre en œuvre, mais la prévision du taux de décès est établie en considérant les données anciennes avec le même poids que les données récentes. Cette considération n'est pas un problème si les données suivent une tendance linéaire. Mais si les données sont chaotiques ou si la tendance n'est pas linéaire (comme c'est le cas pour la population féminine en Ile de France sur le graphique 22), l'utilisation de cette méthode doit être proscrite.

Les défauts principaux des méthodes de prolongement linéaire sont :

- On obtient une donnée précise et non une loi de probabilité ;
- On accorde le même poids à toutes les années, récentes ou non.

Nous verrons plus loin comment la construction de l'EPH (Experimental Probabilistic Hypersurface) permet de remédier à ces inconvénients.

IV. Comparaisons probabilistes entre régions

A. La théorie

Il s'agit de mettre en œuvre le logiciel EvalRisk, développé par la SCM. La théorie est ancienne, et consiste à considérer que le taux de risque (ici le taux de décès) est une variable aléatoire dont il s'agit d'estimer la loi de probabilité à partir du nombre d'accidents (ici le nombre de décès) qui se sont produits. Cette théorie est présentée dans le livre de B. Beauzamy « Méthodes probabilistes pour l'étude des phénomènes réels » [1]. Le logiciel lui-même a été utilisé par le CEA Saclay (risques liés au survol par des avions et au transport de matières dangereuses) et par la Direction de la Sûreté Nucléaire de Défense (risques liés aux manipulations de têtes nucléaires).

Sous sa forme de base, le logiciel permet de déterminer la probabilité de dangerosité d'un produit par comparaison avec un produit de référence :

Si un produit A a causé n accidents pour N utilisations et un produit B a causé m accidents pour M utilisations, quelle est la probabilité que le produit A soit plus dangereux que le produit B ?

Pour déterminer si une région est plus affectée qu'une autre région (ou bien que le pays tout entier) le nombre de décès par cancer ainsi que le nombre d'habitants sont nécessaires. A partir des informations obtenues sur le site de l'Inserm qui donne le taux de décès pour 100 000 habitants ainsi que le nombre de décès, nous retrouvons le nombre d'habitants à partir de la formule :

$$\text{nombre d'habitants} = \frac{\text{nombre de décès}}{\text{taux de décès}} 100000$$

B. Application sur quelques données

Pour la région Champagne, le nombre de décès par cancer en 2005 était de 3564 pour 1 338 340 habitants (soit 266 décès pour 100 000 habitants). A la même période, le nombre de décès pour la France était de 155 407 pour 60 634 800 habitants (soit 256 décès pour 100 000 habitants).

La méthode probabiliste développée dans le logiciel EvalRisk permet d'obtenir la conclusion suivante : la probabilité que la région Champagne soit plus affectée que le pays est de 0,99.

Pour la région Picardie, le nombre de décès par tumeurs en 2005 était de 4 920 pour 1 879 300 habitants (soit 262 décès pour 100 000 habitants). On trouve alors une probabilité de 0,93 que la région de la Picardie soit plus affectée que la France.

Les régions peuvent être comparées entre elles : la probabilité que la Picardie soit plus affectée que la Champagne est alors de 0,23.

Même si les taux de décès pour 100 000 habitants sont proches (266 pour la Champagne et 256 pour la France), les probabilités obtenues sont très différentes (proches de 0 ou de 1) car le nombre d'habitants est grand (60 634 800 pour la France).

L'exemple ci-dessous est réalisé avec un nombre d'habitants plus faible :

Le nombre de décès par cancer en 2005 pour les personnes de plus de 95 ans en Picardie était de 70 pour 2 675 habitants (soit 2617 décès pour 100 000 habitants). A la même période, le nombre de décès chez les personnes de plus de 95 ans en Champagne était de 65 pour 2400 habitants (soit 2708 décès pour 100 000 habitants). La probabilité pour que les personnes de plus de 95 ans en Picardie soient plus affectées que les personnes de plus de 95 ans en Champagne est de 0,42.

C. Utilisation prédictive d'Evalrisk

On peut aussi utiliser Evalrisk pour prédire le nombre de décès dans le futur, pour une région donnée, sous forme probabiliste. La formule est la suivante : sachant que n décès se sont produits pour N habitants, la probabilité d'avoir, dans le futur, n' décès pour N' habitants est :

$$p(n', N'; n, N) = \frac{N+1}{N+N'+1} \frac{\binom{N'}{n'}\binom{N}{n}}{\binom{N+N'}{n+n'}}$$

(voir [1], chapitre 14.)

L'application de cette formule donne le résultat suivant : sachant que pour la Champagne il y a eu, en 2005, 266 décès pour 100 000 habitants, la probabilité d'avoir au moins 300 décès l'année suivante est 0.082.

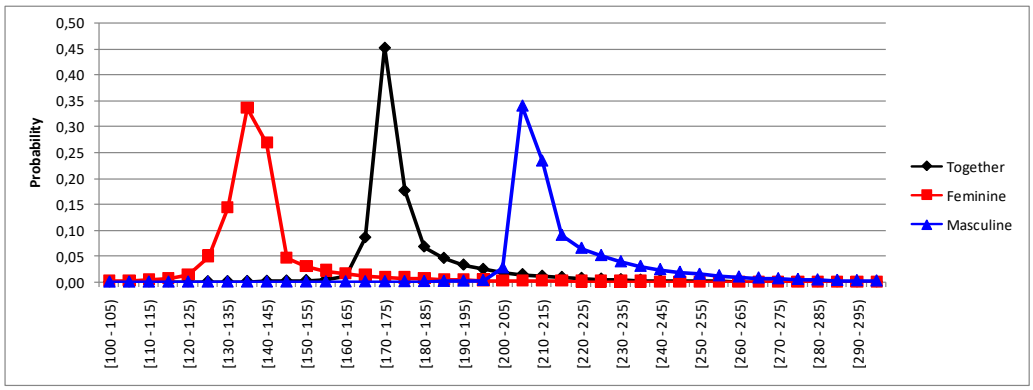
Mais cette estimation ne fait intervenir qu'une seule année, comme c'était déjà le cas pour la comparaison entre régions. Pour faire intervenir tout l'historique, il nous faut un outil plus élaboré, que nous présentons maintenant.

