JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

R. BENINI

Les aspects arbitraires de l'interpolation des séries statistiques

Journal de la société statistique de Paris, tome 45 (1904), p. 374-386

<http://www.numdam.org/item?id=JSFS_1904__45__374_0>

III.

LES ASPECTS ARBITRAIRES DE L'INTERPOLATION DES SÉRIES STATISTIQUES.

§ 1. — Souvent pour la commodité de l'analyse ou par la présomption qu'un phénomène statistique considéré obéit à une loi simple, il convient de mettre en évidence la marche générale de la série et ses principales ondulations en se débarrassant, pour ainsi dire, des nombreuses petites oscillations qui sont ou que l'on suppose produites par de multiples influences perturbatrices. Nous recourons alors à l'interpolation dans le but de substituer à la série observée une série plus simple qui satisfasse à certaines conditions.

Pour les phénomènes périodiques qui présentent des maxima et des minima se répétant à intervalles égaux, il convient d'interpoler une fonction trigonométrique; pour les phénomènes non périodiques, une fonction algébrique entière est mieux indiquée

§ 2. — A ceux qui sont peu familiarisés avec ce genre de calcul, un exemple pratique d'interpolation d'une série périodique ne déplaira pas

On sait que le sinus d'un angle (c'est-à-dire le rapport de la perpendiculaire à l'hypoténuse) prend des valeurs allant de 0 à 1 lorsque l'angle varie de 0° à 90°. Si l'angle est de 30°, la perpendiculaire opposée à l'angle est exactement la moitié de l'hypoténuse et le rapport de ces deux lignes ou sinus $30° = 0{,}5$. Si l'angle est de 45°, la perpendiculaire est un peu plus des sept dixièmes de l'hypoténuse (plus exactement $\sin 45° = 0{,}707107$ ..). Pour un angle de 80°, la perpendiculaire équivaudra à environ 94 p. 100 de l'hypoténuse ($\sin 80° = 0{,}939693$...) et ainsi de suite.

Les valeurs du sinus qui, pour les angles de 0° à 90°, croissent de 0 à l'unité, diminuent symétriquement de l'unité à 0 pour les angles croissant de 90° à 180°, puis diminuent de 0 à — 1 dans le troisième quadrant (c'est-à-dire pour les angles croissant de 180° à 270°, enfin passent de — 1 à 0 dans le quatrième quadrant (angles croissant de 270° à 360°) pour recommencer de même quand le rayon vecteur repasse sur le premier quadrant.

Supposons donc qu'un phénomène statistique donné ait une période de 12 unités de temps (par exemple les mois de l'année). Partageons mentalement la circonférence en 12 parties, chacune de 30°. La première unité de temps sera indiquée sur l'axe des abscisses par 0°, la seconde par 30°, la troisième par 60° et ainsi de suite jusqu'à la douzième qui correspondra à 330°. Si le cycle comprend 24 unités de temps (comme les 24 heures de la journée), on imaginera la circonférence divisée en 24 parties allant de 15 en 15 degrés; s'il comprend 7 unités (les 7 jours de la semaine), les divisions auront chacune la grandeur de 51° 26′.

Cela posé, pour substituer à la série observée une série théorique plus simple et régulière ayant pour équation

$$y = a + b \sin \varphi$$

au moyen de la méthode des moindres carrés, la règle pratique à suivre est très facile.

On fait d'abord la moyenne arithmétique des données fournies par l'observation; cette moyenne est la valeur de a de l'équation précédente; puis on détermine le

coefficient b en multipliant les données de l'observation par les valeurs correspondantes de la variable $\sin\varphi$, on fait la somme de ces produits et on la divise par la somme des carrés de $\sin\varphi$. Le quotient fournit la valeur cherchée de b.

Nous prendrons pour exemple la natalité légitime observée en Italie mensuellement de juillet 1877 à juin 1894. La moyenne annuelle étant ramenée à 12000, la moyenne mensuelle (en supposant que chaque mois correspond exactement au douzième de l'année) est 1 000 (1).

Mois. (a)	Nes vivants legitimes (serie observee). (b)	Variable φ exprimee en degres. (c)	Valeurs de $\sin\varphi$. (d)	Produit $e = b \times d$. (e)	
Juillet	935	0°	0	0	
Août	963	30°	0,5	481,5	
Septembre	1 008	60°	0,866	872,9	
Octobre	966	90°	1	966	+ 3 635,4
Novembre	963	120°	0,866	834	
Décembre	962	150°	0,5	481	
Janvier	1 061	180°	0	0	
Février	1 120	210°	— 0,5	— 560	
Mars	1 096	240°	— 0,866	— 949,1	
Avril	1 044	270°	— 1	— 1 044	— 3 849,9
Mai	972	300°	— 0,866	— 841,8	
Juin	910	330°	— 0,5	— 455	
Moyenne mensuelle. .	1 000		Somme algébrique. .		— 214,5

La somme des carrés de $\sin\varphi$ est 6. Divisant — 214,5 par 6, on obtient le quotient — 35,75.

