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

Traitement des bases de données :

méthodes probabilistes pour l'extraction et le stockage

des informations essentielles

par Bernard Beauzamy

avril 2026

Résumé Opérationnel

Les bases de données recueillies par les entreprises peuvent comporter des dizaines de colonnes (autant que de paramètres enregistrés) et des centaines de milliers de lignes (autant que de mesures).

Nous montrons ici comment des méthodes fondamentalement probabilistes, tenant compte des incertitudes sur les données, permettent de mettre la base de donnée sous un format léger, directement exploitable pour déterminer les paramètres prépondérants et les réglages optimaux.

Ce format est également recommandé pour la conservation de l'information. Ce prétraitement permet d'assurer que les données recueillies sont pertinentes et de bonne qualité. Un exemple est celui de la base de données "maintenances", que nous avons contribué à définir, à la demande de Bouygues Energies & Services : elle concernait le recueil de toutes les inspections réalisées sur le Tribunal Judiciaire de Paris ; le bâtiment est neuf, mais cette base de données permettra d'optimiser les maintenances lorsqu'elles deviendront nécessaires.

I. Présentation du besoin

Une entreprise industrielle surveille en permanence ses process : températures à toutes les étapes, pressions, alimentation en entrée, quantités en sortie, densités, vitesses de déplacement, etc. Il peut y avoir des dizaines de paramètres, mesurés quelquefois toutes les secondes. La base de données est souvent énorme.

L'entreprise voudrait savoir pourquoi, à certains moments, le process ne donne pas satisfaction : la qualité en sortie n'est pas ce qu'elle devrait être, conduisant à des rejets, à des non-qualités. Ou bien, et c'est plus optimiste ! pourquoi à certains moments le process donne satisfaction : est-ce un hasard heureux ? La question ne porte pas sur le réglage d'un paramètre en particulier, mais sur le réglage conjoint d'un grand nombre de paramètres, quelquefois tous les paramètres disponibles.

Cette préoccupation n'est pas propre aux entreprises industrielles : on la rencontre aussi chez les entreprises du secteur tertiaire. Pour une compagnie d'assurance, par exemple, les enregistrements portent sur le portefeuille, sur les sinistres, par catégorie, le paiement des primes, etc. Pourquoi certaines catégories d'assurés régressent-elles ? pertes d'effectif ou baisse du revenu généré ?

Pour le secteur des transports, un exemple fréquent concerne la préoccupation liée à l'état des équipements : état des rails, des chaussées, des véhicules, etc. : on cherche à définir un plan d'inspection pour l'année suivante, de manière à mettre en évidence les situations les plus critiques. Presque toujours, le nombre de paramètres enregistrés est élevé : la nature du matériau, la date de pose, l'usage, les conditions climatiques, etc. On voudrait déterminer les zones les plus critiques (celles à inspecter en premier) et expliquer pourquoi elles le sont (valeurs associées pour les paramètres prépondérants).

Comme on va le voir, ce qui pose problème n'est pas l'abondance des mesures, quelque milliers, millions, voire milliards, c'est le nombre de paramètres que l'on enregistre.

Outre les paramètres qui décrivent l'état du système, on dispose d'une variable d'intérêt, que l'on cherche à optimiser. Suivant les cas, elle peut décrire le taux de rejet des pièces (on voudrait qu'il soit le plus faible possible), le revenu financier (le plus fort possible), la consommation énergétique, la rapidité d'exécution, l'efficacité du maintien en condition opérationnelle, etc.

Enfin, de manière générale, beaucoup d'entreprises recueillent des données par principe, sans bien savoir quel en sera l'usage ; les variables d'intérêt seront définies dans l'avenir, en fonction des besoins ; dans ces conditions, il faut veiller à disposer d'un prétraitement, qui va garantir que les données recueillies sont pertinentes et de bonne qualité. Il faut aussi que la base de données ainsi traitée soit extensible, pour permettre l'ajout de nouvelles données à chaque fois qu'elles deviennent disponibles.

II. Description abrégée de la méthode

On part d'une base de données ordinaire, telle qu'en recueillent communément les entreprises. Dans l'exemple traité ci-dessous, 300 000 lignes, 4 paramètres et une variable d'intérêt. Le traitement se fait en deux étapes.

A. Première étape

Pour chaque paramètre et pour la variable d'intérêt, on définit des "classes d'appartenance" : on ne conserve plus la valeur exacte de la variable, mais le numéro de la classe à laquelle elle appartient. Ceci a deux avantages :

- Cela permet de vérifier que chaque donnée appartient bien à une classe, et donc n'est pas aberrante ;

- Cela permet de tenir compte des incertitudes sur les données.

L'appartenance à des classes est codée sous la forme d'un mot formé de lettres. Par exemple, si X_1 appartient à la 3ème classe, X_2 à la première, X_3 à la 4ème, X_4 à la seconde, on retient le mot CADB.

A la fin de la première étape, on dispose d'une base de données intermédiaire, consistant en la liste des mots et le nombre d'occurrences de chaque mot.

Voici un exemple, tiré du traitement fait plus loin :

mot(X)	Y	nb occurrences
AAHB	9	1
ABGB	1	33
ABGB	2	42

A ce stade, la base de données ne fait plus que 2049 lignes, contre 300 000 initialement.

