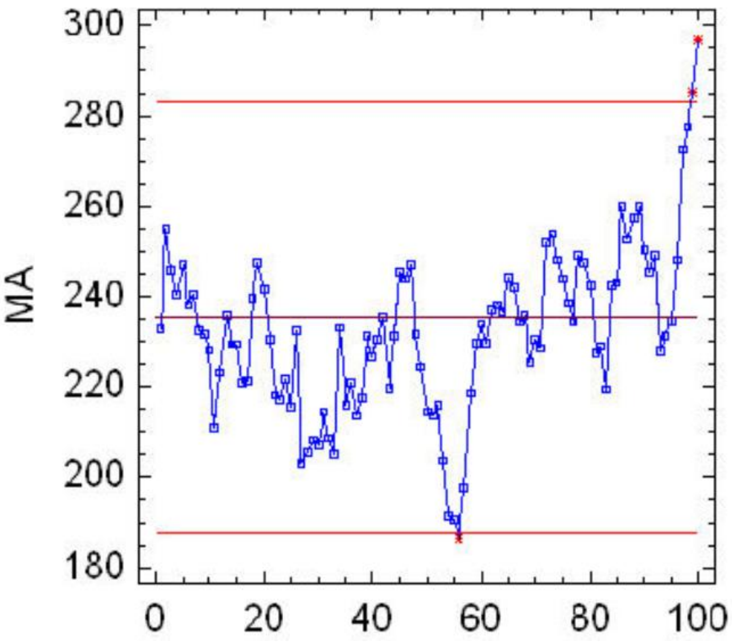
SÉRIE TEMPORELLE
PRINCIPE

PROCESSUS ET SÉRIE TEMPORELLE

Principe : on a un phénomène qui est générateur d'un processus (succession d'états dans le temps)

Processus :

- On considère un ensemble d'états ordonnés chronologiquement
- Chacun de ces états est décrit par une variable aléatoire Xt
- L'ensemble des Xt forme une suite qui peut être représentée graphiquement

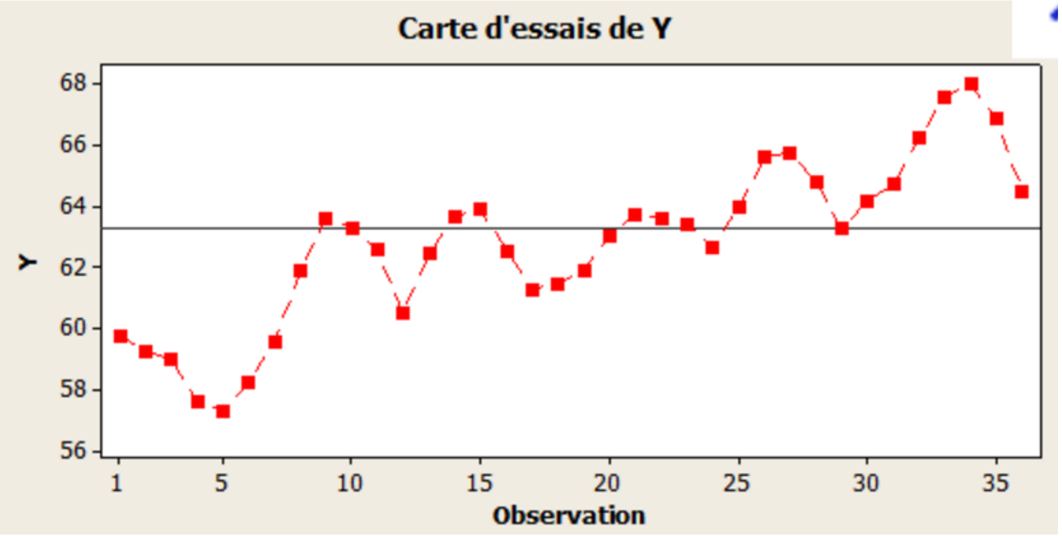


DESCRIPTION D'UN PROCESSUS

Représentation : graphique comme en géométrie

Présence de corrélations (la tendance linéaire) et de fluctuations statistiques,

- La tendance modélisée par une régression linéaire
- La partie stochastique par une loi de distribution



Rappel, la loi de probabilité caractérisée par des paramètres appelés moments (moyenne, variance, …)

Ce sont eux qui vont nous permettre de bien décrire les évolution de notre processus

RAPPELS SUR LES LES MOMENTS

Les deux premiers moments sont importants :

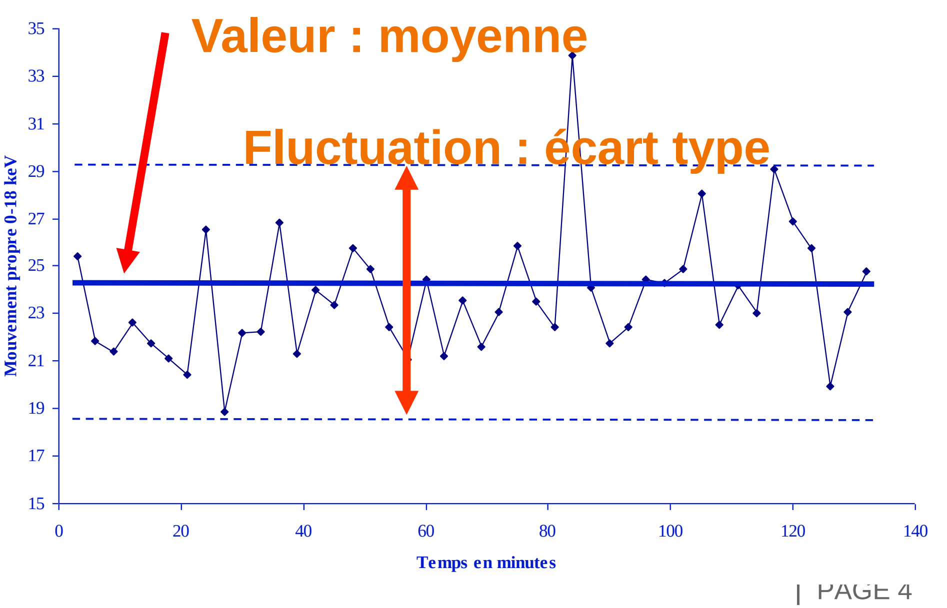
1. Le moment 1 qui est **l'espérance ou la moyenne µ**
   Il donne la tendance générale du processus, son ordre de grandeur

2. Le moment d'ordre 2 qui est la **variance**, elle-même reliée à **l'écart type**
   $Var(X_t) = \sigma^2(X_t)$
   Il donne la « fluctuation » ou dispersion autour de la moyenne

Rappels sur les moments :

$$\mu k = E[(X_t - E(X_t))^k]$$

Ils permettent de décrire de plus en plus finement le processus, un peu comme les dérivées d'ordre k décrivent de mieux en mieux une fonction



POUR RAPPEL (LES 2 PREMIERS MOMENTS)

Tendance des résultats : la moyenne arithmétique

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Fluctuation ou dispersion des résultats : l'écart type

$$\sigma_x = \sqrt{\sum_{i==1}^{n} \frac{\left(x_i - \bar{x}\right)^2}{n-1}}$$

$1^{\text{ÈRE}}$ ÉTAPE VERS LA MODÉLISATION (ANOVA)

Ces deux moments ne suffisent pas

On voudra séparer la partie linéaire (dépendance entre les variables) de la partie purement aléatoire.