L'équation est donc :

$$y = 1\,000 - 35{,}75 \sin\varphi.$$

En appliquant cette formule, nous trouverions des valeurs théoriques insuffisamment approchées de celles que fournit l'observation.

Exemple : le premier terme de la série observée est 935; il résulterait du calcul :

$$1\,000 - 35{,}75 \sin 0° = 1\,000$$

l'écart serait de 65 unités.

Le second terme de la série observée est 963, celui de la série calculée serait :

$$1\,000 - 35{,}75 \sin 30° = 1\,000 - 35{,}75 \times 0{,}5 = 982{,}12$$

avec un écart de 19 unités

Le troisième terme observé est 1 008, le calcul donne :

$$1\,000 - 35{,}75 \sin 60° = 1\,000 - 35{,}75 \times 0{,}866 = 969{,}04$$

avec un écart de 39 unités, etc.

(1) Pour décembre et janvier, il a été fait une correction approximative pour tenir compte de cette circonstance connue qu'un certain nombre d'enfants mâles légitimes nés dans les derniers jours de l'année sont déclarés a l'état civil comme nés en janvier de l'annee suivante pour leur faire gagner un an a la conscription. (*Note de l'Auteur*)

Si l'on veut une plus grande approximation, il faut interpoler la fonction cosinus conformément à l'équation :

$$y = a + b \sin \varphi + c \cos \varphi.$$

Pour cela, on procède d'une manière analogue à la précédente en multipliant les données de l'observation par les valeurs de $\cos\varphi$, c'est-à-dire respectivement par les nombres de la seconde colonne ci-dessous, faisant la somme des produits et divi-

φ.	Valeur du $\cos\varphi$
0°	1
30°	0,866
60°	0,5
90°	0
120°	— 0,5
150°	— 0,866
180°	— 1
210°	— 0,866
240°	— 0,5
270°	0
300°	0,5
330°	0,866

sant par la somme des carrés de $\cos\varphi$ qui est encore 6. Cette opération donne le quotient — 57,75 qui est le coefficient c cherché. L'équation est donc :

$$y = 1\,000 - 35{,}75 \sin \varphi - 57{,}75 \cos \varphi.$$

Cette équation est déjà plus approchée que la précédente; cependant, si on désire une approximation plus serrée des nombres théoriques et de ceux qui sont fournis par l'observation, il conviendra d'interpoler la fonction $\sin 2\varphi$ (sinus du double de l'angle) et $\cos 2\varphi$ (cosinus du double de l'angle). On peut admettre que le phénomène périodique examiné en même temps qu'un cycle principal d'une année comprend deux cycles secondaires d'une demi-année chacun (1).

---

(1) On devra multiplier les données de l'observation par les valeurs indiquées dans les deuxième et troisième colonnes.

2φ	Sin 2φ.	Cos 2φ.
0°	0	1
60°	0,866	0,5
120°	0,866	— 0,5
180°	0	— 1
240°	— 0.866	— 0,5
300°	— 0,866	0,5
0°	0	1
60°	0,866	0,5
120°	0,866	— 0,5
180°	0	— 1
240°	— 0,866	— 0,5
300°	— 0,866	0,5

Faire la somme des produits, la diviser par la somme des carrés de sin 2 φ au cos 2 φ qui est encore 6.

De même, on peut interpoler la fonction $\sin 3\varphi$ et $\cos 3\varphi$ pour mettre en évidence un cycle triple d'un tiers d'année et ainsi de suite. En poussant assez loin, la formule interpolatrice reproduira exactement les données de l'observation, mais alors l'avantage de la simplicité disparaîtra, la formule restant encombrée de termes qui reflètent les ondulations parasites (comme les appelle M. Schiaparelli) de la série qui cachent la vraie loi du phénomène.

Dans le cas précédent de la natalité légitime, au cours des 12 mois de l'année, la formule limitée aux termes *en* 2φ est la suivante :

$$y = 1\,000 - 35{,}75 \sin \varphi - 57{,}75 \cos \varphi + 54{,}85 \sin 2\varphi - 9{,}33 \cos 2\varphi$$

que l'on peut transformer facilement en cette autre exprimée en fonction seulement du sinus de l'angle et du double de l'angle augmentés d'une constante.

