JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

J. MEYRIGNAC

Étude statistique des températures « quotidiennes »

Journal de la société statistique de Paris, tome 112, n° 1 (1971), p. 23-36

<http://www.numdam.org/item?id=JSFS_1971__112_1_23_1>

ÉTUDE STATISTIQUE DES TEMPÉRATURES « QUOTIDIENNES »

L'importance du chauffage au gaz conduit le Gaz de France à étudier les températures quotidiennes, définies le plus souvent comme la moyenne de 8 relevés effectués toutes les 3 heures, parfois comme celle des températures minimales et maximales de la journée. Pour caractériser les consommations de l'ensemble de ses exploitations, le Gaz de France se réfère aux températures de 74 stations météorologiques, pour lesquelles on doit établir les résultats de quelques études, nécessaires partout : le travail est alors exécuté en ordinateur, après un ou deux essais à la main. Certains autres, qui répondent à des besoins plus localisés, doivent être établis rapidement avec les éléments déjà disponibles.

Nous avons donc pris le parti de définir un modèle probabiliste des températures pour ces stations. Une dizaine d'entre elles ont été traitées à partir des séries existantes de températures quotidiennes. Pour les autres nous disposions de moyennes mensuelles.

Dans la suite, nous allons voir :

1. Le modèle adopté pour représenter les températures quotidiennes avec :
   1. 1. Normalité des températures et leurs liaisons, contrôlées sur Nantes et Paris.
   1. 2. Estimation des paramètres pour les stations traitées à partir de données quotidiennes.
   1. 3. Estimation des paramètres pour les stations traitées à partir de données mensuelles.
2. Une étude du nombre de jours donc la température est inférieure à T, exemple de problème traité avec ce modèle.

1. MODÈLE ADOPTÉ POUR REPRÉSENTER LES TEMPÉRATURES QUOTIDIENNES

1. 1. *La normalité des températures et leurs liaisons*

Les données de base sont les températures quotidiennes enregistrées à Nantes de 1924 à 1959, à Paris de 1896 à 1961. On notera t_{ij} la température du jour i de l'année j ($i = 1$ pour le 1er janvier, 365 pour le 31 décembre), n_i le nombre d'observations pour la date i. On calcule

$$m_i = \frac{1}{n_i} \sum_j t_{ij}$$

$$s_i^2 = \frac{1}{n_i - 1} \sum_j (t_{ij} - m_i)^2$$

la variable centrée réduite correspondante à chaque observation

$$x_{ij} = \frac{t_{ij} - m_i}{s_i}$$

et les coefficients de corrélation entre les températures à n jours d'intervalle

$$r_{i,i-n} = \frac{1}{n_i} \sum_j x_{ij}\, x_{i-n,j}$$

Après vérification graphique, on a admis en première approximation :

a) Pour chaque date i, *les températures t_{ij} suivent une loi normale.* Les x_{ij} suivent alors approximativement une loi normale centrée réduite, propriété vérifiée sur les 10, 20 et 30 de chaque mois à Nantes, tous les jours à Paris (graphique 1. 1).

b) *Les coefficients de corrélation $r_{i,i-n}$ à n jours d'intervalle sont indépendants de i.* Dans ce cas, leurs transformées de Fisher

$$z_{i,i-n} = \frac{1}{2} [\mathrm{Log}\,(1 + r_{i,i-n}) + \mathrm{Log}\,(1 - r_{i,i-n})]$$

suivent une loi approximativement normale de moyenne

$$z_n = \frac{1}{2} [\mathrm{Log}\,(1 + r_n) + \mathrm{Log}\,(1 - r_n)]$$

et de variance $1/n_i - 3$. A Nantes et Paris, on a vérifié cette propriété pour les coefficients entre deux jours consécutifs, puis, à Nantes seul, pour les coefficients à 2, 3 et 10 jours d'intervalle (graphique 1. 2).

Cependant, un graphique chronologique faisait soupçonner une légère baisse en octobre et novembre. A Nantes, le coefficient de corrélation moyen, calculé d'après les transformées de Fisher des coefficients quotidiens, était égal à 0,742 en octobre et 0,760 en novembre, alors que la moyenne estimée sur l'ensemble de l'année était 0,78. L'écart n'a pas paru suffisant pour en tenir compte dans le modèle.

c) *Les températures t_{ij} suivent un processus de Markoff*, c'est-à-dire que pour déterminer la loi de probabilité de la température d'un jour i, toute l'information du passé est contenue dans celle du jour $i - 1$, ou du dernier jour connu. On a alors