Il faudra donc **analyser les parties linéaires pour** ne garder (par soustraction de la partie linéaire) que la partie aléatoire qui est utilisée pour les cartes de contrôle.

Outils statistiques pour analyser ces dépendances sont la **covariance et les corrélations.**

Finalement on utilise l''équation fondamentale de l'analyse de la variance ou **ANOVA** :

$$\mathrm{Var}_{\text{total}}(X_t) = \mathrm{Var}_{\text{expliquée}}\ (\text{du modèle}) + \mathrm{Var}_{\text{résiduel}}(\text{du bruit})$$

Pour un groupe de point
Celle d'un ensemble de mesures d'un seul échantillon au temps t

Ce qui reste quand on a enlevé ce qui a été modélisé

QUESTION DE VOCABULAIRE

La partie expliquée à l'aide de régression linéaire constituera le modèle

Ce modèle peut être utilisé différemment suivant le temps :

- Pour une application aux données passées on parlera de lissage ou d'ajustage
- Pour une application aux données du présent on parlera de filtration
- Pour une application aux données futures on parlera de prédiction

RAPPELS SUR L'EXPRESSION DE LA COVARIANCE

Covariance :

Cov (X, Y) = E[(X-μ_X)*(Y-μ_Y)]

Autre notation :

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Interprétation, chacune des deux variables aléatoires est soustrait de sa moyenne (donc centrée)

Puis on estime la moyenne de l'influence de l'une sur l'autre (analyse de leur influence réciproque)

Variables aléatoires **indépendantes**

Cov(X,Y)=0 (équivalent à l'orthogonalité)

Var(X+Y) = Var(X)+Var(Y)

Pour des variables **corrélées** (dépendantes)

Cov(X,Y) mesure l'écart de la variance par rapport à ce qu'elle serait si les variables avaient été indépendantes :

Var(X+Y) = Var(X) + Var(Y) + Cov(X,Y)

Il en résulte qu'il existe une structure dans les suites de valeurs obtenues,

PARAMÈTRE DÉRIVÉ : LA CORRÉLATION

La corrélation est la covariance normalisée (on s'affranchit de facteurs d'échelle)

Normalisation par les variances

Covariance

$$r_p = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \longrightarrow r_p = \frac{\sum_{i=1}^{N}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

Variances$^{1/2}$

La présence de corrélation dans un processus $\rightarrow$ des « forces d'inertie » contrôle le processus,
Il est donc judicieux de les soustraire dès que l'on connait un modèle de relation (du type des régressions linéaires) et d'analyser ce qui reste.

Dans le cas contraire, on prends le risque d'avoir beaucoup de fausses alarmes

PARAMÈTRE DÉRIVÉ : : L'AUTOCORRELATION

Il s'agit d'étudier les relations ou la moyenne des influences réciproques entre les valeurs d'un même paramètre mais décalé dans le temps, Cela revient à rechercher des structures internes dans le processus,

Variable X : X_t
Variable Y : $X_{t \pm h}$

Notation : $\gamma_h = Cov(X_t, X_{t \pm h})$

Graphiquement :

- $Cov(X_t, X_{t \pm h}) > 0$, les écarts par rapport à la tendance sont systématiquement situés du même coté
- $Cov(X_t, X_{t \pm h}) < 0$, ces écarts sont répartis de part et d'autre

VARIOGRAME ASSOCIÉ

On a introduit la covariance :

$$\gamma_h = Cov(X_t, X_{t \pm h})$$

On peut construire le coefficient de corrélation (qui varie entre -1 et 1) :

$$\rho_h = \frac{Cov(X_t, X_{t \pm h})}{Var(X_t)}$$

Le variograme est obtenu en soustrayant la « fluctuation naturelle ou intrinsèque » à la covariance. Puis à analyser l'évolution avec le temps, c'est-à-dire en faisant varier le paramètre h, La forme obtenue permet de déterminer la nature et la portée des interactions.

$$Variograme(h) = Var(X_t) - Cov(X_t, X_{t \pm h}) = \gamma(h)/\rho(h) - \gamma(h)$$

Ambiguïté : la lettre γ sert autant à représenter la covariance que le variograme !!!

NOTION DE STATIONNARITÉ

Il y a stationnarité d'un processus quand les conditions suivantes sont remplies :

1/ **La moyenne** reste **constante** dans le temps

2/ Le processus varie toujours de la même façon dans le temps, Autrement dit **sa variance** ou moment d'ordre 2 est **constante**. Ainsi, une fois retiré du processus la partie modélisable ou explicative, il ne reste qu'un bruit dont la moyenne et la variance sont constante ce qui permet d'envisager une description par une <u>loi de probabilité</u>

3/ Les relations de dépendances entre les variables restent constante. Autrement dit la **covariance et les corrélations ne varient pas,** elle reste toujours du même ordre de grandeur quelque soit le temps. Cela permet d'envisager une modélisation de cette dépendance entre les variables qui ne varient pas (voir le modèle ARIMA).
Ces variables pour un processus sont les valeurs observées pour une même mais à des temps différents

STATIONNARITÉ AU SENS FAIBLE ET TESTS STATISTIQUES

La stationnarité au sens faible décrit les processus qui respectent les deux premières conditions seulement (constance de l'espérance et de la variance) mais pas nécessairement de la troisième (l'influence des valeurs à des temps différents dépend elle-même du temps !!), Ces deux seules conditions permettent de se donner une loi de distribution qui ne varie pas trop

On définira la série temporelle comme un processus stochastique stationnaire au sens faible

L'hypothèse nulle est la stationnarité.

- Test KPSS1
- Test de Leybourne et McCabe2
- Tests de racine unitaire[modifier | modifier le code]

L'hypothèse nulle est la non-stationnarité.

- Test de Dickey Fuller (en)3
- Test augmenté de Dickey Fuller (en) (ADF)4
- Test de Phillips-Perron5(PP)
- Test DF-GLS (ou ERS) 6

REMARQUE SUR LA NON-STATIONARITÉ

La stationnarité permet de garantir une certaine stabilité à sa modélisation, un peu comme le voisinage d'un point en son plan tangent, On s'assure que les choses sont encadrées, restent à la même échelle et n'évoluent pas de façon chaotique.

Une série qui n'est pas stationnaire est dite intégrée

Ces séries possèdent une racine unitaire

Leurs particularité est d'avoir une loi de distribution qu évolue elle aussi avec le temps (alors qu'en stationnarité on s'assure que les moments ne varient pas trop pour pouvoir utilisée cette loi de distribution)
La nature de cette non-sationnarité peut être :

1. déterministe (la moyenne varie)
2. Stochastique (la variance évolue)

MODÉLISATION STOCHASTIQUE ET DÉTERMINISTE

Finalement :

- Le deux premières conditions permettent d'envisager une description stochastique par une loi de probabilité simple, une loi normale (white noise ou bruit blanc)

- La troisième condition permet d'envisager une modélisation déterministe

On retrouve le concept des « erreurs » de type A et de type B

BILAN

- Effectuer un échantillonnage de valeurs périodiquement dans le temps
- Identifier les corrélations qui pourraient exister, et trouver leurs modèle (régression linéaire)
- Soustraire des valeurs mesurées la valeur expliquée pour ne garder que la partie purement aléatoire
- Trouver une loi de distribution de cette partie purement aléatoire
- Trouver une représentation qui permettent de suivre les dérives et de réagir au plus vite dès qu'un écart est observés (aide à la décision)

SÉRIE TEMPORELLE
LA MÉTHODE ARIMA

THÉORÈME DE WAL

Cela rappel la formule de l'ANOVA
Rappel sur la décomposition d'une série temporelle (théorème de Wal) :

