AJUSTEMENT DE COURBES ET SERIES CHRONOLOGIQUES

IUT STID, UE21 M2102

2019-2020

SÉBASTIEN PINEL
sebpinel@gmail.com

Planning du cours

- Cours : 10h
- TD : 10h.
- TP : 10h, 5 séances de TP de 2h.
- Evalutation
    - Evaluation 1 : 10/02
    - Evaluation 2 : 26/02/2020
    - TP : 10/02/2020

Contents

1. AJUSTEMENT LINÉAIRE

Soient X et Y deux variables quantitatives définies sur une même population Ω. On considère le couple de variables (X, Y) dont les modalités sont les couples (x_i, y_i) où x_i = X(ω_i) et x_i = Y (ω_i) pour l'individu ω_i

Si les observations de deux variables statistiques X et Y sont connues individuellement, on commence par les visualiser en les représentant sous la forme d'un nuage de points (cf. figure ci-contre) : dans un repère cartésien, chaque observation (x_i, y_i) est figurée par le point M_i (x_i, y_i) et la forme du nuage donne une information sur le type d'une éventuelle liaison.

Supposons que l'examen du nuage de points conduise à rechercher une droite d'ajustement. Le calcul des coefficients de cette droite va être exposé dans le cas où les observations sont connues individuellement.

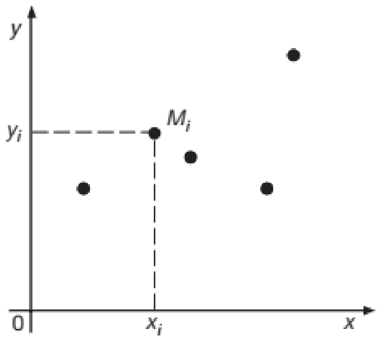


Figure 1. Nuage de points

1.1. Caractéristiques d'un couple de 2 variables quantitatives

<u>Rappels (formule moyenne et écart-type) :</u> Soit X une variable statistique,

Cas de données individuelles :

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{i=N} x_i \qquad \sigma_X = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}(x_i - \bar{x})^2}$$

Cas de données groupées dans un tableau de contingence (covariance pondérée) :

(f_i est la fréquence d'apparition de $x_i$tel que $\sum f_i = 1$)

$$\bar{X} = \sum_{i=1}^{i=N} f_i x_i \qquad \sigma_X = \sqrt{\sum_{i=1}^{i=N} f_i(x_i - \bar{x})^2}$$

<u>Définition/proposition :</u> la covariance entre deux variables statistiques X et Y est définie par

Cas de données individuelles :

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

Cas de données groupées dans un tableau de contingence (covariance pondérée) :

$$\sigma_{XY} = \sum_{i,j=1}^{i=k,l} f_{ij}(x_i - \bar{x})(y_j - \bar{y}) = \sum_{i,j=1}^{i=k,l} f_{ij} x_i y_j - \bar{x}\bar{y}$$

<u>Propriétés de la covariance :</u>

1. $Cov(X,Y) = Cov(Y,X)$;
2. $Cov(X,X) = Var(X)$;
3. $Var(X+Y) = Var(X) + 2Cov(X,Y) + Var(Y)$;
4. Pour a, b, $x_0, y_0 \epsilon \mathbb{R}$, on a $Cov(aX + x_0, bY + y_0) = abCov(X,Y)$
5. Pour a, b, $c \epsilon \mathbb{R}$, on a $Var(aX + bY + c) = a^2Var(X) + b^2Var(Y) + 2abCov(X,Y)$
6. $|Cov(X,X)| \leq \sqrt{Var(X)Var(Y)}$

Définition : En probabilités et en statistique, la **corrélation entre plusieurs variables aléatoires ou statistiques** est une notion de liaison qui contredit leur indépendance.

Définition: Le **coefficient de corrélation entre deux variables statistiques X et Y** est défini par

$$\mathrm{Cor}_{X,Y} := \mathrm{r}_{X,Y} := \rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \times \sigma_Y}$$

Propriétés du coefficient de corrélation linéaire :

Pour a, b, $x_0, y_0 \in \mathbb{R}$, on a

$$\mathrm{r}_{aX+x_0,bY+y_0} = \frac{Cov(aX+x_0, bY+y_0)}{\sigma_{aX+x_0} \times \sigma_{bY+y_0}} = \frac{abCov(X,Y)}{ab \times \sigma_X \times \sigma_Y} = \begin{cases} +\mathrm{r}_{X,Y}, si\ a\ et\ b\ de\ même\ signe \\ -\mathrm{r}_{X,Y}, si\ a\ et\ b\ de\ signe\ opposé \end{cases}$$

Remarques :

1. Numériquement, $\mathrm{r}_{X,Y}$ est compris dans [-1, 1] et est un nombre sans dimension
2. Les deux séries ne sont pas linéairement corrélées⇔ $\mathrm{r}_{X,Y}$=0
3. Les deux séries sont d'autant mieux corrélées que $\mathrm{r}_{X,Y}$ est proche de 1 ou de -1.
4. Ce coefficient de corrélation est extrêmement sensible à la présence de valeurs aberrantes ou extrêmes
5. Une corrélation égale à 0 signifie que les variables ne sont pas corrélées linéairement, elles peuvent néanmoins être corrélées non-linéairement, comme on peut le voir sur la troisième ligne de l'image ci-dessous.

Corrélation	Négative	Positive
Faible	de -0,5 à 0,0	de 0,0 à 0,5
Forte	de −1,0 à −0,5	de 0,5 à 1,0

Remarques :

Il existe d'autre indicateur de corrélation linéaire (Kendall, Spearman) ou non-linéaire.

1.2. Ajustement linéaire d'un nuage de points

1.2.1. La méthode de Mayer

Définition : Soit $(M_i = (x_i;\ y_i))_{1\le i\le n}$ un groupe de n points associe· à une série statistique. Le **point moyen du groupe** est le point G=($\bar{x}$, $\bar{y}$) où $\bar{x}$ et $\bar{y}$ sont les moyennes des séries $(x_i)_{1\le i\le n}$ et $(y_i)_{1\le i\le n}$

Méthode de Mayer : on partage la série en deux groupes, selon les valeurs de x_i en respectant leur ordre croissant, on obtient alors deux groupes de points. On détermine les points moyens, notes G1 et G2, de ces groupes. La droite de Mayer l'unique droite passant par G1 et G2.

Exemple : Le tableau suivant donne l'indice mensuel des dépenses d'assurance maladie d'aout 1994 à juin 1995 (tendances observées à la fin juillet 1995-base 100 janvier 1990, source : Département statistique de la Caisse Nationale de l'Assurance Maladie des Travailleurs Salaries).

Mois	Août 94	Oct 94	Déc 94	Fév 95	Avril 95	Juin 95
Rang du mois $X(i)$	1	3	5	7	9	11
Indice $Y(i)$	$123,4$	$125,9$	$127,5$	$127,9$	129	$131,4$
	premier groupe			second groupe		

On commence par séparer les éléments du tableau en deux groupes.

Puis on calcule les points moyens de chaque groupe :

Soit G1 le point moyen des 3 premiers points du nuage:

x(G1)=(1+3+5)/3= 3

y(G1)=(123; 4+125,9+127,5)/3= 125,6

Soit G2 le point moyen des 3 derniers points du nuage:

x(G2)=(7+9+11)/3= 9

y(G2)=(127,9+129+131,4)/3= 129,43

On place les points G1=(3,125; 6) et G2=(9; 129,43) sur la courbe et on trace la droite passant par ces points.

