CAHIER DU LAMSADE

Laboratoire d'Analyse et Modélisation de Systèmes pour l'Aide à la Décision
(Université de Paris-Dauphine)
Equipe de Recherche Associée au C.N.R.S. N° 656

METHODES ROBUSTES POUR LE TRAITEMENT DE SERIES CHRONOLOGIQUES

Cahier N° 55
Juin 1984

Ph. VALLIN

Sommaire

Résumé

On présente dans ce cahier des outils de traitement de séries chronologiques - détection de saisonnalité et détection de valeurs anormales - fondés sur des méthodes statistiques robustes et non paramétriques. Ces méthodes sont mieux adaptées aux historiques courts, fréquemment rencontrés dans la pratique, que les méthodes statistiques classiques.

Mots clés : méthodes robustes, valeurs aberrantes, saisonnalité.

Robust methods for time series analysis

Abstract

This paper presents some time series analysis tools used to detect saisonnality and to detect abnormal values of the serie. Those tools are based on nonparametric and robust methods. The methods presented here are more adapted than the classical statistical methods dialing with short historical series frequently met in practise.

Key words : robust methods, abnormal values, seasonality.

Introduction

Les méthodes quantitatives utilisées en gestion et, plus particulièrement dans l'analyse des séries chronologiques, reposent, le plus souvent, sur les techniques statistiques classiques paramétriques. Ces techniques, si elles se prêtent bien à l'expression analytique, et donc facilement calculables, fournissent des résultats très sensibles aux variations d'échantillonnage et présupposent un corps d'hypothèses qui ne sont pas toujours vérifiées.

Lors de la mise en place de systèmes de gestion informatisés, la pauvreté des données disponibles dans l'organisation précédente ne permet pas, le plus souvent, d'appliquer avec rigueur ces méthodes.

Désormais, les moyens de calcul dont on dispose permettent d'utiliser des méthodes réclamant plus d'opérations (tri, calcul combinatoire, méthode "Boutstrap") mais s'adaptant mieux au manque de données et à l'abandon des hypothèses de la statistique classique (existence d'une loi caractérisée par ses paramètres).

Dans le cadre d'analyse de séries chronologiques, nous présentons, dans ce qui suit, deux méthodes originales que l'on peut qualifier de robustes - au sens de la stabilité des résultats aux variations d'échantillonnage - et non paramétriques - la loi génératrice de la chronique n'étant pas spécifiée.

Ces méthodes sont inspirées des filtres de lissages robustes et non linéaires (en particulier la médiane mobile) présentées par J.W. Tuckey et P.F. Velleman (Velleman 1980) et des tests non paramétriques de Spearman.

La première méthode traite de la détection automatique des valeurs "anormales" dans une série chronologique. Cette opération de filtrage quasiment indispensable dans la pratique lorsqu'on veut estimer les paramètres d'un modèle est très peu étudiée dans la littérature relative aux méthodes quantitatives appliquées à la gestion sous une forme algorithmique.
Dans une première partie, nous décrirons donc différentes méthodes en partant de techniques statistiques classiques pour arriver à la méthode adaptative et non paramètrique que nous préconisons; quelques résultats informatiques sont présentés pour éclairer la démarche adoptée.

La seconde méthode traite du repérage de séries saisonnières lorsque l'on ne dispose que d'un très court historique, ce qui ne permet pas l'approche économétrique classique.
Ces tests de saisonnalité sont par contre beaucoup mieux connus (cf Usunier-Bourbonnais, 1982) et nous nous contenterons , dans une seconde partie, de présenter la méthode proposée.

I - PREMIERE PARTIE

Méthodes de détection automatique de valeurs aberrantes.

La gestion des stocks, l'élaboration des plans de production, l'organisation de la distribution, nécessitent l'analyse des historiques de ventes. Dans le schéma général du traitement des chroniques, tous les auteurs préconisent la correction d'influences accidentelles ou valeurs "anormales".

Lorsqu'on est conduit à mener une analyse détaillée des chroniques de ventes avec un responsable des ventes, on est systématiquement amené à faire une partition entre :

. un processus d'activité régulière et habituelle qualifiée par l'homme d'expérience de "normale",

. un processus de ventes ou méventes accidentelles qui s'expliquent par des demandes exceptionnelles (tant en intensité qu'en fréquence) ou plus prosaïquement par des erreurs de saisies.

Lorsqu'on veut modéliser la demande pour établir des règles de gestion, il convient de filtrer (ou d'isoler) l'incidence des phénomènes exceptionnels pour ne tenir compte que du premier processus qui, lui, conditionne l'activité courante de l'entreprise.

La perturbation liée aux accidents est particulièrement sensible lorsque la chronique représente des ventes de faible niveau, car les estimations de la moyenne et de la variance sont largement surestiméespar la présence de sorties "éléphants".

L'automatisation du traitement de ces séries nécessite un dépistage automatique de ces valeurs "anormales".

La pauvreté de la littérature sur le sujet (1) et l'absence de méthode théorique vient du fait qu'il n'apparait pas clairement où

---

(1) : Selon J.C. Laboire (1972), il n'existe aucune méthode, sur le plan théorique pour le traitement des valeurs aberrantes. Les propositions qui sont effectuées reposent sur la détermination d'un seuil exprimé en nombre d'écarts-type de la différence entre tendance et observation désaisonnalisée, cet écart étant supposé suivre une loi de Gauss.

se situe la frontière entre la normalité et l'aberration. Cette frontière est d'autant plus délicate à tracer qu'un processus dynamique stable génère des événements que l'on peut qualifier expérimentalement "d'anormal", les longues séries de même couleur à la roulette, par exemple.

Il semble bien que pour le gestionnaire, le critère discriminant soit la fréquence du phénomène que celui-ci provienne d'un processus particulier, étranger au processus "normal" (exportation exceptionnelle, grève, accident météorologique...) ou d'une manifestation exceptionnelle, rare (au sens probabilité) du processus "normal".

L'utilisation de seuils, préconisée par certains auteurs, définit implicitement ce concept de valeur anormale comme un phénomène rare qui, par sa présence dans la chronique étudiée, biaise l'échantillon et doit être éliminé (1).

On retrouve implicitement cette notion en gestion de stocks, lors qu'on définit une qualité de service : on ne peut régler le système de gestion pour faire face à toute éventualité, on choisit de ne pas faire face à certaines situations jugées "exceptionnelles".

Après avoir présenté deux méthodes classiques basées sur des tests d'hypothèses, nous présenterons une méthode non paramétrique, robuste dans le sens où elle ne présuppose pas la connaissance de la loi initiale et fournit des résultats peu sensibles aux variations des paramètres à estimer pour sa mise en oeuvre.