Signal total = Signal expliqué déterministe + Bruit aléatoire

On n'a vu que la partie expliqué du signal pouvait l'être par une régression linéaire
Elle peut l'être aussi par une modélisation périodique (utilisation de la Transformée de Fourrier), C'est en particulier le cas des phénomènes périodiques ou saisonniers

De façon générale on pourra écrire que :

$$X_t = \mu + \underbrace{\boxed{\sum_{i=1}^{p} \varphi_i . X_i}}_{\text{AR(p)}} + \underbrace{\boxed{\sum_{j=1}^{q} \theta_j . \varepsilon_j}}_{\text{MA(q)}}$$

AR(p)
Processus Autorégressif

MA(q)
Processus de la moyenne Mobile

DESCRIPTION DES PROCESSUS AUTORÉGRESSIFS

La valeur qui est relevée au temps t+1, soit X_{t+1}, peut s'exprimer linéairement en fonction des p dernières observations, c'est-à-dire de X_{t-p} à X_t

$$AR(p) = X_t = Cste + \varphi_1 . X_{t-1} + \varphi_2 . X_{t-2} + \varphi_3 . X_{t-3} + \ldots\ldots\ldots + \varphi_p . X_{t-p}$$

Seules les p dernières valeurs sont suffisantes pour modéliser la partie corrélée,

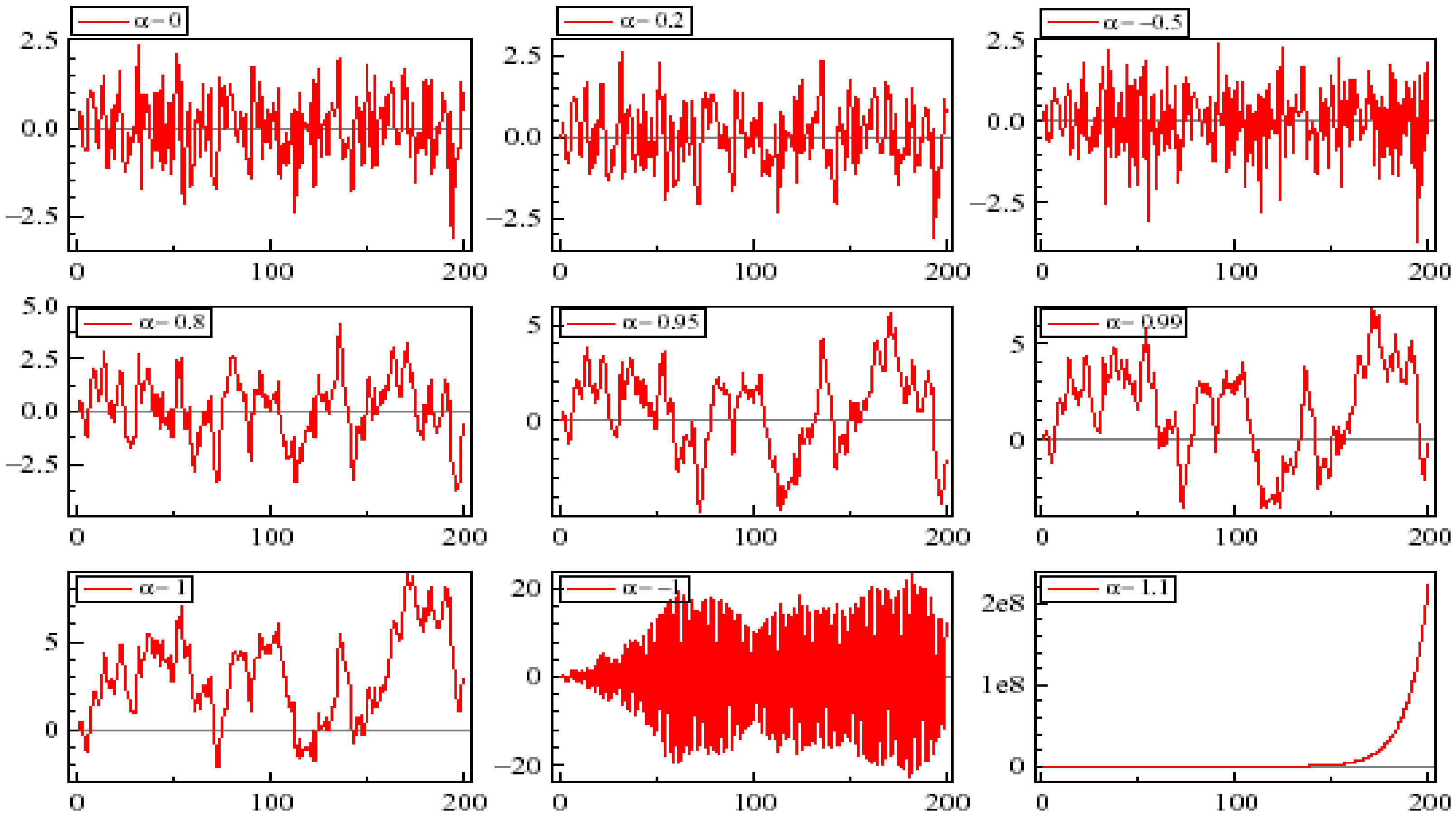
La détermination de l'ordre, p s'effectue à partir des Fonctions d'Autocorrélation (ACF) ou des Fonctions d'Autocorrélation Inverse (AICF) → voir par la suite

L'ittération se fait sur le « rail » du temps

Le processus régressif le plus utilisé est celui pour lequel la valeur mesurée peut être directement en fonction de la valeur précédente, c'est le processus AR(1).

Autrement dit le passé a peu d'influence (peu d'inertie)

PROCESSUS AR(1)



QUELQUES VALEURS

P=1

$$\bar{X}_t = \frac{Cste}{1 - \varphi}$$

$$Var(X_t) = \frac{Var(\varepsilon_t)}{1 - \varphi^2}$$

$$\gamma_j = Cov\left(X_t, X_{t-j}\right) = \varphi^j . Var(X_t)$$

P=2

$$\bar{X}_t = \frac{Cste}{1 - \varphi_1 - \varphi_2}$$

$$Var(X_t) = \frac{Var(\varepsilon_t)}{1 - \varphi_1^2 - \varphi_2^2}$$

$$\gamma_j = Cov\left(X_t, X_{t-j}\right) = \varphi^j . Var(X_t)$$

DESCRIPTION DES PROCESSUS DE MOYENNES MOBILES

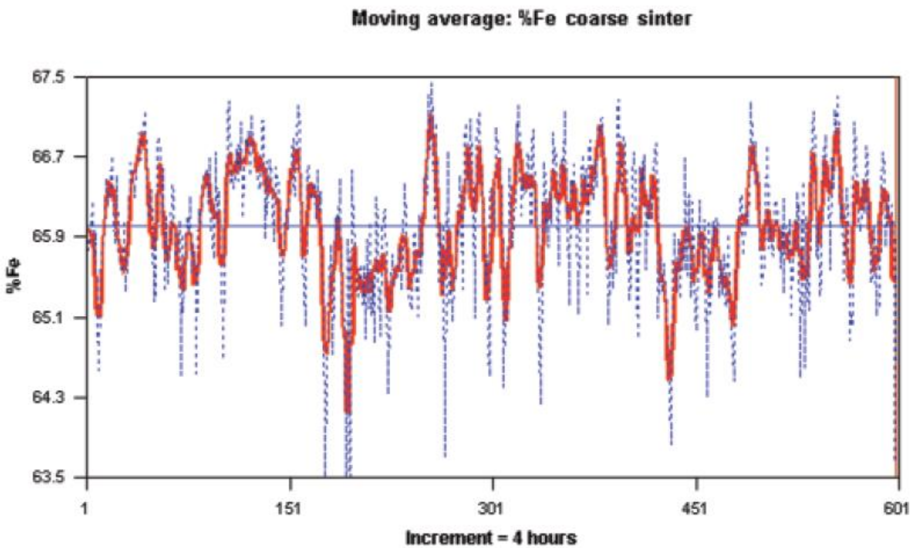
C’est ce qui reste quand on a enlever la partie explicative à l’aide de modèles basés sur une AR(p) et/ou une transformée de Fourrier. Il s’agit donc de modéliser le bruit