$$x_{ij} = r_{i,i-1}\, x_{i-1,j} + \varepsilon_{ij}$$

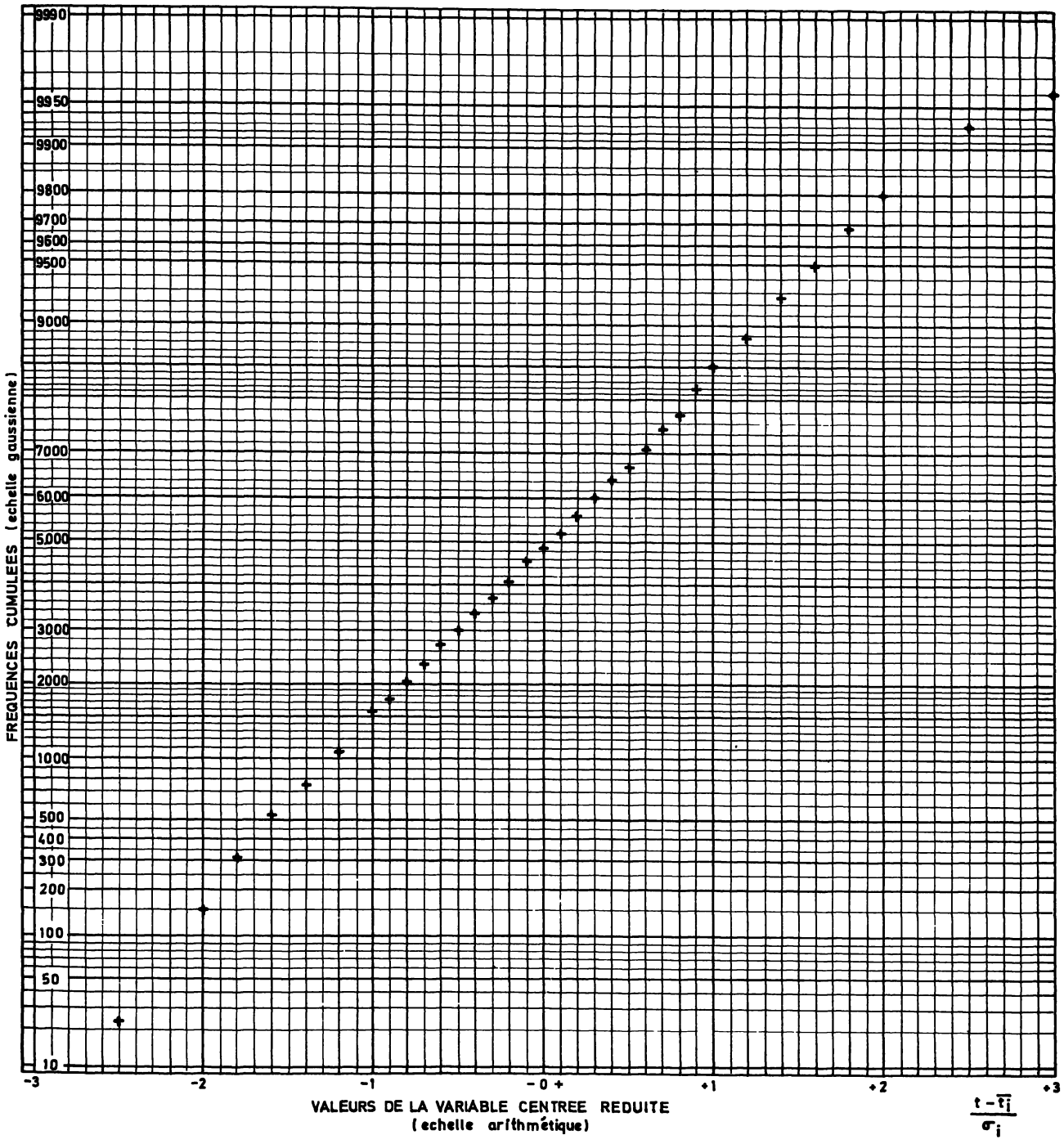
$$= r_{i,i-2}\, x_{i-2,j} + \varepsilon'_{ij}$$

$$x_{i-1,j} = r_{i-2,i-1}\, x_{i-2,j} + \varepsilon_{i-1,j}$$

Graphique 1.1.