---

(1) : Voir la méthode CENSUS II présentée brièvement par Wheelwright et Makridakis (1980).

I.1. Les méthodes paramétriques

a - La méthode de Grubbs (Carletti 1976)

Cette méthode suppose la normalité des observations, une transformation éventuelle des données de base permet d'obtenir cette condition (Carletti 1976).

Pour tester la conformité de l'observation x_i au reste de l'échantillon, on applique alors le test de comparaison entre deux moyennes de deux échantillons,

l'un réduit à un élément : x_i

l'autre composé des autres observations $x_1, x_2, x_{i-1}, x_{i+1}, \ldots x_n$ où n est la longueur de la série.

Si les x_i sont indépendants, alors on sait que :

$$\frac{(x_i - \bar{x}')}{\sqrt{1 + \frac{1}{n-1}}} \Bigg/ \sqrt{\frac{\sum_{j \neq i} (x_j - \bar{x}')^2}{n - 2}} = T$$

suit une loi de Student à n - 2 degrés de liberté.
$\bar{x}'$ représente la moyenne des n - 1 observations du deuxième échantillon.

Si on se fixe α, risque de première espèce, on comparera dans un test unilatéral, la valeur calculée

$$T_{calc} = \frac{|x_i - \bar{x}'|}{\sqrt{\frac{n \sum_{j \neq i} (x'_j - \bar{x}')^2}{(n-1)(n-2)}}}$$ — à la valeur théorique lue

dans la table de Student à n - 2 debrés de liberté $T_{1-\alpha/2n}$

Remarquons que le risque d'écarter une valeur alors qu'elle est normale est fixé ici à $\alpha \times \frac{1}{2} \times \frac{1}{n}$,

$\frac{1}{2}$ car il s'agit du test unilatéral,

$\frac{1}{n}$ car on recommence le test pour les n observations de la chronique, si p est la probabilité de rejet après les n expériences (considérées approximativement comme indépendantes) alors le risque p' à prendre sur un test élémentaire est donné par :

$$p = 1 - (1 - p')^n \simeq np'$$

Pour éviter d'avoir à calculer les $\bar{x}'$ et $\sum_{j \neq i} (x_i - \bar{x}')^2$, pour tout j, on remarque que :

$$x_i - \bar{x}' = \frac{n}{n-1} (x_i - \bar{x})$$

$$\sum_{j \neq i} (x_j - \bar{x}')^2 = \sum_j (x_j - \bar{x})^2 - \frac{n}{n-1} (x_i - \bar{x})^2$$

En posant : $t = \dfrac{|x_i - \bar{x}|}{\sqrt{(n-1)\sum_j (x_j - \bar{x})^2}}$

On obtient : $$\boxed{t = \sqrt{\frac{(n-1)^2}{n} \cdot \frac{T_{calc}^2}{(n-2) + T_{calc}^2}}}$$

Pratiquement, on comparera t à la valeur T_0.

où $T_0 = \sqrt{\dfrac{(n-1)^2}{n} \cdot \dfrac{T^2_{1-\alpha/2n}}{(n-2) + T^2_{1-\alpha/2n}}}$

A partir de n = 10, on peut utiliser une valeur approchée de T_o :

$$T_o = \frac{n-2}{n} V_{1-\alpha/2n}$$

où V correspond à la fonction de répartition, F, de la loi Normale : $F(V_x) = x$.

Cette méthode suppose, pour être exacte que les observations traitées (après transformation éventuelle) suivent une loi normale, ou au moins une loi symétrique. Dans les problèmes de gestion, les demandes de produits sont rarement symétriques, la transformation des données pose alors un problème. Nous allons, par la suite, abandonner cette hypothèse.

b - Test sur les valeurs extrêmes de la série

Nous nous plaçons ici dans le cas de la détection des grandes valeurs (éléphants), cas le plus fréquemment rencontré. Une analyse symétrique pour les valeurs anormalement faibles peut être menée.

On suppose connue la loi des sorties (ventes) par unité de temps. On connait en particulier la tendance et la saisonnalité éventuelles.

On s'interressera alors à la série corrigée des variations saisonnières et de la tendance : $x_1, x_2, \ldots x_i, \ldots x_n$.

On note f(x) et F(x) la densité et la fonction de répartition de ces variables aléatoires considérées comme indépendantes.

On note $z = \max_{i \leq n} \{x_i\}$,

la loi de z est caractérisée par la densité :

$$h(z) = n\, f(z) \times F(z)^{n-1}$$

On peut alors procéder à un test d'hypothèses avec risques α .

Soit Z le seuil défini par : $\int_Z^{\infty} h(z)dz = \alpha$

où $h(z)$ est définie à partir des paramètres estimés sur les n-1 observations les plus faibles.

On considérera l'observation z comme anormale si elle dépasse la valeur Z.

L'observation $x_i = z$ est alors retirée de la chronique.

On peut recommencer le test avec la nouvelle série de longueur n-1. La probabilité de rejeter au moins une observation à tort est alors α .

Le seuil Z calculé pour une chronique de taille n est alors :

$$\int_Z^{\infty} h(z) = \int_Z^{\infty} n f(z) \left(F(z)\right)^{n-1} dz = \int_Z^{\infty} d\left[F(z)\right]^n = 1 - F^n(Z) = \alpha$$

D'où la relation : $F(Z) = (1-\alpha)^{1/n}$

Exemple :

Si une chronique suit une loi Γ de moyenne m = 140, σ = 100, l'un des paramètres caractéristiques de la loi est alors :

$$k = \frac{m^2}{\sigma^2} \simeq 2$$

Pour α = 10%, n = 100, on a :

$$F(Z) \simeq 1 - \frac{\alpha}{n} = 0,999$$

D'où : $Z = 900 \simeq 6,5\ m$

Cette méthode nécessite donc les opérations suivantes :

- i) recherche du maximum de la chronique,
- ii) calcul des paramètres,
- iii) recherche du seuil Z,
- iv) élimination éventuelle :
  si oui, retour en i,
  si non, stop.

La probabilité de rejeter plus de k observations à tort est α^k.

Les méthodes que nous venons de décrire dépendent sensiblement des valeurs de toutes les observations (estimation de paramètres).

Nous proposons dans les paragraphes suivants une méthode robuste (au sens où les valeurs des observations influent beaucoup moins sur le résultat) et une méthode non paramétrique (le type de la loi de probabilité des observations n'a pas besoin d'être spécifié).

I.2. Une méthode robuste paramétrique

Nous proposons ici une méthode d'élimination des valeurs anormales paramétriques puisqu'elle dépend d'une loi de demande identifiée et caractérisée par des paramètres, mais robuste au sens où la procédure d'élimination ne dépend pas de l'amplitude des valeurs anormales, contrairement au cas où on est amené à estimer les paramètres moyenne et variance en intégrant éventuellement les valeurs aberrantes. Cette méthode évitera d'itérer les opérations comme dans les méthodes 1 et 2.

