Modélisation des séries temporelles

Séance 1

Décomposition d'une chronique

Frédéric Sur

`https://members.loria.fr/FSur/enseignement/modseries/`

Plan

Qu'est-ce qu'une série chronologique ?

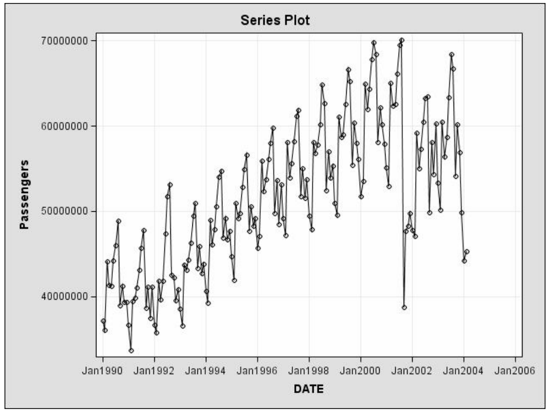
Séries temporelles / Chroniques / *Time series*
→ suite d'observations d'une grandeur au cours du temps.
→ temps *discret*.

Exemples :

- économétrie (taux de chômage),
- finance (cours d'action),
- écologie (pollution),
- démographie (population),
- météorologie (relevé de températures),
- astronomie (fluctuations de la magnitude d'un astre)
- ...

Modèle sous-jacent, intrinsèque, inconnu liant les observations.

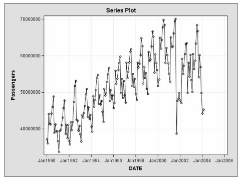
Exemple de série chronologique



Trafic aérien aux USA de 1990 à 2004.

But du cours : décrire l'évolution des chroniques à l'aide de modèles intrinsèques basés sur des propriétés statistiques.

Dans le modèle...



- **Perturbations ponctuelles** : variations de forte amplitude.
  (grève, krach boursier, attentats, mesure aberrante...)
  Cf. « modèles d'intervention ».
- **Tendance $\mathbf{T_t}$** : évolution globale.
- **Variations saisonnières $\mathbf{S_t}$** : fluctuations périodiques.
  « données corrigées des variations saisonnières ».
  → moyenne nulle sur une période.
- **Composante résiduelle (irrégulière) $\mathbf{u_t}$** : fluctuations irrégulières de faible amplitude. Imprévisibles ?
  → moyenne nulle.

Pourquoi étudier / modéliser les chroniques ?

- Description / explication d'un phénomène
chômage : variation tendancielle ou fluctuation saisonnière ?

- Prévision
consommation d'électricité, démographie, chaîne logistique...

- Étude de la dynamique
volatilité boursière

- Impact d'un événement sur une série chronologique
« mesure » de l'impact des politiques publiques

Dans ce cours : techniques statistiques, aspects *temporels*.

Point de vue complémentaire : aspects *fréquentiels*, traitement du signal.

Le cours

Cours en salle A006 (amphi Schwartz)
TP en salles A207, A208, B207.

Équipe pédagogique :

- Marius Albrand – IECL, équipe EDP
- Denis Villemonais – IECL, équipe proba-stat
  `http://www.normalesup.org/~villemonais/`
- Frédéric Sur – LORIA, analyse et traitement d'images
  `https://members.loria.fr/FSur/`

Modalités d'évaluation : test (format TP) le 30 mars 2020.

À faire avant chaque séance...

Polycopié :
– point de vue complémentaire
– démonstrations des résultats évoqués en amphi

⟶ **à lire avant le cours**
(« pour en savoir plus » optionnel)

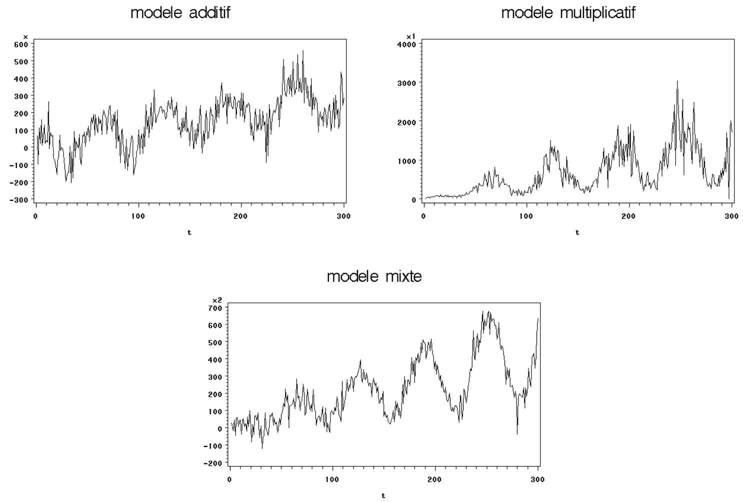
Pour approfondir :

- bibliographie du polycopié disponible à la bibliothèque.
- `members.loria.fr/FSur/enseignement/modseries/`
  → exercices à traiter en TP ;
  → slides + chapitres du polycopié à lire ;
  → suggestions de lectures complémentaires.