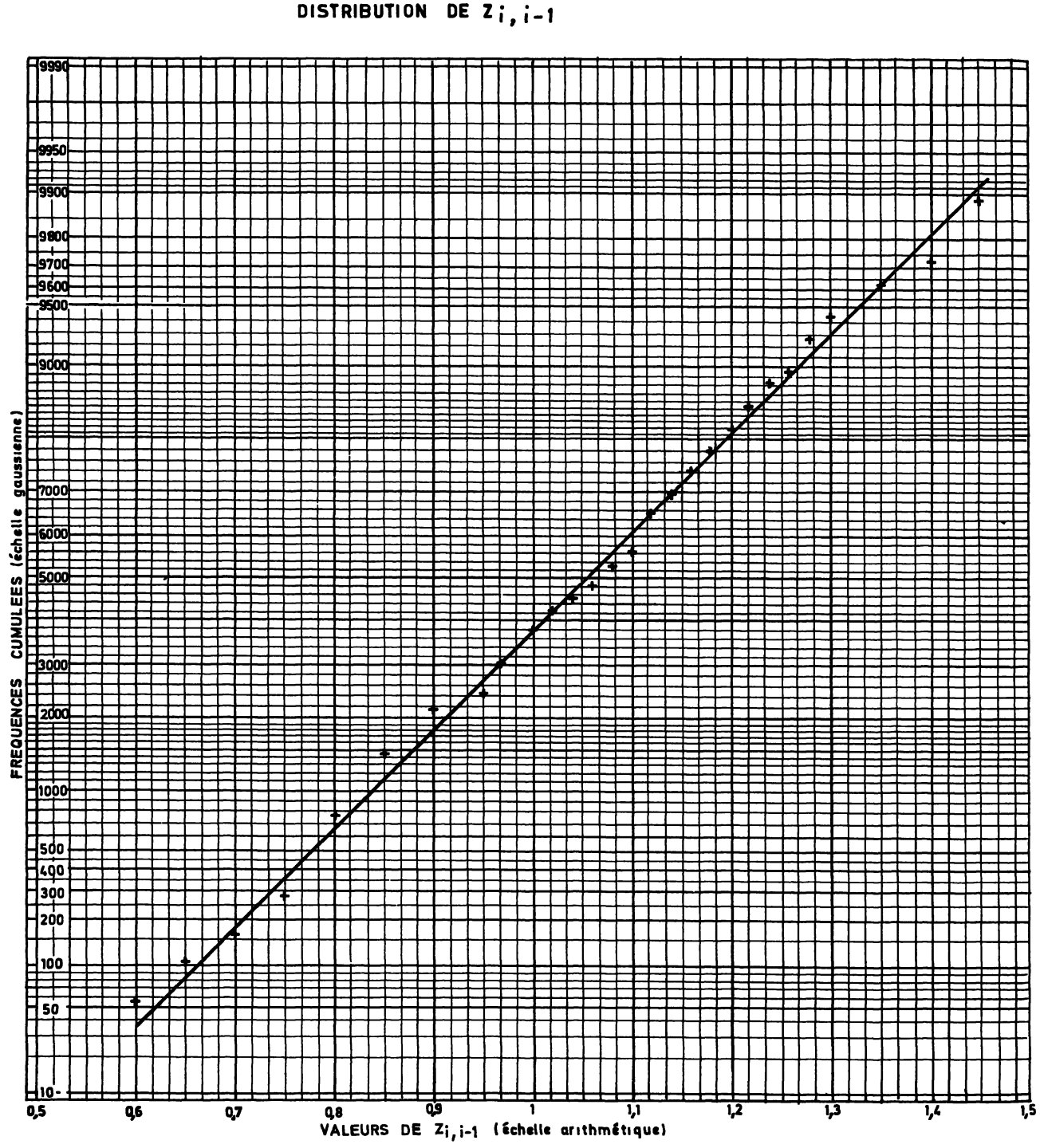
TEMPERATURES DE NANTES
DISTRIBUTION DES VARIABLES CENTREES REDUITES

(AVEC MOYENNES ET ECARTS TYPES OBSERVES POUR CHAQUE DATE)



Graphique 1.2.

CŒFFICIENT DE CORRELATION ENTRE LES TEMPERATURES DE DEUX JOURS SUCCESSIFS A NANTES



d'où

$$r_{t,t-2} = r_{t,\,t-1}\, r_{t-1,\,t-2}$$

et par récurrence

$$r_{t,t-n} = \prod_{k=1}^{n} r_{t,\,t-k}$$

Ici, les $r_{t,t-n}$ sont indépendants de i, donc

$$\boxed{r_n = r_1^n}$$

On a vérifié cette relation pour $n = 2, 3, 4, 5, 10$ à Nantes, tous les n entre 1 et 20 à Paris. Les résultats obtenus sont les suivants :

n	Nantes		Paris	
	r_n	r_n	r_n	r_{1n}
1	0,78	0,78	0,798	0,798
2	0,58	0,61	0,575	0,637
3	0,42	0,47	0,426	0,508
4	0,32	0,37	0,326	0,406
5	0,24	0,29	0,254	0,324
6			0,201	0,258
7			0,163	0,206
8			0,137	0,164
9			0,120	0,131
10	0,084	0,0835	0,106	0,105
11			0,090	0,084
12			0,076	0,067
13			0,066	0,053
14			0,060	0,042
15			0,057	0,034
16			0,056	0,027
17			0,053	0,022
18			0,049	0,017
19			0,047	0,014
20			0,047	0,011

Des calculs analogues ont été effectués dans les quelques villes traitées à partir de données quotidiennes. On a ensuite représenté graphiquement r_n en fonction de n. On a obtenu, comme à Paris, une courbe à concavité tournée vers le haut qui coupe la droite

$$\text{Log } r_n = n \text{ Log } r$$

au point $n = 10$. Le modèle de Markoff n'est donc pas parfaitement conforme à la réalité. On estime cependant qu'il s'en rapproche suffisamment pour l'utiliser lorsqu'il allège sensiblement les calculs.

1.2. *Estimation des paramètres pour les stations traitées à partir de données quotidiennes*