La quasi totalité des distributions des séries économiques de ventes sont dissymétriques. Les modèles multiplicatifs conviennent mieux que les modèles additifs. Il est préférable d'utiliser des lois type log-normale. Nous utiliserons ce type de loi pour caractériser la chronique.

Conformément aux techniques de lissage robuste (cf Paul F. Velleman, 1980), nous nous intéresserons à l'estimation de la médiane, et du mode.

Ces estimateurs sont peu biaisés par la présence éventuelle de valeurs anormales puisqu'ils ne dépendent pas de leurs valeurs, mais de leur fréquence qui, par définition, est faible.

La procédure est simple :

i) estimation de la médiane et du mode,

ii) calcul des paramètres hors influence des valeurs anormales.
Cette phase s'effectue dans le cas de la loi Log-normale grâce à l'existence des relations (cf Callot, 1965) :

médiane $Med = e^{m}$

où m est la moyenne de la loi normale associée,

mode $M_o = e^{m-\sigma^2}$

où σ est l'écart-type.

On a : $m = \text{Log}\ (Med)$, $\sigma^2 = \text{Log}\ (\frac{Med}{M_o})$

La moyenne M et l'écart type S sont donnés par :

$$M = e^{m+\sigma^2/2} = Med \times \sqrt{\frac{Med}{M_o}}$$

$$\mathbb{S} = e^{m+\sigma^2} \times \sqrt{1-e^{-\sigma^2}} = \frac{Med^2}{M_o} \times \sqrt{1-\frac{M_o}{Med}}$$

Le seuil s_o correspondant à un risque de première espèce α = 1%

$\left(\Pr\{X > s_o\} = \alpha\right)$ quand X suit la loi Log normale,

est alors donné par :

$$s_o = e^{m+2,3\sigma} = Med \times e^{2,3\sigma}$$

L'estimation de la médiane ne pose aucune difficulté : l'estimateur empirique converge presque sûrement vers la médiane (Fourgeaud et Fuchs, 1967).

Pour l'estimation du mode, il faut éviter l'usage d'un histogramme des fréquences avec classes pré-déterminées qui introduit une trop grande incertitude, on peut alors procéder comme suit :

Soit F(t) la fonction de répartition empirique des n valeurs observées (x_i)

$$F(t) = \frac{1}{n} \times \text{Card}\left\{x_i / x_i \leqslant t\right\}$$

On note alors : d(t,h) = F(t) - F(t-h)
où h est un paramètre dont la valeur initiale h_o est prise égale au plus petit écart non nul entre deux valeurs successives de l'échantillon.

Plus précisément, si (y_i) représente la suite ordonnée, croissante du sous-ensemble des valeurs distinctes des x_i d'origine,

$$h_o = \min_{i>1}(y_i - y_{i-1}) = y_{i_o} - y_{i_o-1}$$

On recherche alors les valeurs $\hat{y}_1$ de (y_i) tel que $\hat{y}_1 = \max_{y_i} d(y_i, h_o)$

Si il existe plusieurs valeurs $\hat{y}_1$ répondant à la condition, alors on sélectionne parmi celles-ci en augmentant la valeur de h :

$$h_1 = \min_{\substack{i>1 \\ i \neq i_o}} (y_i - y_{i-1})$$

Et ainsi de suite jusqu'à l'obtention d'un seul $\hat{y}_n$.

L'estimation du mode est donc la borne à droite de l'intervalle le plus dense, cette borne à droite étant choisie pour tenir compte de la dissymétrie de la loi génératrice.

Application

Si on s'intéresse dans l'exemple ci-contre à la série des quantités, les méthodes précédentes détectent "l'éléphant" 1635. La moyenne lissée est alors de 58,4, l'écart-type de 64,4.

Par cette méthode, l'estimation de la médiane donne : Med = 40,5, et celle du mode : M_o = 20.

On détermine alors m = 3,70, σ^2 = 0,71

D'où une moyenne $M = Med \times \sqrt{\frac{Med}{M_o}} = 57,6$

Et un écart-type S = 58,4.

Le seuil au delà duquel on considère les valeurs comme anormales étant s_o = 281.

Numéro semaine	Nombre de mouvements	Quantité en 100 m
1	13	20
2	22	229
3	11	57
4	17	34
5	17	123
6	13	35
7	3	5
8	12	49
9	9	21
10	7	20
11	8	20
12	9	33
13	8	17
14	12	60
15	31	269
16	5	41
17	12	18
18	11	62
19	16	66
20	18	1 635*
21	24	114
22	9	40
23	9	44
24	6	13
25	6	57
26	6	13

Par cette méthode, on obtient directement la moyenne corrigée, la prise en compte dans la moyenne de la valeur anormale (ouverture d'un nouveau magasin, en l'occurence) aurait doublé l'estimation :

$$3.095/26 = 119.$$

Interprétation

La détection de cette valeur exceptionnelle signifie que l'événement correspondant est :

. soit un "accident" extérieur qu'il faut analyser (c'est le cas ici),

. soit une manifestation rarissime du phénomène étudié qu'il convient de ne pas prendre en compte pour l'établissement des règles de gestion courantes.

I.3. Une méthode robuste, adaptative et non paramètrique

Ici encore, pour présenter la méthode, nous nous plaçons dans le cas du dépistage des valeurs exceptionnellement grandes par rapport aux valeurs courantes.

a - hypothèses

. les lois de probabilité des observations (X_t) sont quelconques, X_t non stationnaire,

. un phénomène accidentel (Z_t) peut avoir lieu à toute époque d'observation et se réalise suivant la loi :

$$Z_t = \begin{cases} z_t & \text{avec la probabilité } p, \\ 0 & \text{avec la probabilité } 1-p. \end{cases}$$

z_t est grand devant la moyenne $\bar{X}$ (inconnue) de X_t.
p est faible, inférieure à 5%.
p peut donc être interprétée comme la fréquence moyenne d'apparition des "éléphants".

b - principe

La méthode consiste en la mise à jour permanente d'un seuil s au dessus duquel l'observation sera considérée comme exceptionnelle.

Le problème consiste donc à trouver le bon seuil à un instant donné :

- si le seuil est trop bas, trop de valeurs sont considérées comme exceptionnelles : il doit être relevé,
- si il est trop haut, aucune valeur n'est dépistée : il doit être abaissé.

Changement de seuil

A chaque époque d'observation, on dispose d'un seuil s(r), remis à jour à l'époque $r \leqslant t$.

