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

Extrapolation, prédiction, simulation

par Bernard Beauzamy

octobre 2022

A. Qu'est-ce que l'extrapolation ?

C'est une technique qui relève de la prédiction : il s'agit de déterminer quelles sont les valeurs possibles pour une variable dans un domaine qui n'est pas celui où elle a été observée. Typiquement, on dispose d'enregistrements sur une certaine période (mettons 2000-2020 pour fixer les idées) et on voudrait savoir quelles sont les valeurs possibles pour le futur : 2021-2025 par exemple. L'exemple le plus fréquent d'extrapolation est celui d'une série temporelle, mais l'extrapolation peut également être spatiale : on dispose de données dans un domaine et on voudrait se faire une idée de ce qu'elles pourraient être au dehors de ce domaine.

Il ne faut pas confondre ce problème avec celui de la reconstitution de données manquantes qui consiste à reconstruire des données à l'intérieur même de l'ensemble où elles sont normalement connues.

Une précaution s'impose ; elle relève de la modestie et du bon sens. Il ne faut pas vouloir extrapoler sur 20 ans à partir de 10 minutes d'enregistrement. Il y a nécessairement une relation entre les deux durées : celle d'enregistrement (données connues) et celle d'extrapolation (données à prévoir). Notre recommandation est de conserver un facteur $1/2$ entre les deux. Autrement dit, si vous avez 10 ans d'enregistrements, il est légitime de s'intéresser à la prédiction sur les 5 années suivantes. Ce choix de $1/2$ est empirique, mais à l'évidence le facteur doit être <1, sauf arrogance.

La différence avec la prédiction est de nature sémantique : la prédiction utilise toute méthode qui lui convient, y compris la boule de cristal, tandis que l'extrapolation n'utilise que des outils mathématiques appropriés ; elle doit reposer sur des fondements mathématiques solides.

B. Information préliminaire

On voit facilement si l'extrapolation est possible ou non ; deux cas se rencontrent :

1. Cas 1

On distingue, sur les données recueillies, une régularité, une périodicité, que l'on va pouvoir exploiter. On constate que, depuis 1970, la population de la Région parisienne croît linéairement, tandis que celles du Nord-Pas de Calais et celle du Grand Est restent constantes. Il n'est pas utile de se demander pourquoi (et les raisons sont certainement très complexes, car beaucoup de facteurs contradictoires interviennent). On pourra néanmoins conclure que, sur les dix prochaines années, ces tendances vont se poursuivre : c'est, en tout cas, une hypothèse raisonnable.

2. Cas 2

Le débit d'un fleuve est très difficile à extrapoler, à cause de la très grande variabilité, même si on utilise des moyennes mensuelles plutôt que journalières. Les valeurs ne sont pas les mêmes d'un jour à l'autre, d'un mois à l'autre, d'une année à l'autre. Il n'y a aucune périodicité.

Si l'on adopte une approche strictement probabiliste, c'est-à-dire sans connaissance "métier", la solution qui s'impose au problème du prolongement est de choisir l'espérance (la moyenne) des données relevées. Si par exemple on a des données de débit, la prédiction pour le futur sera la moyenne des débits précédemment observés.

Un tel prolongement est satisfaisant en ce sens que c'est la valeur qui, par définition de l'espérance, minimise l'erreur moyenne que l'on peut commettre. En revanche, il ne respecte pas la logique du "signal", ici un débit qui est en général variable avec la saison.

Si on fait un premier prolongement (janvier), suivi d'un deuxième prolongement (février), par la même méthode, on obtient la même valeur, ce qui est très insatisfaisant et paraîtra absurde au spécialiste du sujet.

Une variante consiste à utiliser pour le prolongement, non pas la valeur moyenne, mais la valeur la plus probable. Mais, si on réitère, on obtient la même valeur : c'est la même difficulté que précédemment.

Dans le cas d'un signal ayant une valeur très variable, il n'existe pas de méthode satisfaisante pour l'extrapolation, sauf à introduire des connaissances métier, qu'il faudra évidemment justifier.