B. Seconde étape

On dispose la base de données de telle sorte que, pour un mot donné (réglage des paramètres), on ait en lignes toutes les valeurs possibles de la variable d'intérêt, avec leur probabilité. Voici un exemple :

X1	X2	X3	X4	Y=1	Y=2	Y=3	Y=4	Y=5	Y=6	Y=7	Y=8	Y=9	Y=10	total
1	1	8	2	0	0	0	0	0	0	0	0	1	0	1
1	2	7	2	33	42	25	36	34	32	38	28	28	32	328
1	2	7	3	0	2	2	3	4	1	6	1	1	1	21

A partir de là, il est immédiat de répondre aux questions que se posent les entreprises : le réglage est-il optimal ? est-il robuste ? quels seraient les meilleurs réglages ? Tout ceci est obtenu à partir de la loi conditionnelle de Y connaissant X, qui est explicitement donnée par ce tableau. Dans notre exemple, il n'a plus que 268 lignes.

III. Difficultés rencontrées

Même si la préoccupation est commune, elle est rarement adressée de manière satisfaisante. Les entreprises accumulent d'énormes bases de données, et ces données ne sont jamais validées (comment savoir si les capteurs ont fonctionné correctement ?) ; de plus, elles ne sont jamais exploitées de manière rationnelle, si bien que les décisions d'investissement sont prises de manière irrationnelle, en fonction des sensibilités des uns et des autres, mais sans référence à un historique.

Nous montrons ici comment des méthodes probabilistes très simples permettent l'exploitation des bases de données recueillies, de manière à répondre aux questions posées : que faut-il surveiller dans le process industriel, dans le portefeuille de clients, dans le patrimoine immobilier ou matériel ?

Nous allons procéder de manière extrêmement progressive, en expliquant bien où sont les difficultés et comment les surmonter.

IV. La situation de départ et la situation d'arrivée

Au départ, l'entreprise dispose d'une base de données, où l'on a enregistré des paramètres $X_1, X_2, ...$ et une variable d'intérêt notée Y. La base de données comporte un grand nombre d'enregistrements, ici 300 000 dans le cas qui nous occupe.

A l'arrivée, c'est-à-dire après traitement, la base de données sera considérablement simplifiée ; elle aura le format suivant :

- Un premier groupe de colonnes, autant que de paramètres de réglage (comme dans la base initiale) ;

- Un second groupe de colonnes, décrivant les valeurs possibles de la variable d'intérêt ;

- Pour chaque ligne, le réglage des paramètres et la probabilité de chaque valeur de Y.

La base de données ne contient plus, en général, que quelques milliers de lignes (seulement 250 dans notre exemple) ; chaque ligne permet immédiatement de voir quelle est la valeur du réglage correspondant et de choisir le meilleur.

paramètres de réglage				variable d'intérêt		
X1	X2	X3	X4...	y1	y2	y3...
k1	k2	k3	k4...	p1	p2	p3...

Avant de procéder au traitement des bases de données, commençons par nous interroger sur la notion de réglage : que faut-il entendre par là ?

V. Le cas d'un seul paramètre et d'une variable d'intérêt

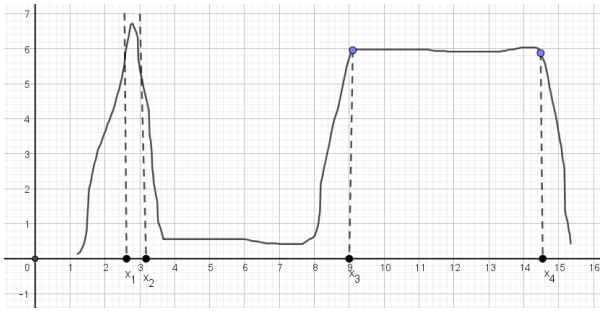
Commençons par la situation la plus simple : on dispose d'un seul paramètre, noté X, et d'une variable d'intérêt Y. Le fonctionnement du process a permis l'enregistrement de N valeurs du couple (x, y), notées (x_n, y_n), $n = 1, ..., N$. On voudrait connaître le réglage de X qui maximise la valeur de Y.

A. Une solution trompeuse

En apparence, la réponse est immédiate et déterministe : parmi tous les couples (x_n, y_n), on choisit celui pour lequel la valeur y_n est maximale, la valeur correspondante x_n est le réglage voulu.

Pourtant, cette solution "optimale" est rarement satisfaisante, parce qu'on n'en connaît ni la probabilité, ni la robustesse. Si le réglage diffère même très faiblement du réglage optimal, que devient la variable d'intérêt ? Elle peut s'effondrer, si le réglage est "pointu". Seconde objection : la détermination d'un réglage présenté comme "optimal" ne nous renseigne pas sur ce que fait le process en général : peut-être est-il, dans beaucoup de cas, très loin de cet optimum.

Le graphique ci-dessous illustre ceci :



Sur l'intervalle $[x_1, x_2]$, Y est très grand, mais cet intervalle est très petit et la valeur s'effondre subitement à gauche et à droite. Sur l'intervalle $[x_3, x_4]$, la fonction est légèrement plus petite, mais cet intervalle est beaucoup plus grand. Il vaudra mieux mettre le réglage optimal dans cet intervalle.

On voit ici que la réponse utile à l'entreprise n'est pas la solution déterministe qui maximise Y ; une analyse probabiliste est nécessaire, même dans le cas d'un seul paramètre de contrôle.

En outre, dans l'immense majorité des cas, on ne dispose pas d'une description fonctionnelle $y = f(x)$ du process en fonction du réglage ; on dispose simplement d'une suite de valeurs enregistrées (x_n, y_n) ; ce n'est pas du tout la même chose. Rien ne dit que, pour une valeur déterminée de x il y aura une valeur unique de y : bien au contraire, en général, il y a une variabilité du process, pour un réglage donné, et cela est dû au fait que beaucoup d'autres paramètres interviennent, qui n'ont pas été mesurés. L'approche à retenir est donc nécessairement probabiliste ; elle se traduit par la loi conjointe du couple (X, Y) comme nous allons maintenant l'expliquer.