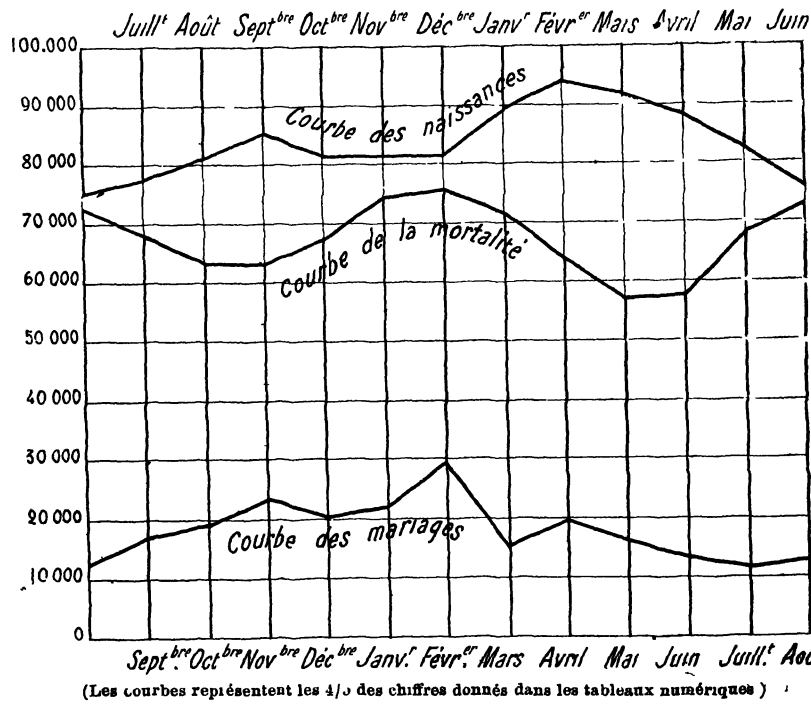
$$y = 1\,000 - 67{,}92 \sin (58^\circ 14' 7'' + \varphi) - 55{,}63 \sin (170^\circ 20' 41'' + 2\varphi).$$

Appliquant la formule, on a les valeurs comparées ci-après :

	Juillet.	Aout.	Septembre.	Octobre.	Novembre.	Decembre.
Observation .	935	963	1 008	966	963	962
Calcul . . .	932,92	974,94	992,34	973,58	955,08	979,96
Différences .	— 2,08	+ 11,94	— 15,66	+ 7,58	— 7,92	+ 17,96

	Janvier.	Fevrier.	Mars.	Avril.	Mai.	Juin.
Observation .	1 061	1 120	1 096	1 044	972	910
Calcul . . .	1 048,42	1 110,72	1 112	1 045,08	959,26	915,70
Différences .	— 12,58	— 9,28	+ 16	+ 1,08	— 12,74	+ 5,70

L'erreur arithmétique moyenne des termes calculés est ± 10 et ne dépasse pas 1 p. 100 des termes observés. Si nous nous étions arrêtés à la première moitié de la formule, l'erreur arithmétique moyenne aurait été 3 fois plus grande



(Les courbes représentent les 1/5 des chiffres donnés dans les tableaux numériques.)

§ 3. — Quand on traite de phénomènes non périodiques, c'est-à-dire de simples successions chronologiques de données ne formant pas un cycle fermé, il est préférable d'interpoler au moyen de fonctions algébriques entières. La formule générale, d'un usage très commode quand la série procède par intervalles égaux de temps, est :

$$y = A + B\psi_1 + C\psi_2 + D\psi_3 + \ldots\ldots$$

La variable ψ_1 est simplement l'intervalle en unités de temps qui sépare chaque terme du point moyen de la série.

Supposons que celle-ci aille de 1892 à la fin de 1902, c'est-à-dire qu'elle comprenne 11 années, la valeur de ψ_1 correspondant à 1897, qui est l'année moyenne, sera 0, les valeurs correspondant à 1896, 1895, 1894 et celles correspondant à 1898, 1899, 1900 seront respectivement égales à -1, -2, -3 et à $+1$, $+2$, $+3$. Les valeurs que prend la variable ψ_2 ne sont que les carrés des ψ_1 diminués de la moyenne arithmétique de ces mêmes carrés et les valeurs de ψ_3 ne sont que les cubes des ψ_1 diminués des termes en progression arithmétique que l'on obtient en interpolant une droite entre ces cubes, etc.

La détermination des constantes A, B, C... se fait par un procédé analogue à celui qui a servi pour les séries périodiques. On calcule avant tout la moyenne arithmétique des quantités observées, ce qui donne la valeur de A. Puis on détermine B en multipliant les données de l'observation par les valeurs de ψ_1, on fait la somme des produits, on la divise par la somme des carrés de ψ_1 ; on détermine C en multipliant les données de la série par les valeurs de ψ_2 et en divisant la somme des produits par la somme des carrés de ψ_2 et ainsi de suite.

Nous donnons à titre d'exemple un tableau des valeurs des variables ψ_1, ψ_2, ψ_3, ψ_4 pour le cas d'une série de 10 termes et d'une de 11 termes lorsqu'on veut pousser l'interpolation jusqu'à une courbe du quatrième degré.

Pour ces calculs, on s'aide de tables comme celles qu'a dressées récemment M. Pareto : tables pour faciliter l'application de la méthode des moindres carrés (1).