A l'instant t, notons :

$$Xsup = \left\{ x_i \;/\; x_i > s(r),\ r \leqslant i \leqslant t \right\}$$
$$Xinf = \left\{ x_i \;/\; x_i < s(r),\ r \leqslant i \leqslant t \right\}$$

et :
$$\bar{y} = \min \left\{ x \;/\; x \in Xsup \right\}$$
$$\underline{y} = \max \left\{ x \;/\; x \in Xinf \right\}$$

A chaque époque t, on dispose donc de s, $\bar{y}$, $\underline{y}$.

Deux critères seront utilisés pour la remise à jour du seuil :

. longueur d'une chronique sans "éléphants" repérés,

. nombre d'éléphants repérés dans une chronique de longueur fixée.

Règle :

Le seuil, considéré comme trop élevé, sera abaissé au niveau $\underline{y}$,

. si un nombre N_a de mouvements consécutifs observés sans apparition d'éléphants est supérieur à n_a,

. ou si le nombre d'éléphants repérés dans une chronique de longueur fixe L est inférieur à un seuil s_a.

Le seuil est relevé au niveau $\bar{y}$,

. si le nombre N_e de mouvements observés entre deux "éléphants" est inférieur à n_e,

. si le nombre d'éléphants repérés dans une chronique de longueur L est supérieur à un seuil s_e.

Ce double critère permet une plus grande sensibilité pour la remise à jour du seuil nécessaire lors d'un changement de tendance.

c - détermination des paramètres

Notons e_i la variable aléatoire de Bernouilli, valant :

. 1 si le mouvement i est un mouvement éléphant,

. 0 sinon.

Pour une chronique de longueur n : x_t, x_{t+1}, ... x_{t+n-1}

la variable $E_n = e_t + e_{t+1}, \ldots e_{t+n-1}$ suit une loi binomiale de paramètres n, p.

La condition d'abaissement du seuil est :

$$\{N_a \geqslant n_a\} \quad \text{ou} \quad \{E_L \leqslant s_a\}$$

Si le seuil est bien réglé, on l'abaisse donc à tort avec la probabilité :

$$\beta = Pr\{(N_a \geqslant n_a) \quad \text{ou} \quad (E_L \leqslant s_a)\}$$
$$= 1 - Pr\{(N_a < n_a) \text{ et } (E_L > s_a)\}$$

où N_a suit une loi de Pascal $Pr\{N_a \geqslant k\} = (1-p)^k \qquad k > 0$

E_L suit une loi binomiale $B(L,p)$.

Ces deux variables n'étant pas indépendantes.

De même, la condition de relèvement du seuil est :

$$\{N_e \leqslant n_e\} \quad \text{ou} \quad \{E_L \geqslant s_e\}$$

On relève donc à tort un seuil bien réglé avec la probabilité :

$$\alpha = 1 - Pr\{(N_e > n_e) \text{ et } (E_L < s_e)\}$$

N_e suit la loi : $Pr\{N_e = k\} = (1-p)^k p$

De façon approchée, on supposera, pour régler les paramètres, que les deux événements qui interviennent sont équivalents, c'est à dire que la réalisation de l'un entraine la réalisation de l'autre..

Plus précisément, on doit résoudre :

$$Pr\{A \text{ et } B\} = 1 - \alpha \text{ (ou } 1 - \beta)$$

Sous l'hypothèse simplificatrice d'équivalence, la loi du couple est alors :

		B	
		vrai	faux
A	vrai	x	0
	faux	0	y

On est donc amené à résoudre :

$$\Pr\{A\} = 1 - \alpha \text{ (ou } 1-\beta)$$
$$\Pr\{B\} = 1 - \alpha \text{ (ou } 1-\beta)$$

En particulier pour $\alpha = \beta = 0{,}05$, p = 5%, on pourra prendre :

$$L = 50,\ s_a = 0,\ s_e = 6,\ n_a = 58,\ n_e = 1.$$

d - application

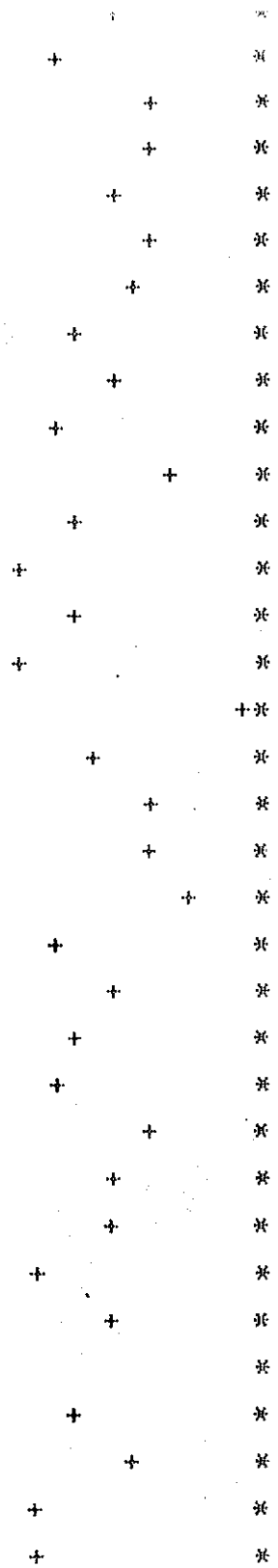
On montre dans les figures 1 et 2 ci-après le comportement du seuil devant une chronique générée artificiellement.

Chronique générée : (+) suivant un loi de Poisson de moyenne m = 10.
Eléphants (E) de valeur 5m environ, avec une fréquence de 5%.

La figure n°1 correspond à une simulation effectuée avec une valeur initiale du seuil égale à 30, la trace du seuil (*) reste un moment à sa valeur initiale (apparition de valeurs anormales), puis au bout d'un certain temps, est réajustée à la baisse. L'estimation de la moyenne est après filtrage de 9,58, alors qu'elle aurait été de 13,5 sans filtrage.

Figure 1

Légende : E : "éléphant",
* : valeur du seuil,
+ : valeur de la chronique.



E

Figure 1 (suite)

MOYENNE SERIE 13.50413

MOYENNE LISSEE 9.583333

La figure n°2 montre :

l'adaptation du seuil avec une initialisation à 0, le seuil monte régulièrement au début pour atteindre au bout de 10 valeurs un niveau de croisière à 15. Ici aussi, tous les éléphants ont été éliminés sans aucune hypothèse sur la loi de la chronique.

On présente en annexe 3 les éditions du programme de détection appliqué sur un cas réel. Ce programme a été introduit dans le cadre d'une gestion des stocks d'un grossiste, gérant 20.000 références. A chaque révision des paramètres de gestion, des sorties d'articles saisies sur les bons de livraison, sont filtrées.

Sur cette édition, figure la référence (2), le nombre de mouvements (3), la valeur du seuil en fin de traitement (4), la moyenne hors lissage (5) et après filtrage (6), le nombre d'éléphants" détectés (7), la valeur cumulée des "éléphants" (8).