B. Loi conjointe

Les valeurs possibles de X sont décomposées en classes ; notons-les $I_1,...,I_{N_X}$. Par exemple, s'il s'agit d'âges, on aura les classes 0-10 ans, 10 ans – 20 ans, etc. S'il s'agit de vitesses, on aura les classes 0 – 10 m/s, 10 – 20 m/s, etc. La définition des classes est à la discrétion de l'utilisateur ; elle est en général facile. Ensuite, on oublie les valeurs précises pour X et on ne retient que le numéro de la classe où elles sont tombées : cela s'appelle réaliser un histogramme. On fait de même pour Y avec les classes $J_1,...,J_{N_Y}$. Avec ces notations, N_X est le nombre de classes pour X et N_Y est le nombre de classes pour Y.

La loi conjointe consiste en la réalisation d'un tableau rectangulaire, comportant $N_X \times N_Y$ classes ; dans chaque cellule, on met le nombre de fois où le couple (x_n, y_n) tombe dans cette cellule.

Par exemple, dans le tableau suivant, nous avons un total de 15 mesures, réparties ainsi selon les classes (les x sont en abscisse, les y en ordonnée) :

y3	1	0	1	0
y2	2	1	2	4
y1	0	0	3	1
	x1	x2	x3	x4

Pour chaque valeur de x, il y a trois valeurs possibles de y, dans ce cas. La question est : quelle valeur choisir pour x ? La réponse, de nature probabiliste, est la valeur de x qui maximise l'espérance conditionnelle. Expliquons ceci.

Lorsqu'on donne à x une valeur quelconque, x_1, x_2, x_3, x_4, on ne sait pas a priori quelle valeur pour Y le process va donner : il y a trois valeurs possibles. Mais toutes ne sont pas équiprobables. Par exemple, si nous choisissons x_1, y_1 n'apparaît pas, y_2 apparaît deux fois et y_3 une fois. On dira donc que la loi de probabilité de Y sachant $X = x_1$ est :

valeur	y1	y2	y3
proba	0	2/3	1/3

L'espérance de cette loi est $E(Y \mid X = x_1) = \frac{2y_2}{3} + \frac{y_3}{3}$. On l'appelle "espérance conditionnelle".

La réponse à la question posée est alors évidente : nous allons choisir la valeur de x pour laquelle cette espérance conditionnelle est la plus forte possible. Traitons un exemple numérique, avec des valeurs précises pour Y.

30	1	0	1	0
20	2	1	2	4
10	0	0	3	1
Y/X	x1	x2	x3	x4

Y\X	x1	x2	x3	x4
30	1	0	1	0
20	2	1	2	4
10	0	0	3	1
somme valeurs Y	3,00	1,00	6,00	5,00

La loi conditionnelle s'obtient en divisant chaque colonne par la somme des valeurs de Y correspondante ; on obtient :

Y\X	x1	x2	x3	x4
30	0,33	0,00	0,17	0,00
20	0,67	1,00	0,33	0,80
10	0,00	0,00	0,50	0,20

L'espérance de gain correspondant à chaque situation est le produit de la valeur de Y par sa probabilité ; on obtient le tableau suivant :

Y\X	x1	x2	x3	x4
30	10,00	0,00	5,00	0,00
20	13,33	20,00	6,67	16,00
10	0,00	0,00	5,00	2,00

et l'espérance conditionnelle est la moyenne par colonne :

Y\X	x1	x2	x3	x4
30	10,00	0,00	5,00	0,00
20	13,33	20,00	6,67	16,00
10	0,00	0,00	5,00	2,00
espérance conditionnelle	7,78	6,67	5,56	6,00

On constate donc, dans cet exemple, que c'est la valeur x_1 qu'il faut retenir pour X. Si cette valeur est retenue, nous pouvons espérer pour Y une valeur moyenne égale à 7.78 : c'est mieux qu'avec les autres choix possibles pour X.

C. Le process est-il bien réglé ?

Nous constatons, dans le traitement de cet exemple, que la probabilité de x_1 n'intervient pas, or nous avons conclu que x_1 était le meilleur réglage possible. Il y a ici un moyen de déterminer si, dans l'état actuel des choses, le process est bien réglé ou non.

Commençons par déterminer la probabilité de chaque x_i : c'est le nombre de fois où il revient, divisé par le nombre total d'occurrences, ici 15. Nous avons le tableau suivant :

	x1	x2	x3	x4
y3	1	0	1	0
y2	2	1	2	4
y1	0	0	3	1
somme	3	1	6	5
proba	0,20	0,07	0,40	0,33

La valeur x_1 revient 3 fois sur 15 et a donc probabilité 0.20 (voir tableau ci-dessus). Nous en concluons que ce process est horriblement mal réglé, puisque la valeur de réglage qui conduit à un optimum de fabrication ne revient qu'une fois sur cinq !!!

VI. Réalisation d'un histogramme

Nous allons maintenant voir comment procéder lorsque le process dépend d'un nombre plus élevé de paramètres. La première chose à faire est de construire la loi conjointe. Pour cela, rappelons la démarche à suivre pour réaliser un histogramme.