En définitive c’est ce qui nous intéresse le plus car c’est sur ce reste stochastique que sont établies les cartes de contrôles

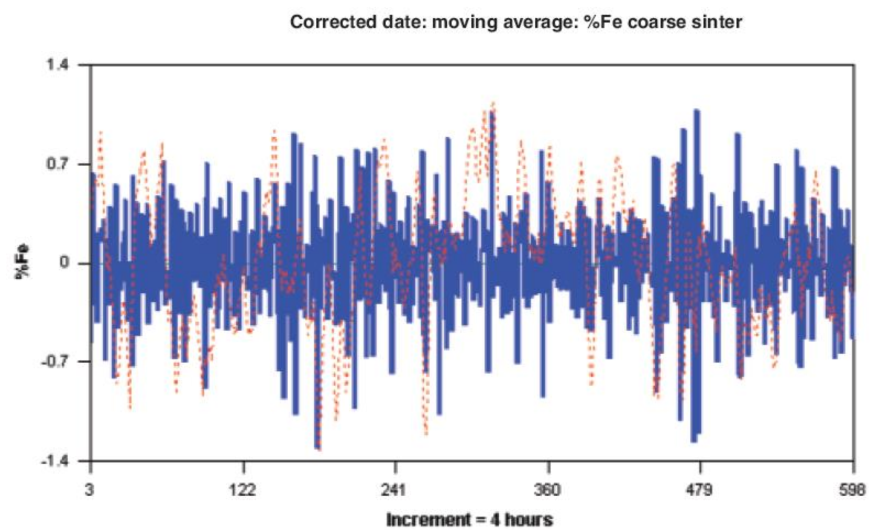
Le processus de moyenne mobile suit :

$$AM(q) = X_t = Cste + \theta_1.\varepsilon_{t-1} + \theta_2.\varepsilon_{t-2} + \theta_3.\epsilon_{t-3} + \ldots\ldots\ldots + \theta_p.\varepsilon_{t-q}$$

Remarque, la constante est nulle



Mean = 6.60E+01 Median = 6.61E+01 Std. = 1.05E-02



Mean = 6.62E+01
Targeted average = 6.61E+01

CARTES DE CONTRÔLE

PRINCIPALES CARTES

On étudiera que 3 cartes (correspond à la quasi-totalité des cartes utilisées)

- La carte de SHEWART
- La carte de EWMA
- La carte de CUSUM

PRINCIPE

Collecter un ensemble de valeurs ordonnées chronologiquement

« Filtrer » la partie déterministe à l'aide d'un modèle (régression linéaire, transformée de Fourrier, …), en repérant éventuellement la présence d'un processus récursif AR(p)

Le reste ou résidu est normalement un bruit, et, très souvent un bruit blanc modélisable par une loi normale centrée réduite

C'est à partir de ce « reste stochastique » que sont établies les cartes de contrôle. En effet ce bruit servira à déterminer une cible. Les valeurs mesurées sont ensuite comparée à cette cible. Des critères d'aide à la décision permettront d'affirmer si le processus suit son cours normal ou si un écart ou une dérive est en train de se manifester

PRINCIPAUX PARAMÈTRES

Principe :

- *La **valeur suivie** choisie est en général le <u>résultat</u> de la mesure*

- *La **valeur cible** choisie est en général la tendance du processus (estimé par la <u>moyenne</u> des résultats)*

- *La **comparaison** se fait en prenant en compte de la fluctuation acceptée pour ce processus (estimé par l'<u>écart type</u> des résultats)*

- *La **décision** (normal / écart) se fait par rapport à un <u>seuil</u> à définir*

LES DIFFÉRENTS TYPES DE CARTES

La carte **Shewhart** 1931

Principe : comparaison de la valeur mesurée avec la valeur moyenne
Conséquence : on considère que le comportement passé du processus n'apporte aucune information pour pouvoir décider si le comportement est normal.
Détection : les **dérives importantes**, ie les variations « brusques » et locales

La carte **EWMA** 1959

Principe : ce qui est suivi c'est le comportement de la moyenne dans le passé avec celui de la moyenne réactualisée par la mesure qui vient d'être faite. L'importance du comportement passé et réactualisé est ajusté par un poids λ.
Conséquence : l'influence du comportement passé de la moyenne décroit exponentiellement avec λ : c'est donc l'influence du passé récent qui est pris en compte.
Détection : des **dérives modérées** mais persistantes

La carte **CUSUM** (CUMmlative SUMmation) 1954

Principe : la valeur suivie S est la somme cumulée des écarts entre la valeur expérimentale observée x_i et la valeur cible de référence X°.
Conséquence : l'ensemble des écarts avec la moyenne est pris en compte
Détection : des **très faibles dérives** ou variations (qui par sommation des écarts passés deviennent significatives)

EWMA :

Filtration des hautes fréquences pour ne garder que la tendance générale (nécessite donc un échantillonnage important)

Le poids λ caractérise le poids du passé ou l'inertie du système.

Si une anomalie est déjà dans les valeurs récemment mesurées, elle a moins de chance d'être détectée, elle est donc moins adaptée aux échantillonnages fréquents

Très utile pour les détections saisonnières

Pour des dérives modérées

Elle permet de détectée des dérives persistantes

Les valeurs de λ sont en général comprises entre 0,25 et 0,5. Si λ tend vers 1 mieux vaut utiliser une carte Shewart, au contraire, si λ tend vers 0 mieux vaut utiliser une carte CUSUM

CUSUM

Détection des faibles changements dans la moyenne, c'est-à-dire des dérives petites

Les valeurs de h sont en général comprises entre 4 et 5

LA CARTE CUSUM : PRINCIPE

Principe : la valeur suivie S est la somme des écarts entre la valeur mesurée x_i et la cible (valeur de référence X°)

- Valeur cible : une estimation de la moyenne
- Valeur suivie : la somme des écarts entre les valeurs observées et la valeur cible
- Écart type : une estimation de l'écart type expérimental
- Intervalle de décision : $\pm$ h.σ (en général h $\approx$ 5)

Les principaux paramètres se déduisent de l'écart à la cible maximal acceptable :

$$K = \frac{\delta}{2}.\sigma$$

$$\delta = \frac{\left| x_{\max acceptable} - x_{ref} \right|}{\sigma}$$

Il faut donc avoir une idée de la précision voulue

Exemple : on veut être vigilent pour des écarts de 2 °C sur une valeur moyenne estimée à 25°C sachant que la variabilité est de 0.5 °C

$$K = 1, \quad \delta = 4$$

MÉTHODOLOGIE

$1^{ère}$ étape : le calcul de la moyenne μ et de son écart type σ

$2^{ème}$ étape : se donner un seuil d'alarme δ (en unité σ), l'alarme est « déclenchée » quand les « fluctuations » dépassent $\delta.\sigma$

en gral $\delta \sim 1$

$3^{ème}$ étape : construction de la carte

Souvent basée sur deux paramètres : k et h

Deux méthodes pour identifier ces paramètres :