Pour $n = 10$.				Pour $n = 11$.			
ψ_1.	ψ_2.	ψ_3.	ψ_4.	ψ_1.	ψ_2.	ψ_3.	ψ_4.
$-4,5$	$+12$	$-25,2$	$+43,2$	-5	$+15$	-36	$+72$
$-3,5$	$+4$	$+8,4$	$-52,8$	-4	$+6$	$+7,2$	-72
$-2,5$	-2	$+21$	$-40,8$	-3	-1	$+26,4$	-72
$-1,5$	-6	$+18,6$	$-7,2$	-2	-6	$+27,6$	-12
$-0,5$	-8	$+7,2$	$+43,2$	-1	-9	$+16,8$	$+48$
$+0,5$	-8	$-7,2$	$+43,2$	0	-10	0	$+72$
$+1,5$	-6	$-18,6$	$+7,2$	$+1$	-9	$-16,8$	$+48$
$+2,5$	-2	-21	$-40,8$	$+2$	-6	$-27,6$	-12
$+3,5$	$+4$	$-8,4$	$-52,8$	$+3$	-1	$-26,4$	-72
$+4,5$	$+12$	$+25,2$	$+43,2$	$+4$	$+6$	$-7,2$	-72
				$+5$	$+15$	$+36$	$+72$
$\Sigma(\psi_1)^2 = 82,5$		$\Sigma(\psi_2)^2 = 528$		$\Sigma(\psi_1)^2 = 110$		$\Sigma(\psi_2)^2 = 858$	
$\Sigma(\psi_3)^2 = 3\,088,8$		$\Sigma(\psi_4)^2 = 16\,473,6$		$\Sigma(\psi_3)^2 = 6\,177,6$		$\Sigma(\psi_4)^2 = 41\,184$	

(1) Communication présentée à l'assemblée annuelle des statisticiens officiels et de la Société suisse de statistique tenue a Lausanne en 1898.

Si la série observée a un caractère statique, il suffira de s'arrêter à la moyenne arithmétique, c'est-à-dire de faire $y = A$; si elle a un caractère dynamique et une allure plutôt rectiligne, on utilisera les deux premiers termes de la formule $y = A + B\psi_1$, ce qui équivaut à interpoler une droite plus ou moins inclinée sur l'axe des temps ; si la série a un caractère dynamique et une allure plutôt parabolique avec une seule inflexion principale, on fera $y = A + B\psi_1 + C\psi_2$, c'est-à-dire qu'on interpolera une parabole ordinaire, etc.

Un examen facile des deux exemples ci-dessus montre comment les variables ψ_1, ψ_2, ψ_3 constituent respectivement une série linéaire ou une parabole du deuxième degré ou une parabole cubique ; en d'autres termes, de telles séries ont respectivement constantes les différences premières de leurs termes, ou les différences secondes (différences des différences), ou les différences troisièmes, etc. La somme algébrique pour chacune des séries est toujours zéro.

De ce que nous avons dit des variables ψ_2, ψ_3, etc., en tant qu'elles sont fonctions particulières de ψ_1, on comprend que l'équation

$$y = A + B\psi_1 + C\psi_2 + D\psi_3 + \ldots\ldots$$

peut se mettre convenablement sous la forme d'un développement suivant les puissances entières de ψ_1 ou (pour nous servir de la notation la plus usitée) de x

$$y = a + bx + cx^2 + dx^3 + \ldots\ldots$$

sachant que $\psi_1 = x$; $\psi_2 = x^2 - \frac{n^2 - 1}{12}$; $\psi_3 = x^3 - \frac{3n^2 - 7}{20}x$; etc. . . . .

n est le nombre des termes de la série.

§ 4. — Avec l'interpolation, nous avons donc pour but, dans les séries empiriques, de séparer la partie constante ou régulièrement variable qui exprime la vraie loi du phénomène, de la partie variable sans loi suffisamment définie qui atteste l'intervention de causes secondaires perturbatrices. Mais il est facile de voir que des critériums absolus pour opérer cette séparation n'existent pas et que la solution du problème présente beaucoup d'aspects arbitraires.

Limiter l'arbitraire voudrait dire rapprocher, au point de vue de l'exactitude, la loi empirique de la loi naturelle, c'est-à-dire faire faire un pas immense aux sciences ou parties de science qui emploient les méthodes statistiques. Sur les aspects arbitraires des procédés d'interpolation, M. Schiaparelli a écrit jadis avec une grande compétence (1) et la question a été reprise récemment par M. Pareto (2).

Nous, qui nous proposons seulement de vulgariser, plus qu'elles ne le sont parmi ceux qui étudient la statistique, les méthodes d'interpolation, nous mettrons du nôtre quelques considérations qui ne seront peut-être pas inutiles.

Avant tout, le choix de la fonction-type paraît arbitraire, rien ne justifie la pré-

(1) Voir : Ephémérides astronomiques de Milan pour l'année 1867. Appendice : *Sur le moyen de dégager la véritable expression des lois de la nature des courbes empiriques.* — Voir aussi : *Sur les variations périodiques du climat de Milan.* Memoire de G. V. Schiaparelli et G. Celoria, astronomes de l'observatoire royal de Brera, à Milan.

(2) « Quelques exemples d'application des methodes d'interpolation a la statistique ». (*Journal de la Societé de Statistique de Paris*, numero de novembre 1899.)

férence donnée aux fonctions algébriques entières ou aux fonctions périodiques du sinus, du cosinus de l'angle ou des multiples exacts de l'angle ; rien, disons-nous, sauf la plus grande facilité et la rapidité du calcul. C'est, comme dit M. Schiaparelli, un avantage subjectif pour le calculateur ; la nature, en établissant ses lois, ne s'est certainement pas arrêtée devant la complication des fonctions.

