Société de Calcul Mathématique SA
*Outils d'aide à la décision*
*depuis 1995*

# Données fines ou données grossières ?

Newsletter "mathématiques du réel", no 8

Bernard Beauzamy, 26/07/2024

La plupart des gens – et la totalité des ingénieurs – répondront sans hésitation : plus les données sont fines et précises, mieux cela vaut. Cette conviction provient en particulier de la citation de Lord Rutherford (1871-1937) : "Qualitative is nothing than poor quantitative".

Il s'agit pourtant d'une complète erreur, sur tous les plans : scientifique, économique et social. Voyons cela grossièrement !

## *A. Obtenir l'information*

L'information grossière s'obtient facilement, croit-on ; l'information fine requiert du travail.

Est-ce vrai ? Une information fine et précise est-elle toujours plus difficile à obtenir qu'une information grossière ?

Une information fine et précise, relative à un signal que l'on cherche à connaître, concerne la valeur de ce signal, pendant une durée bien spécifiée.

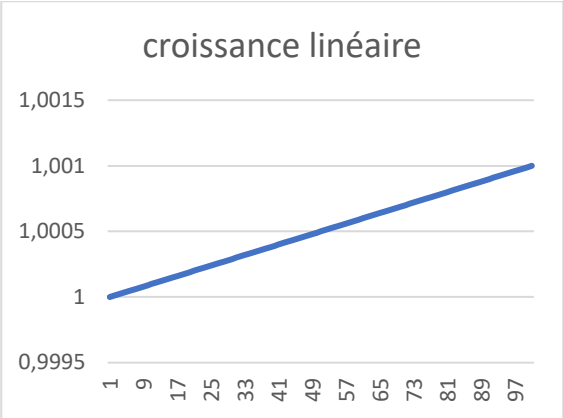
A l'inverse, une information grossière concerne une tendance (hausse ou baisse), à horizon de temps plus ou moins précis : un mois, un an, etc.

La plupart des gens considèrent que l'information fine est plus difficile à obtenir, du fait du niveau de détail. L'information grossière ne serait qu'un substitut lorsque l'information fine ne peut être obtenue, comme le dit Rutherford.

Ce point de vue est radicalement erroné : l'information fine porte généralement en soi une cohérence qui facilite l'identification, tandis que l'information grossière est dépourvue de cette cohérence. Nous allons illustrer ceci par des exemples simples, mais très significatifs.