A. Définition des classes

Soit X une variable quelconque, pour laquelle les valeurs $x_n,\ n=1,...,N,$ ont été enregistrées. La réalisation de l'histogramme consiste à répartir ces valeurs en classes. Pour simplifier, on dira que la première classe commence en 0 (c'est le cas le plus fréquent) et on notera w (width) la largeur des classes : toutes les classes ont la même largeur.

Soit x une valeur quelconque ; elle appartient à la $k^{ème}$ $(k=1,2,...)$ classe si :

$$(k-1)w \le x < kw,$$

(par convention, les classes sont des intervalles fermés à gauche et ouverts à droite).

Les inégalités ci-dessus sont équivalentes à :

$$k-1 \le \frac{x}{w} < k,$$

ce qui se traduit par :

$k-1=\text{int}\left(\frac{x}{w}\right)$, où "int" désigne la partie entière. Le nombre $k,$ caractérisant la classe de $x,$ est donc défini par :

$$k = \mathrm{int}\left(\frac{x}{w}\right) + 1 \quad (1)$$

B. Construction de l'histogramme

Soit K le nombre de classes pour X ; K se calcule immédiatement à partir de la plus petite valeur, de la plus grande et de la taille des classes. On définit, en VBA sous Excel :

dim histo(1 to K) as long
for n = 1 to Ntot 'le nombre total de mesures
$k = \mathrm{int}\left(\frac{x_n}{w}\right) + 1$
histo(k)=histo(k)+1
next n

et à la fin, pour chaque k, $histo(k)$ indique le nombre de fois où l'on est tombé dans la $k^{ème}$ classe.

La construction est la même si nous avons deux variables X_1, X_2 :

dim $histo(1\ to\ K_1, 1\ to\ K_2)\ as\ long$
for n = 1 to Ntot 'le nombre total de mesures
$k_1 = \mathrm{int}\left(\frac{x_{1,n}}{w_1}\right) + 1, k_2 = \mathrm{int}\left(\frac{x_{2,n}}{w_2}\right) + 1,$
$histo(k_1, k_2) = histo(k_1, k_2) + 1$
next n

et à la fin, pour chaque couple (k_1, k_2), $histo(k_1, k_2)$ indique le nombre de fois où l'on est tombé dans le rectangle formé par la $k_1^{ème}$ classe pour X_1 et la $k_2^{ème}$ classe pour X_2.

Cette construction s'étend immédiatement à un nombre quelconque de variables ; on définit :
dim $histo(1\ to\ K_1, ..., 1\ to\ K_p)\ as\ long$ s'il y a p variables $X_1, ..., X_p$.

La réalisation informatique de l'histogramme est facile : il suffit de "balayer" la base de données initiale, une seule fois, et d'incrémenter l'indicateur correspondant à chaque mesure. La question est ensuite : que faire avec cet histogramme ? et d'abord : comment le représenter ?

C. Représentation de l'histogramme

Si nous n'avons qu'un seul paramètre X et une variable d'intérêt Y, la représentation est facile : l'histogramme est un rectangle de dimensions $K_X \times K_Y$ (K_X est le nombre de classes pour X, K_Y le nombre de classes pour Y). Dans la cellule k_X, k_Y, on met le nombre de fois, parmi les

N mesures, où l'on a eu simultanément $X = k_X, Y = k_Y$; la somme de ces nombres doit faire N. Ensuite, divisant tous ces nombres par N, on obtient la probabilité $P\{X = k_X, Y = k_Y\}$. Cela s'appelle la loi conjointe du couple (X, Y).

Si nous avons deux paramètres X_1, X_2 et une variable d'intérêt Y, la représentation est un parallélépipède de dimensions $K_1 \times K_2 \times K_Y$. Dans la cellule (k_1, k_2, k_Y), on met le nombre de fois, parmi les N mesures, où l'on a eu simultanément $X_1 = k_1, X_2 = k_2, Y = k_Y$; la somme de ces nombres doit faire N. Ensuite, divisant tous ces nombres par N, on obtient la probabilité $P\{X_1 = k_1, X_2 = k_2, Y = k_Y\}$. Cela s'appelle la loi conjointe du triplet (X_1, X_2, Y).

En dimension supérieure, c'est-à-dire si le nombre p des paramètres est ≥ 3, le concept est le même, mais la visualisation n'est plus possible. Formellement, l'histogramme est un "hyper-cube" dans un espace de dimension $p+1$ (p paramètres, plus la variable d'intérêt Y). La loi conjointe des $p+1$ paramètres se définit de la même manière, mais il n'y a plus d'interprétation visuelle. Nous y reviendrons plus loin.

VII. Utilisation de l'histogramme dans le cas de deux paramètres

Voyons maintenant le cas où notre process dépend de deux réglages (X_1, X_2) et produit une variable d'intérêt Y. On recherche le réglage du couple (X_1, X_2) qui maximise Y. On voudrait aussi savoir : dans notre process, tel qu'il est actuellement, ce réglage optimal est-il fréquent ou est-il rarissime ? En d'autres termes, sommes-nous proches ou éloignés du réglage optimal, en moyenne ?

Nous réalisons l'histogramme du triplet (X_1, X_2, Y) selon les règles indiquées plus haut. Il y a K_1 classes pour X_1, K_2 classes pour X_2, K_Y classes pour Y. La représentation visuelle est donc un parallélépipède en dimension 3, avec X_1 en abscisse, X_2 en ordonnée et Y en cote (vertical).