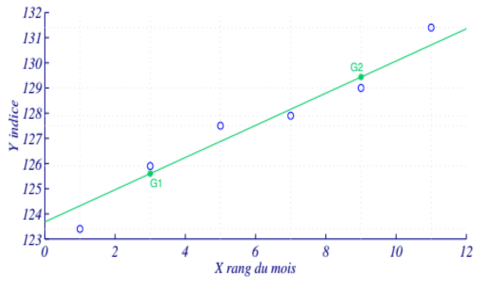


Figure 2. Droite de Mayer passant par le nuage de points

1.2.2. La méthode des moindres carrés

Nous allons chercher la meilleure droite au sens des moindres carrés, c.à.d. telle que $\sum_{i=1}^{i=n}|M_iH_i|^2$ soit minimum

<u>Remarque :</u> Les distances sont comptées parallèlement à l'un des axes des coordonnées ; nous avons choisi ici l'axe des ordonnées

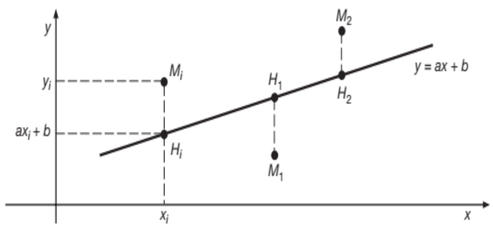


Figure 3. Interprétation géométrique de la droite des moindres carrés

Il s'agit de déterminer la droite d'équation $\{y = b_0 + b_1 x\}$ telle que :

$$d(b_0, b_1) = \sum_{k=1}^{k=N} (y_k - (b_0 + b_1 x_k))^2$$

Les inconnues sont b_0 et b_1. b_1 est le coefficient directeur de la droite et b_0 son ordonnée à l'origine.

3. La fonction d a pour variables b_0 et b_1. Les dérivées partielles seront ainsi calculées suivant ces deux variables et les équations définissant le point critique résolues par rapport à ces deux variables. Ici les inconnues ne sont pas les variables X et Y mais les variables b_0 et b_1.

<u>Définition/Proposition :</u> La **droite $D_{Y/X}$ de régression de** Y par rapport à X est la droite d'équation $y = \beta_0 + \beta_1 x$ qui rend la fonction $d(b_0, b_1)$ minimale.

Les valeurs des nombres réels β_0 et β_1 en sont déduites: $\beta_1 = \frac{Cov(X,Y)}{Var(X)}$ $\beta_0 = \bar{y} - \beta_1 \bar{x}$

<u>Remarque :</u>

Cette méthode s'appelle la méthode des moindres carrés ordinaires car la fonction d représente une distance quadratique (c.-à-d. d'ordre 2). Le minimum de cette fonction d positive et différentiable sur $\mathbb{R}^2$ est réalisé en un unique point critique, c'est-à-dire en un point où les deux dérivées partielles de cette fonction sont nulles.

Remarques :

1. La droite de régression obtenue par la méthode des moindres carrés, passe par le centre de masse $(\bar{x}, \bar{y})$
2. En échangeant les rôles des variables X et Y. nous obtenons la régression linéaire de …… en……… qu'on note ……..En général les deux droites de régression sont distinctes.

1.3.Décomposition de la variation et coefficient de détermination

Définition : Le but d'un ajustement linéaire est d'expliquer une partie de la variation de la variable Y appelée la **variable expliquée ou à expliquer** du fait de sa dépendance linéaire à la variable X, qui est appelée la **variable explicative**.

- Toute variable admet une certaine variation, généralement mesurée en terme de variance. Pour mesurer la variation de la variable Y, nous considérerons les différences entre les observations y_k et le moyenne $\bar{y}$.
  La somme des carrés $\mathrm{SCT} := \sum_{k=1}^{k=N}(y_k - \bar{y})^2$est appelée **la somme des carrés totale ou variation totale**.
- Si la variable Y dépend de X, et que nous la mesurons sur des individus avec différences valeurs de X, nous observerons une variation en conséquence. Ici nous ne mesurons que la **variation expliquée** par la régression linéaire de Y en X obtenue par la méthode des moindres carrés.
  En notant $\widehat{y_k}$ la valeur prédite par cet ajustement, la **variation expliquée par le modèle** est évaluée par la somme des carrés $\mathrm{SCE} := \sum_{k=1}^{k=N}(\widehat{y_k} - \bar{y})^2$ qui est appelée **la somme des carrés due à la régression** ou **variation expliquée par le modèle**.
- Généralement, lorsque nous mesurons Y sur des individus avec une même valeur de X, nous observons encore une certaine variation. Il s'agit de la **variation inexpliquée** par la régression linéaire de Y en X obtenue par la méthode des moindres carrés.
  La somme des carrés $\mathrm{SCR} := \sum_{k=1}^{k=N}(y_k - \widehat{y_k})^2$est appelée **la somme des carrés des résidus** ou la **variation résiduelle**.

Définition : toutes ces caractéristiques de la régression peuvent se résumer dans un tableau appelé **tableau de l'analyse de la variance**.

Sources de variation	Sommes des carrés
Expliquée par la régression	SCE
Résiduelle	SCR
Totale	SCT

Théorème : (décomposition de la variation totale) La décomposition de la variation totale se traduit mathématiquement par $\sum_{k=1}^{k=N}(y_k - \bar{y})^2 = \sum_{k=1}^{k=N}(\widehat{y_k} - \bar{y})^2 + \sum_{k=1}^{k=N}(y_k - \widehat{y_k})^2$

Remarque : La formule de décomposition de la variance correspond à l'égalité suivante:

Variation totale= Variation expliquée+ Variation inexpliquée⇔SCT=SCE+ SCR

Définition : $\widehat{y_k} - \bar{y}$ représente la différence expliquée par l'ajustement.

Le **résidu de l'ajustement** est $r_k := (\widehat{y_k} - \bar{y})$ et représente la différence inexpliquée par l 'ajustement

Le pourcentage de la variation totale qui est expliquée par l'ajustement est évalué par le coefficient de détermination défini ci-dessous.

Définition : Le **coefficient de détermination**, noté R^2 , est défini par $R^2=\frac{Variation\ Expliquée}{Variation\ Totale} = \frac{SCE}{SCT}$

Remarques : Ce coefficient donne une idée du pourcentage de variabilité de la variable à modéliser, expliqué par la variable explicative.

Le modèle explique donc R^2% de la variance des données (rendant ou non le modèle pertinent).

1. Le coefficient de détermination R^2 prend ses valeurs entre …….. et ………..

2. Plus ce coefficient est proche de 1, …………… est le modèle et plus les données sont alignées sur la droite de régression.