Toutefois, l'arbitraire est moins grand qu'il ne paraît à première vue. Dans la nature dominent les nombres simples. Les lois de la vibration des cordes, de la chute des corps, des oscillations du pendule, de la réflexion et de la réfraction de la lumière, etc. ; les lois des proportions définies et des proportions multiples en chimie et celles de la périodicité des corps simples (1) sont là pour l'attester.

Mais même si nous entrons dans certains ordres de faits collectifs, dans lesquels on considère comme type la moyenne des cas observés ou la classe de plus grande fréquence, on trouve des phénomènes qui varient en raison simple (directe ou inverse) d'un autre ou en raison des carrés, des racines carrées, etc.

Telles sont, en anthropométrie, les relations entre le poids et la stature, entre la stature et la fréquence du pouls, etc.

En démographie, le nombre des possesseurs de patrimoines paraît inversement proportionnel à l'importance de ces patrimoines ; si nous considérons le revenu (revenu total), le carré du nombre des possesseurs sera en raison inverse des cubes des revenus-limites considérés. Certes, il s'agit de lois empiriques d'une approximation encore large, mais qui peuvent être invoquées comme indices en faveur de l'emploi de fonctions plus simples dans l'interpolation. De plus, il arrive souvent en démographie que deux phénomènes qui sont nécessairement facteurs d'un troisième se déroulent chacun pendant quelque temps d'une manière disons plutôt rectiligne ;

---

(1) La loi de périodicite de Mendelejeff, qui sert a determiner par voie de prévision les proprietés des corps simples non encore decouverts, s'énonce ainsi : les corps simples dont les poids atomiques sont liés par des rapports numériques simples (lnéaires) sont tres semblables par leurs propriétes. Ainsi, dans la famille de l'azote, de l'oxygène, des metaux alcalins, les poids atomiques vont en croissant comme suit :

		Poids atomique.
	Famille de l'azote.	
Azote	14	= 14
Phosphore	14 + 17	= 31
Arsenic	14 + 17 + (1 × 44)	= 75
Antimoine	14 + 17 + (2 × 44)	= 119
Bismuth	14 + 17 + (4 × 44)	= 207
	Famille de l'oxygène.	
Oxygène	16	= 16
Soufre	16 + (1 × 16)	= 32
Selenium	16 + (4 × 16)	= 80
Tellure	16 + (7 × 16)	= 128
	Famille des métaux alcalins.	
Lithium	7	= 7
Sodium	7 + (1 × 16)	= 23
Potassium	7 + (2 × 16)	= 39
Rubidium	7 + (5 × 16)	= 87
Cesium	7 + (8 × 16)	= 135

De même, dans la famille des métaux alcalino-terreux (magnesium, calcium, strontium, baryum), les poids atomiques s'obtiennent en ajoutant a la constante 24 les produits de 16 par 0, 1, 4 et 7.

il est alors clair que le troisième phénomène produit par les deux premiers ne pourra que suivre à peu près une parabole ordinaire.

En effet, on a :

$$(a + bx)(c + dx) = ac + (ad + bc)x + cdx^2.$$

Tel est le cas de la natalité légitime dans un pays où la nuptialité va en croissant en progression arithmétique et où la fécondité, par la diffusion des habitudes improprement appelées malthusiennes, décroît lentement suivant une autre progression arithmétique. Quand l'un des phénomènes parcourt une parabole et l'autre une droite inclinée sur l'axe des temps, le phénomène résultant se mouvra sur une courbe du troisième degré. Ainsi l'interpolation par voie de fonctions algébriques entières est suggérée très souvent par la nature même du cas examiné.

§ 5. — La méthode des moindres carrés ne peut mériter d'autres raisons de préférence que la simplicité des calculs.

Pourquoi, dans l'infinité de séries linéaires ou paraboliques que l'on peut substituer aux séries observées avec la même moyenne et un champ de variabilité moindre, adoptons-nous justement celle qui réalise la condition du minimum de la somme des carrés des écarts entre ses termes et les termes correspondants de l'observation? On répond : pour limiter l'arbitraire. Mais l'arbitraire, s'il n'existe plus pour les calculateurs *uti singuli,* subsiste pour eux *uti universi,* comme collectivité, s'ils conviennent de s'attacher à ce principe.

La nature des problèmes pourrait suggérer de rendre minima la somme des cubes (supposés tous positifs) ou celle des quatrièmes puissances, etc., ou d'égaler le plus grand des écarts positifs au plus grand des écarts négatifs. La méthode des moindres carrés fait déjà peser notablement, dans la série, les variations exceptionnelles ; elle oblige d'une certaine manière la courbe interpolée à se pencher vers elles, retirant quelque chose à l'approximation de tous les autres termes.

