REVUE DE STATISTIQUE APPLIQUÉE

R. SNEYERS

Sur un critère de sélection de séries multidimensionnelles types à usage normalisé

Revue de statistique appliquée, tome 27, n° 2 (1979), p. 69-74

<http://www.numdam.org/item?id=RSA_1979__27_2_69_0>

SUR UN CRITÈRE DE SÉLECTION DE SÉRIES MULTIDIMENSIONNELLES TYPES A USAGE NORMALISÉ

R. SNEYERS

Institut royal météorologique de Belgique

1. INTRODUCTION

Au cours des dernières années, les services climatologiques ont été confrontés de plus en plus fréquemment avec la demande, d'un type nouveau, relative à des séries chronologiques multidimensionnelles devant permettre l'étude comparative objective de divers systèmes dépendant directement des conditions météorologiques comme, par exemple, les sytèmes de captation d'énergie solaire ou les systèmes de conditionnement d'air.

Le souhait des utilisateurs est de recevoir pour des époques spécifiées de l'année des séries d'observations représentatives des conditions météorologiques moyennes, c'est-à-dire, fournissant le meilleur résumé climatologique de l'époque de l'année considérée et respectant aussi bien les liaisons naturelles entre valeurs simultanées que celles entre les valeurs successives.

En d'autres termes, ces séries doivent conduire à des valeurs empiriques aussi proches de la normale que possible :

1) de la moyenne et de la variance de chaque élément ;

2) des covariances entre les divers éléments ;

3) des corrélations sériales qui régissent l'évolution de ces éléments dans le temps.

Diverses solutions à ce problème ont déjà été proposées ou sont à l'étude, notamment dans le cadre du sous-projet "conditions climatiques et année de référence" inclu au projet OTAN-CCMS "L'emploi rationnel de l'Energie" [1], [2], [3], [4]. On doit malheureusement les considérer comme inadéquates en raison soit de leur manque de généralité, soit de la complexité de leur méthodologie, soit encore du manque d'uniformité du résultat auquel elles conduisent.

Ces écueils peuvent toutefois être évités si l'on adopte une méthode statistique de sélection fondée sur la valeur prise par la forme quadratique qui est à la base de la méthode des moindres carrés généralisés. Une telle méthode offre, en outre, l'avantage de permettre les sélections les plus diverses, puisqu'elle s'applique aussi bien dans le cas du choix du meilleur résumé climatologique naturel qu'à celui du choix de l'ensemble naturel caractéristique des conditions les plus extrêmes.

Le but de cette note est de présenter la statistique de sélection et d'en donner un exemple d'application.

Mots-clés : **Sélection ; séries climatologiques.**

2. LE CRITERE DE SELECTION

Soit la série aléatoire simple de valeurs du vecteur x à k composantes :

$$x_i(x_{ij}) \quad , \quad i = 1, 2, \ldots, n \; ; \; j = 1, 2, \ldots, k \tag{1}$$

représentant l'évolution d'une année à l'autre d'une variable météorologique à k dimensions. En outre, soit $v \equiv ||v_{jj'}||$ la matrice (non singulière) des variances covariances des éléments du vecteur x, c'est-à-dire, qu'on a :

$$v_{jj'} = \text{cov}(x_j, x_{j'}) \; , \quad j, j' = 1, 2, \ldots, k \tag{2}$$

Si $\mu(\mu_j) \equiv E[x(x_j)]$ est la moyenne du vecteur x, la statistique définie par le produit matriciel :

$$X_i = (x_i - \mu)' \, v^{-1} (x_i - \mu) \tag{3}$$

possédera une répartition de χ^2 à k degrés de liberté dès que les composantes du vecteur x sont réparties conjointement selon une loi normale ; cette répartition ne sera qu'approchée dans le cas contraire.

Cela étant, l'élément de la série x_i qui s'approchera le mieux de la moyenne μ sera celui qui conduit à la valeur de X_i la plus proche de zéro. De plus, grâce à la loi de χ^2, on pourra faire une estimation exacte ou approchée de la probabilité associée à chaque valeur de X_i, c'est-à-dire à chaque élément x_i de la série chronologique.