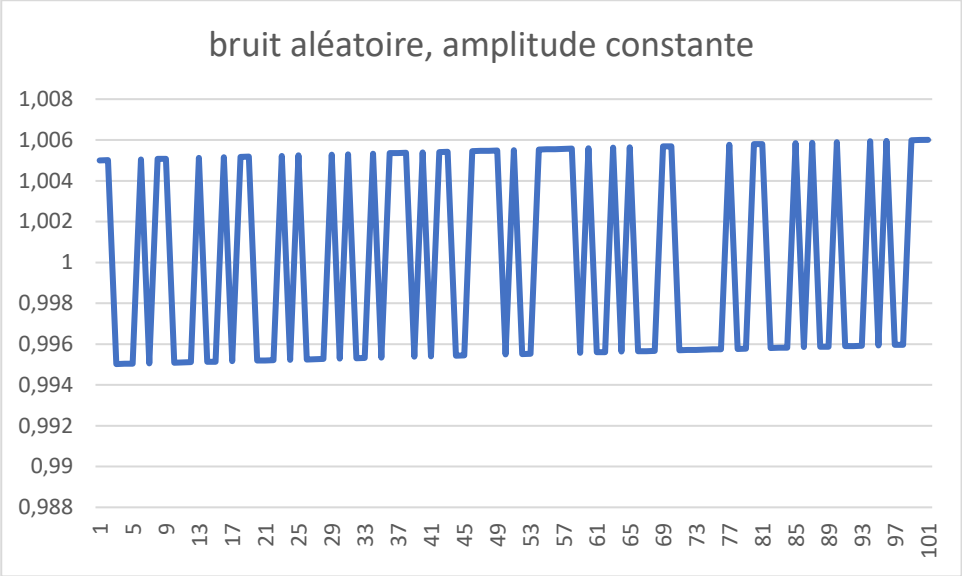
### 1. Exemple 1 : information fine

Un signal croît linéairement pendant 100 pas de temps ; il passe de 1 à 1.001. Un tel signal est très facile à identifier et ses valeurs à chaque instant sont très précises.




## 2. Exemple 2 : information grossière

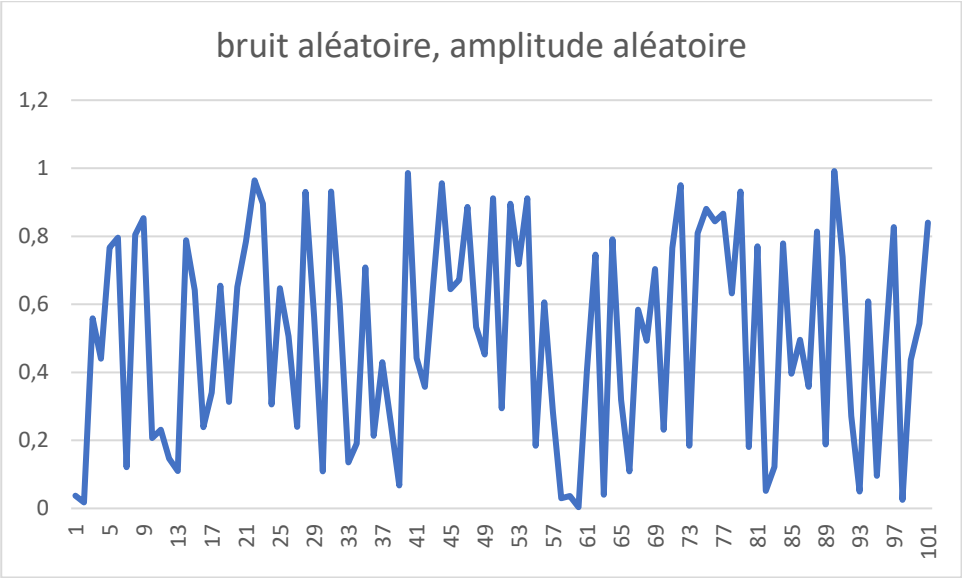
Le signal précédent, à chaque instant, est affecté par un bruit aléatoire, toujours de même amplitude ; ce bruit est positif avec probabilité $1/2$ et négatif avec probabilité $1/2$. Voici le résultat :




Les périodes de croissance et de décroissance du signal se suivent sans ordre apparent, et il est très difficile de prévoir la valeur du signal à horizon de temps donné.