3. R^2=1⇔SCR=0

⇔ Les valeurs observées (x_i, y_i) sont sur une droite

⇔ la liaison entre Y et X est parfaitement linéaire

⇔$y_i = \widehat{y_i}$

4. R^2=0⇔SCE=0⇔ La droite des moindres carrés est horizontale. Il n'y a aucune liaison linéaire observée entre Y et X. il n'y pas de liaison linéaire, mais possibilité d'une liaison d'un autre type

Proposition : le coefficient de détermination R^2 est égal au carré du coefficient de corrélation empirique entre x et y, noté $\rho_{x,y}$, :

$$R^2 = \rho_{x,y}^2 = \left(\frac{Cov(x,y)}{\sigma_x \times \sigma_y}\right)^2 = \left(\frac{\sum_{i=1}^{i=N}(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x \times \sigma_y}\right)^2 = \frac{(\overline{xy}-\bar{x}\bar{y})^2}{Var(x)\times Var(y)} = \frac{\left(\frac{1}{N}\sum_{i=1}^{i=N}(x_i y_i)-\bar{x}\bar{y}\right)^2}{Var(x)\times Var(y)}$$

/ !\

Si les deux variables sont totalement indépendantes, alors leur corrélation est égale à 0.

La réciproque est cependant fausse !

En effet, le coefficient de corrélation indique uniquement une dépendance linéaire. D'autres phénomènes, par exemple, peuvent être corrélés de manière exponentielle, ou sous forme de puissance

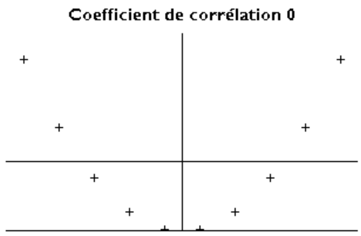


Figure 4. Soit X~U([-1 ;1]) et Y=X² X et Y ne sont pas indépendants, mais leur corrélation vaut 0

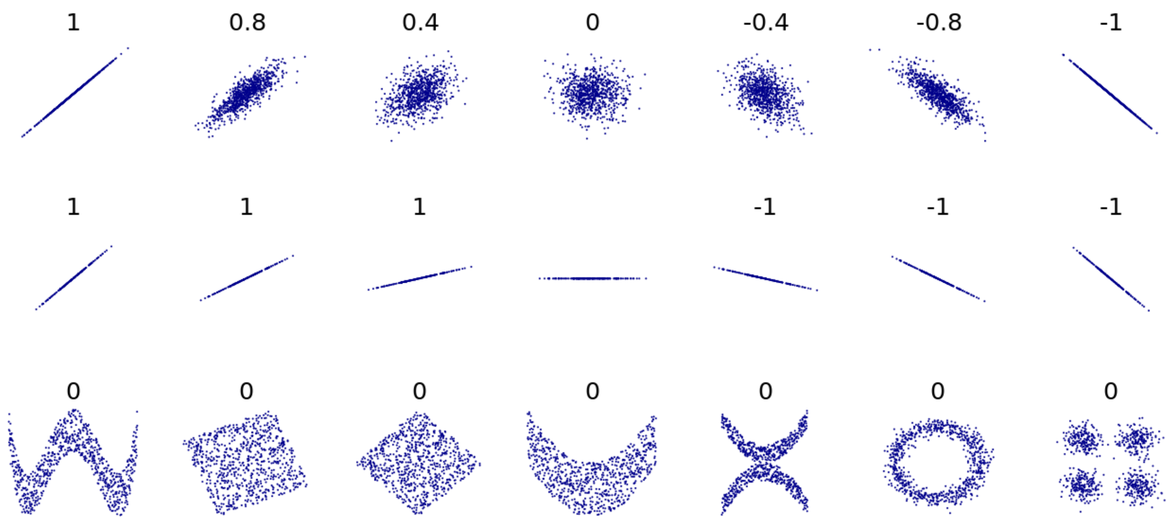


Figure 5. Exemples de coefficients de corrélation : corrélation linéaire pour les deux premières lignes, non linéaire pour la troisième ligne.

2. SERIES CHRONOLOGIQUES

2.1. Introduction

Une série chronologique est constituée de l'ensemble des observations d'une grandeur effectuées à intervalles réguliers au cours du temps. Les exemples dans le monde économique et social sont donc nombreux : inflation, cours boursiers, chômage, productions, exportations, natalité, immigration, scolarisation, logement, ... Ces grandeurs sont souvent mesurées à l'aide d'indices ou de taux publiés dans des revues spécialisées. Les exemples ne manquent pas non plus à l'intérieur même de l'entreprise: chiffre d'affaires, stocks, ventes, prix, clientèle, ... La plupart des disciplines scientifiques sont amenées à traiter des données temporelles : météorologie, médecine, physique, ...

La spécificité de l'analyse d'une série chronologique, qui la distingue d'autres analyses statistiques, est précisément dans l'importance accordée à l'ordre dans lequel sont effectuées les observations. En séries chronologiques la dépendance temporelle entre les variables constitue la source principale d'information

2.2. Définition et exemples

Définition : Une **série chronologique** (ou série temporelle, ou encore chronique) est une série statistique double (bidimensionnelle) dans laquelle la première composante du couple est le temps. La deuxième composante est alors la succession des valeurs (quantitatives ou qualitatives ordinales) que prend au cours du temps une grandeur) de définition constante.

Soient $t_1, t_2, \dots, t_n$ les instants d'observation. On note y_{t_i} l'observation de la série chronologique à l'instant t_i.

$$t_1 \to y_{t_1}$$
$$t_2 \to y_{t_2}$$
$$\cdots$$
$$\cdots$$
$$t_n \to y_{t_n}$$

Remarques :

1. la principale caractéristique d'une série chronologique (celle qui la distingue d'un simple échantillon aléatoire) est que ses observations sont fonction du temps.

2. Par convention, les valeurs du temps sont rangées dans l'ordre chronologique, et les points du nuage statistique correspondant sont reliés entre eux par des segments de droite.

But de l'étude des séries:

- Comprendre le passé : analyser et expliquer les valeurs observées décrire ;
- Prédire le futur : bâtir des prévisions pour les valeurs non encore observées ; ´
- Etudier le lien avec d'autres séries chronologiques.

Méthodologie :

1. Représentation graphique des données.

2. Choisir un modèle de composition (additif ou multiplicatif)

3. Lissage de la courbe par la méthode des moyennes mobiles afin d'isoler la tendance et estimation de cette dernière. Estimer la tendance Ct

3. Isolement et estimation des variations saisonnières.

4. Prédiction à court terme via (au choix)

- La reconstitution de la série avec ses trois composantes modélisées ;
- une méthode de prédiction par lissages exponentiels.

Exemple 1 :

Trafic aérien international, que nous dénommerons désormais « Airlines »

(Source: http://go.to/forecasting/).

Tableau 1. Trafic aérien international de janvier 1949 à décembre 1960 (milliers)

	1949	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960
Janvier	112	115	145	171	196	204	242	284	315	340	360	417
Février	118	126	150	180	196	188	233	277	301	318	342	391
Mars	132	141	178	193	236	235	267	317	356	362	406	419
Avril	129	135	163	181	235	227	269	313	348	348	396	461
Mai	121	125	172	183	229	234	270	318	355	363	420	472
Juin	135	149	178	218	243	264	315	374	422	435	472	535
Juillet	148	170	199	230	264	302	364	413	465	491	548	622
Août	148	170	199	242	272	293	347	405	467	505	559	606
Septembre	136	158	184	209	237	259	312	355	404	404	463	508
Octobre	119	133	162	191	211	229	274	306	347	359	407	461
Novembre	104	114	146	172	180	203	237	271	305	310	362	390
Décembre	118	140	166	194	201	229	278	306	336	337	405	432

Exemple 2 :

Population totale depuis 1920 en milliers

(Source : INSEE)

