



**Use of the  
Experimental Probabilistic Hypersurface  
in Epidemiology**

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

Document adressé au CEA

*Direction de la Protection et de la Sûreté Nucléaire*

(à l'attention de M. Thierry de Bruyne)

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

12 mars 2008

## Table of contents

I. Introduction .....	3
II. Traditional methods .....	3
III. General statements concerning EPH .....	4
IV. Description of the problem .....	5
V. Description of the EPH .....	6
A. <i>Specific construction</i> .....	6
B. <i>Computing the parameter <math>\lambda</math></i> .....	7
VI. Applications of EPH to epidemiological problems .....	7
VII. Study of increments .....	9
References .....	11

# I. Introduction

We present here an application of the Experimental Probabilistic Hypersurface (EPH) to epidemiology, and, more exactly, to the prediction of the future evolution of number of deaths from a certain disease.

The obtained statement is of the following type: given a certain disease and a certain region, given a certain number of deaths in the past, for instance over a range of years 1980-2000, what can we expect in the future ?

# II. Traditional methods

Traditional methods are of two kinds:

## A. *Linear regression*

Most commonly, one builds a linear regression using known data, and this leads to some forecast for the future. This method was for instance used by the European Environment Agency in order to predict the evolution of pollution in rivers (see [1]).

The drawbacks of this method are well-known: first, the predicting tool is simply linear, which is very inefficient in the case where the data in the past oscillate; second, the result is just a number, with no uncertainty at all upon it.

## B. *Probability law*

Another possibility is to consider the data from the past as results of a random experiment, build the law of probability of this random variable from these data (the histogram of the law) and take the expectation of the law as the guess in the future. This is possible only if the data from the past are numerous enough and if nothing has changed in the general conditions of the experiment. See our report, part 1, for a more detailed investigation of this question. Usually, as it is the case here for cancers, there are so many changes in the general environment that this stationarity assumption cannot be fulfilled.

Conversely, EPH will give a way of obtaining results which are of probabilistic nature (the result is a probability law, not a number), and which rely more on the recent past than on the ancient past.

Let us first recall some general statements about EPH. See [2] for a general construction.

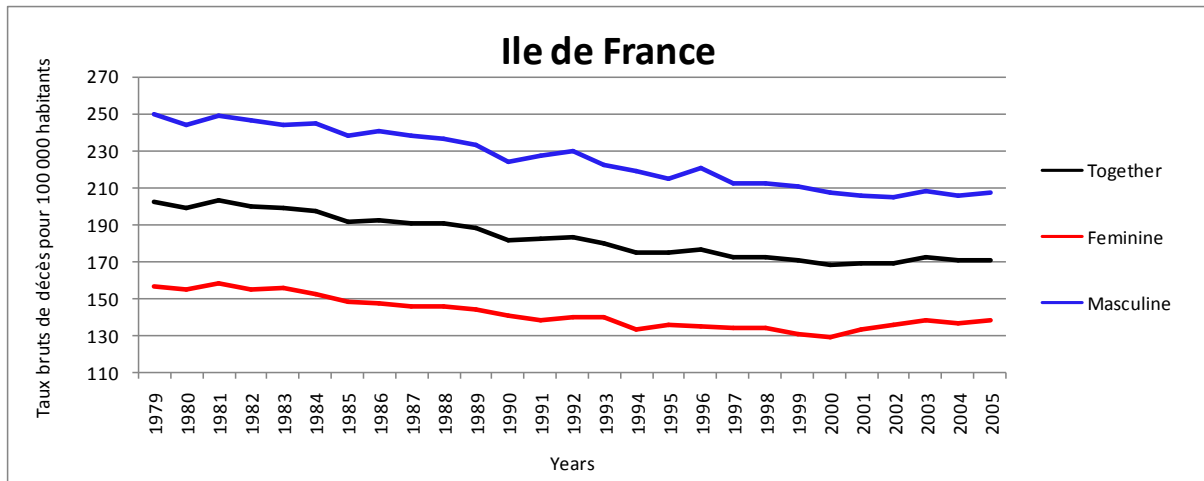
### III. General statements concerning EPH

- The EPH provides a way of "storing" information, under a probabilistic form: if an experiment, or a computational code, has been executed  $N$  times, we have at our disposal  $N$  results, each depending on the specific values given to the parameters. In the case of epidemiology, the "results" are the number of deaths (annual mortality), during  $N$  years, from a certain disease.
- The EPH consists in a collection of densities of probability, one above each point of the configuration space. Here, the configuration space is simply indexed by years.
- The density is a Dirac measure (certainty) if the measure has been made at that one point, otherwise, the density is less and less concentrated when we move further away from measure points. This means that our predictions will be less and less precise when we move to the future, which is quite normal.
- The construction given in [2] is based upon a general principle of maximal entropy (minimal information). No specific or artificial assumption is made.
- The key point in the construction is the propagation of information from a measure point to any other point.
- The "speed" of this propagation depends from one parameter which is tuned with the rule of maximum entropy.