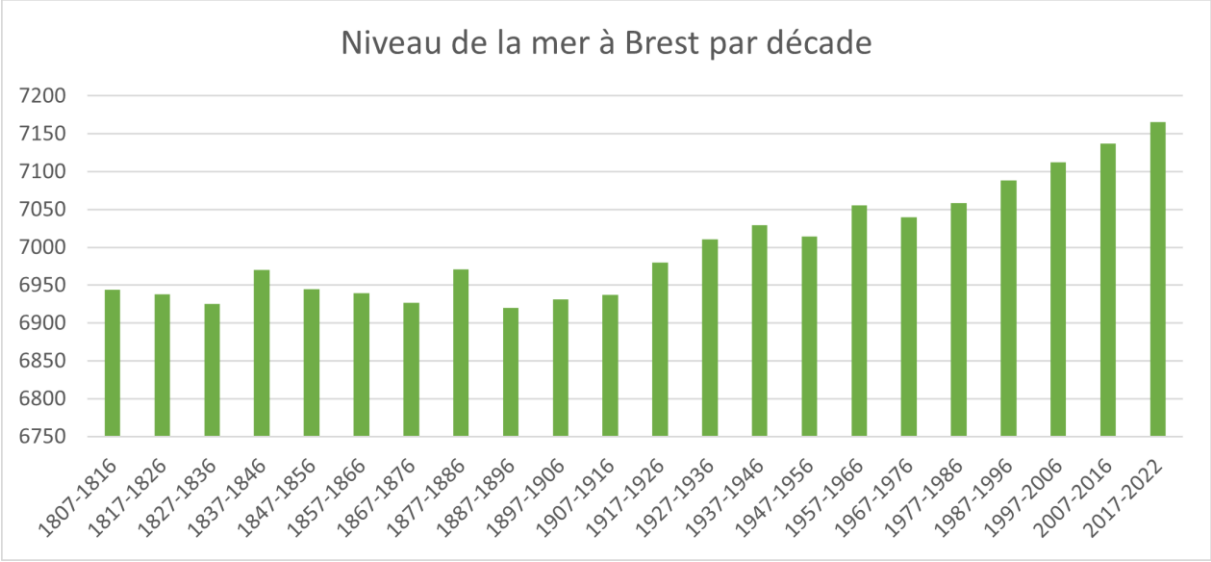
## 3. Exemple 3 : information très grossière

Le bruit aléatoire de l'exemple précédent ne l'est plus seulement en direction, mais aussi en intensité :




Sur ce dernier exemple, il paraît difficile d'établir une tendance, à la hausse ou à la baisse, sur quelque horizon de temps que ce soit.

Une illustration vraiment frappante est donnée par le niveau moyen de la mer à Brest :




(source Service Hydrographique et Océanographique de la Marine)

A Brest, les résultats diffèrent d'une année sur l'autre et même d'une décade à l'autre. La très grande variabilité de la moyenne annuelle, d'une année sur l'autre, est vraisemblablement liée à la variabilité du climat. Lorsque la pression atmosphérique baisse, le niveau de la mer s'élève (la pression de la colonne d'air est plus faible).

On trouve ainsi des périodes de 30 ans pendant lesquelles le niveau moyen baisse. L'extrême variabilité du graphique ci-dessus montre qu'il n'est pas possible de faire une prévision fiable sur dix ans : dans les dix années qui viennent, le niveau peut aussi bien monter que baisser.

Notons bien que, lorsqu'on parle de variation du niveau de la mer (sous-entendu par rapport à la terre), on ne sait pas si c'est la mer qui monte ou la terre qui s'enfonce (ou les deux). Les relevés du SHOM utilisent des marégraphes, qui existent depuis 200 ans environ. Grâce à ces marégraphes, la détermination instantanée du niveau de la mer (donnée fine) est facile. Mais la tendance sur 30 ans ne résulte pas d'un ajustement statistique des données fines (comme par exemple une droite de régression).

Retenons ceci sous une forme très frappante : quoique l'on sache mesurer le niveau de la mer à Brest depuis très longtemps (200 ans) et avec une très grande précision, on est incapable de dire quel sera le niveau moyen dans dix ans.

## *B. Exploitation de l'information : fine ou grossière ?*

### 1. La connaissance d'un terrain

Vous êtes un riche propriétaire foncier et vous venez d'acquérir un terrain de plusieurs km2 dans les pampas. Vous vous demandez : vais-je y faire de la culture ou de l'élevage ?

L'information fine, pour autoriser la culture, consiste en une analyse des sols. Il faudra donc des prélèvements. Si vous voulez être très bien informés, il faudra un prélèvement tous les mètres : le terrain peut ne pas être homogène. Et si les dimensions sont 5km x 5 km, il faudra 25 millions de prélèvements.