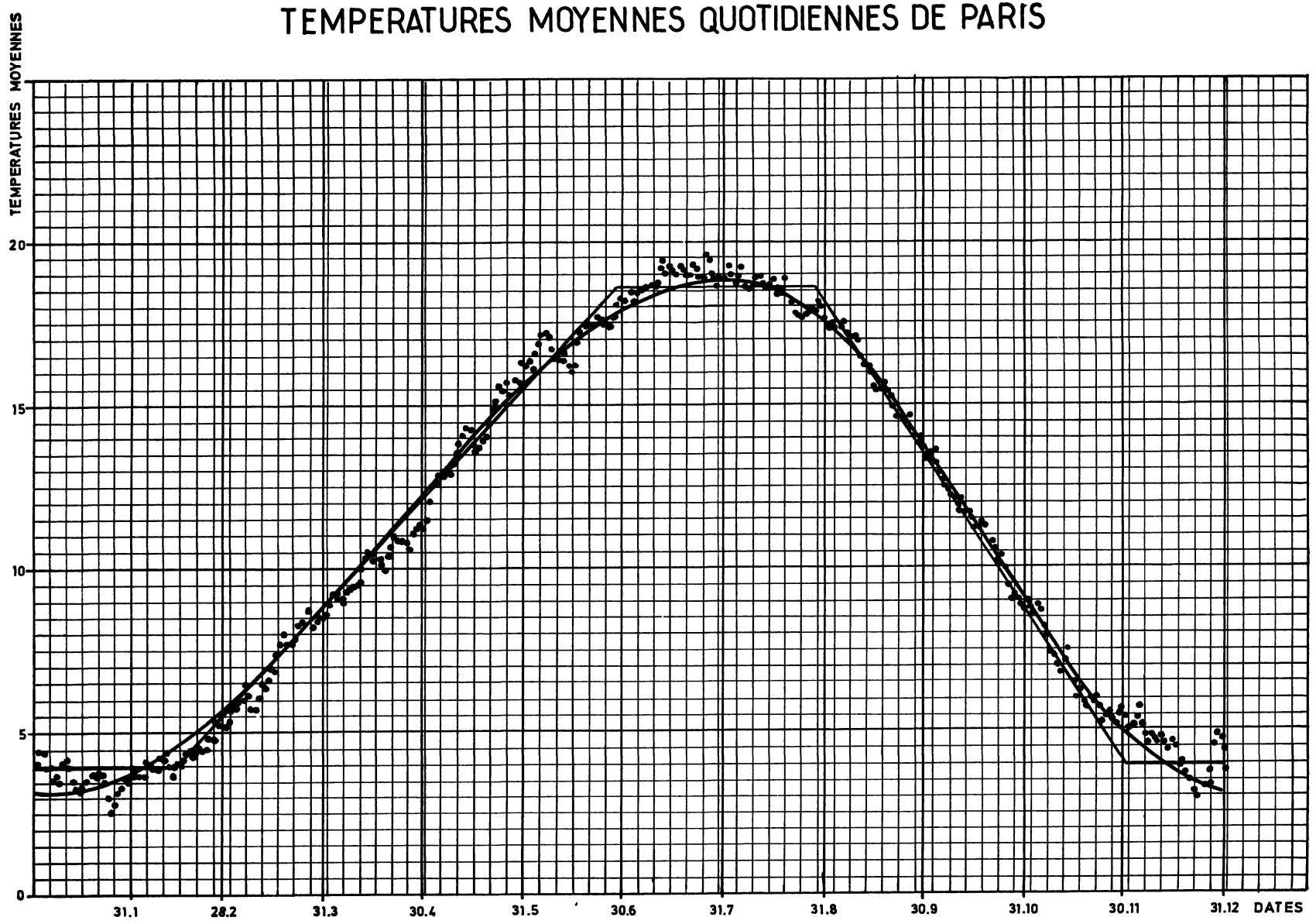
1.2.1. *Ajustement des moyennes*

a) *Forme adoptée* : Compte tenu de la dissymétrie de la courbe observée graphiquement, on a ajusté aux moyennes 2 arcs de sinusoïdes avec passage de l'un à l'autre aux jours le plus froid ($i = i_3$) et le plus chaud ($i = i_4$), en moyenne sur longue période.

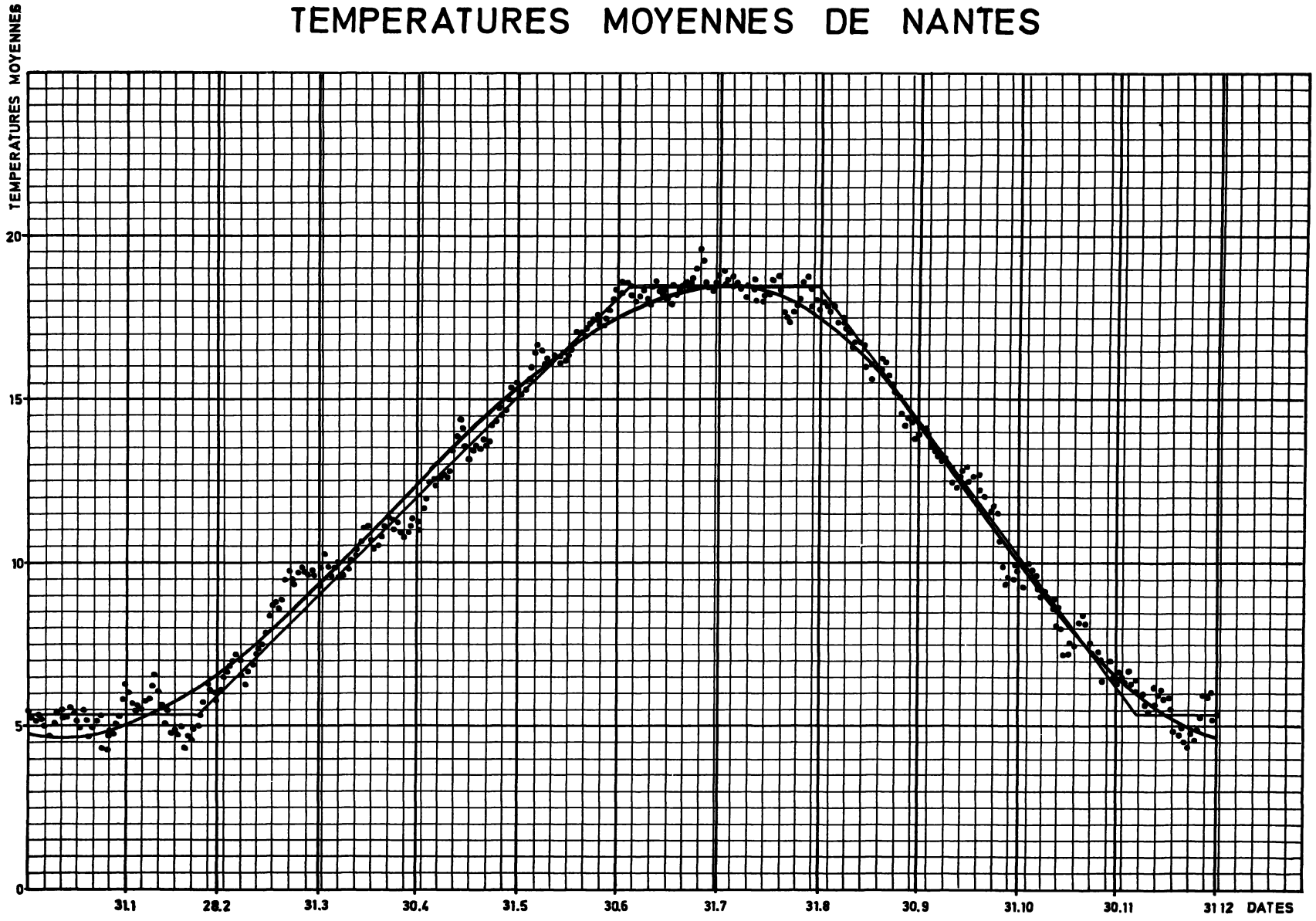
On a en outre imposé à la courbe globale d'être périodique de période 365 jours, continue, sans point anguleux. Les moyennes ajustées s'écrivent alors

$$m_t = m_0 + a \sin h_1 (i - i_1) \text{ pour } i_3 \leqslant i \leqslant i_4$$
$$m_t = m_0 + a \sin h_2 (i - i_2) \text{ pour } i < i_3 \text{ et } i > i_4$$

Graphique 2.1.