1. A partir de tables à double entrée : $\delta.n^{1/2}$ et l'ARL
2. A partir des risques α (fausse alarme) et β (alarme négligée)

en gral $k \sim 0{,}5$ et $k \sim 4\text{-}5$

$4^{ème}$ étape : on trace la cible Si

CALCUL DE LA CIBLE ET DES LIMITES

Calcul de la somme (valeur de suivi) :

Estimation de la moyenne

$S_1 = x_1 - X°$
$S_2 = S_1 + (x_2 - X°)$
$S_3 = S_2 + (x_3 - X°)$

C'est S_i qui est suivi graphiquement

etc…

Valeur observée

$S_i = S_{i-1} + (x_i - X°)$

Graphiquement, les limites de contrôle sont donnée par les relations :

$C_i^+ = Max[0, (C_{i-1}^+ - K) + (x_i - X°)]$
$C_i^- = Max[0, (C_{i-1}^- - K) - (x_i - X°)]$

Avec pour valeurs initiales :
$C_0^+ = C_0^- = 0$

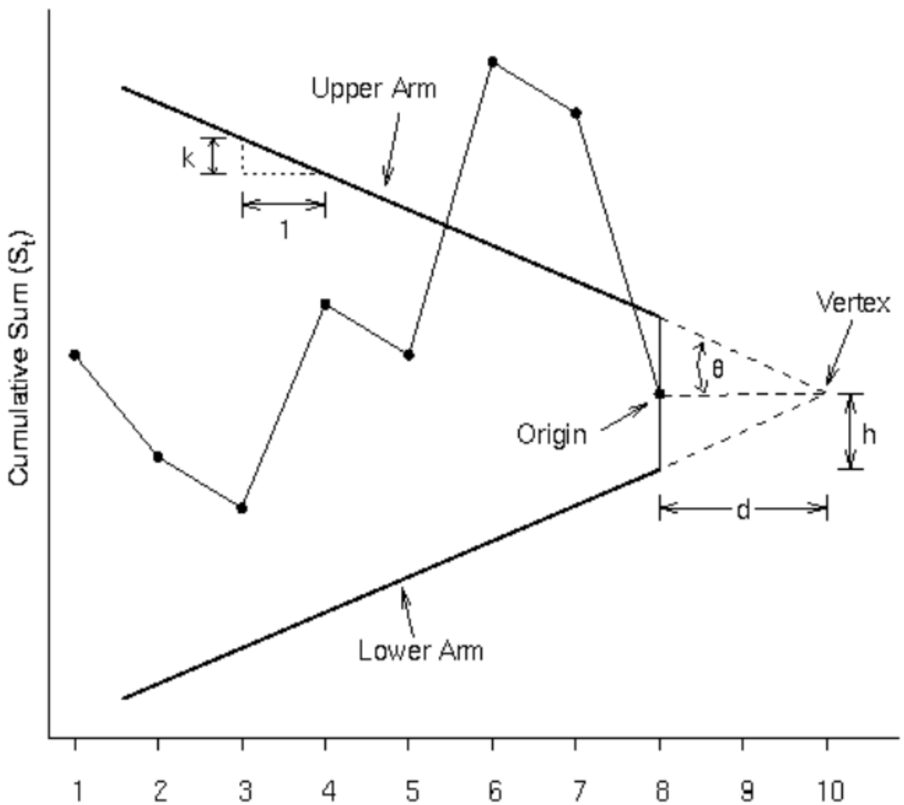
LA CARTE CUSUM : POUR UNE APPROCHE GRAPHIQUE

Deux paramètres doivent être calculés :

La fenêtre d'observation d $$d = \frac{2}{\delta^2} \mathrm{Ln} \frac{1-\beta}{\alpha}$$

La pente des droites de contrôle $$\mathrm{tg}\,\theta = \frac{h.\sigma}{d}$$

Paramètres liés à k et h estimés lors de la 3ème étape



Dans l'exemple précédent :

- d = 1,8 soit approximativement 2
- tg θ = 1,25 ($\theta \approx 50°$)

On remarque sur le graphe que S_i reste proche de zéro … sauf quand une anomalie apparaît

LA CARTE EWMA

Principe : la valeur suivie Z_i est un équilibre entre le passé (représentées par la moyenne valeurs observées) et le présent (représentées par la dernière valeur observée). La proportion entre le passé et le présent est ajusté par un paramètre λ.

- Valeur suivie : la valeur observée.
- Valeur cible : une moyenne pondérée par un paramètre λ.
- Écart type : ne estimation de l'écart type expérimental
- Intervalle de décision : $\pm$ h.σ (en général h = 3)

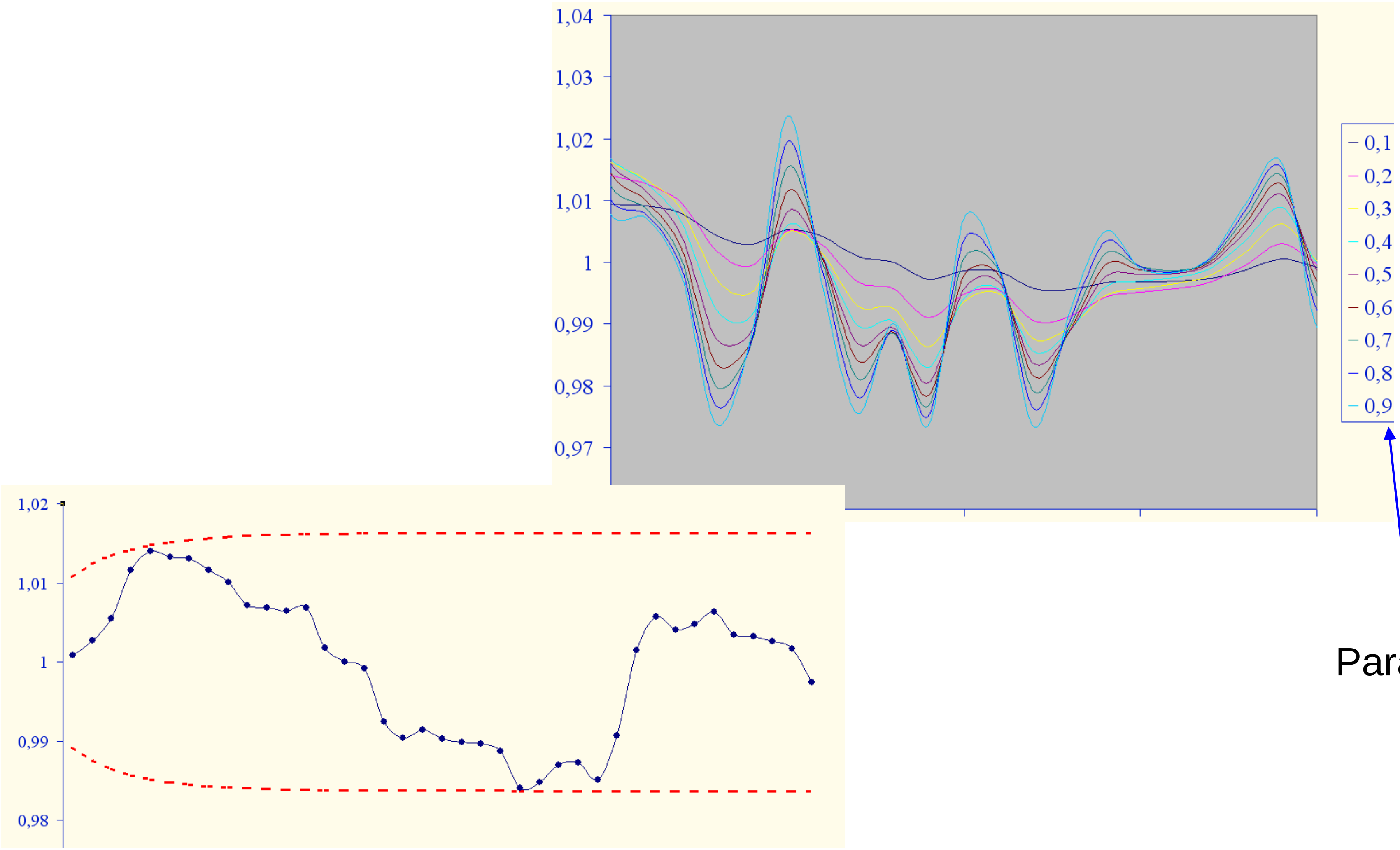
Calcul de la valeur de suivi (moyenne pondérée)