## IV. Description of the problem

Most of the time, epidemiology is dealing with a kind of data which represent the evolution of a certain disease with time. Then an important question concerns the forecast of such an information for the future.

The graph below describes the number of deaths from tumor for the region Ile de France, computed for 100 000 habitants. The age of people falls into interval  $[0, 85]$ . The information is given separately for men, women, and for both sexes together.



Graph 1: Numbers of deaths from tumors for the region Ile de France, computed for 100 000 habitants.

The X-axis represents the years, the Y-axis is a number of deaths.

The period of observation represents 27 years : from 1979 till 2005, and we would like to know what will happen in 2006, 2007, 2008 and so on.

In order to answer to this question, first we give the brief description of the EPH and then we show how to apply it to epidemiological problems.

## V. Description of the EPH

We refer the reader to the general construction [2] for justifications. Here, we deal only with one-dimensional space (the only parameter is the time), and, more precisely, with a discrete space (we deal only with years). The discretization parameter is denoted by  $j$ .

### A. Specific construction

Let  $A_n$  be the measure points ; here they are just years : 1979 to 2005. Let  $\theta_n$  be the value of the measure at each point, that is the number of deaths of that year.

At any point  $x$  of the parameter space, the EPH gives a probability law  $p_j(x)$ , with  $\sum_j p_j(x) = 1$ , by the following formula :

$$p_j(x) = \frac{1}{\sum_{i=1}^N 1/d_i} \left( \frac{1}{d_1} p_{1,j} + \dots + \frac{1}{d_N} p_{N,j} \right),$$

where  $p_{n,j}(x)$  is the density sent by the  $n$ -th measure, that is :

$$p_{n,j}(x) = \frac{c_n \tau}{\sigma_n \sqrt{2\pi}} \cdot \exp \left( -\frac{(t_j - \theta_n)^2}{2\sigma_n^2} \right)$$

with :

$$\sigma_n = \frac{\tau e^{\lambda d_n}}{\sqrt{2\pi e}}$$

and  $d_n = d(x, A_n) = |x - A_n|$ ,  $n = 1, \dots, N$ , represents the distance (that is the number of years) between the year  $x$  and the year  $A_n$ .

Since we want to build the prediction for 2008, then this parameter can be bounded as  $Years \in [x_{\min}, x_{\max}]$  with  $x_{\min} = 1979$  and  $x_{\max} = 2008$ .

The parameter  $d_n$  represents the distance between the examined year (for example  $x = 2008$ ) and the year  $A_n$  when the measure was made.

Since this distance appears as an exponential, we see that the last observations will give more influence to the final result than the early ones. This principle here suits completely with the feature of Epidemiology where the last observations have more importance.

The parameter  $t_j$  represents a number of deaths with the following discretization :

$$t_j = t_{\min} + \frac{j}{\nu} (t_{\max} - t_{\min}), \quad j = 0, \dots, \nu$$

and, as we mentioned, the parameter  $\theta_n$  represents the observed values of the number of deaths in the corresponding year.

### B. Computing the parameter $\lambda$

As we mentioned, the construction of EPH is based upon a general principle of maximal entropy (minimal information). In practice it means that we do not add any information such as choice of arbitrary laws. See [2] for the general construction and the Minimal Information Lemma.

The coefficient  $\lambda$  represents a “speed” of propagation of information, depending on  $N$  and on  $\nu$ . It is given by the formula:

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

where :

$$d_{\max} = \max_{n=1,\dots,N} \max \left\{ |x_{\min} - A_n|, |x_{\max} - A_n| \right\}$$