Tableau 2. Population totale de la France depuis 1920 en milliers

année	population	année	population	année	population	année	population	année	population
1920	39 000	1935	41 940	1950	41 829	1965	48 758	1980	53 880
1921	39 240	1936	41 910	1951	42 156	1966	49 164	1981	54 182
1922	39 420	1937	41 930	1952	42 460	1967	49 548	1982	54 492
1923	39 880	1938	41 960	1953	42 752	1968	49 915	1983	54 772
1924	40 310	1939	40 000	1954	43 057	1969	50 318	1984	55 026
1925	40 610	1940	39 000	1955	43 428	1970	50 772	1985	55 284
1926	40 870	1941	37 800	1956	43 843	1971	51 251	1986	55 547
1927	40 940	1942	37 700	1957	44 311	1972	51 701	1987	55 824
1928	41 050	1943	37 000	1958	44 789	1973	52 118	1988	56 118
1929	41 230	1944	36 500	1959	45 240	1974	52 460	1989	56 423
1930	41 610	1945	38 000	1960	45 684	1975	52 699	1990	56 735
1931	41 860	1946	40 287	1961	46 163	1976	52 909	1991	57 055
1932	41 860	1947	40 679	1962	46 998	1977	53 146	1992	57 372
1933	41 890	1948	41 112	1963	47 816	1978	53 376		
1934	41 950	1949	41 480	1964	48 310	1979	53 606		

Exemple de domaine d'application de séries statistiques :

- Economie
  I.N.S.E.E., Chambre de Commerce et d'Industrie, Bourse, ...
- Météorologie
  pluie, débit des cours d'eau, température, ensoleillement, ...
- Biologie végétale et animale :
  plantes, cellules, animaux : comportement, évolution, ...
- Phénomène aléatoire dans le temps :
  trafic téléphonique, trafic routier, utilisation d'un guichet bancaire, …

2.3. Corrections des données

Avant de se lancer dans l'étude d'une série chronologique, il est souvent nécessaire de traiter, modifier les données brutes.

2.3.1. Transformation de toute la série:

Changements de repère, d'échelle temporelle, standardisation, transformation de la variable observée

Transformation Logarithme

Notons y_t=log(x_t). Alors pour de petites variations de $y_{t+1} - y_t \cong \frac{x_{t+1}}{x_t}$

Autrement dit, l'accroissement sur la courbe transformée est approximativement le pourcentage d'accroissement sur la courbe initiale.

La transformation logarithmique fait partie de la famille des transformations de Box-Cox. On peut citer également la transformation logistique adaptée aux séries qui varient dans un intervalle constant de temps.

Exemple : Trafic aérien international : effet de la transformation logarithme

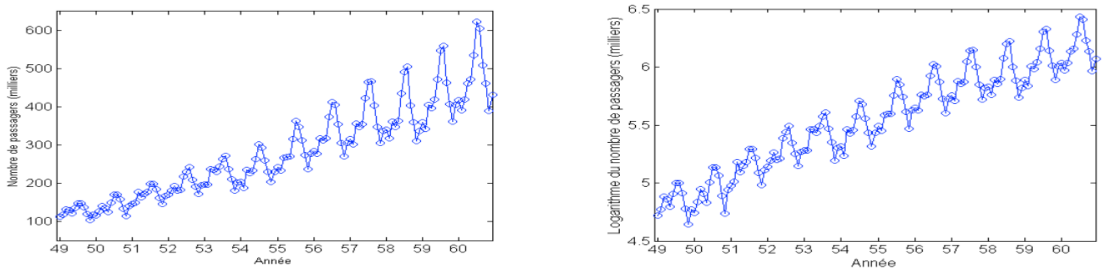


Figure 6. Trafic aérien international : effet de la transformation logarithme

On constate que la transformation a bien l'effet escompté de rendre la variance à peu près constante.

2.3.2. Transformation ponctuelle

Traitement des données manquantes (interpolation)

Traitement de données accidentelles et aberrantes (filtre)

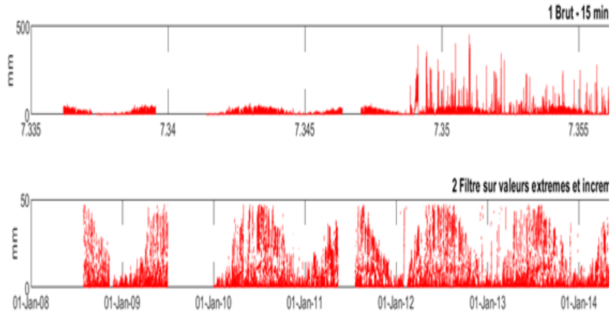


Figure 7. Application d'un filtre éliminant les valeurs aberrantes pour des données d'évaporation échantillonnées à 15 minutes

2.4. Décomposition d'une chronique

L'examen de séries graphiques longues laisse apercevoir un certain nombre de composantes fondamentales d'un mouvement d'ensemble.

On considère de manière classique qu'une série chronologique est la "résultante" des trois composantes décrites ci-dessous.

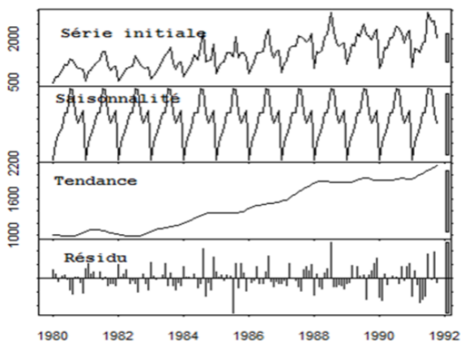


Figure 8. Vente de vins australiens

2.4.1. Tendance à long terme

La **tendance à long terme** m_t ou **tendance générale** ou **tendance séculaire** ou "**trend**" correspondant à une tendance persistante de variation dans un sens déterminé qui se maintient pendant un long intervalle de temps. Elle traduit le comportement "moyen" de la série.

Exemple : hausse des prix depuis l'euro, hausse des températures au cours de l'anthropocène

2.4.2. Mouvement périodique

Le **mouvement périodique** s_t (**mouvement saisonnier** ou **variation saisonnière**) correspondant à un phénomène qui se répète à intervalles de temps réguliers

Exemple : chute de production pendant les vacances d'été, boom des ventes des grands magasins en décembre) correspond à un phénomène qui se répète à intervalles.

Propriété : La variation saisonnière est stable : si *p* est la période, $s_{t+p} = s_t$, ∀*t*,

Remarques :

1. La première partie de la propriété traduit la répétition d'un même phénomène dans le temps et par période p; $s_t = s_{t+p} = s_{t+2p} =$ ….

Ainsi un facteur saisonnier est donc totalement déterminé par p coefficients saisonniers : $s_1, s_2, \dots, s_p$

En pratique, pour le mouvement saisonnier de période p, on fera l'hypothèse d'une compensation exacte sur une période entre les variations saisonnières positives et les variations saisonnières négatives. Cette contrainte est apportée pour donner une valeur unique au phénomène saisonnier, la constante pouvant toujours être incluse dans le mouvement tendanciel. On peut se ramener à une tendance de moyenne nulle en effectuant une translation $s_t \longrightarrow s_t^* - \bar{s}$ avec $\bar{s} = \frac{1}{p}\sum_{i=1}^{i=p} s_i$

2.4.3. Composante résiduelle

La **composante accidentelle** ou **composante résiduelle** e_t de par sa nature qui souvent n'exprime plus que du "bruit". Ces fluctuations sont supposées de faible amplitude et de moyenne nulle sur un petit nombre d'observations consécutives. Cette composante peut parfois devenir importante. La composante résiduelle (ou bruit ou résidu) correspond à des fluctuations irrégulières, en général de faible intensité mais de nature aléatoire. On parle aussi d'aléas.