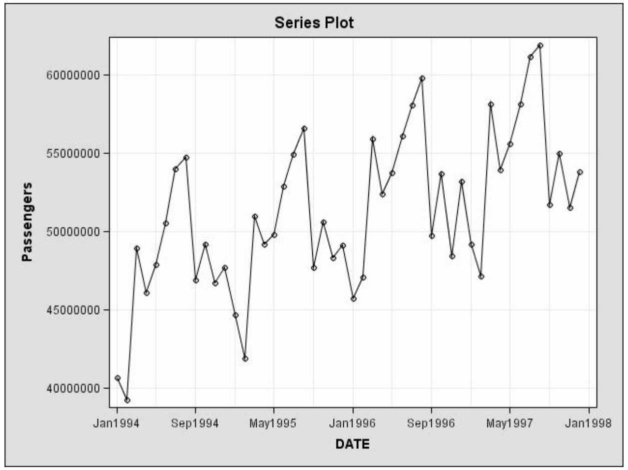
Séance 1

Modèles de décomposition

- Modèle additif : $X_t = T_t + S_t + u_t$
- Modèle multiplicatif : $X_t = T_t \cdot S_t \cdot u_t$
- Modèle mixte : $X_t = T_t \cdot S_t + u_t$



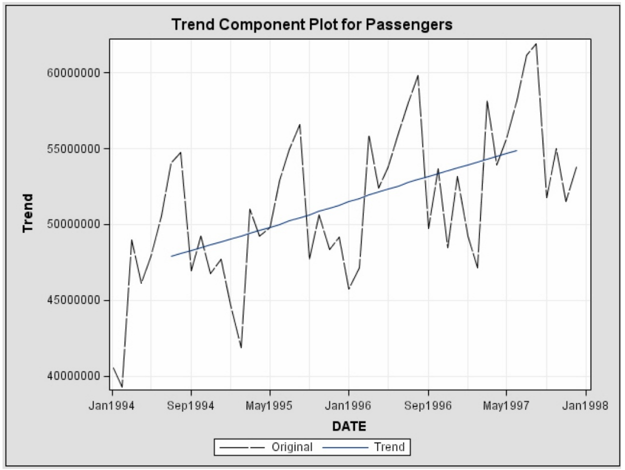
Exemple de décomposition



Trafic aérien aux USA de 1994 à 1997.

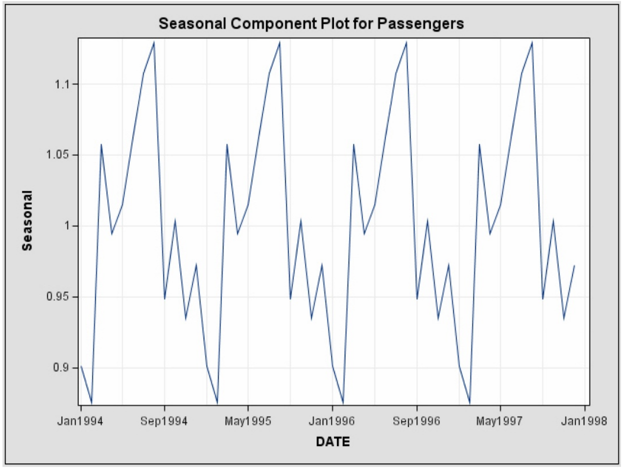
$\rightarrow$ modèle mixte.