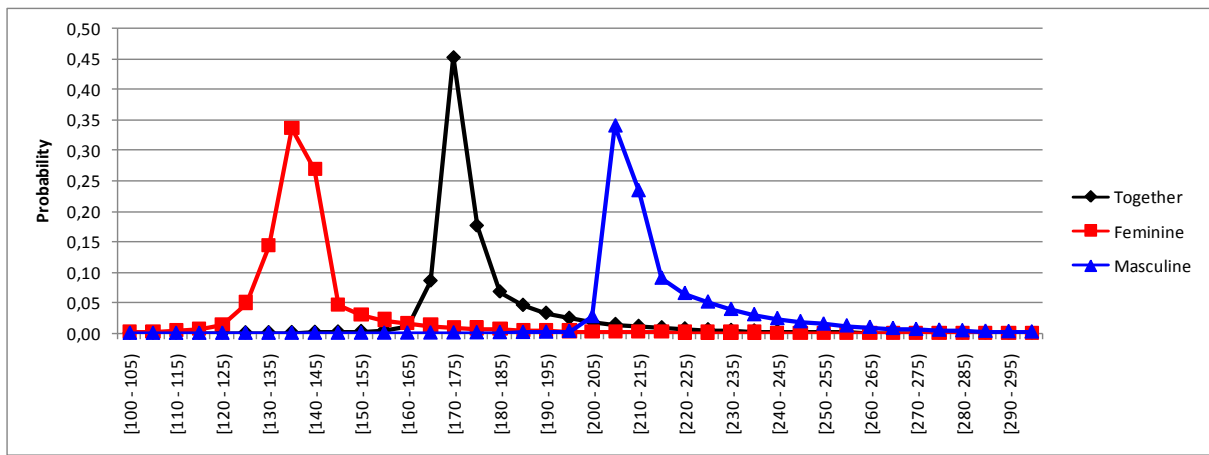
## VI. Applications of EPH to epidemiological problems

As we said, we want to build a prediction of number of deaths for 2006 -2008 using the data from the past.

Using rules of EPH we proceed as follows :

1. We define the boundaries for the result, that is the biggest interval which may contain the results. Here, we can take, for example,  $[100; 300]$
2. We define the boundary for the subdivision :  $\tau = 5$  This means that our precision will be 5 deaths (in other terms, we are not interested by units).
3. We define the boundary for the parameter (years) :  $x \in [1979, 2008]$ .
4. We compute the densities of probability  $p_j(x)$  for years 2008 separately for man, woman and both sexes together using the construction of EPH and the formulas above.

Here are the results :



Graph 2 : The forms of probability laws for 2008 for Ile de France

Here each probability law consists of “bumps”, which are sums of Gaussian laws. The “size” of each bump depends on the distance from 2008 to the year when the observation was made. The result is not symmetric, because all previous values (from the past) are above the last recorded value (2005).

For example for women, the probability law gets its strongest influences from the year 2005 when were registered 138.2 deaths from tumour, as one can see easily on the graph.

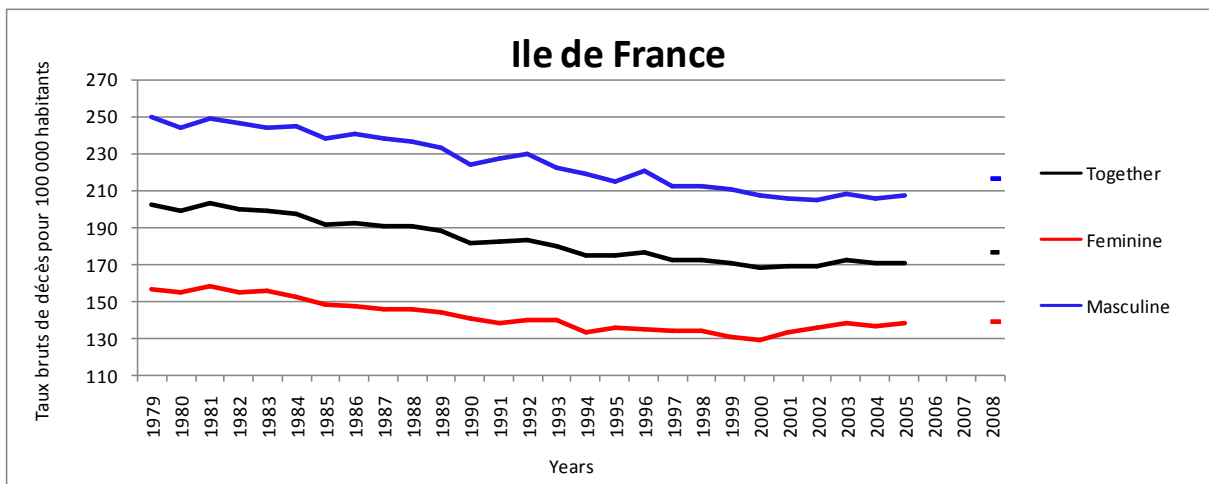
The advantage of the EPH method is that we do not have just one precise value as a result of prediction (for example, as in the case of linear regression). From the graphs above, we can derive what is the probability that the number of deaths will fall into some interval, for example :