Graphique 2.2.



avec

$$h_1 = \pi/(i_4 - i_3)$$
$$h_2 = \pi/(i_3 + 365 - i_4)$$
$$i_1 = (i_3 + i_4)/2$$
$$i_2 = (i_3 + 365 + i_4)/2$$

b) *Estimation de m_0 et a*

Si l'on applique la méthode des moindres carrés

$$m_0 = \sum m_i/365$$
$$a = \frac{2}{365}\left[\sum_{i_3}^{i_4} m_i \sin h_1 (i - i_1) + \sum_{i_4+1}^{i_3+364} m_i \sin h_2 (i - i_2)\right]$$

cette estimation a été retenue pour m_0. Pour a, on remarque que

$$\sum_{i_3}^{i_4} (m_i - m_0)^2 = \sum_{i_3}^{i_4} [m_i - m_0 - a \sin h_1 (i - i_1)]^2 + a^2 \sum_{i_3}^{i_4} \sin^2 h_1 (i - i_1)$$
$$+ 2a \sum_{i_3}^{i_4} [m_i - m_0 - a \sin h_1 (i - i_1)] \sin h_1 (i - i_1)$$

or

$$\mathrm{E}\,[m_i - m_0 - a \sin h_1 (i - i_1)]^2 = \mathrm{Var}\, m$$

Var m étant la variance résiduelle des m_i, qui paraît négligeable.

$$\mathrm{E}\,[m_i - m_0 - a \sin h_1 (i - i_1)] = 0$$

par définition.

$$\sum_{i_3}^{i_4} \sin^2 h_1 (i - i_1) = \frac{1}{2} \sum_{i_3}^{i_4} [1 - 2 \cos 2 h_1 (i - i_1)] = \frac{i_4 - i_3 + 1}{2}$$

donc

$$\mathrm{E}\left[\sum_{i_3}^{i_4} (m_i - m_0)^2\right] = \left(i_4 - i_3 + 1\right)\left(\mathrm{Var}\, m + \frac{a^2}{2}\right)$$

de même

$$\mathrm{E}\left[\sum_{i_4+1}^{i_3+364} (m_i - m_0)^2\right] = (i_3 + 364 - i_4)\left(\mathrm{Var}\, m + \frac{a^2}{2}\right)$$

d'où

$$\boxed{\mathrm{E}\left[\sum (m_i - m_0)^2\right] = 182{,}5\, a^2}$$

et une estimation de a.

A Nantes et Paris, on a appliqué les deux méthodes d'estimation. Les résultats sont les suivants :

Estimation de a	Nantes	Paris
par les moindres carrés	6,92	7,86
par a^2	6,96	7,89

Les deux résultats ont paru suffisamment voisins pour traiter les autres stations par la 2e méthode, plus légère et indépendante de i_3 et i_4.

c) *Estimation de i_3 et i_4*

En ajustant deux arcs de sinusoïde aux moyennes observées, on admet qu'elles ont pour espérance mathématique

$$m_0 + a \sin h_1 (i - i_1) \text{ pour } i_3 \leqslant i \leqslant i_4$$
$$m_0 + a \sin h_2 (i - i_2) \text{ pour } i > i_4 \text{ et } i < i_3$$

Donc si l'on ajuste deux droites à ces moyennes par la méthode des moindres carrés, la 1er D_1 entre i_3 et i_4, la 2e D_2 entre i_4 et $i_3 + 364$, les pentes obtenues auront pour espérance mathématique

$$a \sum_{i_3}^{i_4} (i - i_1) \sin h_1 (i - i_1) \Big/ \sum_{i_3}^{i_4} (i - i_1)^2$$
$$a \sum_{i_4+1}^{i_3+364} (i - i_2) \sin h_2 (i - i_2) \Big/ \sum_{i_4+1}^{i_3+364} (i - i_2)^2$$

Si l'on assimile ces sommes à des intégrales, ces expressions deviennent

$$24\, a/\pi^2 (i_4 - i_3)$$
$$-\, 24\, a/\pi^2 (i_3 + 355 - i_4)$$

Par ailleurs, les points de D_1 et D_2 d'abscisses respectives i_1 et i_2 ont pour ordonnée moyenne m_0. Ce sont donc des estimations de

$$i_1 = (i_3 + i_4)/2$$
$$i_2 = (i_3 + 365 + i_4)/2$$

Chacune de ces droites, construite à partir d'une première évaluation de i_3 et i_4 donne donc une estimation de $i_3 + i_4$ et $i_4 - i_3$, donc de i_3 et i_4. A Paris et à Nantes, l'écart entre les deux estimations obtenues restait inférieur à 2 jours. Nous avons retenu l'entier le plus proche de la moyenne de ces deux estimations.