V. Utilisation de l'EPH (Experimental Probabilistic Hypersurface)

Comme expliqué plus haut, la construction générale de l'EPH est donnée dans [2] et son application spécifique à l'épidémiologie est donnée sur le document joint.

Voici deux exemples de résultats obtenus.

Dans le graphique ci-dessous, on trouve les lois de probabilité prédites pour 2008 pour l'Ile de France (taux de cancers pour 100 000 habitants), pour la tranche d'âge 0-85. On a distingué hommes, femmes, et ensemble.



Graphique 25 : les lois de probabilité prédites pour l'Ile de France pour 2008

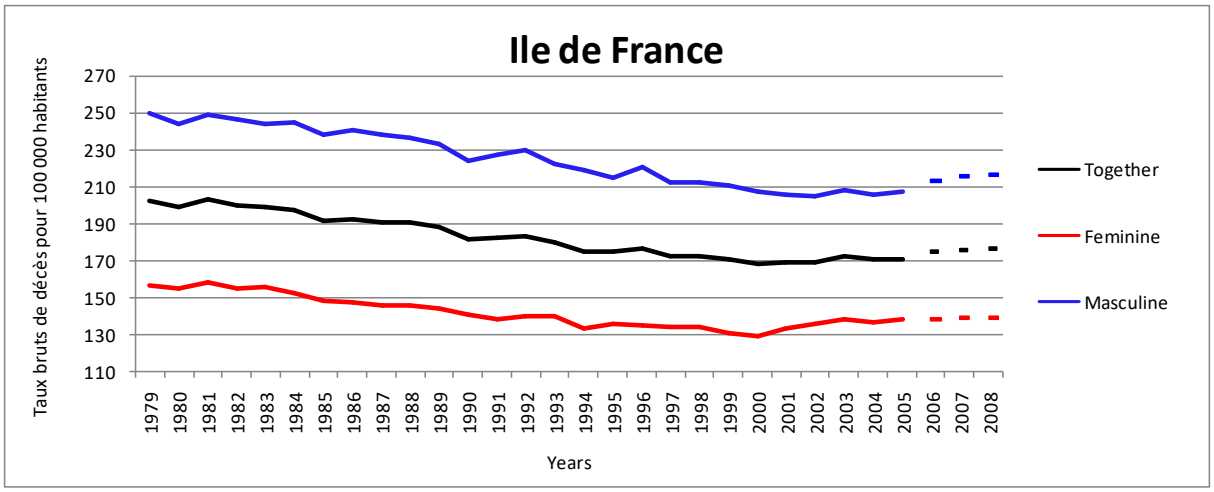
On en déduit immédiatement des résultats du type :

$$\mathrm{P}_{Together}\ \ Number\ of\ deaths \geq 210\ \ = 0.05$$

$$\mathrm{P}_{Feminine}\ \ Number\ of\ deaths \geq 210\ \ = 0.01$$

$$\mathrm{P}_{Masculine}\ \ Number\ of\ deaths \geq 210\ \ = 0.62$$

A partir de ces lois, on obtient des prédictions ponctuelles ; voici l'évolution prévue des taux de décès pour l'Ile de France, pour les années 2006, 2007, 2008, à partir des données connues jusqu'en 2005.



Graphique 26 : prédiction des taux de décès pour l'Ile de France

Ces prédictions sont obtenues en prenant les espérances des lois de probabilité calculées précédemment.

On note que la prédiction est d'une légère remontée. Ceci est logique, dans la mesure où la méthode exploite les données anciennes, qui avaient une valeur plus élevée. Mais la remontée est faible, dans la mesure où les valeurs anciennes ont un poids plus faible que les valeurs récentes. Aucune méthode probabiliste, appliquée aux données elles-mêmes, ne peut conclure à une baisse constante des taux, dans la mesure où des valeurs plus basses que les valeurs actuelles n'ont jamais été observées. Pour conclure à une baisse continue des taux, il faudrait appliquer la méthode aux variations de taux d'une année sur l'autre, et non aux taux eux-mêmes. Ce choix est artificiel et nous ne le retenons pas.

Références

[1] Bernard Beauzamy : Méthodes probabilistes pour l'étude des phénomènes réels. SCM SA, 2004.

[2] Méthodes probabilistes pour l'analyse des incertitudes liées à la sûreté des réacteurs nucléaires. L'Hypersurface Probabiliste : Construction Générale et Applications. Rapport rédigé par Olga Zeydina, Ingénieur de Recherche, Société de Calcul Mathématique S. A. en préparation de sa thèse de doctorat "Méthodes probabilistes pour la Sûreté Nucléaire". Thèse préparée à l'Université de Bretagne Sud, Laboratoire de Mathématiques et Applications Thèse codirigée par Emile Le Page et Bernard Beauzamy. Rapport no 4 adressé à l'Institut de RadioProtection et de Sûreté Nucléaire, en application de la commande R50/11026029 du 29 novembre 2006. Avril 2007.