Prenons une cellule quelconque (k_1, k_2) du plan horizontal : le plan (X_1, X_2). Au-dessus de cette cellule, nous avons une colonne, représentant toutes les valeurs possibles de Y lorsque $X_1 = k_1$ et $X_2 = k_2$ et, dans chaque case, le nombre de fois où cela s'est produit. Notons N_{k_1,k_2} le nombre total de fois, parmi N, où l'on a observé simultanément $X_1 = k_1$ et $X_2 = k_2$. En divisant le nombre figurant dans chaque cellule de la colonne par N_{k_1,k_2}, nous obtenons une loi de probabilité de Y sachant $X_1 = k_1$ et $X_2 = k_2$; elle se note "loi de $Y \mid (X_1 = k_1, X_2 = k_2)$". C'est une loi "conditionnelle", puisqu'on fait l'hypothèse $X_1 = k_1$ et $X_2 = k_2$. Elle décrit la probabilité respective des diverses valeurs de Y dans cette configuration.

A partir de cette loi conditionnelle de Y sachant (X_1, X_2), on calcule l'espérance conditionnelle, selon la formule habituelle permettant de calculer l'espérance d'une loi : $E = \sum_i p_i y_i$, où les y_i sont les valeurs prises et les p_i sont les probabilités correspondantes.

Le réglage optimal sera la valeur du couple (X_1, X_2) pour laquelle cette espérance conditionnelle est maximale.

Si on s'intéresse à la robustesse du réglage du process, on commence par se demander : quelle est la probabilité de cette valeur précise du couple (X_1, X_2) ? Cette probabilité se définit comme le nombre de fois où elle revient, au sein des N mesures qui ont été faites.

On peut étendre cette définition de robustesse : on se fixe un seuil, mettons 75% de l'optimum et on calcule la probabilité totale (au sens précédent) des réglages qui dépassent ce seuil. En d'autres termes, si $y_{\max}$ est l'optimum recherché, combien de fois, parmi nos N mesures, avons-nous dépassé $0.75 y_{\max}$? Si cela revient souvent, le process est bien réglé ; si cela revient très rarement, les bons réglages relèvent du miracle.

VIII. Cas d'un nombre élevé de paramètres

Une difficulté apparaît si le nombre de paramètres est élevé (nombre de colonnes dans la base de données) : il peut être difficile de construire la loi conjointe de l'ensemble, parce que la mémoire est insuffisante, et cette loi conjointe est en définitive très peu utile.

Si on veut suivre l'approche précédente, avec p paramètres $X_1, ..., X_p$ et une variable d'intérêt Y, on définira :

$$histo(1\ to\ K_1, 1\ to\ K_2, ..., 1\ to\ K_p, 1\ to\ K_Y)\ as\ long$$

Pour la représentation, on parle d'hypercubes ; c'est un produit $I_1 \times I_2 \times \cdots \times I_{p+1}$; $p+1$ est ici le nombre de colonnes de la base de données. Le mot "cube" est trompeur : la largeur des intervalles n'est pas la même pour toutes les dimensions. Avec cette définition, nous aurons créé $K_1 \times K_2 \times \cdots \times K_p \times K_Y$ hypercubes.

Par exemple, si on dispose de 11 paramètres et d'une variable d'intérêt, chacun prenant 10 valeurs possibles, l'espace des configurations a taille 10^{12} ; il y a 10^{12} produits élémentaires (cellules) $I_1 \times I_2 \times \cdots \times I_{12}$.

Une première difficulté tient au fait que, lorsque le nombre de colonnes est élevé, pour des raisons de mémoire, Excel n'accepte pas une définition du type :

$$histo(1\ to\ K_1, 1\ to\ K_2, ..., 1\ to\ K_p, 1\ to\ K_Y)\ as\ long$$

Cela dépend évidemment de la mémoire disponible et des valeurs respectives de $K_1,...,K_p,K_Y$.

Une seconde difficulté, encore plus sérieuse, est que, si on dispose d'un million de données, la plupart des configurations seront vides. Si on dispose de 10^6 données, et que l'on crée 10^{12} configurations, la plupart des configurations seront vides et on les a créées pour rien.

En d'autres termes, on aura créé une énorme quantité d'hypercubes pour rien ; l'exploration sera impossible en pratique. Il est essentiel que l'exploration soit en relation avec la liste des données (de l'ordre du million, exploration facile) et non avec la structure des hypercubes.

IX. Traitement détaillé d'un exemple concret

Nous allons suivre pas à pas le traitement d'un exemple concret, pour illustrer la méthode.

A. Présentation de l'exemple

Dans cet exemple, nous disposons de 300 000 lignes (y compris l'entête) ; chaque mesure porte sur 4 paramètres et une variable d'intérêt. Les premières lignes sont :

X1	X2	X3	X4	Y
0,194	1	1986	7	9
0,194	7	1986	6	2
0,194	7	1986	5	1
0,194	7	1986	4	10

X_1, X_2, X_4 sont des caractéristiques, X_3 est une date (mise en service de l'équipement) et Y est une note de qualité, entre 1 et 10. Plus précisément, voici le minimum et le maximum pour chaque paramètre et pour Y :

	X1	X2	X3	X4	Y
min	0,002	1	1875	0	1
max	52,46	9	2021	30	10

B. Mise en mémoire des données

On commence par mettre l'ensemble des données en mémoire, ce qui évite d'ouvrir le classeur Excel à chaque fois.

```
Dim Ntot As Long
Ntot = Sheets(1).Range("A1").End(xlDown).Row
'le nombre total de lignes, ici 300 000
Dim data As Variant '
data = Sheets(1).Cells(1, 1).Resize(Ntot, 5)
Dim w1 As Double
w1 = 10 'taille classe 1er param
Dim w2 As Double
```

```
w2 = 1 'taille classe 2eme param
Dim w3 As Double
w3 = 15 'taille classe 3eme param
Dim w4 As Double
w4 = 5 'taille classe 4eme param
Dim w5 As Double
w5 = 1 'taille classe 5eme param
```

Les tailles des classes sont déterminées en fonction des valeurs min et max de chaque paramètre. Pour Y, il y a une classe pour chaque note entre 1 et 10.