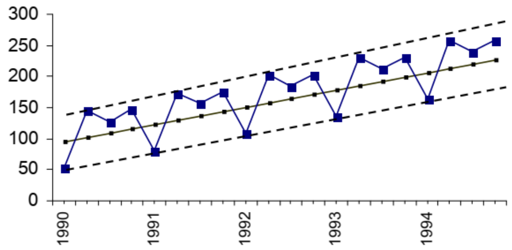
Exemple : grèves, conditions météorologiques exceptionnelles, crash financier

2.5. Les modèles de composition de ces 3 composantes.

2.5.1. Le modèle additif

Dans un modèle additif, on suppose que les 3 composantes : tendance, variations saisonnières et variations accidentelles sont indépendantes les unes des autres.

La série x_t s'écrit alors comme la somme de ces 3 composantes : $x_t = m_t + s_t + e_t$



Graphiquement, l'amplitude des variations est constante autour de la tendance et Les 2 droites tracées sont à peu près parallèles entre elles

2.5.2. Le modèle multiplicatif

La série x_t s'écrit alors de la manière suivante : $x_t = m_t \times s_t \times e_t$ qu'on ramène à $x_t = m_t \times (1 + s_t) \times (1 + e_t)$

Graphiquement, l'amplitude des variations n'est pas constante autour de la tendance et Les 2 droites tracées ne sont pas parallèles entre elles

Remarque :

Dans le cas d'une série (x_t) à valeurs positives, le modèle multiplicatif se ramène à un modèle additif en considérant la série (z_t) :=(ln(x_t)) où ln(x_t)=ln(m_t)+ln(s_t)+ln(e_t).

Remarque : il existe d'autres décomposition comme par exemple le modèle mixte : $x_t = m_t \times s_t + e_t$

2.5.3. Choix du modèle

Méthode de la bande : On utilise le graphe de la série et la droite passant par les minima et celle passant par les maxima.

- Si ces 2 droites sont à peu près parallèles : le modèle est additif.
- Si ces 2 droites ne sont pas parallèles : le modèle est multiplicatif.

Méthode du profil : On utilise le graphique des courbes superposées

- Si les différentes courbes sont à peu près parallèles : le modèle est additif.
- Sinon (les pics et les creux s'accentuent) : le modèle est multiplicatif.

Tableau 3. Consommations trimestrielles en électricité d'une entreprise de vente par Internet pendant les trois premières années

Trimestres \ Années	1997	1998	1999
1	4,5	5,5	7,2
2	4,1	4,9	6,4
3	3,7	4,4	4,8
4	5,1	6,5	6,8

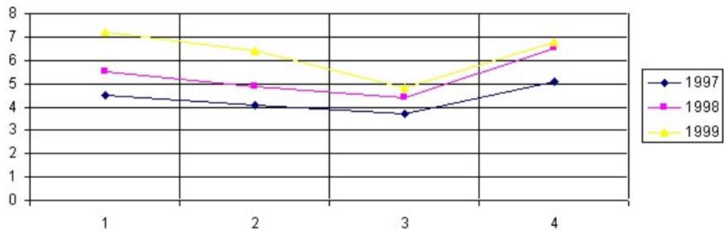


Figure 9. Graphes superposés correspondant à la série issue du tableau ci-contre

Méthode du tableau de Buys et Ballot :

1.On calcule, pour chacune des années, la moyenne et l'écart type.

2. On trace les points d'abscisse la moyenne et d'ordonnée l'écart type de la même année.

3. On trace la droite des moindres carrés de ces points.

Si l'écart type est indépendant de la moyenne (c.à.d. la pente (a) de la droite des moindres carrés est très proche de 0), le modèle est additif.

Si l'écart type est fonction de la moyenne (c.à.d. La pente (a) de la droite des moindres carrés n'est pas nulle), le modèle est multiplicatif.

3. ANALYSE DE LA TENDANCE

Leur objet est de créer, à partir de la série initiale x_t, une autre série m_t ne contenant que la tendance. Quand la tendance est enlevée, ne restent plus que les composantes saisonnière et accidentelle. On verra des techniques d'estimation de la composante saisonnière, puis comment on analyse la composante accidentelle.

Tendance paramétrique

On cherche à savoir si, en faisant abstraction des variations saisonnières, la série statistique a une fonction analytique déterminée par certains paramètres. La méthode des moindres carrés (méthode qui minimise les écarts quadratiques entre modèle et observations permet de fixer paramètres. Citons quelques tendances : tendance linéaire: $m_t = \mathrm{a} + \mathrm{b} \times \mathrm{t}$ (c.à.d. la série suit grossièrement une droite); quadratique: m_t= a+bt+ct^2 ; logarithmique: $m_t = \mathrm{a} + \mathrm{b} \times \ln(\mathrm{t})$; exponentiel $m_t = \lambda e^{a+bt}$; puissance $m_t = \lambda t^a$; logistique ; Gompertz ; …

Problème : Ces méthodes analytiques sont simples, mais reposent sur l'hypothèse d'une tendance évoluant selon une fonction analytique déterminée, hypothèse qu'on ne peut pas fréquemment faire, même à la suite d'une transformation de variable

Tendance non-paramétrique

$$X \longrightarrow \boxed{\textit{filtre}} \longrightarrow Y$$

En l'absence de référence à un modèle précis pour la tendance, on préférera utiliser une méthode non-paramétrique qui filtre la tendance en éliminant le facteur saisonnier tout en réduisant les irrégularités. Dans la suite, un filtre est une sorte de « boîte noire » régularisant une chronique X en la transformant en une chronique Y qui est une approximation de la composante tendancielle de la chronique X.

Nous étudierons deux des principaux filtres linéaires qui sont la moyenne mobile et le lissage exponentiel simple. Un filtre linéaire est une application linéaire de l'ensemble des chroniques dans lui-même transformant la chronique X en une nouvelle chronique Y de la façon suivante :

$$y_t = \sum_{k \epsilon K} \alpha_k x_{t+k} \text{ avec } K \subset \mathbb{Z} \text{ et } \sum_{k \epsilon K} \alpha_k = 1$$

Le choix du filtre linéaire approprié à certains objectifs se fait par l'intermédiaire du choix de ses coefficients α_k

3.1. Moyennes mobiles simples et centrées

Soit $(x_t)_{1 \le t \le N}$ une série chronologique de taille N, on suppose que cette série est découpée en périodes avec p observations par période, c.-à-d. que cette série est p-périodique ou possède une p-saison (c.à.d. $x_{t+p=}x_t$).

3.1.1. Moyennes mobiles simples et centrées

Définition : On appelle série des **moyennes mobiles centrée d'ordre p** la série des valeurs $y_{p,t}$ calculées de la façon suivante:

Si p est impair, alors p=2k+1 et $y_{p,t} = \sum_{i=-k}^{i=k} \frac{1}{p} x_{t+i} = \frac{x_{t-k} + x_{t-k+1} + \ldots + x_t + \ldots + x_{t+k-1} + x_{t+k}}{p}$

(p termes au numérateur avec x_t au milieu)

Si p est pair, alors p=2k et

$$y_{p,t} = \frac{1}{p}\left(\frac{x_{t-k}}{2} + \sum_{i=-k+1}^{i=k-1} x_{t+i} + \frac{x_{t+k}}{2}\right) = \frac{\frac{x_{t-k}}{2} + x_{t-k+1} + \ldots + x_t + \ldots + x_{t+k-1} + \frac{x_{t+k}}{2}}{p}$$

(p+1 termes au numérateur avec x_t au milieu).