1.2.2. *Ajustement des écarts-types*

La forme du nuage incite à ajuster aux écarts-types une courbe de la forme

$$s_i = s_0 + a_1 \sin h (i - i_1) + a_2 \sin 2 h (i - i_2)$$

avec

$$h = \frac{2\pi}{365}$$

L'estimation des coefficients par la méthode des moindres carrés donne

$$s_0 = \sum s_i/365$$

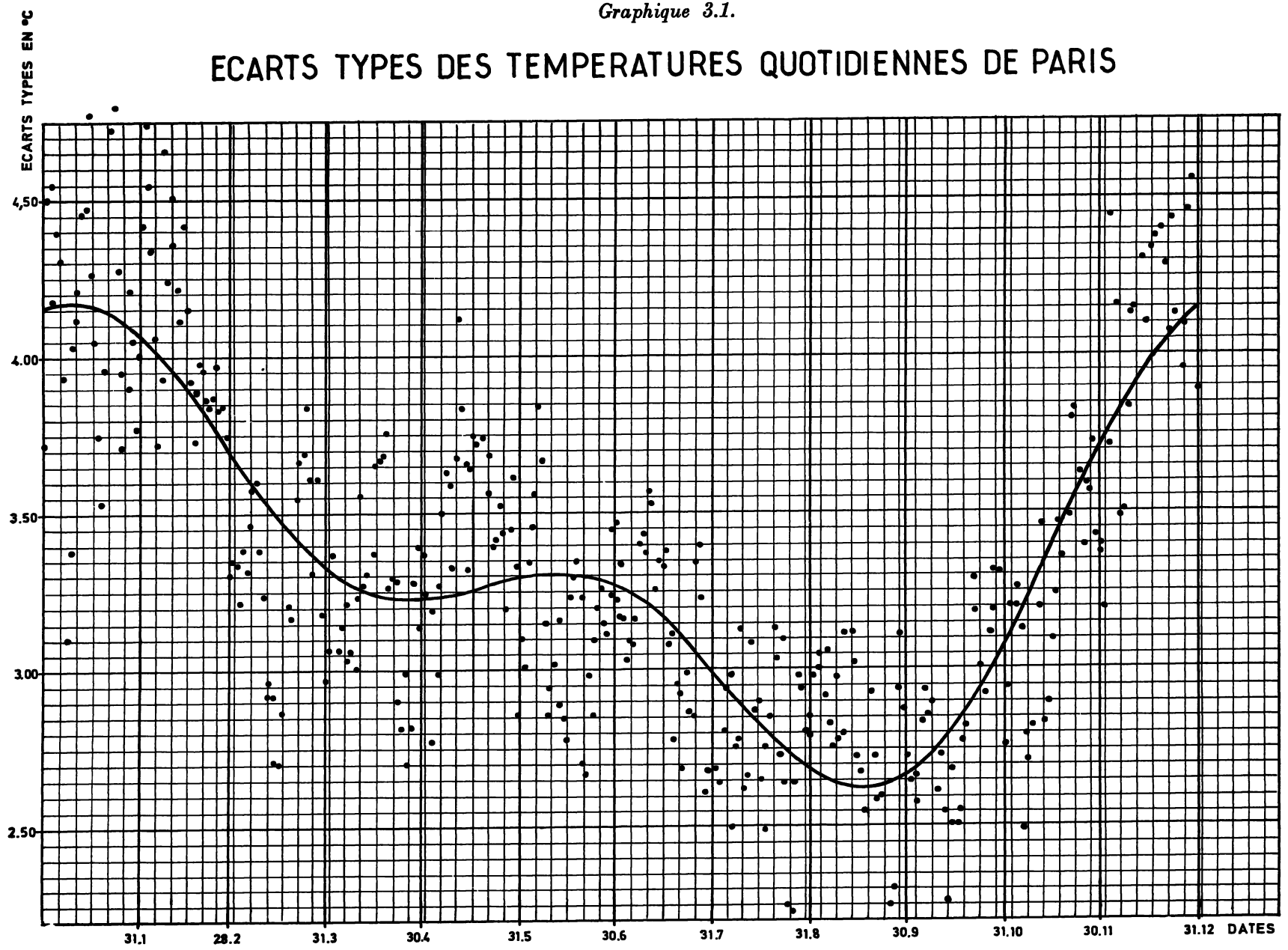
$$a_1 = \frac{2}{365} \sum s_i \sin h (i - i_1)$$