On notera simplement X à la place de (X_1, X_2, X_3, X_4).

C. Loi conjointe

On définit :

```
Dim loiXY(1 To 6, 1 To 9, 1 To 10, 1 To 7, 1 To 10) As Integer
'le nombre de fois où XY tombe dans chaque classe ; cela s'appelle la "loi conjointe" de X, Y.
Dim k1 As Integer
Dim k2 As Integer
Dim k3 As Integer
Dim k4 As Integer
Dim k5 As Integer

Dim n As Long
For n = 2 To Ntot
k1 = Int(data(n, 1) / w1) + 1 'le numéro de la classe pour X1
k2 = Int(data(n, 2) / w2)
k3 = Int((data(n, 3) - 1875) / w3) + 1
k4 = Int(data(n, 4) / w4) + 1
k5 = Int(data(n, 5) / w5)
data(n, 6) = k1 'ou Chr(k1 + 65) si on préfère une lettre plutôt qu'un chiffre
data(n, 7) = k2 'Chr(k2 + 65)
data(n, 8) = k3 'Chr(k3 + 65)
data(n, 9) = k4 'Chr(k4 + 65)
data(n, 10) = k5 ' Chr(k5 + 65)
loiXY(k1, k2, k3, k4, k5) = loiXY(k1, k2, k3, k4, k5) + 1
Next n
```

Le tableau loiXY dépend de 5 paramètres et, dans chaque case, indique le nombre de fois (parmi Ntot) où cette case est atteinte.

Il est bon de vérifier :

```
Dim sum As Double
For k1 = 1 To 6
For k2 = 1 To 9
```

```
For k3 = 1 To 10
For k4 = 1 To 7
For k5 = 1 To 10
sum = sum + loiXY(k1, k2, k3, k4, k5)
Next k5
Next k4
Next k3
Next k2
Next k1
MsgBox sum
```

et on trouve bien sum = 299 999, comme prévu.

D. Une première réduction de la base de données

On va maintenant constituer un tableau (sur sheets(2)) qui contient les classes, pour chaque paramètre, et le nombre d'occurrences. Ce tableau représente une étape essentielle : à partir de là, il n'est plus utile d'entrer dans la combinatoire, vue plus haut, résultant du nombre de paramètres et des valeurs prises par chacun. Voici les premières lignes du tableau :

X1	X2	X3	X4	Y	nb occurrences	mot(X)	
1	1	8	2	9	1	AAHB	*
1	2	7	2	1	33	ABGB	
1	2	7	2	2	42	ABGB	
1	2	7	2	3	25	ABGB	
1	2	7	2	4	36	ABGB	
1	2	7	2	5	34	ABGB	
1	2	7	2	6	32	ABGB	
1	2	7	2	7	38	ABGB	
1	2	7	2	8	28	ABGB	
1	2	7	2	9	28	ABGB	
1	2	7	2	10	32	ABGB	*
1	2	7	3	2	2	ABGC	

Les quatre premières colonnes contiennent le numéro de la classe où se tient chaque paramètre, la 5ème est la valeur prise par Y en ce cas, la 6ème est le nombre d'occurrences de cette situation dans la base de données. La 7ème est un "condensé" des 4 premières : au lieu de la liste 1,1,8,2, on écrit le mot AAHB : chaque chiffre est remplacé par la lettre de rang correspondant. Ceci va simplifier le tri et la recherche.

Voici le code :

```
Dim data2 As Variant
data2 = Sheets(2).Cells(1, 1).Resize(Ntot, 8)

Dim m As Integer
m = 2
For k1 = 1 To 6
```

```
For k2 = 1 To 9
For k3 = 1 To 10
For k4 = 1 To 7
For k5 = 1 To 10
If loiXY(k1, k2, k3, k4, k5) > 0 Then
data2(m, 1) = k1
data2(m, 2) = k2
data2(m, 3) = k3
data2(m, 4) = k4
data2(m, 5) = k5
data2(m, 6) = loiXY(k1, k2, k3, k4, k5)
data2(m, 7) = Chr(k1 + 64) & Chr(k2 + 64) & Chr(k3 + 64) & Chr(k4 + 64) 'le mot associé à chaque X
m = m + 1
End If
Next k5
Next k4
Next k3
Next k2
Next k1
Sheets(2).Cells(1, 1).Resize(m, 8) = data2
```

L'intérêt évident est que ce tableau ne comporte plus que 2 049 lignes (entête inclus) contre 300 000 précédemment. On vérifie que la somme des nombres en colonne 6 fait bien 299 999.

```
Dim Ntot2 As Long
Ntot2 = Sheets(2).Range("A1").End(xlDown).Row
```

Dans la suite, on oublie complètement la base de données initiale et on ne travaille plus que sur ce tableau.