Exemple de décomposition



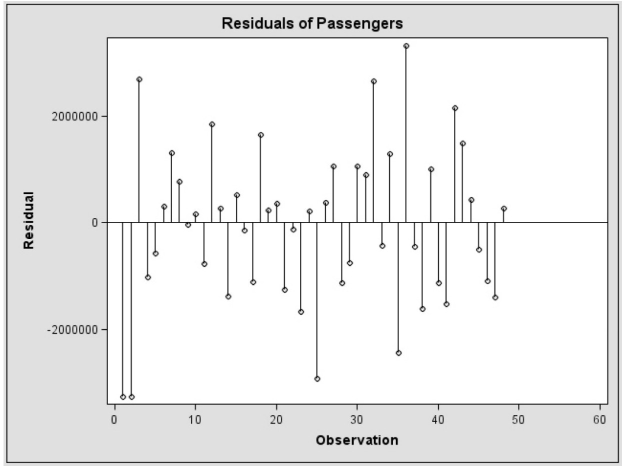
Tendance T_t.

Exemple de décomposition



Composante saisonnière S_t.
(modèle multiplicatif, donc S_t centré sur 1 et pas 0)

Exemple de décomposition



Résidus u_t.
(pas de structure apparente)

Composante tendancielle T_t

Exemples de composante tendancielle T_t *paramétriques* :

- linéaire : $T_t = at + b$
- quadratique : $T_t = at^2 + bt + c$
- exponentielle : $T_t = T_0 e^{at}$
- ...

avec les paramètres a,b,c,T_0... à déterminer.

Composante saisonnière S_t

S_t périodique, de période p.

$\longrightarrow$ éventuellement plusieurs composantes périodiques superposées.

Exemple : comp. annuelle + comp. trimestrielle...

Question : comment trouver les périodes ?

$\longrightarrow$ « ça se voit », expertise.
$\longrightarrow$ analyse fréquentielle (périodogramme - analyse de Fourier).

Comment trouver la (les) période(s) ?

$\rightarrow$ Transformée de Fourier Discrète

Pour une chronique $X_0, X_1, \ldots, X_{T-1}$, TFD :

$$\forall k \in \{0, 1, \ldots, T-1\}, \; \widehat{X}_k = \sum_{t=0}^{T-1} X_t e^{-2i\pi kt/T}$$

TFD inverse : (formule de reconstruction)

$$\forall t \in \{0, 1, \ldots, T-1\}, \; X_t = \frac{1}{T} \sum_{k=0}^{T-1} \widehat{X}_k e^{2i\pi kt/T}$$

X_t réel : $\widehat{X}_1 = \overline{\widehat{X}_{T-1}}, \widehat{X}_2 = \overline{\widehat{X}_{T-2}} \ldots$ donc : (T impair)

$$\forall\, t, \; X_t = \frac{\widehat{X}_0}{T} + \frac{1}{T} \sum_{k=1}^{(T-1)/2} 2|\widehat{X}_k| \cos(2\pi kt/T + \phi)$$

Conclusion : $2|\widehat{X}_k|$ est l'amplitude de la composante de fréquence k/T (ou période T/k).

Le périodogramme

Définition : périodogramme

$$\forall\, 0 < k \leqslant (T-1)/2, \quad J_{T/k} = \left|\widehat{X}_k\right|^2$$

$\rightarrow$ représente l'amplitude de la composante de période $\frac{T}{k}$ (fréquence $\frac{k}{T}$).
Le graphe de J est appelé *périodogramme*.
Le graphe de $\left|\widehat{X}_k\right|^2$ est le *spectre de puissance*.

$\longrightarrow$ les "pics" dans le périodogramme permettent d'identifier la période des composantes saisonnières.

Remarque : les $(\widehat{X}_k)$ ont tendance à décroître avec la fréquence.
(cf Riemann-Lebesgue & co)

Composante résiduelle u_t (le bruit)

$\rightarrow$ pas de tendance ou de phénomène périodique dans u_t

Modélisation : (u_t) réalisation d'un processus aléatoire (U_t) *stationnaire*.
(i.e. propriétés statistiques *invariantes au cours du temps*)

Définition : (U_t) stationnaire au second ordre

- $\forall t, \; \mathbb{E}(U_t) = m$ (moyenne constante)
- $\forall t, s, \; \mathrm{cov}(U_t, U_s) = \gamma(|t-s|)$
  (autocovariance symétrique, invariante par translation)

En particulier $\forall t$, $\mathrm{var}(U_t) = \gamma(0)$ (variance constante).