Avantage de la méthode

Cette procédure s'applique à n'importe quelle chronique, quelle que soit sa (ou ses) loi(s), ne nécessite aucun historique au départ. En particulier, elle pourra être appliquée pour les produits nouveaux.

Elle est adaptative et permet de tenir compte d'évolution dans le temps de la chronique, évite le calcul a priori de moyenne et d'écart-type.

Elle est parfaitement adaptée aux sorties par lots (colisage minimum) alors qu'un seuil calculé analytiquement à partir d'une moyenne et d'un écart-type peut tronquer systématiquement un multiple de colisage.

Figure 2

Légende : E : "éléphant"
* : valeur du seuil,
+ : valeur de la chronique,
(+*) : valeurs identiques (décalées pour le graphisme)

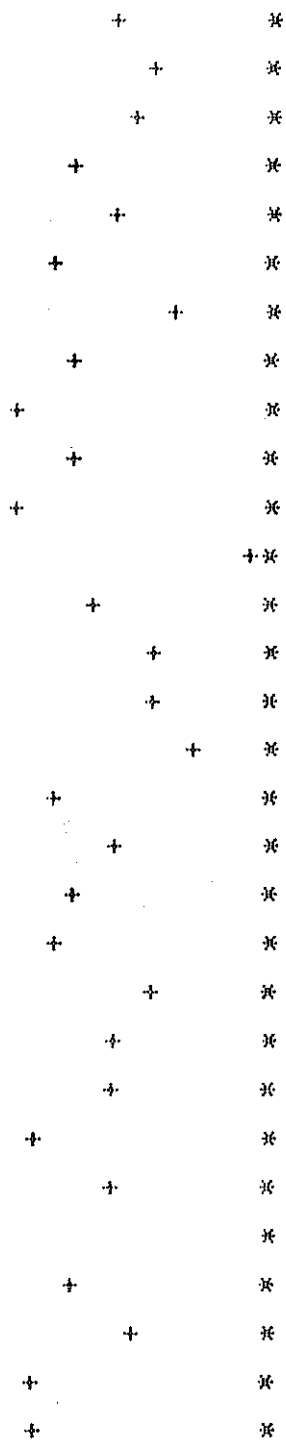


Figure 2 (suite)

E

MOYENNE SERIE 13.50413

MOYENNE LISSEE 9.428572

Sur le plan informatique, elle se prête au temps réel puisque, dès les premiers mouvements et en cas de changement de tendance, on peut l'utiliser sous forme de signal dont le gestionnaire tiendra compte ou non, la réponse du gestionnaire, traitée, permettant de positionner le seuil plus rapidement.

II - DEUXIEME PARTIE

Repérage de séries saisonnières sur la base d'un historique annuel.

Dans le traitement fréquent d'un très grand nombre de séries (certaines sociétés traitent périodiquement plusieurs milliers, voire plusieurs dizaines de milliers de séries), le lissage exponentiel est une méthode économique en temps de calcul, mais mal adapté aux séries saisonnières.

Il s'avère donc indispensable de pouvoir diagnostiquer une éventuelle saisonnalité, avant la mise en place d'un système de prévision.

Nous proposons, par la suite, un test simple, permettant d'accepter ou de rejeter l'hypothèse de stabilité de la chronique lorsqu'on ne dispose que d'un court historique (12 mois, par exemple).

Cette méthode est voisine de celle proposée par J.C. Laloire (72) pour repérer une tendance éventuelle. Cet auteur propose des tests non paramètriques, fondés sur les travaux de Olmstead (47) et de Sved et Eisenhart (43), mais fournit des résultats assymptotiques lorsque le nombre d'observations dépasse 30.

II.1. Principe

Soit $\{x_t\}$, t = 1,2,...12, la série chronologique disponible sur 12 mois. On désire tester l'hypothèse H : $x_t = m + E_t$, où E_t est une variable aléatoire continue ; E_t indépendantes entre elles.

Notons que, dans le cas d'existence d'une tendance, on pourra remplacer m par m(t), après avoir calculé la tendance sur les 12 observations disponibles.

Pour plus de clarté, conservons le cas où m_t est constante.

Considérons la suite y_t, où $y_t = 0$ si $x_t < m$

$y_t = 1$ si $x_t > m$

L'événement $x_t = m$ ayant une probabilité nulle, nous le négligeons.

On s'intéresse alors à la suite des y_t où, sous l'hypothèse H, les y_t suivent une loi de Bernouilli, de paramètre $P = Pr\{x_t > m\}$

Dans le cas de l'existence d'une saisonnalité, les y_t sont fortement corrélées, et cette corrélation doit apparaitre dans la longueur des sous-suites de mêmes valeurs (appelées séquences).

Par exemple : une série saisonnière dont les fortes ventes se situent au dernier trimestre, conduira à une suite y_t du type :

000000000111 2 séquences

alors qu'une série purement aléatoire présentera certainement un profil différent :

011000101110 7 séquences.

Le principe de base repose sur la réduction du nombre de séquences dans le cas de l'existence d'une saisonnalité, car, dans la pratique, les périodes de forte vente sont limitées à une ou deux.

Remarquons d'autre part que, par construction, on obtient un minimum de 2 séquences puisque m sera estimé par :

$$\hat{m} = \frac{\sum_{t=1}^{12} x_t}{12} \qquad \text{donc } \min\{x_t\} < \hat{m} < \max\{x_t\}$$

On est donc ramené à un test unilatéral sur le nombre de séquences : si s représente le nombre de séquences observées dans un historique,

$s \leqslant s_0$, conduira au rejet du modèle stationnaire, donc à la présomption d'une saisonnalité.

$s > s_0$, conduira la l'acceptation de la saisonnalité.

s_0 étant le seuil à calculer compte-tenu du risque de première espèce : $\Pr \{ s \leqslant s_0 \ / \ H \text{ vraie} \} = \alpha$

II.2. Détermination du seuil de décision

Soit $\{y_t\}$ une suite associée à une série.

Nous noterons n_1 le nombre de variables y_t nulles,
n_2 le nombre de variables y_t égales à 1.

n_1 et n_2 sont des variables aléatoires, $n_1 + n_2 = 12$.
S est le nombre de séquences de la série.

Pour k pair, on aura :

$$(1) \qquad \Pr \{s=k\} = \sum_{i=\frac{k}{2}}^{12-\frac{k}{2}} \Pr \{n_1 = i\} \times \Pr \left\{ \frac{s=k}{n_1=i} \right\}$$

Pour k impair, on aura :

$$(2) \qquad \Pr \{s=k\} = \sum_{i=\frac{k-1}{2}}^{12-\frac{k-1}{2}} \Pr \{n_1 = i\} \times \Pr \left\{ \frac{s=k}{n_1=i} \right\}$$

où $\Pr \left\{ \frac{s=k}{n_1=i} \right\}$ est la probabilité d'obtenir k séquences avec une suite de n_1 `0' et $n_2$ `1'.