$$P_{Together} \{ \text{Number of deaths} \geq 210 \} = 0.05$$

$$P_{Feminine} \{ \text{Number of deaths} \geq 210 \} = 0.01$$

$$P_{Masculine} \{ \text{Number of deaths} \geq 210 \} = 0.62$$

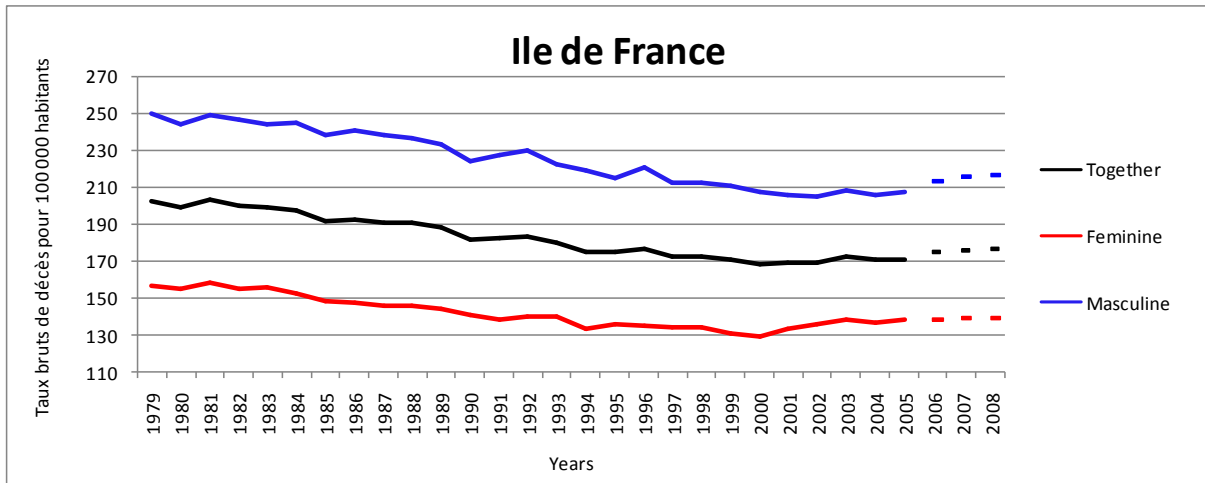
When the probability laws are obtained, the values chosen in order to predict are the mathematical expectations of these laws :



Graph 3 : Numbers of deaths from tumor with prediction for 2008

So, in 2008, we predict 139.2 deaths from tumor for women, 216.2 for men and 176.8 for both sexes accordingly (for 100 000 ha).

Using the same methods, we can calculate the predictions for 2006 and 2007 :



Graph 4 : Numbers of deaths from tumor with prediction for 2006-2008

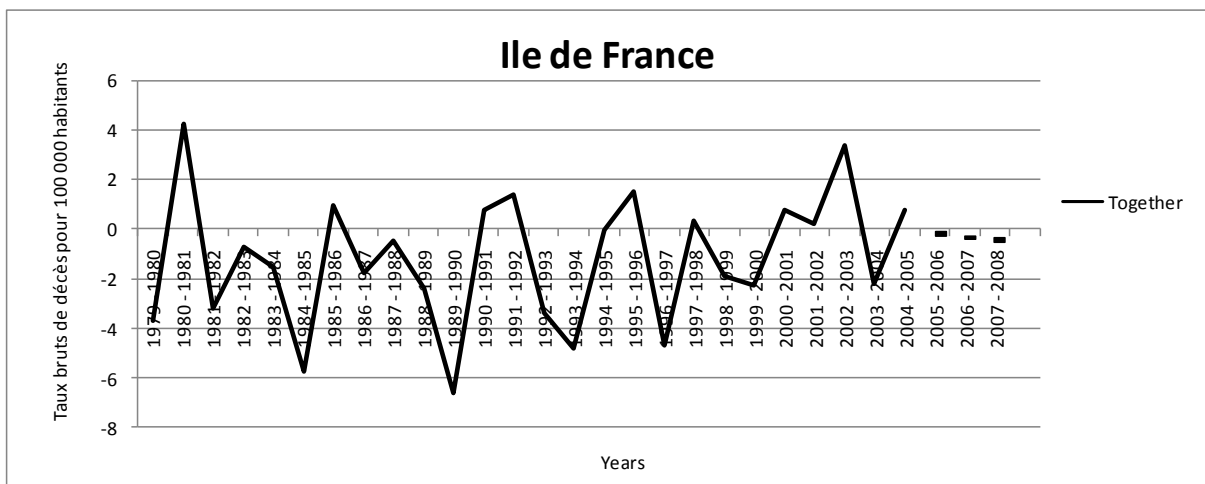
We observe that the prediction is that of a slight increase, which seems to contradict the trend from the past. However, the EPH builds new data from old ones (with weights) and old data are all larger than the recent ones, so their influence is felt.