Exemple :

$$y_{4,3} = \frac{\frac{x_1}{2} + x_2 + x_3 + x_4 + \frac{x_5}{2}}{4}$$

$y_{3,5}$ =……………………………………………………

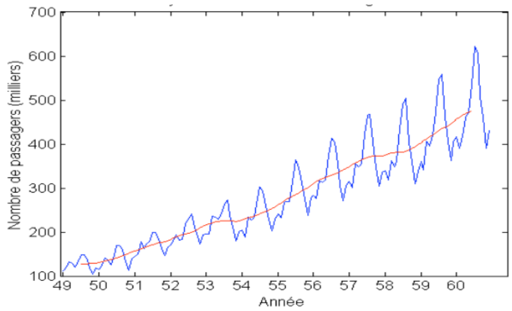


Figure 10. Moyenne mobile centrée d'ordre 12, données de trafic aérien international

Remarques :

1. La série lissée est plus courte que l'originale puisque k (où k tel que p=2k ou p=2k+1 selon la parité de p) valeurs sont manquantes à chaque extrémité de la période d'observation. La série des moyennes mobiles n'est définie que pour t tel que k+1 ≤ t ≤ N – k. Il y a (N – p + 1) moyennes mobiles centrées de longueur impaire p et (T – p) moyennes mobiles centrées de longueur paire p.

Comme les lissages utilisés suppriment les valeurs du bord. Ils ne sont donc pas à recommander pour la prévision.

2.A p fixé, alors t seulement varie dans la série $z_{p,t}$.

3. Lorsqu'on découpe l'année en 4 trimestres, p est pair.

Les propriétés de la moyenne mobile d'ordre p fixé entraînent que son application

1. Conserve une tendance linéaire : si x_t a une tendance linéaire, $y_{p,t}$ aura la même tendance ;
2. Supprime toute saisonnalité de période p : si x_t a une composante saisonnière, $y_{p,t}$ n'en aura pas;
3. Atténue de façon optimale le bruit : le bruit de $y_{p,t}$ sera plus faible que celui de x_t.

Remarque :

3. on dit la moyenne mobile « lisse » la chronique, en ce sens que la série Y est moins dispersée que la série initiale X. Mais les nouvelles irrégularités qui sont corrélées entre elles, peuvent faire apparaître des oscillations parasites qui ne figuraient pas dans la série initiale (effet de Slutsky-Yule)

Remarque : Choix pratique de l'ordre d'une moyenne mobile

Le but d'un lissage par moyenne mobile est de faire apparaître l'allure de la tendance, en gommant les variations saisonnières. On fait disparaître la composante saisonnière de période p avec une moyenne mobile d'ordre p. En pratique on doit trouver le meilleur compromis pour le choix de l'ordre p de lissage optimal

Exemple : On considère la série chronologique représentant les ventes d'huitres en tonnes par trimestres

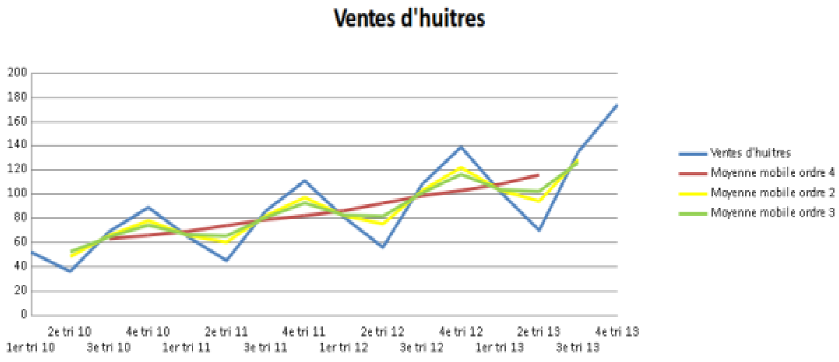


Figure 11. Influence de l'ordre d'une moyenne mobile

Tableau 4. Ventes d'huitres en tonnes par trimestres

Ventes d'huitres	
Période	Tonnes
1er tri 10	52
2e tri 10	36
3e tri 10	69
4e tri 10	89
1er tri 11	65
2e tri 11	45
3e tri 11	86
4e tri 11	111
1er tri 12	81
2e tri 12	56
3e tri 12	108
4e tri 12	139
1er tri 13	102
2e tri 13	70
3e tri 13	135
4e tri 13	174

Si la série des moyennes mobiles fait clairement apparaître une tendance linéaire croissante, alors nous pouvons calculer la droite de régression à partir de la série des moyennes mobiles.

3.1.2. Alternatives à la moyenne mobile simple

Inconvénients de la méthode des moyennes mobiles

Un changement de niveau ou de pente de la tendance à une date t entraîne une mauvaise approximation de cette composante pendant toute une période précédant et suivant cette date (figure ci-contre). C'est la raison pour laquelle on fait l'hypothèse d'une tendance monotone à faible courbure.

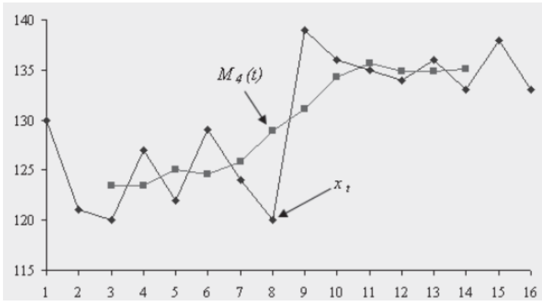


Figure 12. Limite de la méthode des moyennes mobiles

Extension du concept de moyenne mobile

Les moyennes mobiles simples sont des moyennes équi-pondérées.

Les moyennes mobiles centrées d'ordre pair accordent 2 fois moins de poids aux 2 valeurs extrêmes.

De façon générale, on peut définir des moyennes mobiles pondérées par des poids p_i pour accorder plus d'importance à certaines observations, souvent aux observations centrales. La somme des poids doit être égale à 1. Ainsi on parle par exemple de moyenne mobile gaussienne.

Médiane mobile simple

Cette méthode est moins sensible aux valeurs aberrantes. La technique est identique à la méthode des moyennes mobiles, on remplace simplement la moyenne par la médiane. Cette médiane est affectée au centre de l'intervalle sur lequel elle est calculée.

On remarque que la médiane mobile centrée d'ordre 2 est égale à la moyenne mobile centrée d'ordre 2.

3.2. La droite de tendance (ou trend)

Soit $(x_t)_{1\leq t\leq N}$ une série chronologique de taille N, et $(m_{p,t})_{t \in T_p}$ la série des moyennes mobiles (définie sur un sous ensemble T_p de {1, …, N}, lorsque la moyenne mobile est calculable)

Lorsque la série des moyennes mobiles fait apparaître une tendance linéaire, nous pouvons calculer la droite $D_{m_{p,t}/t}$ de régression à partir de la série des moyennes mobiles.