Lorsque n_1 (et donc n_2) est fixé toute suite $\{y_t\}$ est équiprobable, sous l'hypothèse H, de probabilité $(1-p)^{n_1} \times p^{n_2}$, si on note respectivement ch(k,i), ch(i), le nombre de suites présentant k séquences sachant que $n_1 = i$ et le nombre de suites pouvant être construites avec i '0', alors :

$$\Pr\left\{\frac{s=k}{n_1=i}\right\} = \frac{ch(k,i)}{ch(i)} .$$

En suivant la méthode proposée dans (1), on peut montrer que, pour k pair :

$$ch(k,i) = 2 \binom{i-1}{\frac{k}{2}-1} \binom{12-i-1}{\frac{k}{2}-1}$$

pour k impair :

$$ch(k,i) = \binom{i-1}{\frac{k-1}{2}} \binom{12-i-1}{\frac{k-3}{2}} + \binom{i-1}{\frac{k-3}{2}} \binom{12-i-1}{\frac{k-1}{2}}$$

où : $\binom{x}{y} = 0$ si $y > x$.

On donne en annexe n°1 les valeurs de ch(k,i) pour i = 1,2,..., 6 sachant que ch(k,i) = ch(k, 12-i) pour $i > 6$.

La variable n_1 suit une loi binomiale conditionnelle, car n_1 est toujours supérieure à 1, d'où :

$$\Pr\{n_1=i\} = q^i \, p^{12-i} \times \binom{12}{i} / 1-q^{12}-p^{12} \text{ avec } q = 1-p$$

Comme q est généralement de l'ordre de 0,5, $1-q^{12}-p^{12} = 0,9995$, on peut négliger l'influence de la condition.

---

(1) : Statistiques et Informatique Appliquée. Lebart et Fénelon (71).

MOIS 1 104
MOIS 2 104
MOIS 3 162
MOIS 4 190
MOIS 5 189
MOIS 6 181
MOIS 7 190
MOIS 8 135
MOIS 9 118
MOIS 10 120
MOIS 11 108
MOIS 12 125

LA MOYENNE EST: 145.8333
NOMBRE DE SEQUENCES 3

Série n°1

MOIS 1 945
MOIS 2 895
MOIS 3 845
MOIS 4 845
MOIS 5 915
MOIS 6 1015
MOIS 7 960
MOIS 8 875
MOIS 9 895
MOIS 10 995
MOIS 11 1120
MOIS 12 1300

LA MOYENNE EST: 967.0833
NOMBRE DE SEQUENCES 4

Série n°2

MOIS 1 901
MOIS 2 815
MOIS 3 900
MOIS 4 646
MOIS 5 883
MOIS 6 811
MOIS 7 766
MOIS 8 584
MOIS 9 761
MOIS 10 863
MOIS 11 773
MOIS 12 751

LA MOYENNE EST: 804.5
NOMBRE DE SEQUENCES 4

Série n°3

Analyse de saisonnalité des séries n°1, 2 et 3.

On peut alors calculer :

$$\Pr\{s \leq s_0\} = \sum_{k=1}^{s_0} \sum_{i(k)} q^i \, p^{12-i} \, \binom{12}{i} \times \frac{ch(k,i)}{ch(i)}$$

La variation de i dépend de k, comme nous l'avons vu dans les formules (1) et (2).

Pratiquement, si le résidu suit une loi presque symétrique $q \simeq p \simeq 0{,}5$, on obtient alors, pour $s_0 = 4$:

$$\Pr\{s \leq 4\} = 11{,}3\%.$$

D'où la règle :

Avec un risque de se tromper de l'ordre de 10%, on peut, pour dépister une éventuelle saisonnalité, vérifier si le nombre de séquences est inférieur ou égal à 4.

Ce test a l'avantage de ne pas dépendre de la variance de la série et d'éviter tout calcul d'intervalle de confiance autour de la moyenne, ce qui est pratiqué habituellement.

II.3. Application

Les trois premières séries sont présentées par les auteurs comme des exemples de séries saisonnières, l'analyse ci-contre donne les résultats.

Série n°1	saisonnière s = 3	(Bourbonnais p. 116 - 12 derniers mois).
Série n°2	saisonnière s = 4	(Calot p. 382 - chronique de l'année 1956)
Série n°3	saisonnière s = 4	(Giard p. 39 - chronique de l'année 1975)

Les séries n° 4 et 5 suivantes ont été générées par un tirage aléatoire.

Série n° 4 :

Génération de 12 valeurs de la variable $x_t = 20 + 10v$ où v suit une loi normale centrée réduite.

x_t	5,2	25,8	5,2	24,6	3,4	27,1	9,2	4,6	22,2	33,2	1,0	6,0
y_t	0	1	0	1	0	1	0	0	1	1	0	0

$\bar{x} = 13,96$ et $S = 9$

On rejettera l'hypothèse de saisonnalité.

Série n°5 :

Génération de 12 réalisations d'une variable binomiale de paramètres $n = 10$, $p = 0,1$.

x_t	1	0	2	1	2	0	0	1	1	1	0	0
y_t	1	0	1	1	1	0	0	1	1	1	0	0

$\bar{x} = 0,75$ et $S = 6$

Ici, l'hypothèse de symétrie n'est pas vérifiée et tend à atténuer la puissance du test. On rejette néanmoins l'hypothèse de saisonnalité.

II.4. Généralisation au cas de familles d'articles

Dans le cas, fréquent en pratique, où l'on dispose de plusieurs références, d'une même famille de produits qui, par hypothèse,

doivent avoir le même comportement, il y a avantage à étudier les historiques de cette famille simultanément.

On pourra en particulier tester sur la variable :

$$M = \max_{i=1,m} \{S_i\}$$

où S_i représente le nombre de séquences données de la référence i de la famille de m éléments.

Si $F_i(s)$ représente la fonction de répartition de S_i, la loi de M est alors :

$$\Pr\{M \leq S\} = \prod_{i=1}^{m} F_i(s)$$

sous l'hypothèse H généralisée à toutes les références.

Sous l'hypothèse d'un même type de loi de la demande des références, en particulier si les lois sont toutes symétriques, les $F_i(s)$ sont identiques et on a :

$$\Pr\{M \leq S\} = F^n(s)$$

Donc, pour deux références, on pourra refuser l'hypothèse H avec un risque de première espèce de 7,5% si M est inférieure ou égale à 5.