$$z_i = \lambda . X° + (1 - \lambda) . z_{i-1}$$

Les limites de surveillance sont données par les expressions :

$$LS = X° +/- h . \frac{s}{\sqrt{n}} . \sqrt{\frac{\lambda}{2-\lambda}(1-(1-\lambda)^{2i})}$$

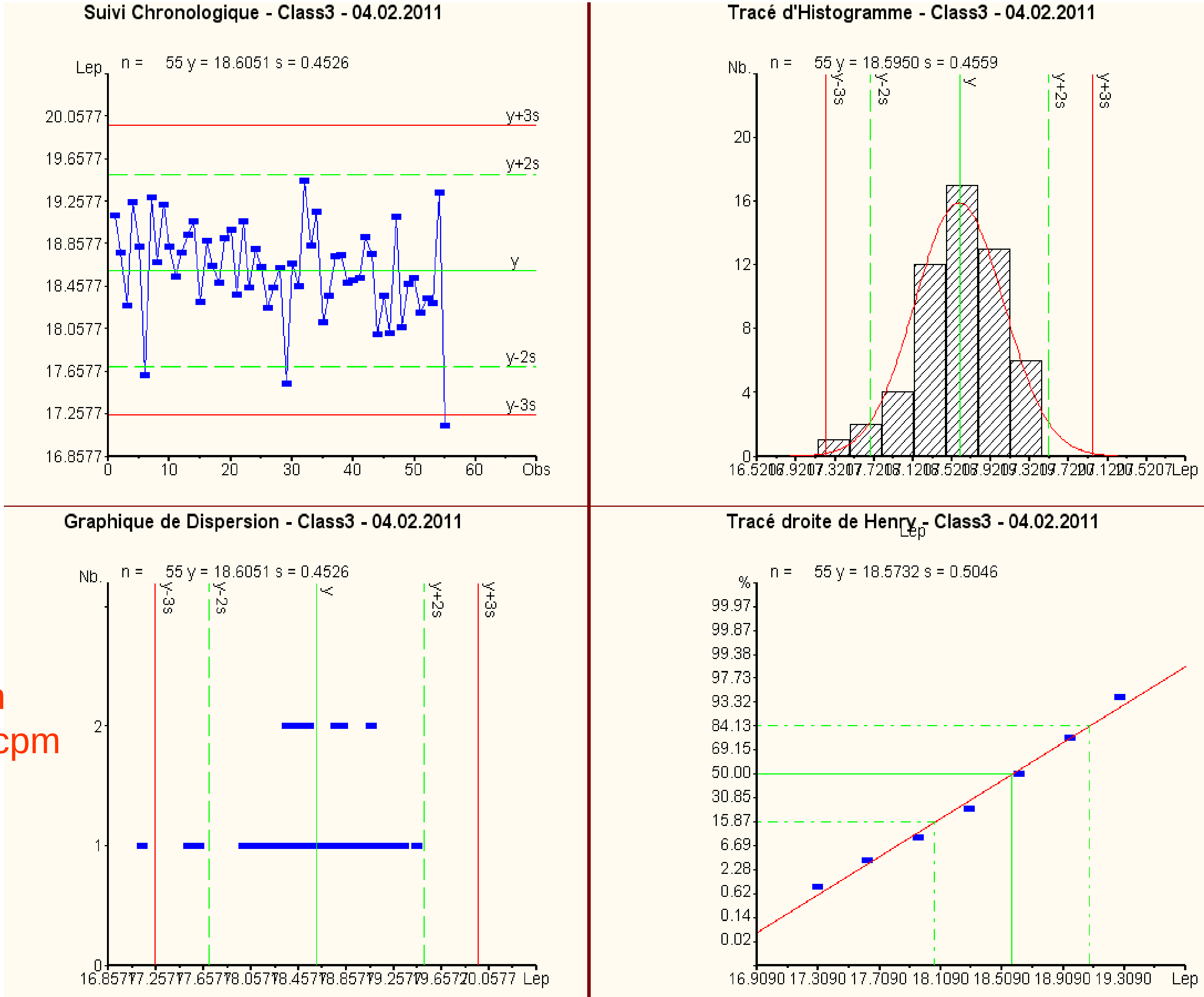
EXEMPLES



Paramètre λ

ETUDE DU BRUIT DE FOND

VARIATION DU BRUIT DE FOND (LEPTON, 100 MINUTES)