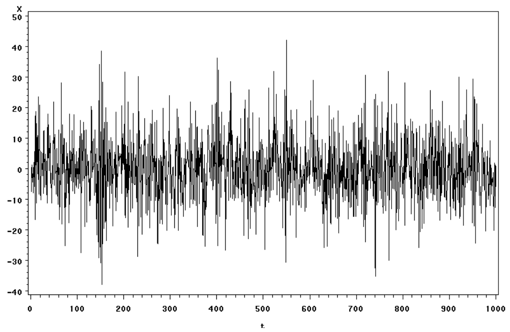
$\rightarrow$ les moments d'ordre 1 et 2 ne varient pas au cours du temps.

Cas particuliers de processus stationnaires

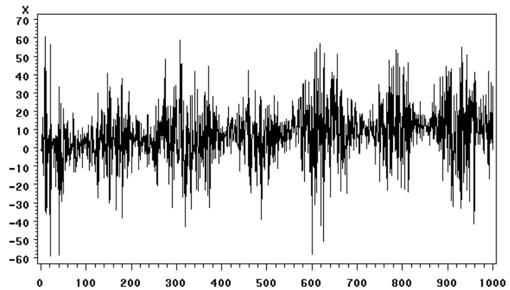
- Bruit blanc faible : $m = 0$ et $\forall h > 0, \; \gamma(h) = 0$.
  $\rightarrow$ *absence d'autocorrélation temporelle.*
- Bruit blanc fort : $m = 0$ et (ε_t) i.i.d.
  $\rightarrow$ *indépendance temporelle.*
- Bruit blanc gaussien : (ε_t) i.i.d. $\sim \mathcal{N}(0, \sigma^2)$.
  $\rightarrow$ *cf résidus dans la régression.*

Exemples

bruit blanc gaussien



processus non stationnaire



Caractérisation des processus stationnaires

Les processus stationnaires considérés dans ce cours seront de la forme :

$$U_t = m + \sum_{j=0}^{+\infty} a_j \varepsilon_{t-j}$$

avec (ε_t) **bruit blanc gaussien**

Remarque : bien entendu, en pratique on se débrouillera pour que le processus ne dépende pas d'une infinité de termes a_j... Cf séance 2.

Par définition, (U_t) caractérisé par

- moyenne m,
- fonction d'autocovariance γ ou d'autocorrélation ρ.

$\rightarrow$ comment estimer m et γ/ρ à partir de T observations ?

Estimation de la moyenne m

Estimateur de m : $\quad \overline{U}_T = \frac{1}{T}\sum_{t=1}^{T} U_t$

Propriété : asymptotiquement, si U_t stationnaire et ε_t b.b.g. :

$$\overline{U}_T \sim \mathcal{N}(m, C/T)$$

(où C dépend des covariances $\gamma(h)$)

Si μ valeur estimée :

$\rightarrow$ intervalle de confiance : $\quad \mu \pm 2\sqrt{\frac{C}{T}}$

$\rightarrow$ test de significativité :
calcul de la p-valeur $\Pr(|\overline{U}_T| \geq \mu)$ sous hypothèse $m = 0$
si p-valeur $< 5\%$, alors la moyenne est significative
(i.e. *significativement non nulle*)

Estimation de l'autocovariance γ / autocorrélation ρ

Définition : fonction d'autocorrélation

$$\forall h \geqslant 0,\ \rho(h) = \frac{\mathrm{Cov}(U_t, U_{t+h})}{\sqrt{\mathrm{Var}(U_t)\mathrm{Var}(U_{t+h})}} = \frac{\gamma(h)}{\gamma(0)}.$$

Fonction d'autocorrélation empirique :

$$\widehat{\rho}(h) = \frac{T}{T-h}\,\frac{\sum_{t=h+1}^{T}(U_t - \overline{U}_T)(U_{t-h} - \overline{U}_T)}{\sum_{t=1}^{T}(U_t - \overline{U}_T)^2}$$

Graphe de $\widehat{\rho}(h)$: *corrélogramme*
(**ACF** = *auto-correlation function*).

Propriété : asymptotiquement, si U_t stationnaire et ε_t b.b.g. :

$$\widehat{\rho}(h) \sim \mathcal{N}(\rho(h), C'/T)$$

(où C' dépend des covariances $\gamma(h)$)

$\rightarrow$ intervalle de confiance, test de significativité

Intérêt pratique du corrélogramme

(fondamental pour la prochaine séance)

Proposition

Si $U_t = m + \sum_{j=0}^{k} a_j \varepsilon_{t-j}$, alors $\forall h > k,\ \rho(h) = 0$.