Exemple : pour les deux références (cf annexe 2) d'une même famille de matériel électronique, le maximum de séquences est celui de la référence 12 : 5 séquences.

On conclut donc à l'existence d'une saisonnalité, ce qui est confirmé par l'allure générale des courbes qui présentent des maxima et minima aux mêmes périodes.

Conclusion

Le type de méthodes que nous venons de présenter est en général simple à mettre en oeuvre bien que difficile à analyser profondément sur le plan théorique.

Nous voyons dans la simplicité et la robustesse de ces méthodes des qualités primordiales quant à leur utilisation dans les systèmes de gestion informatisés à un double titre :

. dans les gros systèmes centralisés traitant d'énormes volumes d'information où le temps réservé au calcul scientifique ne doit pas prédominer sur les autres fonctions,

. dans les mini ou micro systèmes où l'on ne dispose généralement que de moyens restreints de traitement de fichiers (historiques nécessairement limités).

Dans la majorité des entreprises, l'informatisation récente de la gestion ne permet pas d'avoir accès à des historiques homogènes, exploitables rapidement par moyens informatiques ; ce qui freine l'emploi de méthodes économétriques ou statistiques classiques où la plupart des hypothèses sous-jacentes ne peuvent être confirmées.

L'approche non paramétrique semble donc mieux adaptée au contexte actuel ; les exemples présentés ici ont d'ailleurs été conçus pour résoudre des problèmes concrets rencontrés lors de la mise en place de systèmes de gestion informatisés.
Sur le plan théorique, ces méthodes conduisent à des analyses intéressantes dans des domaines variés tels que :

. marchés aléatoires,
. études des propriétés des estimateurs de mode et médiane,
. calculs combinatoires et de probabilités.

Puisque, pratiquement et par simulation, ces méthodes donnent des résultats satisfaisants, il serait intéressant d'approfondir leurs propriétés théoriques afin de mieux les exploiter.

Annexes 1, 2 et 3

Annexe 1

Calcul du nombre de chemins

(Pour des raisons de symétrie nous nous restreignons à $n_1 \leq 6$)

n_1	s	$ch(s;n_1)$	$ch(s,n_1)/ch(n_1)$ (%)
1	2	2	16,6
	3	10	83,4
2	2	2	3,9
	3	10	19,6
	4	18	35,3
	5	21	41,2
		51	100,0
3	2	2	0,9
	3	10	4,5
	4	32	14,5
	5	64	29,1
	6	56	25,5
	7	56	25,5
		220	100,0
4	2	2	0,4
	3	10	2,0
	4	42	8,5
	5	84	17,0
	6	126	25,5
	7	126	25,5
	8	70	14,1
	9	35	7,1
		495	100,0
5	2	2	0,2
	3	10	1,1
	4	48	5,4
	5	120	13,4
	6	180	20,1
	7	260	29,0
	8	160	17,9
	9	80	8,9
	10	30	3,3
	11	6	0,7
		896	100,0
6	2	2	0,2
	3	10	1,1
	4	50	5,4
	5	100	10,8
	6	200	21,6
	7	200	21,6
	8	200	21,6
	9	100	10,8
	10	50	5,4
	11	10	1,1
	12	2	0,2
		924	100,0

Annexe 2 : ventes mensuelles (en unités) de deux articles d'une même famille.

(EC : en commande).