C. Simulations

Nous venons de voir que le prolongement par espérance, qui est le plus satisfaisant sur le plan probabiliste, était inacceptable en pratique puisqu'il ne respecte pas les spécificités du signal ; il en est de même du prolongement au moyen de la valeur la plus probable.

Beaucoup de gens diront : mais enfin, nous avons des données en très grand nombre, par exemple 10 000 ou davantage, il devrait être possible de se servir de cet amas de données pour prédire la suivante !

Eh bien non ! Sauf cas particuliers évoqués plus haut (comportement linéaire, périodique, etc.), il n'est pas possible de prévoir la prochaine donnée. Par contre, il est tout à fait possible de la simuler, ce qui est entièrement différent.

La simulation va permettre de faire croire au commun des mortels qu'on a une parfaite maîtrise du phénomène, qu'on le comprend dans ses moindres détails. Elle va permettre de générer des données qui, sur un futur aussi long que l'on voudra, vont ressembler parfaitement aux données connues. Celui qui regarde les données simulées ne verra aucune différence avec les données réelles, et dira que c'est absolument admirable.

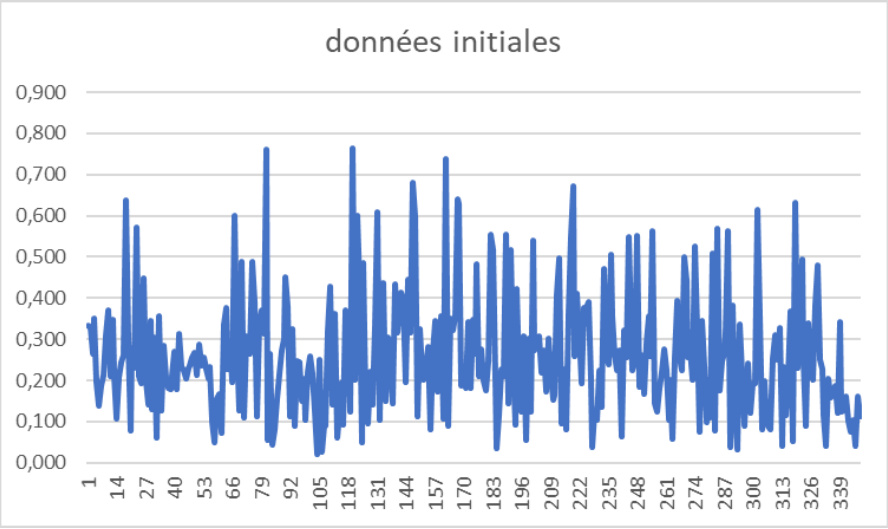
A ceci près que les données simulées n'ont absolument aucune valeur prédictive ; elles sont issues de tirages aléatoires, comme nous allons le voir. Donc, insistons bien sur ceci :

Règle. - Un processus d'aide à la décision ne doit jamais reposer sur des données simulées.

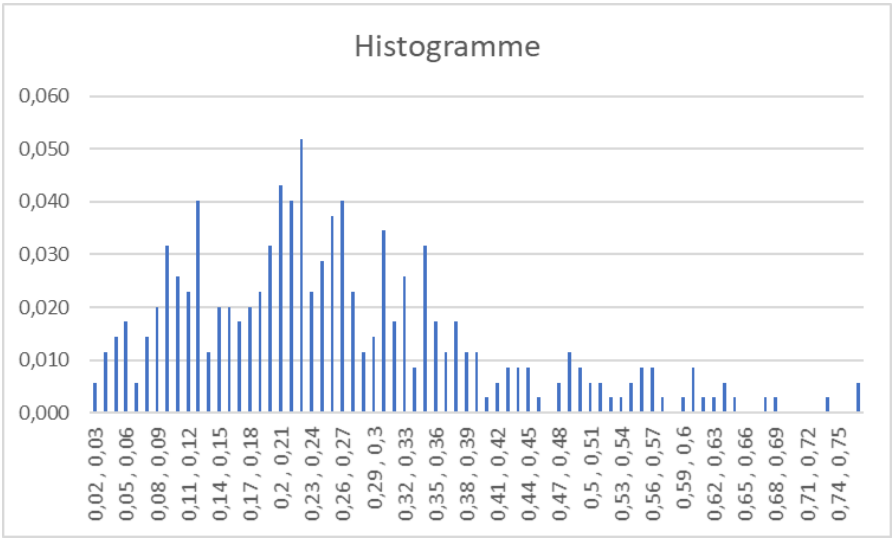
On comprendra mieux ceci lorsque nous aurons expliqué comment simuler des données, à partir d'un historique. Deux étapes sont nécessaires :