Dans la pratique, le vecteur μ et la matrice ν ne sont pas connus ; aussi convient-il de les remplacer dans (3) par leurs estimations :

$$\hat{\mu} = \bar{x} \quad \text{avec} \quad \bar{x} = (\Sigma \, x_i)/n$$

$$\hat{v} \equiv ||\hat{v}_{jj'}|| \quad \text{avec} \quad \hat{v}_{jj'} = \sum_i (x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'})/(n-1) \tag{4}$$

La statistique de sélection devient ainsi :

$$\hat{X}_i = (x_i - \hat{\mu})' \, \hat{v}^{-1} \, (x_i - \hat{\mu}), \tag{5}$$

sa répartition selon une loi de χ^2 n'étant plus qu'approchée dans tous les cas.

Notons toutefois que si $\hat{X}'_l$, $l = 1, 2, \ldots, n$ désigne la série ordonnée des valeurs de $\hat{X}_i$ et si $F(\hat{X})$ est la fonction de répartition de la statistique $\hat{X}_i$, comme on a (cf [5] p. 19) :

$$E[F(\hat{X}'_l)] = l/(n+1), \tag{6}$$

en posant :

$$\hat{F}(\hat{X}'_l) = l/(n+1) \tag{7}$$

on peut procéder à une estimation non paramétrique de la probabilité associée aux valeurs de la statistique $\hat{X}_i$.

3. APPLICATION A LA SELECTION DE SERIES CLIMATOLOGIQUES TYPE.

Dans la pratique, les séries demandées sont des séries de valeurs journalières d'un ou de plusieurs éléments météorologiques appartenant à un même mois. Si N est le nombre de jours du mois considéré et si n est le nombre d'années d'observa-

tion dont on dispose, la sélection doit conduire au choix d'une série de N observations journalières d'un même mois parmi n séries semblables.

Dans ces conditions, si y_i, $s_i^2(y)$ et $r_i(y)$, $i = 1, 2, \ldots, n$, sont pour l'élément météorologique y, les valeurs empiriques respectives de la moyenne, de la variance et de la corrélation sériale fournies par les valeurs journalières de chacun des n mois de la période d'observation, la sélection devra porter sur les valeurs d'un vecteur x comportant un nombre de triplets $[y, s^2(y), r(y)]$ égal au nombre d'éléments météorologiques considérés simultanément.

Il est clair toutefois que le nombre de composantes du vecteur x peut être réduit selon les nécessités pratiques, par exemple, en abandonnant un ou plusieurs coefficients de corrélation sériale.

Par ailleurs, afin de pouvoir déterminer correctement la part qui revient à chaque élément dans la valeur trouvée pour la statistique $\hat{X}_i$, il peut être utile de passer aux composantes centrées réduites définies par la relation :

$$z_i = \frac{x_i - \hat{\mu}_i}{\sqrt{v_{ii}}} \tag{1}$$

Dans ce cas, la matrice des variances-covariances des composantes z_i se ramène à la matrice de corrélation $\hat{\rho} \equiv ||\hat{\rho}_{jj'}||$ entre les divers éléments du vecteur x.

4. EXEMPLE. SELECTION DES MOIS CONSTITUANT L'ANNEE TYPE DANS LE CAS DE LA VARIABLE CONJOINTE TEMPERATURE DE L'AIR-RAYONNEMENT A UCCLE (PERIODE 1958-1975).

Les composantes de la variable multidimensionnelle considérée ici sont les moyennes mensuelles de la température de l'air et du rayonnement reçu sur une surface horizontale ainsi que les variances empiriques déduites des valeurs journalières de ces éléments calculées pour chaque mois de la période 1958-1975. Ceci conduit, pour le mois de janvier, aux valeurs t_i, $v_i(t)$, g_i et $v_i(g)$ du tableau 1, ainsi qu'aux valeurs réduites x_1, x_2, x_3 et x_4 calculées selon 3.(1). Les statistiques $\hat{X}$ ont ensuite été déterminées pour les couples (x_1, x_2), (x_3, x_4) ainsi que pour la variable à 4 dimensions (x_1, x_2, x_3, x_4).