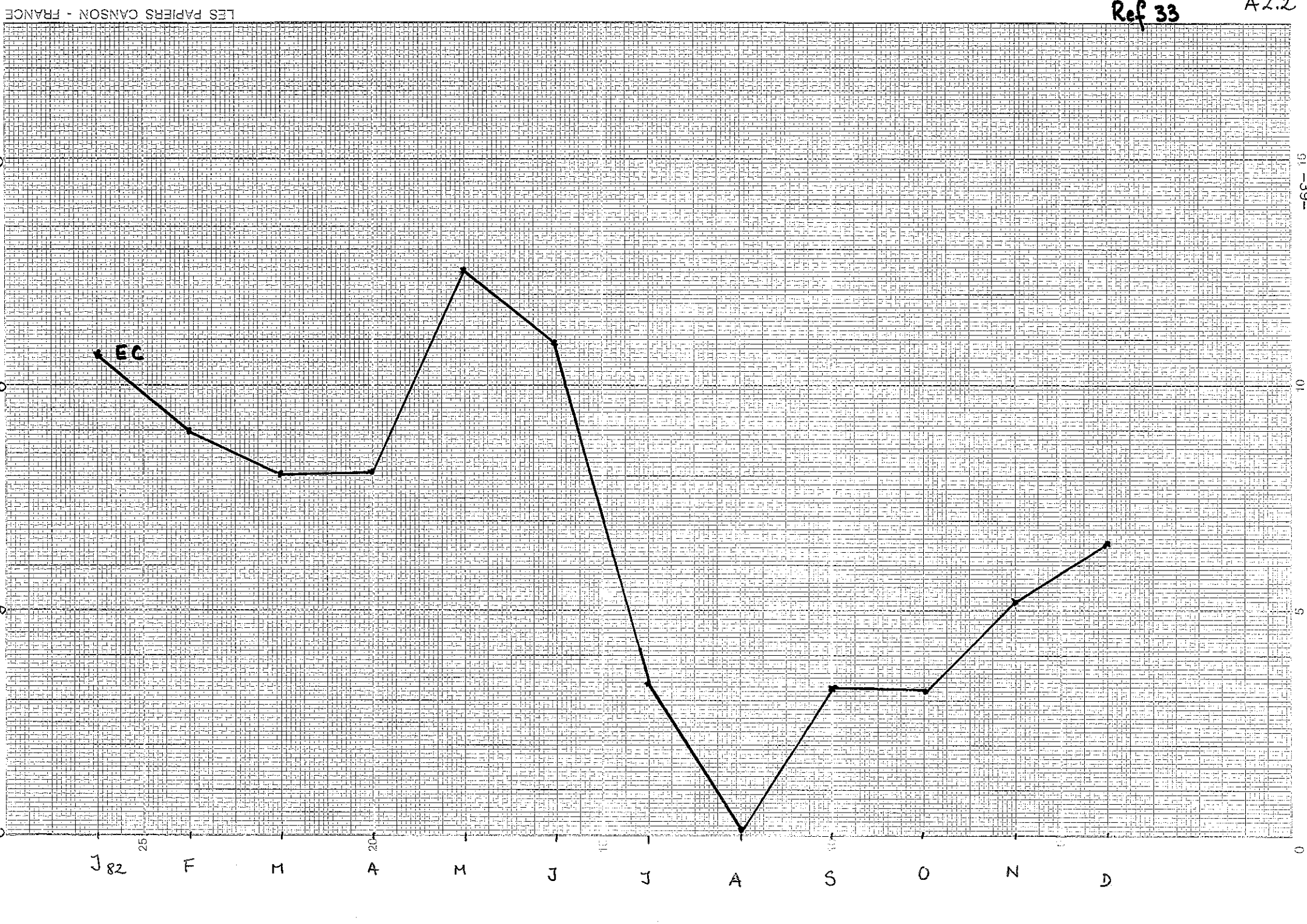
Ref 12

m: 226

J F M A M J J A S O N D

J82

LES PAPIERS CANSON - FRANCE


LES PAPIERS CANSON - FRANCE
EC
25
20
15
10
5
0
J 82
F
M
A
M
J
J
A
S
O
N
D

ARTICLE	NB DE MOUVEMENT	SEUIL	MOYENNE INITIALE	MOYENNE LISSEE	NB D"ELEPHANTS	CUMUL ELEPHANTS	PAGE 1
12.2501.0	272	3000,0	394,480	359,250	2	10.300,000	
12.2501.1	402	3000,0	450,740	397,480	4	23.000,000	
12.2501.2	181	3000,0	425,960	395,000	1	6.000,000	
12.2501.3	216	3000,0	438,880	391,580	2	11.000,000	
12.2501.5	237	2000,0	359,490	339,830	1	5.000,000	
12.2501.6	190	3000,0	412,630	388,350	1	5.000,000	
12.2501.7	474	5000,0	520,460	460,080	3	30.000,000	
12.2501.8	119	2488,8	384,870	345,760	1	5.000,000	
12.2505.0	422	2000,0	502,550	486,650	1	7.200,000	
12.2505.1	454	5000,0	510,130	457,070	2	25.000,000	
12.2505.5	82	3000,0	367,070	339,500	1	2.600,000	
12.2505.7	494	8000,0	559,870	505,240	2	28.000,000	
12.2520.0	135	1200,0	190,370	162,120	3	4.300,000	
12.3252.1	182	800,0	90,100	85,080	1	1.000,000	
12.3443.2	279	1000,0	108,240	105,030	1	1.000,000	
12.3444.2	293	600,0	94,190	91,090	1	1.000,000	
12.3830.5	54	2000,0	290,740	258,490	1	2.000,000	
13.6021.1	79	1000,0	145,560	134,610	1	1.000,000	
13.6041.0	71	1000,0	172,220	160,400	1	1.000,000	
13.6041.1	224	977,7	158,980	151,400	2	2.000,000	
201.0003.9	29	357,0	64,860	37,550	2	867,000	
204.0032.1	315	1875,0	148,470	144,170	1	1.500,000	
204.0033.1	393	1020,0	102,340	95,690	2	2.805,000	
204.0034.1	113	1237,3	113,760	103,350	1	1.275,000	
204.2033.1	329	2000,0	311,550	277,600	3	12.000,000	
204.2034.1	188	1777,7	236,960	208,370	3	6.000,000	
204.3552.1	67	1155,5	217,910	190,900	1	2.000,000	
204.3553.1	106	7000,0	510,370	467,610	1	5.000,000	
204.3555.1	38	1000,0	194,730	172,970	1	1.000,000	
204.4002.1	58	2036,4	423,980	378,780	1	3.000,000	
206.0418.9	221	110,0	13,620	12,370	2	300,000	
206.3003.5	28	435,5	195,670	73,290	1	3.500,000	
206.6396.9	112	2600,0	343,920	301,980	1	5.000,000	
206.7020.1	44	3911,1	975,000	765,110	1	10.000,000	
206.7103.1	402	20000,0	1.192,030	692,920	6	204.800,000	
206.7104.1	269	10000,0	1.114,310	896,780	5	63.000,000	
206.7105.1	155	9000,0	938,700	726,970	3	35.000,000	
206.8550.1	159	5333,3	444,650	403,160	1	7.000,000	
206.8570.1	179	4088,8	352,170	326,060	1	5.000,000	
206.8632.1	204	2500,0	369,110	348,010	2	5.000,000	
206.8642.1	39	2000,0	296,150	225,000	1	3.000,000	
206.9356.9	69	4000,0	626,080	561,760	1	5.000,000	
206.9357.9	75	3555,5	622,660	577,020	1	4.000,000	
207.5422.5	94	26,6	5,450	4,480	2	100,000	
207.8400.9	125	200,0	16,400	15,330	1	150,000	
208.0601.9	34	700,0	136,000	72,620	2	2.300,000	
208.0638.9	33	311,1	56,420	47,680	1	336,000	
208.0696.9	267	100,0	8,100	7,750	1	100,000	
210.1127.9	38	59,5	16,650	14,400	1	100,000	
211.0424.9	154	39,1	4,550	4,180	1	60,000	
211.1002.9	64	40,0	7,670	6,200	1	100,000	
216.1802.9	111	184,0	22,630	18,300	1	500,000	
216.1803.9	227	200,0	21,320	20,090	1	300,000	
216.1804.9	283	100,0	21,270	19,580	1	500,000	
219.2450.2	31	9,7	2,450	1,900	1	19,000	
231.2704.0	77	101,3	19,460	17,980	1	132,000	

2 3 4 5 6 7 8

Bibliographie

L. LEBART ET JP FENELON :
Statistique et Informatique Appliquées, Dunod, Paris, 1971.

JC USUNIER ET R. BOURBONNAIS :
Pratique de la Prévision à Court Terme, Conception de Systèmes de Prévision, Dunod, Paris, 1982.

V. GIARD :
Gestion de la Production, Calcul Economique, Economica, Paris, 1981.

P.F. VELLEMAN :
Définition and Comparaison of Robust non Linear Data Smoothing Algorithms, J.A.S.A., 1980, Volume 75, p. 609-615.

G. CALLOT :
Cours de Statistique Descriptive, Dunod, Paris, 1964.

G. CARLETTI :
Detection Automatique de Valeurs Anormales, Revue de Statistique Appliquée, 1976 , Vol XXIV, n°3, p. 61-69.

J. OLMSTEAD :
Distribution of Sample Arrangement for Runs Up and Down. Tables Annals of Mathematical Statistics, 1947 , vol 17, p. 24-33.

J.C. LALOIRE :
Méthodes de Traitement des Chroniques Statistiques et Prévision de Ventes, Dunod, Paris, 1972.

C. FOURGEAUD ET A. FUCHS :
Statistique, Dunod, Paris, 1967.

F.S. SVED ET C. EISENHART :
Tables for Testing Randonness of Grouping in a sequence of Alternatives. Annals of Mathematical Statistics, 1943, vol 14, p. 66-87.

S.C. WHEELWRIGHT AND S. MAKRIDAKIS :
Forecasting Methods for Management, 3ème édition, John Wiley and sons, New York, 1980.