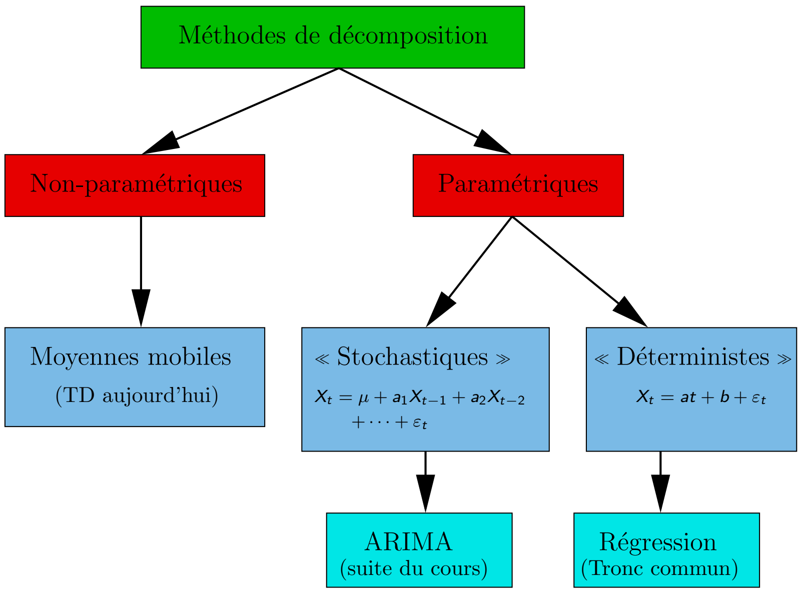
Preuve : $\text{Cov}(U_t, U_{t+h}) = \sum_{i=0}^{k}\sum_{j=0}^{k} a_i a_j \text{Cov}(\varepsilon_{t-i}, \varepsilon_{t+h-j})$.

Or $\text{Cov}(\varepsilon_{t-i}, \varepsilon_{t+h-j}) = 0$ si $h \neq j - i$ (car (ε_t) b.b.)

Donc si $h > k$, $\forall\, i, j \in \{0, \ldots, k\}$, $\text{Cov}(\varepsilon_{t-i}, \varepsilon_{t+h-j}) = 0$.

Séance 1

Décomposition des chroniques : typologie



Modèles paramétriques déterministes

(X_t) : chronique mensuelle (trimestrielle dans le poly.)

Modèle paramétrique :

$$X_t = \underbrace{a_1 + a_2 t}_{T_t} + \underbrace{\sum_{i=1}^{12} b_i \delta_i(t)}_{S_t} + u(t)$$

δ_i : indicatrice du mois i.

Estimation des paramètres aux moindres carrés des résidus :

$$\begin{cases} \min\limits_{a_i, b_j} \sum_{t=1}^{T} \left(X_t - a_1 - a_2 t - \sum_{i=1}^{12} b_i \delta_i(t) \right)^2 \\ \text{t.q. } \sum_{i=1}^{12} b_i = 0 \end{cases}$$

$\rightarrow$ tables de Buys-Ballot dans le poly.

Filtrage par moyennes mobiles

Définition : moyenne mobile

Soit (X_t) une chronique, et
$Y_t = M(X_t)$ la chronique telle que

$$\forall t,\ Y_t = \sum_{i=-m_1}^{m_2} \theta_i X_{t+i}$$

(Y_t) est obtenue de (X_t) par *filtrage par moyenne mobile*.

Remarque : M est un opérateur linéaire.

Décomposition d'une chronique par filtrage

Premier objectif : décomposition de $X_t = T_t + S_t + u_t$.

On cherche un filtre tel que

- $M(T_t) = T_t$;
- $M(S_t) = 0$;
- $M(u_t)$ « aussi petit que possible ».

de sorte que : $M(X_t) \simeq T_t$

Les moyennes mobiles arithmétiques

Filtre symétrique : $m_1 = m_2\ (= m)$, $\theta_i = \theta_{-i}$
$M(X_t) = \sum_{i=-m}^{m} \theta_i X_{t+i}$.

Propriété

Une moyenne mobile symétrique telle que $\sum_{i=-m}^{m} \theta_i = 1$ conserve les chroniques affines $X_t = at + b$ (i.e. $\forall\, t,\ M(X_t) = X_t$)

Propriété

Parmi ces moy. mob., celles qui « minimisent » $M(u_t)$ avec (u_t) réalisation d'un b.b. faible (ε_t) vérifient : $\forall i, \theta_i = \frac{1}{2m+1}$.