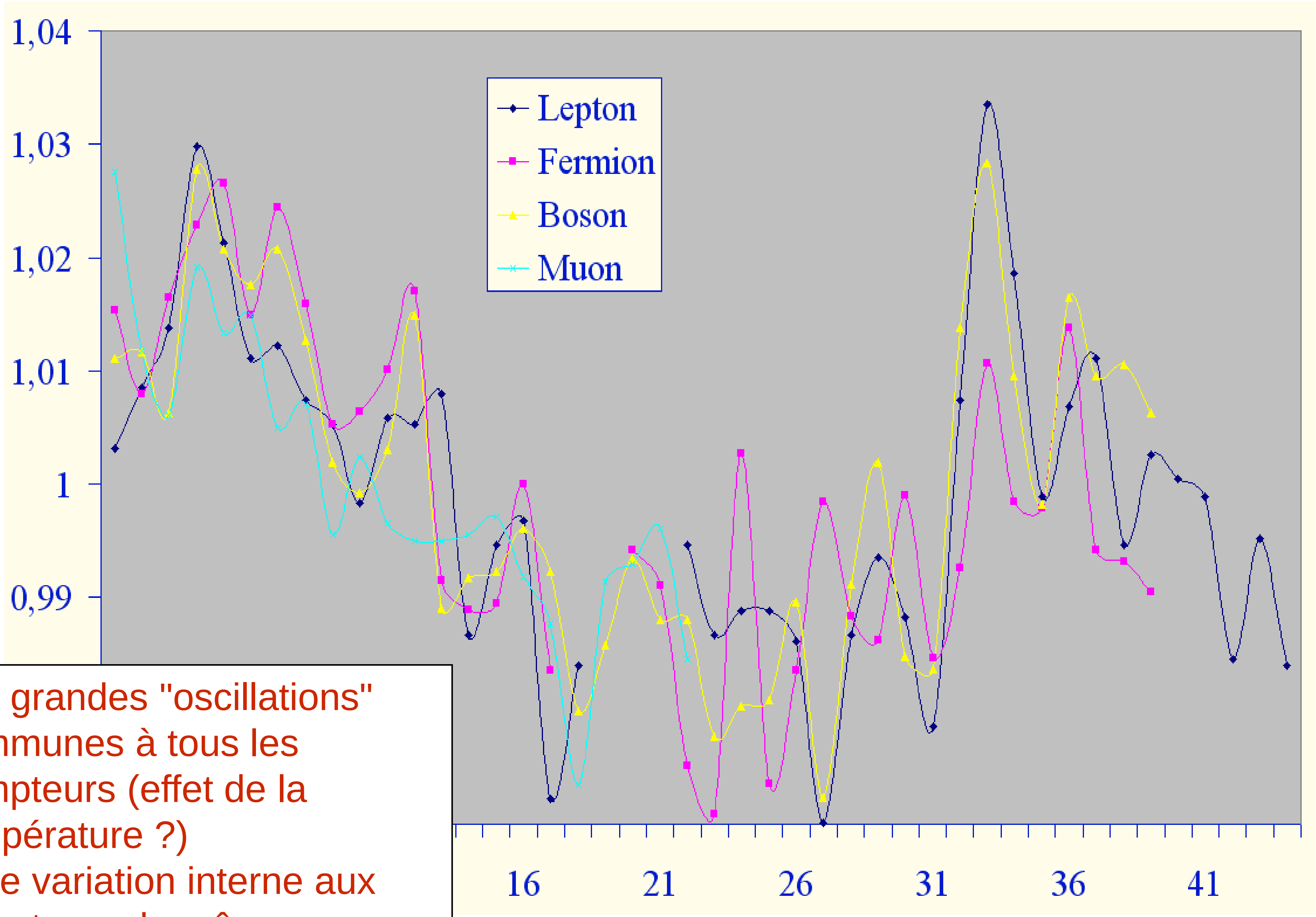


x_{moy} =18,6 cpm

σe_{xp} = 0,4526 cpm

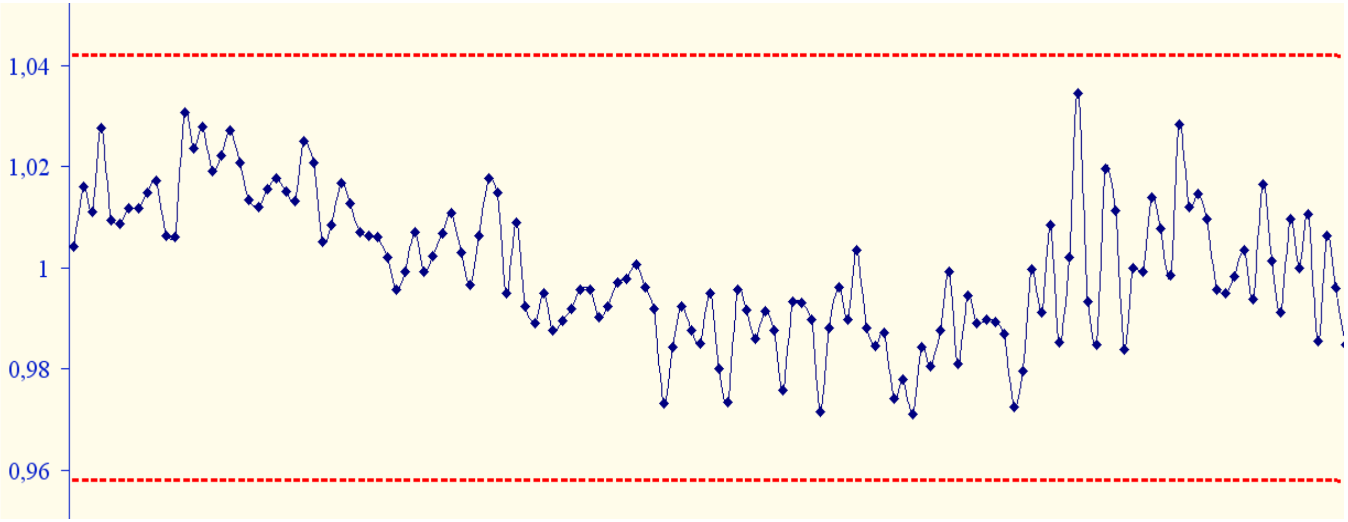
n = 55

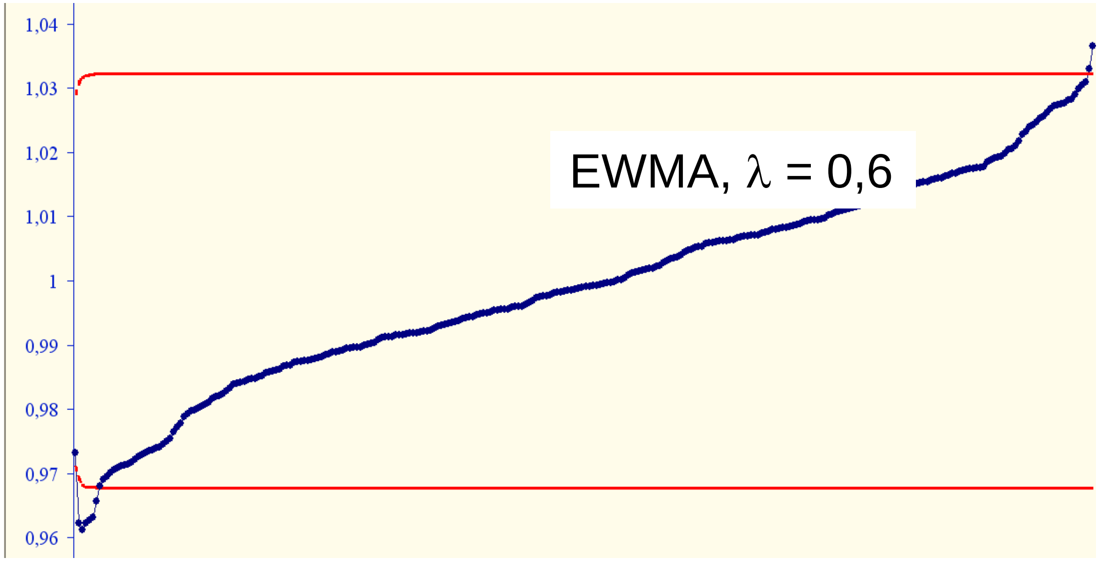
ÉVOLUTION DU BRUIT DE FOND (TOUTES LES 17 HEURES)



•De grandes "oscillations" communes à tous les compteurs (effet de la température ?)
•Une variation interne aux compteurs de même amplitude