Etape 1. – On détermine la loi de probabilité du phénomène. Ceci se fait tout simplement en construisant un histogramme.

Voici tout d'abord les données brutes caractérisant le phénomène ; il s'agit d'un process industriel, enregistré sur une certaine durée (348 mesures) :

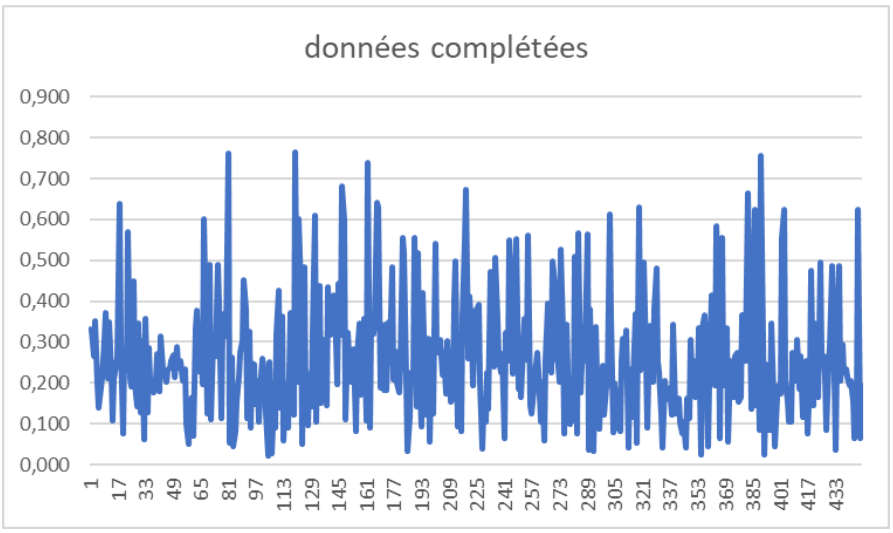


Voici l'histogramme associé à ces données ; il comporte 75 classes, de largeur 0.01 :

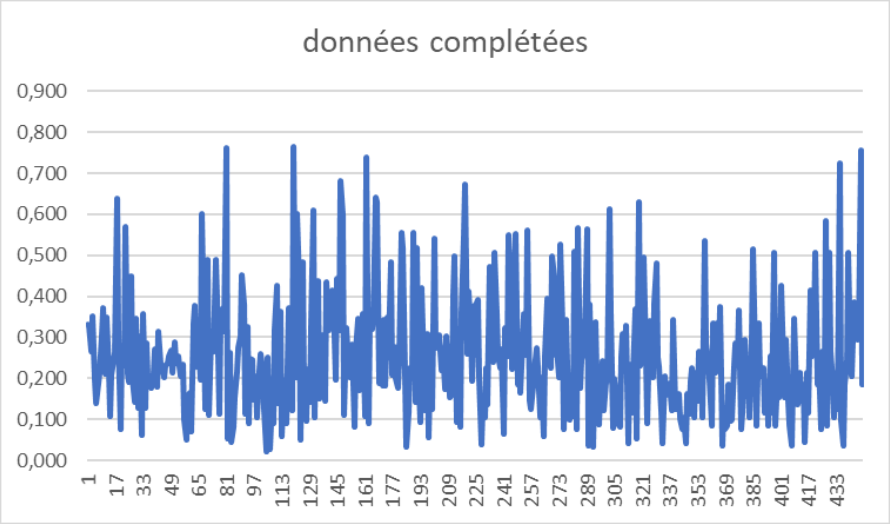


A chaque classe est associée une probabilité : c'est la hauteur du bâton dans la figure ci-dessus.