Il apparaît de la sorte que pour la température seule, deux mois de janvier, ceux de 1961 et de 1970, s'approchent le mieux de la normale, tandis que pour le rayonnement seul, la statistique $\hat{X}$ conduit à janvier 1960. Par contre, pour les deux éléments pris conjointement, le mois de janvier à retenir est celui de 1967.

Il ressort, en outre, de ces résultats que les conditions extrêmes observées en janvier 1963 ont été caractérisées par un déficit très important de la température et par un excès un peu moins important à la fois de la moyenne et de la variabilité du rayonnement.

L'adéquation d'une loi de χ^2 à 4 degrés de liberté à la répartition de la statistique $\hat{X}$ dans le cas des variables conjointes x_1, x_2, x_3 et x_4 a été vérifiée, par ailleurs, en examinant les probabilités que cette loi assigne aux valeurs extrêmes et à la 9ê valeur (médiane inférieure) de la statistique obtenues de cette manière pour chacun des douze mois de l'année (tableau 2).

A cet effet, on note que si F est la loi de répartition des éléments d'une série aléatoire simple d'effectif égal à 18, la loi de répartition du maximum de la

TABLEAU 1

Valeurs empiriques des moyennes t_i et g_i de la température de l'air et du rayonnement à Uccle et variances correspondantes des valeurs journalières $v_i(t)$ et $v_i(g)$ pour les mois de janvier de 1958 à 1975. Valeurs réduites x_1, x_2, x_3 et x_4 et valeurs de la statistique $\hat{X}_i$ pour les couples (x_1, x_2), (x_3, x_4) et pour le vecteur (x_1, x_2, x_3, x_4).

	Température t (0,1°C)					Rayonnement g (J/cm²)					t + g
Année (i)	t_i	$v_i(t)$	x_1	x_2	$\hat{X}_i$	g_i	$v_i(g)$	x_3	x_4	$\hat{X}_i$	$\hat{X}_i$
1958	37	848	−0,02	−0,63	0,41	202	26 693	−1,01	0,79	2,90	3,10
59	26	866	−0,45	−0,60	0,69	300	28 159	2,24	1,04	5,02	6,92
60	37	2 858	0,00	1,97	4,03	245	21 779	0,40	−0,04	0,22	4,04
61	35	988	−0,07	−0,45	0,22	247	18 775	0,50	−0,54	0,96	1,49
62	49	1 577	0,53	0,32	0,46	251	34 583	0,60	2,12	4,64	9,51
63	36	1 256	−3,09	−0,10	10,02	287	33 507	1,80	1,94	4,87	10,71
64	20	616	−0,70	−0,93	1,65	198	19 658	−1,13	−0,39	1,28	3,75
65	37	683	−0,01	−0,84	0,74	208	18 703	−0,80	−0,55	0,70	1,46
66	23	3 202	−0,60	2,42	5,88	256	22 874	0,77	0,15	0,64	5,99
67	43	1 930	0,27	0,77	0,77	230	18 140	−0,08	−0,65	0,48	1,17
68	34	1 809	−0,11	0,62	0,38	210	21 850	−0,73	−0,02	0,64	1,43
69	62	783	1,07	−0,71	1,42	215	13 502	−0,59	−1,43	2,05	2,43
70	40	1 655	0,15	0,42	0,22	200	19 564	−1,06	−0,41	1,12	1,73
71	45	1 683	0,35	0,45	0,40	257	15 159	0,80	−1,15	3,45	3,52
72	25	1 392	−0,49	0,08	0,24	220	28 030	−0,42	−1,02	1,98	2,02
73	41	446	0,19	−1,15	1,32	197	19 681	−1,17	−0,39	1,38	2,22
74	69	877	1,37	−0,59	1,99	242	17 160	0,31	−0,81	1,21	3,13
75	75	529	1,61	−1,04	3,15	219	18 088	−0,44	−0,66	0,46	3,38

TABLEAU 2

Probabilités F_{max}, F_{min} et F_9 que la loi du χ^2 à 4 degrés de liberté assigne à la plus grande valeur de la statistique $\hat{X}$, à sa plus petite valeur et à la 9e valeur dans la série ordonnée de 18 valeurs obtenues pour chacun des douze mois