## VII. Study of increments

Another way would be to build the EPH upon the increments (that is, consecutive differences between one year and the next) . Let us denote them as :

$$\text{Increment}_n = \theta_{n+1} - \theta_n \text{ with } n = 1, \dots, N - 1$$

We present here the graph of increments obtained for Ile de France (men and women together) :

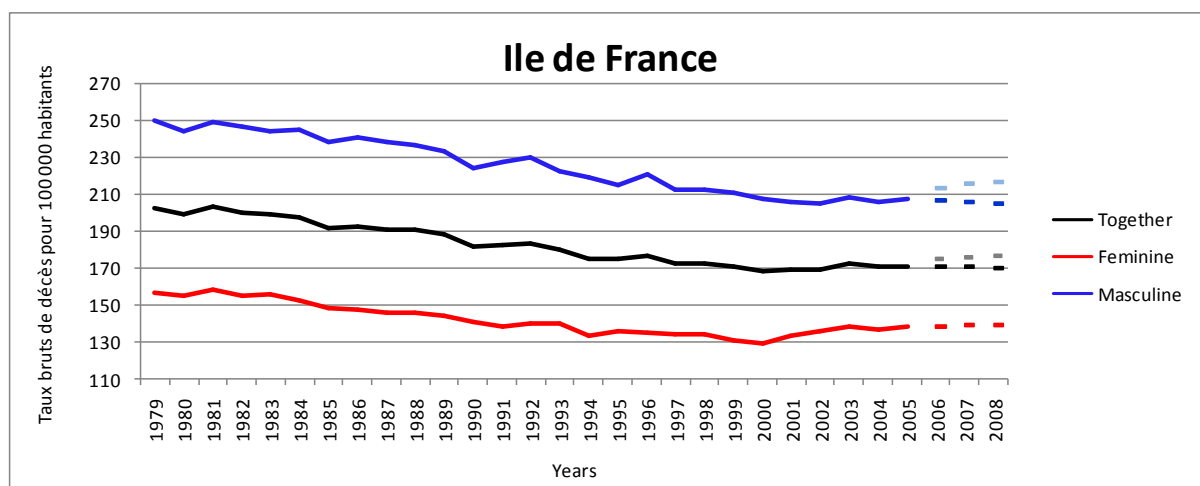


Graph 5 : Increments with prediction for 2006-2008

The dotted line indicates the prediction (expectation) for the future using the EPH.

If the increment is negative, it means a diminution of the number of deaths in comparison with the previous year.

From the graph above, we can easily see the predominance of negative increments upon the positive ones. This was reflected on the predictions : the number of deaths will decrease in the future. This result differs from the previous one, when we built our forecast only from the observed data (not increments). Here is the graph which presents two results at the same time :



Graph 6 : Numbers of deaths with prediction for 2006-2008 based upon observed data and increments.

Dot lines in dark blue, black and dark red represent the forecast based upon the increments. Dot lines in light blue, grey and light red represent the forecast based upon the observed data (previous result).

As we can see from the graph, lines in dark red and light red are identical, which means that the forecast for women based upon increments and observed data are the same here; but for men and both sexes together the characters of lines are different : the forecast based upon increments shows a diminution of number of deaths in the future and the prediction based upon observed data indicates an increase.

More generally, we see that applying the EPH to the phenomenon itself, or to its first derivative, or to its second derivative, and so on, do not provide the same results. This is quite natural : trying to predict the speed of a mobile, or trying to predict its position, usually do not lead to the same results.

Which one should we choose ? We think that the choice should be made upon the concept which has a physical meaning. If, for a mobile, the speed is governed by some actuators, then the speed should be the object for prediction.

In our case here - epidemiology - there is no justification for anything else than the raw data themselves ; the differences from one year to the next do not have any physical meaning, so our claim is here that EPH should be used upon raw data.

## References

- [1] Charline Carlier : Méthodes probabilistes pour l'Environnement, SCM SA, 2007
- [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.