Etape 2. – On génère 100 nombres aléatoires, suivant la loi de probabilité associée à l'histogramme ; on ajoute ces 100 données après celles déjà enregistrées, et voici ce qu'on obtient :



Le résultat est saisissant : on a l'impression que les données générées s'harmonisent parfaitement avec les données recueillies. Ce n'est pas étonnant, puisqu'elles ont été générées selon la même loi de probabilité. Mais insistons une fois encore sur le fait que ce sont des données générées de manière aléatoire ; on peut dire qu'elles incorporent une certaine "connaissance métier", puisqu'elles sont fabriquées à partir des données d'origine. Par contre, elles ne permettent en aucune façon une prédiction fiable. Du reste, si on recommence le processus de création aléatoire, on tombera sur un aspect différent.



La simulation de données n'est nullement interdite, pourvu qu'elle soit présentée comme telle. Elle a peu de valeur si elle réalisée une seule fois, mais peut devenir un outil d'investigation si on la répète un nombre de fois suffisant. En effet, si on effectue un grand nombre de simulations, on a un bon panorama de ce que peuvent être les données dans l'avenir et ceci est à la base de la méthode dite "de Monte-Carlo". Ce qui est répréhensible, c'est de n'en faire qu'une et de la présenter comme une certitude.

La question qui se pose, lorsqu'on fait des simulations, est d'évaluer la "représentativité" de ces simulations : a-t-on exploré tous les cas possibles, une majorité, ou une proportion infime ? Cette question est très complexe. Dans le cas présent, nous avons 75 classes (chacune avec probabilité différente) ; si nous voulons générer une seule donnée, il y a donc 75 possibilités. Si nous voulons en générer deux, il y aura $75\times 75=5\,625$ possibilités et ainsi de suite : le nombre croît très vite. Si nous voulons réaliser des simulations crédibles en nombre raisonnable, il sera nécessaire de réduire le nombre de classes.

Si nous sommes en présence des résultats financiers d'une entreprise, la première classe signifie "résultats très mauvais" et a probabilité (mettons) 1/10. Nous serons inquiets si les résultats sont très mauvais trois fois de suite. Simuler une telle situation (trois résultats successifs) se fait avec probabilité $p=\left(\frac{1}{10}\right)^3=\frac{1}{1\,000}$. A chaque simulation, nous avons probabilité $1-p$ de ne pas voir la situation dramatique ; au bout de N simulations, cette probabilité est $(1-p)^N$; la probabilité de voir la situation dramatique est donc $1-(1-p)^N$ et elle est supérieure à 1/2 dès que $N\geq\frac{Log(1/2)}{Log(1-p)}\approx 693.$

Une entreprise sait très bien quelles sont les situations défavorables pour son bilan, mais les simulations "à l'aveugle" sont systématiquement utilisées pour les analyses de risques faites par les banques (Bâle III) et les compagnies d'assurance (Solvabilité II).

De manière générale, les simulations sont très utiles pour estimer une quantité qui dépend d'une valeur moyenne, comme l'intégrale d'une fonction : c'est une application directe de la loi des grands nombres. Mais leur intérêt est fortement réduit s'il s'agit d'identifier les situations à risque, comme dit plus haut, parce que si ces situations sont de faible probabilité (comme c'est le cas en général), les simulations ne les trouveront pas.

Si un code de calcul comporte 40 paramètres, chacun pouvant prendre 10 valeurs, l'espace des configurations comporte 10^{40} possibilités. Faire un milliard de runs paraît énorme, mais ne représente qu'une proportion de $\frac{10^9}{10^{40}}=10^{-31}$ de la réalité. Autrement dit, si une situation dangereuse existe quelque part, on ne la trouvera jamais de cette manière.