D'après le 1^er^ chapitre, cette droite $D_{m_{p,t}/t}$ a pour équation $y = \beta_0 + \beta_1 x$ avec

$$\beta_1 = \frac{Cov(m_{p,t},t)}{Var(t)} \qquad \text{et} \qquad \beta_0 = \overline{m_{p,t}} - \beta_1 \bar{t}$$

où $Cov(m_{p,t}, t)$ est la covariance de $(m_{p,t})_{t \in T_p}$ et de $(t)_{t \in T_p}$, $\bar{t}$ et $\overline{m_{p,t}}$ les moyennes respectives de $(t)_{t \in T_p}$ et de $(m_{p,t})_{t \in T_p}$

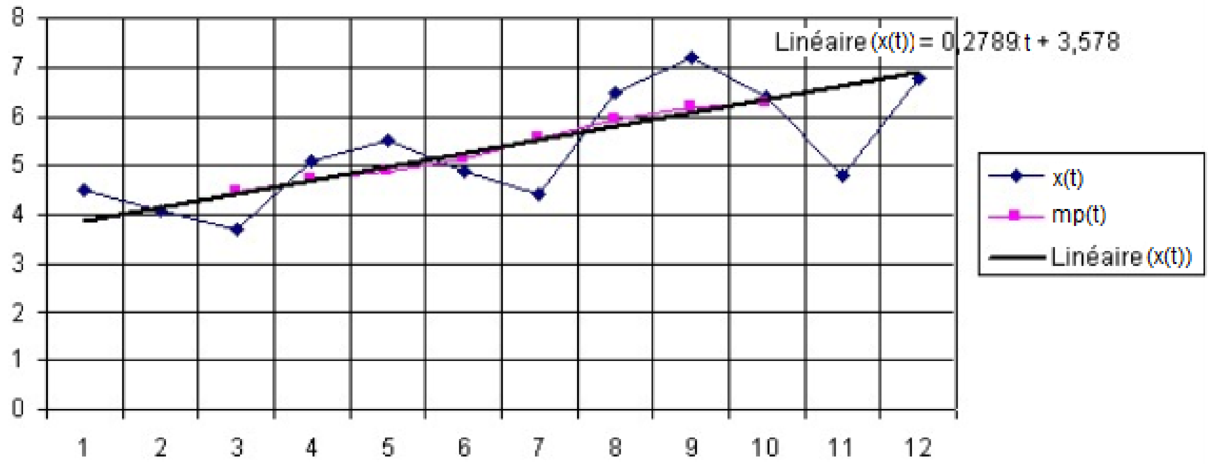


Figure 13. Série chronologique, lissage par moyenne mobile et droite de tendance

<u>Remarque :</u>

1. Une autre méthode plus grossière aurait été de calculer $D_{x_t/t}$. On montre que $|\rho_{x_t,t}| \leq |\rho_{m_{p,t},t}|$ (coefficients de corrélation). Les variations saisonnières affectent le coefficient de correlation $\rho_{x_t,t}$

Il est clair que le modèle linéaire n'est pas satisfaisant pour décrire l'évolution de $(x_t)_{1\leq t\leq N}$ puisque beaucoup de valeurs s'écartent de la droite mais on va s'en servir comme première approximation pour construire un modèle plus précis tenant compte des variations saisonnières.

4. ETUDE DES COMPOSANTES SAISONNIERES

Si on étudie une chronique avec variations saisonnières, l'évaluation de la tendance à chaque date t par la moyenne mobile centrée de longueur adéquate, conduit pour chaque coefficient saisonnier à plusieurs valeurs qu'il faut résumer. Pour bien comprendre toutes les opérations successives pour la détermination des coefficients saisonniers et de la série corrigée des variations saisonnières, on se reportera à l'exemple traité en fin de chapitre

Reprenons l'exemple du trafic aérien mensuel entre 1949 et 1960. Le modèle choisi est multiplicatif $x_t = m_t \times s_t \times \varepsilon_t$

S'agissant de filtrer la saisonnalité, il est recommandé de choisir la taille de la fenêtre égale à la périodicité (filtre d'ordre 12)

On constate que l'augmentation régulière du trafic est entrecoupée – semble-t-il-de deux paliers, un correspondant grosso modo à l'été 1953, l'autre à l'hiver 1958. Il serait intéressant à ce stade d'avoir des informations supplémentaires pouvant (peut-être) expliquer ces paliers.

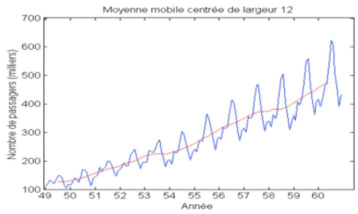


Figure 14. Moyenne mobile centrée d'ordre 12, données de trafic aérien international

La série résiduelle $\frac{x_t}{m_t} = s_t \times \varepsilon_t$

Si le lissage utilisé a bien filtré toute la saisonnalité, alors dans la série x_t/m_t il ne reste que la saisonnalité "bruitée" par un terme aléatoire.

Malgré ce bruit, certaines caractéristiques semblent apparaître :

- D'août à novembre, la décroissance du trafic semble régulière ; en revanche, la croissance du trafic de novembre à juillet-août semble être perturbée par deux légers creux en février et en avril;
- concernant les séries saisonnières (c.à.d. les séries du type s_{t+12} où t décrit un mois donné) :
  celles des mois de juillet-août semblent indiquer une croissance de la fluctuation estivale (bien entendu : indépendamment de la tendance haussière)
  tandis que celle de mars (2ème "pic") semble indiquer une baisse du trafic pour ce mois-ci au cours du temps.

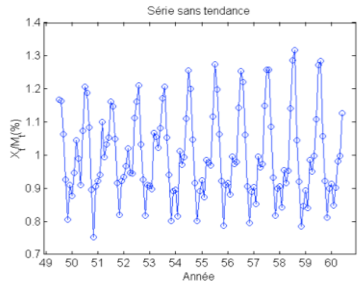


Figure 15. série sans tendance x_t/m_t, données de trafic aérien international

4.1. **Correction des variations saisonnière pour le modèle additif**

<u>Modèle considéré :</u> $x_t = m_t + s_t + \varepsilon_t$.

Soient $(x_t)_{1\le t\le N}$ une série chronologique de taille T de période p et n tel que $p \times n$ =T

On approxime m_t par $\widehat{m}_t \coloneqq y_{p,t}$ (La tendance m_t est estimée par ajustement ou lissage (moyenne ou médiane mobile)). On va maintenant estimer les variations saisonnières s_t pour un modèle additif.

Les coefficients saisonniers étant périodiques de période p, on dispose pour chacun des p coefficients saisonniers de (n–1) valeurs qui sont (n–1) différences $\{x_t - y_{p,t}\}$.

On résume ces (n–1) valeurs par

- leur moyenne arithmétique ;
- ou leur médiane ;

- ou leur moyenne arithmétique après élimination de la valeur la plus faible et de la valeur la plus élevée.

On stocke ces valeurs dans une série ($\hat{s}_t$) périodique de période p.

Si la somme des coefficients saisonniers n'est pas nulle sur une période, on corrige les coefficients saisonniers obtenus de façon à avoir une somme nulle : $\hat{s}_t \longrightarrow s_t^* - \bar{\hat{s}}\ avec\ \bar{\hat{s}} = \frac{1}{p}\sum_{i=1}^{i=p} \hat{s}_i$

Définition : La **série désaisonnalisée** ou série **corrigée des variations saisonnières**, notée CVS, la série chronologique x_t^* : =$x_t - s_t^*$ à laquelle on a enlevé les variations saisonnières.