Dans l'exemple ci-dessous, la simple vue montre que la série des y serait bien représentée par une droite descendant vers l'axe des temps (x). En effet, la variation exceptionnelle du deuxième terme ne peut modifier l'impression visuelle qui résulte de l'ensemble des autres termes.

x.	y.
— 3	53
— 2	39
— 1	51
0	51
+ 1	49
+ 2	49
+ 3	48
Moyenne. . .	48,71

Au contraire, l'équation de la droite interpolée par la méthode des moindres carrés qui est :

$$y = 48{,}71 + 0{,}07\,x$$

montre un mouvement plutôt ascendant et cela en raison de l'importance que prend

dans le calcul la petitesse du second terme de la série; mais il n'y a aucun doute que l'impression visuelle ne soit la plus juste dans le cas en question.

En effet, si l'événement auquel se rapporte le second terme est de ceux qui se produisent tous les vingt ou trente ans et se trouve par hasard compris dans une série ou partie de série de sept ans seulement, il ne devrait entrer dans le calcul que pour les valeurs extrêmes que l'on rencontre d'ordinaire dans les périodes de semblable brièveté. En outre, toute variation brusque est un indice non douteux de l'intervention d'une cause nouvelle, transitoire ou non, et, comme nous le verrons tout à l'heure, quand un fait nouveau survient dans une série commencée, il n'est plus rigoureux de procéder par une interpolation unique pour toute la série, mais on doit procéder par voie d'interpolation partielle dans les parties séparées de la variation extraordinaire.

§ 6. — La théorie des faits nouveaux est un des points magistralement discutés par M. Schiaparelli. Il dit en substance qu'à la formation des constantes de l'équation concourent et collaborent tous les termes de la série, tandis que cette collaboration ou cette solidarité n'a souvent pas de raison d'être. Quand une cause nouvelle survient au milieu ou au tiers de la série, pourquoi la supposer opérante avant le temps où elle a réellement commencé à se manifester ? Cependant, cette supposition est implicitement contenue dans la marche même du calcul des coefficients de la variable. M. Schiaparelli prend l'exemple de la variation diurne de la température. « En vingt-quatre heures elle passe par quatre stades différents, séparés l'un de l'autre par une solution de continuité, à la vérité peu accusée, mais cependant visible. En effet, durant la nuit, il n'existe d'autre cause que les mouvements du calorique dans l'atmosphère par rayonnement et par conductibilité. Mais à partir de l'aube entre en jeu une nouvelle cause, la réflexion des rayons calorifiques du soleil dans l'atmosphère, analogue à la réflexion de la lumière qui est la cause du crépuscule. Finalement, au lever du soleil, entre en jeu la radiation directe de l'astre. Il est évident, d'après cela, que la température diurne ne peut, en théorie, être représentée par une formule unique valable pour les vingt-quatre heures, de même que pour des raisons semblables, on ne peut représenter, par une formule unique, la variation diurne de l'éclairement d'un point exposé à un ciel complètement libre. Là encore, il y a quatre stades différents : nuit complète, éclairement solaire dans le jour, éclairement crépusculaire du matin et du soir. »

Ces réflexions sont très justes, la portée en est la suivante : si on connaît le moment où le fait nouveau intervient dans la série, il conviendra d'en étudier séparément l'influence et au besoin de procéder par voie d'interpolation partielle ; s'il n'est pas connu, on devra considérer le fait comme agissant pendant tout l'intervalle.

Un tracé continu unique pour figurer le développement de notre commerce avec la France de 1882 à 1902 ne serait pas à approuver ; l'application des tarifs différentiels en 1888, leur remplacement en 1892 par le tarif maximum et le traité de 1899 constituent des faits nouveaux — pour ne parler que des principaux — qui brisent la série et obligent à des interpolations partielles.

Autre exemple : on sait qu'en 1885 les spéculateurs prévinrent les augmentations des droits de douane sur les cafés et sur les sucres en faisant de grandes provisions de ces denrées. Pour savoir comment se sont probablement répartis dans la consommation des diverses années ces approvisionnements exceptionnels, on fera une inter-

polation et nous ne pourrons interpoler la série qu'à partir de 1885 et non pour les années antérieures ; autrement, si la série embrassait la période 1881-1900 et si nous faisions une interpolation unique, nous répartirions sur les années 1881-1884 l'effet des approvisionnements extraordinaires faits de 1885, alors qu'il est clair que ces approvisionnements seraient entrés dans la consommation des années postérieures à 1885 et non des années antérieures.

§ 7. — A quel degré d'approximation porterons-nous l'interpolation pour distinguer l'accidentel du constant ?

Voici une autre matière d'appréciations subjectives variables d'individu à individu. Nous arrêterons-nous aux deux ou trois premiers termes de la formule analytique ?

Interpolerons-nous une droite ou une parabole ordinaire, ou irons-nous aux courbes du troisième, quatrième ou cinquième degré ?

Il est difficile de donner une règle *a priori ;* cela dépend beaucoup de l'intuition de l'observateur. Il conviendra de voir si l'étendue et le caractère alternatif des écarts positifs et négatifs entre les écarts des valeurs empiriques et des valeurs calculées dans la marche de l'interpolation sont de nature à être expliqués par des causes purement accidentelles. Il faudra faire attention aux approximations inégales qui se produisent aux différents stades du calcul. Quand la formule interpolatrice conduit plus ou moins rapidement à l'absurde ou à l'improbable, on se gardera de la considérer comme valable pour l'avenir ; si elle y conduit lentement, nous nous reposerons dans la tranquille confiance que des faits nouveaux, quels qu'ils soient, interviendront certainement pour modifier la marche de la courbe ; mais si elle y conduit très vite, nous devons préciser quels faits nouveaux on peut raisonnablement attendre dans un avenir rapproché pour empêcher la course à l'impossible ou à l'improbable.