$$a_2 = \frac{2}{365} \sum s_i \sin 2 h (i - i_2)$$

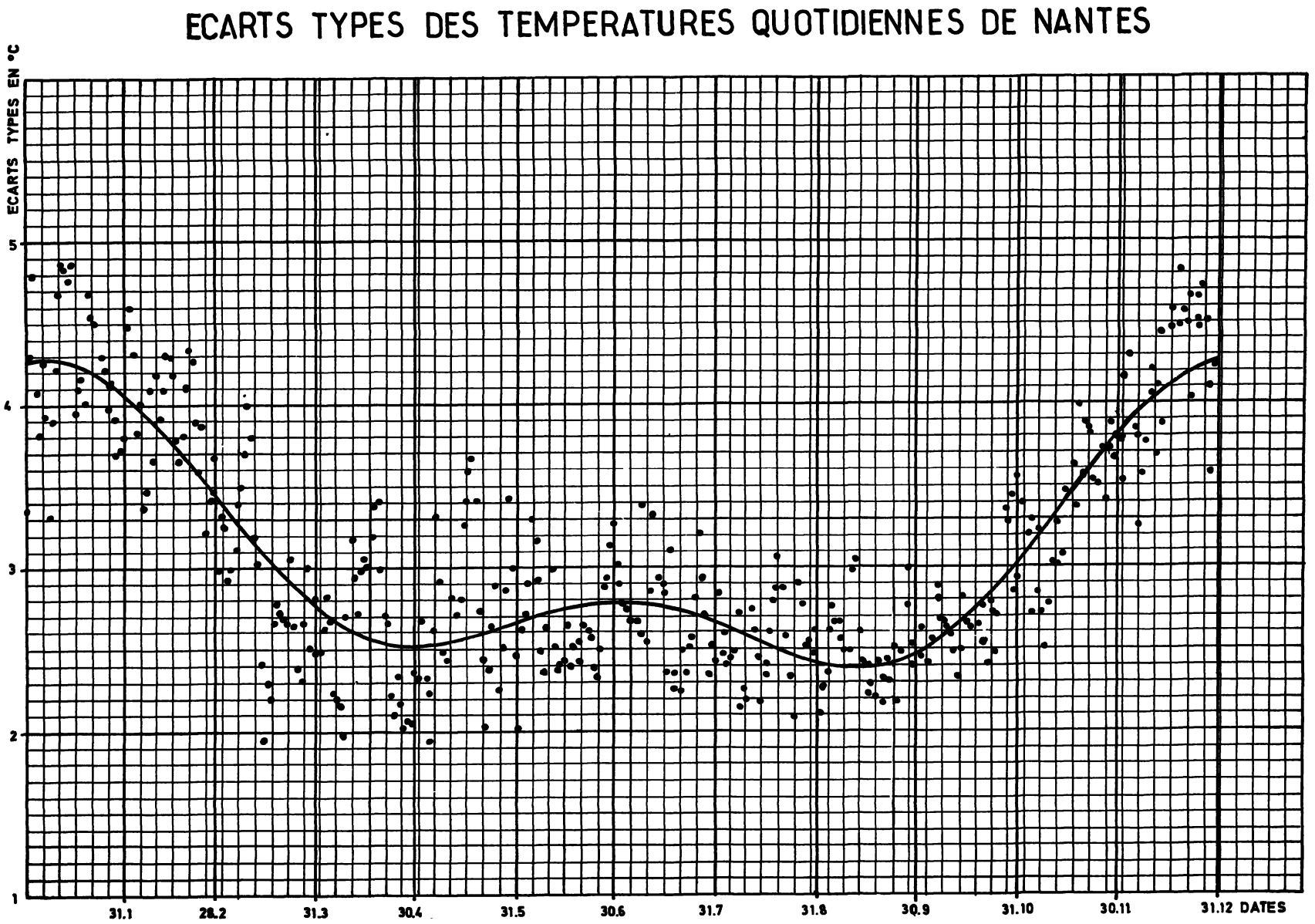
$$\operatorname{tg} hi_1 = -\sum s_i \cos hi \Big/ \sum s_i \sin hi$$

$$\operatorname{tg} 2 hi_2 = -\sum s_i \cos 2 hi \Big/ \sum s_i \sin 2 hi$$

Graphique 3.1.



Graphique 3.2.



1.3. *Estimation des paramètres pour les stations traitées avec des séries de températures mensuelles*

Les données de base étaient ici les températures moyennes mensuelles θ_{mj} du mois m de l'année j. On a calculé leur moyenne

$$\bar{\theta}_m = \frac{1}{n} \sum_j \theta_{mj}$$

et leur variance

$$\sigma_m^2 = \frac{1}{n-1} \sum_j (\theta_{mj} - \bar{\theta}_m)^2$$

Nous avons reconduit à ces stations le modèle admis pour celles que nous avions traitées à partir de données quotidiennes.

Les $\bar{\theta}_m$ sont les moyennes mensuelles de températures moyennes quotidiennes « observées ». Mais la variance mensuelle s'écrit, en fonction des écarts-types quotidiens « observés »

$$\sigma_m^2 = \frac{1}{n_m^2} \left[\sum_{i \in m} s_i^2 + 2 \sum_{\substack{i,j \in m \\ i<j}} \sum r^{j-i} s_i s_j \right]$$

n_m étant le nombre de jours du mois m.
Si l'on assimile tous les s_i à leur moyenne s_m du mois m cette relation devient

$$n_m^2 \sigma_m^2 = s_m^2 \, \mathrm{A}(n_m)$$

avec

$$\mathrm{A}(n_m) = n_m + \sum_{i=1}^{n_m - 1} \sum_{j=i+1}^{n_m} r^{j-i} \# n_m + \frac{2r}{1-r} \left(n_m - \frac{1}{1-r} \right)$$

Sur les stations traitées avec des données quotidiennes les coefficients de corrélation à 1 jour d'intervalle r varient de 0,75 à 0,80. Pour celles que nous traitions en mensuel, on a admis

$$r = 0{,}77$$
$$\mathrm{A}(n_m) = 7{,}70\, n_m - 66{,}92$$

La moyenne mensuelle des écarts-types de températures quotidiennes s'écrit alors

$$s_m = \sigma_m \, n_m / \sqrt{\mathrm{A}(n_m)}$$

Pour estimer les paramètres des moyennes et écarts-types ajustés, on a affecté à tous les jours du mois m la température $\bar{\theta}_n$ et l'écart-type quotidien s_m et utilisé les programmes établis pour les stations traitées avec des données quotidiennes.

2. ÉTUDE DU NOMBRE DE JOURS DONT LA TEMPÉRATURE EST INFÉRIEURE A T

2.1. *Forme de la distribution adoptée*

Pour une période donnée (hiver de septembre à juin, mois ou quinzaine), on notera n (T) le nombre de jours plus froids que la température T fixée. Lorsqu'on considère les diverses années, ce nombre est une variable aléatoire. Pour répartir au mieux ses investissements entre équipements de base et de pointe, le Gaz de France s'est intéressé à la distribution de ces n (T) pour tout T et en particulier

— à leur moyenne N (T),

— aux valeurs N_2 (T), N_{50} (T), N_{98} (T) telles que n (T) a respectivement une probabilité 2 %, 50 % et 98 % d'être inférieur à N_2 (T), N_{50} (T) et N_{98} (T).