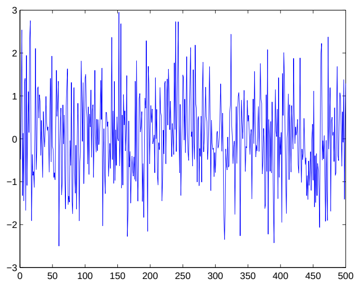
Preuve : $\mathrm{Var}(M(\varepsilon_t)) = \mathrm{Var}(\sum_{i=-m}^{m} \theta_i \varepsilon_{t+i}) = \sigma^2 \sum_{i=-m}^{m} \theta_i^2$,
à minimiser, sous contrainte $\sum_{i=-m}^{m} \theta_i = 1$.

Définition : moyenne mobile arithmétique

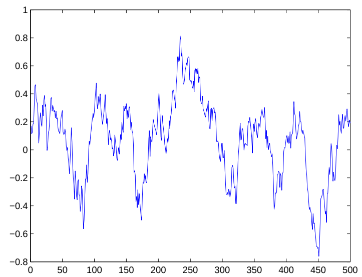
$$\forall i \in \{-m, \dots, m\},\ \theta_i = \frac{1}{2m+1}.$$

Effet de Slutsky-Yule

Attention à l'apparition éventuelle d'une périodicité « artificielle ».



bruit blanc gaussien, $\sigma = 1$



lissage arithmétique, $2m + 1 = 19$

Remarque : l'écart-type du bruit est divisée par $\sqrt{2m+1}$ (ici $\simeq 4.36$)

Saisonnalité et moyennes mobiles arithmétiques

$$M_{2m+1}(X_t) = \frac{1}{2m+1}\left(X_{t-m} + X_{t-m+1} + \cdots + X_{t+m}\right).$$

Propriété

Les composantes saisonnières S_t de période $2m+1$ et de moyenne nulle sur une période sont éliminées par M_{2m+1} : $M_{2m+1}(S_t) = 0$.

Remarque : composantes saisonnières de période $2m$:

$$M_{2m}(X_t) = \frac{1}{2m}\left(\frac{1}{2}X_{t-m} + X_{t-m+1} + \cdots + X_{t+m-1} + \frac{1}{2}X_{t+m}\right).$$

Récapitulatif

Hypothèses : (X_t) chronique à décomposer sous la forme :

$$X_t = T_t + S_t + u_t,$$

avec

- T_t linéaire
- S_t de période p et moyenne nulle sur la période.

Alors :

- $M_p(X_t) = M_p(T_t) + M_p(S_t) + M_p(u_t) \simeq T_t$.
- $M_k(S_t + u_t) \simeq S_t$ si $k << p$
  Problème : pas forcément de moyenne nulle sur une période.
- $M_k(S_t + u_t) - M_p\big(M_k(S_t + u_t)\big)$ est de moyenne nulle.

Un algorithme de décomposition par filtrage

Décomposition de $X_t = \mathbf{T_t} + \mathbf{S_t} + \mathbf{u_t}$.

1. périodogramme, connaissances $\rightarrow$ période p de S_t
2. estimation de la tendance : $\mathbf{T_t} = M_p(X_t)$
3. estimation de $\Sigma_t = S_t + u_t$ par $\Sigma_t = X_t - T_t$
4. estimation de la composante saisonnière : $\mathbf{S_t} = M'(\Sigma_t)$
   avec $M'(\Sigma_t) = M_k(\Sigma_t) - M_p M_k(\Sigma_t)$ et k « petit »
5. estimation de la série *corrigée des variations saisonnières* (cvs) : $X'_t = X_t - S_t$
6. composante résiduelle $\mathbf{u_t} = X'_t - T_t$

Généralisation

- moyennes mobiles non-arithmétiques pour tendance de degré > 1.
- filtrage itératif

Remarque : c'est l'idée du programme **Census X11** (1965).

$\rightarrow$ tendance, saisonnalité, données c.v.s., résidus.

Séance 1

Conclusion

La décomposition (additive / multiplicative / mixte ?) d'une chronique permet d'estimer :

- tendance,
- composante saisonnière (période ?),
- composante résiduelle / irrégulière (processus stationnaire).

Dans ce cours :

- filtrage (moyennes mobiles, *Census X11*)
  → en TD aujourd'hui
  discussion détaillée du filtrage : cf traitement du signal
- processus ARIMA
  → objet du reste du cours