Tout cela doit être précédé d'un travail préparatoire ayant pour but de dégager la série des variations dues à des causes spéciales connues ou que l'on peut connaître. Ainsi, si je soupçonne ou puis prouver que les variations de la natalité légitime (ou mieux des conceptions) dépendent dans une certaine mesure de la variation des mariages — ce qui n'a pas besoin d'une longue démonstration — et, dans une autre mesure, des variations de la mortalité (puisqu'aux les époques de grave mortalité — ou morbidité — dans beaucoup de familles menacées ou frappées, les rapports sexuels des époux deviennent plus rares), au moyen d'un calcul de corrélation double, on déterminera la série de la natalité telle qu'elle se serait comportée si les mariages et les décès étaient restés une quantité constante.

La série étant ainsi dégagée des variations dues à des causes connues ou qu'on peut connaître, le reste donnera la loi propre de la série encore troublée par des causes inconnues, accidentelles, secondaires, que l'interpolation tente cependant d'éliminer. Nous ne passerons pas toutefois sous silence que le calcul des corrélations comprend encore celui de l'interpolation ; il semble ainsi que nous tournons dans un cercle vicieux.

En réalité, les deux méthodes sont distinctes ; dans la première, les variations du phénomène donné sont rapportées aux variations concomitantes des autres phénomènes ; dans la seconde, les variations restantes sont rapportées à la simple succession des temps.

Un autre travail préparatoire consiste à rechercher si, par aventure, le phénomène donné ne peut se décomposer en groupes spéciaux ayant diverses lois de développement et de périodicité. Comme il y a pour l'individu ou pour des groupes spéciaux d'individus des phénomènes discontinus qui deviennent continus dans l'ensemble, ainsi il y a des phénomènes périodiques pour l'individu ou pour des groupes choisis qui perdent ou changent pour l'ensemble leur caractère de périodicité. Il suffit de penser au cas dans lequel, au maximum de la fonction pour un groupe, correspond le minimum de la fonction pour un autre groupe, et *vice versa ;* dans l'ensemble, la périodicité tend à disparaître. Ainsi, une marche plutôt rectiligne d'un phénomène considéré dans son ensemble pourra résulter de développements paraboliques l'un concave, l'autre convexe par rapport à l'axe des temps des deux principaux groupes qui composent l'ensemble. Il est en outre sous-entendu que tout dépend de l'intuition de celui qui étudie les faits.

§ 8. — L'intéressante question qui se présente maintenant est celle de savoir si les différents termes du développement d'une fonction algébrique entière ou d'une fonction périodique peuvent être considérés comme l'expression et la mesure de causes distinctes ou d'un groupe de causes.

Pour mieux nous faire comprendre nous prendrons un exemple. Si on prend l'équation de la courbe décrite par un projectile lancé dans le vide sous la forme $y = bx + cx^2$, personne ne doute que le terme bx ne dépende de la force initiale qui agit en fonction simple du temps, tandis que le terme cx^2 tient à la pesanteur qui agit en raison du carré des temps. Dans d'autres cas, il est plus difficile de juger lorsque la formule d'approximation est plus ou moins approchée. L'astronome Carlini en a donné un élégant exemple à propos de la variation diurne du baromètre (1).

Il avait reconnu que la variation diurne peut être assez bien représentée par une formule avec des termes en φ et 2φ, c'est-à-dire par la somme de deux variations distinctes, l'une ayant une période de vingt-quatre heures et l'autre une de douze heures.

$$y = 751^{mm},793 + 0^{mm},524 \sin(163° 9' + \varphi) + 0^{mm},227 \sin(129° 48' + 2\varphi).$$

M. Carlini attribue la première variation aux effets multiples de la chaleur solaire dans l'atmosphère (flux physique), qui a précisément un cycle principal unique de vingt-quatre heures, et la seconde à l'attraction du soleil sur l'océan atmosphérique (flux dynamique) qui a dans la journée deux périodes de douze heures chacune.

Son explication toutefois ne satisfait pas MM. Schiaparelli et Celoria qui opposent des arguments dans lesquels nous ne voulons pas entrer; ils considèrent comme très vraisemblable « que la séparation des termes de la formule périodique est comme en général et même toujours, en ce qui concerne les phénomènes météorologiques, un pur fait analytique dépendant de la forme arbitraire de la fonction choisie comme type et non l'expression d'un fait physique. Il est possible que la variation totale du baromètre soit la somme de deux actions distinctes comme celles que suppose Carlini; mais il ne paraît pas acceptable de conclure que les effets de ces actions se

(1) Francesco Carlini : « Sur la loi des variations horaires du baromètre », *Mémoire de la Société italienne des sciences,* tome XX, Modène, 1828, cité par Schiaparelli et Celoria, dans le mémoire relatif à cette question.

manifestent séparément dans les deux premiers termes d'une formule empirique, lorsque l'observation montre l'existence de termes ultérieurs non négligeables. »

Par ces derniers mots, les deux astronomes font allusion à un terme en 3φ, à une onde triple de la variation diurne qui est spécialement sensible au solstice d'hiver et dont on ne saurait à quoi rapporter la période de huit heures, sinon au même flux physique ou au flux dynamique.