Mois	F_{max}	F_{min}	F_9
1	0,969	0,119	0,459
2	0,932	0,008	0,475
3	0,861	0,108	0,551
4	0,966	0,017	0,311
5	0,889	0,115	0,524
6	0,839	0,039	0,538
7	0,958	0,134	0,518
8	0,935	0,014	0,388
9	0,953	0,096	0,405
10	0,979	0,048	0,366
11	0,906	0,117	0,423
12	0,975	0,014	0,256

série est F^{18} (cf [5] p. 19) et il s'ensuit que les 12 valeurs de F^{18} qu'on en déduit seront réparties selon une loi uniforme, c'est-à-dire que la statistique de Fisher correspondante :

$$-2\sum_{1}^{12}\ln(F^{18}) = -36\sum_{1}^{12}\ln F \qquad (1)$$

possède une répartition de χ^2 à $12 \times 2 = 24$ degrés de liberté.

Comme les valeurs de F_{max} et de $(1 - F_{min})$ du tableau 2 conduisent respectivement à 31,8 et à 31,5 pour une valeur critique de 36,4 au niveau 0,05, on peut conclure à un accord acceptable.

De même, pour F_9, on a (cf [5] p. 20) :

$$E(F_9) = 9/19 \quad , \quad \text{var } F_9 = 9/19\,[1 - (9/19)]/20 \qquad (2)$$

d'où, pour la moyenne $\overline{F}_9$ des douze valeurs de F_9 :

$$\text{var } \overline{F}_9 = (\text{var } F_9)/12 = 0{,}001039. \qquad (3)$$

Avec, $\overline{F}_9 = 0{,}434$, il vient ainsi :

$$[E(F_9) - \overline{F}_9]^2/\text{var } \overline{F}_9 = 1{,}54 \qquad (4)$$

soit à nouveau une valeur inférieure à la valeur critique 3,84 au niveau 0,05 de la variable χ^2 à 1 degré de liberté, et ici aussi un bon accord avec l'hypothèse nulle.

5. CONCLUSIONS

En résumé, le critère de sélection proposé présente effectivement les qualités de rationalité, de simplicité et d'objectivité indispensables au choix normalisé de séries type pour les applications dont il a été fait mention au début de cette note.

Accessoirement, on aura remarqué la bonne adéquation de la répartition des valeurs de la statistique $\hat{X}$ avec une loi de χ^2 et ce bien qu'aucune précaution n'ait été prise pour s'assurer de la normalité de la répartition conjointe des éléments qui composent le vecteur x.

Il va de soi qu'on peut souhaiter réunir au maximum les conditions nécessaires à cette normalité en procédant aux transformations appropriées. Rappelons simplement à ce propos qu'ici cette transformation consiste en une fonction puissance (ici la racine carrée, puisque l'on a substitué l'écart-type empirique à la variance empirique) et que dans le cas où il aurait été fait appel au coefficient de corrélation sériale $\hat{\rho}$, cette transformation est donnée par la relation :

$$x_\rho = \log \frac{1 + \hat{\rho}}{1 - \hat{\rho}}$$

BIBLIOGRAPHIE

[1] LUND H. – Report A 3, Second symposium on the Use of Computers for Environmental Engineering Related to Buildings, Paris, June 1974.

[2] SAITO H. and MATSUO Y. – Standard Weather Data of Shase Computer Program of Annual Energy Requirements and Exemple Results of Hourly Load for 10 Years in Tokyo, Paper A 4, Second Sympsoium on the Use of Computers for Environmental Engineering Related to Buildings, Paris, June 1974.

[3] ANDERSEN B., EIDORFF S., LUND H., PEDERSEN E., ROSENORN S., VALBYORN O. – Meteorological Data for Design of Building and Installation : A Reference Year. Status Byggeforskningsinstitut, SBI – Rapport 89 – Ko benhavn 1974.

[4] KORSGAARD V., LUND H. – Test reference Year (TRY), Final report NATO - CCMS Project on "Rational Use of Energy" - Sub-Project" Climatic Conditions and Reference Year", 1977.

[5]/ SNEYERS R. – Sur l'analyse statistique des séries d'observations, Note technique n° 143, OMM, Genève 1975.