A première vue, la distribution des n (T) paraît comparable à une loi binomiale (forme en J pour les températures très froides, puis glissement du mode de gauche à droite quand les températures T se réchauffent). La définition serait effectivement celle d'une loi binomiale si les lois quotidiennes de températures étaient toutes identiques et les températures quotidiennes indépendantes les unes des autres.

Elle pourrait sans doute être encore approchée par une loi de Poisson pour les températures T très froides, une loi normale pour les températures moyennes, malgré les lois quotidiennes de températures variables avec la date. Mais, en accroissant la variance de ces n (T) à moyenne donnée, les liaisons entre températures de 2 jours consécutifs ne permettent pas de représenter la distribution par une loi de Poisson.

On doit donc rechercher une forme de distribution :

— dont l'histogramme évolue comme celui d'une loi binomiale lorsque les températures étudiées varient des plus froides aux plus courantes,

— dont la variance V $[n(T)]$ puisse différer de la moyenne.

Une loi Γ généralisée

$$n\,(T) = a\,\Gamma_p$$

répond à cette double condition.

Dans une pré-étude effectuée sur Nantes, on a simulé 500 années de températures suivant un modèle voisin de celui qui est retenu ici. Sur cet échantillon, on a effectué 2 tests χ^2 :

— le 1er avec $T = -2$ a donné

$$\chi^2 = 19{,}27 \text{ avec } 13 \text{ degrés de liberté}$$

c'est à-dire une valeur qui a plus de 10 chances sur 100 d'être dépassée.

— le 2e avec $T = 3$ a donné

$$\chi^2 = 39{,}70 \text{ avec } 41 \text{ degrés de liberté.}$$

Nous avons donc retenu, au moins provisoirement, cette forme de distribution.

2.2. *Exécution des calculs*

2.2.1. *Calcul de la moyenne* N (T) *et de la variance* V [n(T)]

L'établissement de ces formules est calqué sur l'étude d'une loi binomiale. On associe à chaque jour i une variable

$$z_i = 0 \quad \text{si} \quad t_i > T$$
$$= 1 \quad \text{si} \quad t_i \leqslant T$$

on a alors

$$n\,(T) = \sum_i z_i$$

la somme étant étendue à tous les jours de la période étudiée (quinzaine, mois, ou hiver de septembre à juin).

D'où

$$N\,(T) = \sum_i F\,(x_i)$$

x_i étant la variable centrée réduite

$$x_i = \frac{T - m_i}{s_i}$$

et F la fonction de répartition de la loi normale centrée réduite

$$V[n(T)] = \sum_i F(x_i)[1 - F(x_i)] + 2\sum\sum_{i<j}[P(x_i, x_j, r_{j-i}) - F(x_i)F(x_j)]$$

$P(x_i, x_j, r_{j-i})$ étant la fonction de répartition de la loi normale à 2 variables x_i et x_j.

Le coût de ces calculs était trop élevé pour les 74 stations. Nous avons donc divisé la France en 7 zones climatiques, et exécuté intégralement les calculs dans la « station principale » de chaque zone. Pour les autres, on a recherché, au niveau de chaque quinzaine, la température T_0 telle que $n(T_0)$ ait à peu près la même distribution, dans la station principale, que $n(T)$ dans la station secondaire étudiée. Pour cela, on a évalué, par quinzaine,

m_{q_0} = température moyenne de la quinzaine q dans la station principale,
s_{q_0} = moyenne des écarts-types ajustés de la quinzaine q dans la station principale,
m_q = température moyenne de la quinzaine q dans la station à étudier,
s_q = moyenne des écarts-types ajustés de la quinzaine q dans la station à étudier,

et égalé les variables centrées réduites.

$$\frac{T_0 - m_{q_0}}{s_{q_0}} = \frac{T - m_q}{s_q}$$

On admet alors, au niveau de chaque quinzaine

$$N(T) = N_0(T_0)$$
$$= \text{moyenne de } n(T_0) \text{ pour la station principale}$$
$$V[n(T)] = V_0[n(T_0)]$$
$$= \text{variance de } n(T_0) \text{ pour la station principale}$$

et on en déduit les totaux mensuels et annuel. Les couples de jours (i, j) qui chevauchent 2 quinzaines ont été intégrés dans la variance totale avec le T_0 du jour i.

2.2.2. *Calcul de* $N_2(T)$, $N_{50}(T)$, $N_{98}(T)$

On rappelle que

$$n(T) = a\Gamma_p$$

Γ_p suivant une loi Γ de paramètre p dont la moyenne $p + 1$ est égale à la variance. Par suite,

$$a(p + 1) = N(T)$$
$$a^2(p + 1) = V[n(T)]$$

d'où

$$a = \frac{V[n(T)]}{N(T)}$$
$$p = \frac{N^2(T)}{V[n(T)]} - 1$$

une table de loi Γ associe pour chaque p, la fonction de répartition à la variable réduite

$$u = \frac{n(T)}{\sigma[n(T)]}$$

on en déduit

$$N_2(T) = u_2(p)\,\sigma[n(T)]$$
$$N_{50}(T) = u_{50}(p)\,\sigma[n(T)]$$
$$N_{98}(T) = u_{98}(p)\,\sigma[n(T)]$$

J. MEYRIGNAC
(*Gaz de France*)