La question est bien loin d'être close; on peut néanmoins admettre que des groupes de causes différentes ont leur expression analytique sinon dans un terme *isolé* de la formule algébrique ou trigonométrique, du moins dans des groupes spéciaux de termes.

Telle me semble être la pensée de Pareto (1).

En démographie, le cas de l'onde double dans la variation annuelle de la natalité légitime est analogue à celui de Carlini.

Comme nous l'avons vu au début de cet article, l'équation de la courbe de la natalité légitime pour les douze mois de juillet à juin est

$$y = 1000 - 35{,}75 \sin\varphi - 57{,}75 \cos\varphi + 54{,}85 \sin 2\varphi - 9{,}39 \cos 2\varphi$$

les termes en 3φ, 4φ pouvant être négligés.

Les causes du phénomène doivent, comme on sait, être recherchées dans les variations annuelles de la nuptialité et de la mortalité. La première est exprimée par la formule :

$$y = 1\,000 + 251{,}99 \sin\varphi - 135{,}25 \cos\varphi$$

le début de la période étant reporté au mois de septembre, c'est-à-dire au dixième mois avant les naissances. Les termes suivants de la formule ou sont négligeables par rapport au premier ou reflètent les deux maxima spéciaux des mariages des époques précédentes, le carême et l'avent.

La variation annuelle de la mortalité par l'influence déjà notée sur les conceptions (2) doit aussi être rapportée au dixième mois avant la natalité et être exprimée par

$$y = 1000 + 54{,}89 \sin\varphi - 19{,}02 \cos\varphi - 100{,}86 \sin 2\varphi + 50{,}72 \cos 2\varphi$$

Les termes suivants en 3φ, etc., peuvent être négligés.

Eh bien ! il y a beaucoup de vraisemblance dans l'hypothèse que la première partie de la formule de la natalité qui se rapporte à une période annuelle (termes

(1) *Quelques exemples d'application, etc.*, p. 5. Lorsqu'on applique cette formule ($y = A + B\psi_1 + C\psi_2 + \ldots\ldots$), on observe en général que les courbes simples qu'on obtient successivement ne vont pas en se rapprochant d'une manière uniforme de la courbe réelle; la precision commence par augmenter rapidement; ensuite, il y a une période où elle augmente lentement; de nouveau elle augmente rapidement et ainsi de suite. Ces périodes, pendant lesquelles la précision augmente lentement, séparent les grands groupes des sinuosités, en d'autres termes, *elles séparent des groupes d'influences de plus en plus particulières qui s'exercent sur le phénomène.*

(2) Voir « De certains points obscurs de la démographie ». août 1896, *Giornale degli Economisti*. Les formules calculées ci-dessus l'ont été sur des chiffres mensuels relatifs à un total de 12 000 cas dans l'année; pour calculer sur des chiffres absolus, il suffira de prendre au lieu de la constante mensuelle 1 000 les moyennes suivantes : pour la natalité, 84,180; pour la mortalité, 67,084; pour la nuptialité, 18,742, en augmentant proportionnellement les coefficients respectifs des variables.

en φ) dépend de la première partie de la formule de la mortalité renforcée dans une certaine mesure par l'onde unique de la nuptialité ; tandis que la seconde partie de la formule relative aux naissances légitimes avec des termes en 2φ et une double période dérive de la seconde partie de la formule de la mortalité ; les petites divergences peuvent s'expliquer par les influences diverses que la mortalité exerce sur la contrainte morale des époux, suivant les catégories d'âge qui en sont frappées de préférence dans les diverses saisons.

Nous avons voulu rouvrir la question et rien de plus, car nous ne trouvons pas assez justifiée l'opinion pessimiste de M. Schiaparelli et de M. Celoria que la représentation des phénomènes météorologiques, démographiques, etc , par une formule analytique « ne fait pas faire un pas vers la connaissance de la vraie loi de ces phénomènes ». Au contraire, nous trouvons très suggestives ces méthodes par lesquelles on essaie de remonter, dans la dynamique de la série, de la résultante aux composantes et qui nous invitent à chercher derrière le voile des symboles analytiques les causes spécifiques des phénomènes. Certes, la théorie statistique mathématique de l'interpolation a besoin d'être perfectionnée, mais en attendant il est bon de constater que dans ces dix dernières années les procédés supérieurs de la statistique ont trouvé accueil jusque dans les sciences biologiques, où ils sont appelés à contribuer à la solution des problèmes ardus qui concernent l'évolution des espèces.

R. Benini.

(*Giornale degli Economisti*, janvier 1904)

Traduit avec l'autorisation de l'auteur par :

P. des Essars.