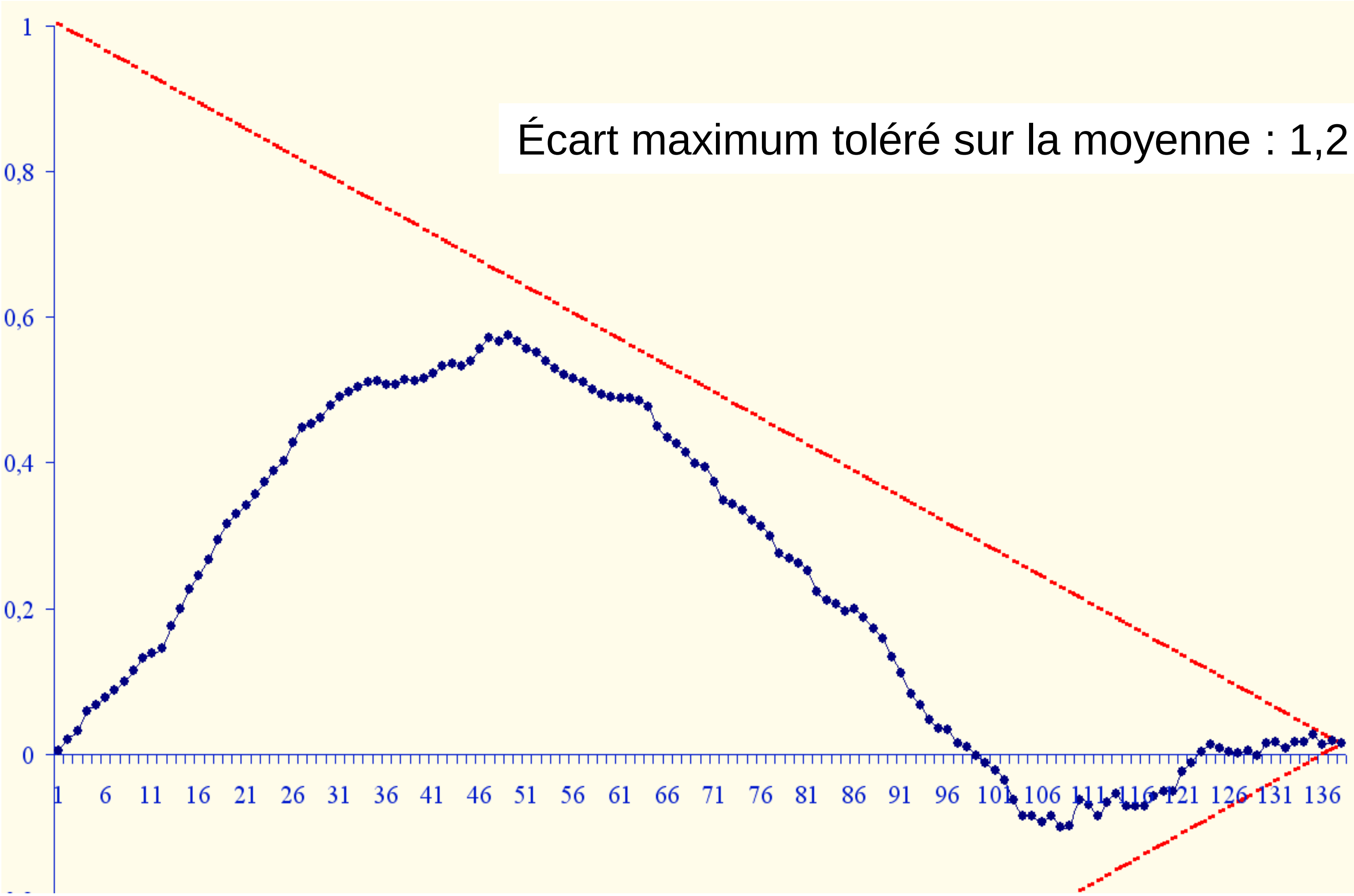
CARTES SHEWHART ET EWMA





CARTE CUSUM

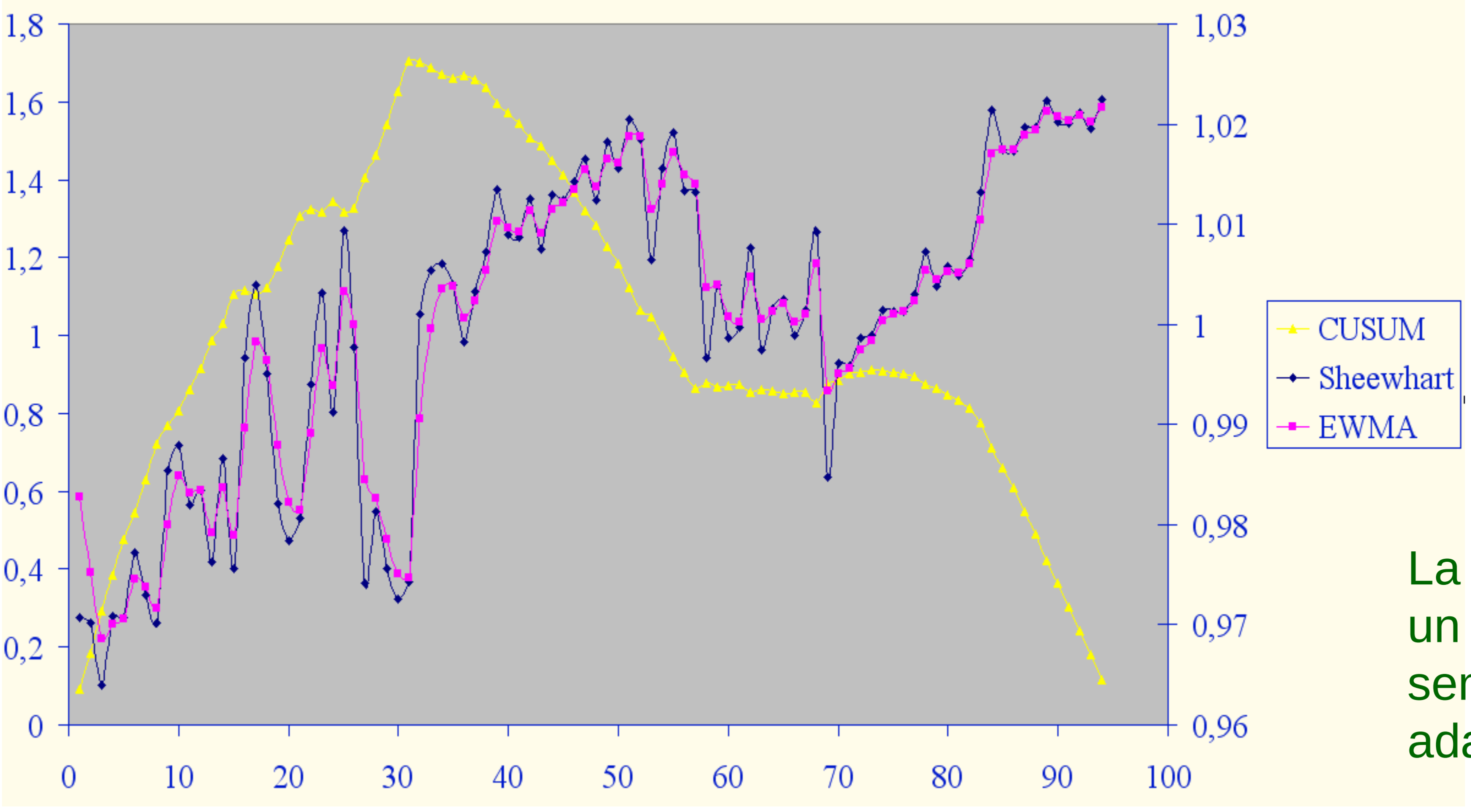
Écart maximum toléré sur la moyenne : 1,2 %



EFFICACITÉ TRITIUM

Analyse des résultats du compteur "Meson" (une centaine de données) :

- La variable efficacité ne suit pas une loi normale
- Il y a une tendance chronologique (vieillissement du PM)
- Moyenne : 65,77 % - Écart type expérimental : 7



La carte EWMA avec un paramètre de 0,6 semble la plus adaptée