Voici la répartition selon le nombre d'occurrences :

k	Proba(Y>k)		k	Proba(Y=k)
0	1,000		1	0,100
1	0,900		2	0,100
2	0,800		3	0,099
3	0,701		4	0,100
4	0,601		5	0,100
5	0,501		6	0,099
6	0,401		7	0,100
7	0,302		8	0,101
8	0,201		9	0,100
9	0,101		10	0,101

Le programme est le suivant :

```
Dim k As Integer
```

```
For k = 0 To 9
For n = 2 To Ntot2
If data2(n, 5) > k Then
count = count + data2(n, 6)
End If
Next n
Sheets(2).Cells(k + 2, 15) = count / (Ntot - 1)
Sheets(2).Cells(k + 2, 14) = k
count = 0
Next k
Sheets(2).Cells(1, 14) = "k"
Sheets(2).Cells(1, 15) = "Proba(Y>k)"

Dim LoiY(1 To 10) As Long
For n = 2 To Ntot2
LoiY(data2(n, 5)) = LoiY(data2(n, 5)) + data2(n, 6)
Next n

For k = 1 To 10
Sheets(2).Cells(k + 1, 18) = LoiY(k) / (Ntot - 1)
Sheets(2).Cells(k + 1, 17) = k
Next k
sum = 0

Sheets(2).Cells(1, 17) = "k"
Sheets(2).Cells(1, 18) = "Proba(Y=k)"
```

On détermine la loi de $X = (X_1, X_2, X_3, X_4)$:

```
Dim LoiX(1 To 6, 1 To 9, 1 To 10, 1 To 7) As Long 'la loi de X = (X1,X2,X3,X4)

For n = 2 To Ntot2 'nombre d'occurrences
LoiX(data2(n, 1), data2(n, 2), data2(n, 3), data2(n, 4)) = LoiX(data2(n, 1), data2(n, 2), data2(n,
3), data2(n, 4)) + data2(n, 6)
Next n

'pour vérifier
For k1 = 1 To 6
For k2 = 1 To 9
For k3 = 1 To 10
For k4 = 1 To 7
sum = sum + LoiX(k1, k2, k3, k4)
Next k4
Next k3
Next k2
Next k1
```

Dans le tableau ci-dessus, certaines lignes comportent un "*" en dernière colonne. Ce sont les lignes où le réglage change. Par exemple, en ligne 12, on a ABGB, puis en ligne 13 ABGC. On s'intéresse aux valeurs possibles de Y à l'intérieur d'un même réglage ; il est donc nécessaire de savoir quand celui-ci commence et cesse. C'est l'objet de la liste suivante, déduite des "*" du tableau :

liste	debut	fin
2	2	2
12	3	12
21	13	21
31	22	31
41	32	41
51	42	51
61	52	61
64	62	64
74	65	74
77	75	77
87	78	87
90	88	90

Voici le code :

```
m = 1
Dim liste As Variant
liste = Sheets(2).Cells(1, 11).Resize(Ntot2, 3)
Dim i As Integer
For i = 2 To Ntot2
If data2(i, 7) <> data2(i + 1, 7) Then
data2(i, 8) = "*"
Sheets(5).Cells(m, 10) = i
liste(m, 1) = i
m = m + 1
End If
Next i
Sheets(2).Cells(1, 1).Resize(Ntot2, 8) = data2
Sheets(2).Cells(2, 10).Resize(Ntot2, 1) = liste
Dim Ntot3 As Integer
Ntot3 = m - 1
Sheets(2).Cells(1, 10) = "liste"
'MsgBox Ntot3
Sheets(2).Cells(1, 11) = "debut"
Sheets(2).Cells(1, 12) = "fin"
Sheets(2).Cells(2, 11) = liste(1, 1)
liste(1, 2) = liste(1, 1)
liste(1, 3) = liste(1, 1)
Sheets(2).Cells(2, 12) = liste(1, 1)
For i = 1 To Ntot3 - 1
Sheets(2).Cells(i + 2, 11) = liste(i, 1) + 1
```

```
liste(i + 1, 2) = liste(i, 1) + 1
Sheets(2).Cells(i + 2, 12) = liste(i + 1, 1)
liste(i + 1, 3) = liste(i + 1, 1)
Next i
Sheets(5).Cells(1, 1).Resize(Ntot3, 3) = liste
sum = 0
```

E. Le tableau sous forme finale

Voici maintenant la forme finale du traitement de la base de données d'origine :

X1	X2	X3	X4		Y=1	Y=2	Y=3	Y=4	Y=5	Y=6	Y=7	Y=8	Y=9	Y=10	total
1	1	8	2		0	0	0	0	0	0	0	0	1	0	1
1	2	7	2		33	42	25	36	34	32	38	28	28	32	328
1	2	7	3		0	2	2	3	4	1	6	1	1	1	21
1	2	8	1		60	78	67	76	76	71	71	61	64	70	694
1	2	8	2		116	118	115	130	122	129	139	127	139	125	1 260
1	2	9	1		171	170	164	165	184	176	168	162	199	186	1 745
1	2	9	2		9	10	4	9	15	10	11	11	11	13	103
1	2	9	3		0	1	0	0	2	0	1	0	0	0	4
1	2	10	1		196	199	162	183	175	191	200	191	199	203	1 899

Le tableau comporte 268 lignes seulement. Les 4 premières colonnes reçoivent les divers réglages pour les 4 paramètres : il y a une seule ligne par réglage. Les 10 colonnes suivantes comportent le nombre d'occurrences de chaque valeur de Y pour ce réglage. Par exemple, en ligne 3, pour le réglage 1,2,7,2, la valeur Y=33 est prise 33 fois. La dernière colonne est la somme des 10 précédentes ; elle indique le nombre total d'occurrences de Y, toutes valeurs confondues, pour ce réglage.