Pour l'information grossière, vous vous contenterez du survol en avion : quelques minutes de location d'un appareil, pilote inclus. Un habitué de ces questions vous dira instantanément s'il est préférable de faire de l'élevage ou de la culture. Il voit cela à l'aspect du sol, à la présence d'eau, au type de végétation, etc.

### 2. Codes de calcul et démonstrations de sûreté

En particulier pour la filière nucléaire, les démonstrations de sûreté utilisent nécessairement des codes de calcul, car l'expérimentation est en général impossible. Mais ces codes de calcul sont d'une extrême complexité : imaginez un code qui dépend de 50 paramètres, chacun prenant dix valeurs ; l'espace des configurations est de taille $10^{50}$ et aucune exploration déterministe n'est concevable sur ordinateur. On a souvent la tentation de faire une exploration aléatoire, mais un article récent de Giovanni Bruna et Bernard Beauzamy :
https://www.scmsa.eu/archives/Methodes_probabilistes_demonstrations_surete_BB_GB_2024_05_20.pdf
insiste sur les erreurs à ne pas commettre. Une démonstration de sûreté n'est possible, là encore, qu'à partir d'un code "grossier", qui comprendra peu de paramètres et qui donnera des majorations et non des valeurs extrêmement précises. L'utilisation d'un code fin, pour une démonstration de sûreté, est impossible du fait de la taille de l'espace à explorer.

### 3. Le point de vue du chef d'entreprise

Chaque chef d'entreprise croule littéralement sous des tonnes d'information, généralement totalement indigestes : voici, au mm près, la taille des saucisses fabriquées dans notre usine de Strasbourg pendant la journée du 24/05/2022. Nous-mêmes, à la SCM, avons fait, à plusieurs reprises, des "tableaux de bord" pour dirigeants : ces tableaux visaient à synthétiser l'information pour qu'elle devienne compréhensible, donc à la rendre grossière. Le relevé de tous les paramètres, enregistrés toutes les secondes, sur tous les process de fabrication de chaque usine, n'intéresse pas le chef d'entreprise.

## *C. En conclusion*

Nous ferons deux recommandations, qui paraissent être du bon sens :

1. Si vous mettez en place un nouveau système d'information, commencez par une approche grossière, en vous demandant "à quoi vont servir les données recueillies ?". Une fois ceci fait, en fonction des résultats, vous pourrez affiner ou non.

Nous mentionnerons ici une anecdote, qui illustre bien ce point. Il y a quelques années, nous avons travaillé pour Air Liquide, sur des questions de fiabilité des équipements. Air Liquide a commencé par se demander : sur quels continents avons-nous des difficultés ? puis : dans quelles villes ? etc. C'est une approche "top – down". Ils n'ont pas, comme on voit trop souvent, constitué un énorme système d'information relevant la taille de toutes les rondelles en caoutchouc et tous les cotons-tiges.

L'idée selon laquelle on va créer un énorme système d'information, en mettant des capteurs partout parce que des subventions sont disponibles pour cela, est généralement malsaine. On croule sous les données, presque toujours de mauvaise qualité, et on ne sait pas quoi en faire. Nous avons rencontré cette situation un grand nombre de fois. Il y a quelques années, Airbus nous avait consultés : voici une énorme quantité de données, générées automatiquement ; que peut-on en faire ? Notre réponse a été : absolument rien, ce qui confortait l'impression initiale des responsables et tout le monde a beaucoup ri.

2. Il y a aussi une difficulté, de nature psychologique : les gens qui gèrent des données extrêmement précises ont tendance à développer une certaine forme d'arrogance ; ils se considèrent comme spécialistes et vont rejeter par principe quiconque n'a pas une compréhension fine de leur activité. On demande souvent à la SCM de porter des jugements sur la manière dont les spécialistes traitent leurs propres données et nos conclusions sont : si vous étiez un peu moins spécialistes, vous y verriez un peu plus clair. On dit cela sous la forme "il ne faut pas avoir le nez sur le guidon".