Pour obtenir la loi de probabilité, pour chaque ligne, il suffit de diviser le nombre d'occurrences par le nombre total (en dernière colonne) ; voici les deux premières lignes :

X1	X2	X3	X4		Y=1	Y=2	Y=3	Y=4	Y=5	Y=6	Y=7	Y=8	Y=9	Y=10	total
1	1	8	2		0	0	0	0	0	0	0	0	1	0	1
1	2	7	2		0,101	0,128	0,076	0,110	0,104	0,098	0,116	0,085	0,085	0,098	1

Le total des probabilités, pour chaque ligne, fait 1.

A partir de là, on détermine immédiatement l'espérance conditionnelle de Y dans chaque configuration : c'est la somme $E(Y)=\sum_{i=1}^{10} y_i p_i$. Voici la liste des espérances conditionnelles :

X1	X2	X3	X4	E(Y|X)
1	1	8	2	9,00
1	2	7	2	5,37
1	2	7	3	5,57
1	2	8	1	5,46
1	2	8	2	5,63
1	2	9	1	5,59
1	2	9	2	5,88
1	2	9	3	4,75
1	2	10	1	5,57

F. Réponses aux questions posées

Cette présentation finale permet de répondre de manière très simple à toutes les questions posées :

1. Combien de fois avons-nous la valeur maximale $Y=10$?

C'est la somme de toutes les valeurs de la colonne Y=10 dans le tableau ; on trouve 30 231 sur 299 999 mesures.

2. Quelles sont les réglages qui favorisent Y=10 ?

Il suffit de faire un tri décroissant selon la colonne Y=10. On trouve :

X1	X2	X3	X4	Y=1	Y=2	Y=3	Y=4	Y=5	Y=6	Y=7	Y=8	Y=9	Y=10	total
1	3	9	1	1 738	1 656	1 649	1 743	1 730	1 606	1 707	1 703	1 697	1 738	16 967
1	4	9	1	1 225	1 205	1 204	1 201	1 263	1 294	1 162	1 246	1 262	1 193	12 255
1	5	10	1	1 150	1 113	1 086	1 133	1 162	1 121	1 142	1 177	1 200	1 174	11 458
1	5	9	1	1 123	1 118	1 134	1 140	1 100	1 135	1 155	1 165	1 183	1 168	11 421
1	3	10	1	975	1 041	1 018	942	993	1 018	1 001	1 003	997	1 047	10 035
1	6	9	1	1 007	953	966	1 035	989	1 003	1 005	955	978	1 013	9 904
1	8	10	1	924	929	907	897	908	968	983	980	904	925	9 325
1	6	10	1	967	892	861	891	871	899	845	903	895	910	8 934
1	4	10	1	795	774	779	795	774	744	789	768	790	842	7 850

C'est le réglage 1,3,9,1 le plus favorable.

3. Combien de fois avons-nous des valeurs $Y \geq 7$?

Il suffit de faire le cumul sur les colonnes correspondantes. On trouve 120 413 fois sur 299 999 ; probabilité 0.4

4. Quel est le réglage qui donne la meilleure espérance conditionnelle ?

Il suffit de faire un tri décroissant sur le tableau des espérances conditionnelles ; on trouve :

X1	X2	X3	X4	E(Y|X)
3	3	9	2	10,00
4	2	10	1	9,33
1	1	8	2	9,00
2	9	5	3	9,00
2	9	10	1	9,00
1	8	1	5	8,50
1	5	3	4	8,00
1	5	3	5	7,50
2	7	7	2	7,50

On ne trouve qu'un seul réglage pour lequel l'espérance conditionnelle est maximale : c'est le réglage 3,3,9,2, qui n'intervient qu'une seule fois et donne la valeur $Y=10$. Pour tous les autres réglages des paramètres, il y a plusieurs valeurs possibles pour Y et de ce fait l'espérance conditionnelle est plus faible ; voir tableau ci-dessus.

X. En conclusion

Le format de données vu plus haut :

X1	X2	X3	X4	Y=1	Y=2	Y=3	Y=4	Y=5	Y=6	Y=7	Y=8	Y=9	Y=10	total
1	1	8	2	0	0	0	0	0	0	0	0	1	0	1
1	2	7	2	33	42	25	36	34	32	38	28	28	32	328
1	2	7	3	0	2	2	3	4	1	6	1	1	1	21

consistant à énumérer d'abord les réglages des paramètres (sous forme de classes), puis les valeurs possibles de la variable d'intérêt, répond bien aux préoccupations exposées au début :

- Ce format est extrêmement compact : le nombre de lignes est considérablement réduit ;
- On a procédé à une validation préliminaire de toutes les données : elles s'organisent en classes ;
- On peut constamment ajouter de nouvelles données.

Des compléments sur les méthodes probabilistes utilisées peuvent être trouvés dans les livres suivants :

XI. Références

[IEPE] Bernard Beauzamy : Introduction à l'étude des Probabilités Expérimentales. Ouvrage édité et commercialisé par la Société de Calcul Mathématique SA, ISBN 979-10-95773-02-3, ISSN 1767-1175. Relié, 192 pages, janvier 2023.

[MPPR] Bernard Beauzamy : Méthodes Probabilistes pour l'étude des phénomènes réels. Ouvrage édité et commercialisé par la Société de Calcul Mathématique SA, ISBN 2-9521458-0-6. ISSN 1767-1175, broché, 369 pages. Seconde Edition, juin 2016.