Pour toutes les dates pour lesquelles on dispose de la valeur de la moyenne mobile, et donc d'une évaluation de la tendance, on peut calculer l'écart entre le modèle et l'observation :

$$e_t = x_t - y_t - s_t^* = x_t^* - y_t$$

Si le modèle est adapté, les valeurs absolues des écarts ne doivent pas être élevées, et leur somme voisine de zéro.

Remarque :

La particularité de la série CVS est que les données de x_t^* sont directement comparables : on a enlevé l'effet des saisons (donc le caractère propre de chaque mois on peut donc par exemple comparer les données d'un mois de janvier et celle d'un mois de juillet)

4.2. **Correction des variations saisonnière pour le modèle multiplicatif**

Modèle considéré : $x_t = m_t \times s_t \times \varepsilon_t$.

Soient $(x_t)_{1 \le t \le N}$ une série chronologique de taille T de période p et n tel que $p \times n$ =T

On approxime m_t par $\widehat{m}_t \coloneqq y_{p,t}$ (La tendance m_t est estimée par ajustement ou lissage (moyenne ou médiane mobile)). On va maintenant estimer les variations saisonnières s_t pour un modèle multiplicatif.

Les coefficients saisonniers étant périodiques de période p, on dispose pour chacun des p coefficients saisonniers de (n – 1) valeurs qui sont (n – 1) quotients $\{\frac{x_t}{y_{p,t}}\}$. On résume ces (n–1) valeurs par

- leur moyenne arithmétique ;
- ou leur médiane ;
- ou leur moyenne arithmétique après élimination de la valeur la plus faible et de la valeur la plus élevée.

On stocke ces valeurs dans une série ($\hat{s}_t$) périodique de période p

Si la somme des (1 + $\hat{s}_t$) n'est pas égale à p sur une période, on fait une correction proportionnelle :

$$1 + \hat{s}_t \longrightarrow 1 + s_t^* = \frac{1 + \hat{s}_t}{1 + \bar{\hat{s}}} avec\ \bar{\hat{s}} = \frac{1}{p}\sum_{i=1}^{i=p} \hat{s}_i$$

Définition : La **série désaisonnalisée** ou série **corrigée des variations saisonnières**, notée CVS, la série chronologique x_t^* : =$\frac{x_t}{1+s_t^*}$ à laquelle on a enlevé les variations saisonnières.

Le modèle multiplicatif prédit ainsi des valeurs et il est alors naturel, pour toutes les dates auxquelles on dispose de la valeur de la moyenne mobile, et donc d'une évaluation de la tendance, de considérer les résidus et sous la forme

$$e_t = \frac{x_t}{y_{p,t}(1+s_t^*)}-1=\frac{x_t^*}{y_{p,t}} - 1$$

Enfin, les écarts entre le modèle et les observations sont égaux à :

$$x_t - y_{p,t}(1 + s_t^*) = y_{p,t}(1 + s_t^*)e_t$$

4.3. **Exemple de décomposition d'une série chronologique**

Pour déterminer la tendance et les coefficients saisonniers d'une chronique, on peut actuellement utiliser un logiciel (R, SAS, SPSS, …) ou un tableur. Néanmoins, une bonne compréhension des méthodes demande de les avoir appliquées.

Tableau 5. Ventes en France d'essence aviation (en milliers de tonnes, source : Comité Professionnel du Pétrole)

Ici nous détaillons les étapes successives du traitement de la chronique des ventes trimestrielles en France d'essences aviation

Trimestre / Année	1er trimestre	2e trimestre	3e trimestre	4e trimestre	Moyenne annuelle
2005	3,6	7,0	7,6	3,7	5,5
2006	3,6	6,7	7,4	3,9	5,4
2007	3,7	6,4	7,1	4,1	5,3
2008	3,6	5,7	7,1	3,7	5
Moyenne trimestrielle	3,7	6,5	7,6	3,9	5,3

Nous pouvons commencer par représenter la série.

Une saisonnalité de période 4 (nombre de trimestres dans l'année) apparaît sur la représentation graphique (cf. figure ci-contre), ce qui explique que la suite des moyennes mobiles de longueur 4 filtre la tendance.

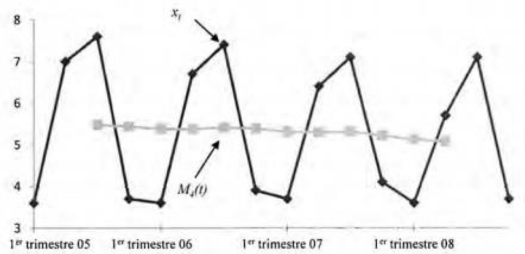


Figure 16. Chronique du tableau précédent et suite des moyennes mobiles de longueur 4.

Pour une décomposition de cette chronique, nous allons envisager successivement le modèle additif et le modèle multiplicatif.

4.3.1. Schéma additif

Nous indiquons ici les étapes nécessaires pour obtenir la série CVS et la série des résidus. Dans cet exemple, la synthèse des coefficients saisonniers (obtenus via un tableur) a été réalisée par la moyenne.

Tableau 6. Décomposition de la chronique du tableau précèdent avec le schéma additif

Colonne C : moyennes mobiles de longueur 4 évaluant la tendance

colonne D : différence entre valeurs observées et tendance=$x_t - \widehat{m}_t (= s_t + \varepsilon_t)$

colonne E : Calcul des $\hat{s}_t$

E4=(D4+D8+D12)/3 E5=(D5+D9+D13)/3

E6=(D6+D10+D14)/3 E7=(D7+D11+D15) / 3

⇒ 1ères valeurs des 4 coefficients saisonniers qu'on reporte sur toute la colonne

	A	B	C	D	E	F	G	H
1	t	x_t	$M_4(t)$	x_t-$M_4(t)$	s_t	s^*_t	CVS	Ecart
2	1	3,6			-1,64	-1,61	5,21	
3	2	7			1,02	1,04	5,96	
4	3	7,6	5,48	2,13	1,97	1,99	5,61	0,13
5	4	3,7	5,44	-1,74	-1,45	-1,42	5,12	-0,32
6	5	3,6	5,38	-1,78	-1,64	-1,61	5,21	-0,16
7	6	6,7	5,38	1,33	1,02	1,04	5,66	0,28
8	7	7,4	5,41	1,99	1,97	1,99	5,41	0,00
9	8	3,9	5,39	-1,49	-1,45	-1,42	5,32	-0,07
10	9	3,7	5,31	-1,61	-1,64	-1,61	5,31	0,00
11	10	6,4	5,30	1,10	1,02	1,04	5,36	0,06
12	11	7,1	5,31	1,79	1,97	1,99	5,11	-0,20
13	12	4,1	5,21	-1,11	-1,45	-1,42	5,52	0,31
14	13	3,6	5,13	-1,53	-1,64	-1,61	5,21	0,09
15	14	5,7	5,08	0,62	1,02	1,04	4,66	-0,42
16	15	7,1			1,97	1,99	5,11	
17	16	3,7			-1,45	-1,42	5,12	

colonne F : calcul des coefficients saisonniers « centrés » : $s_t^* - \bar{\hat{s}}$

colonne G : calcul de la série CVS ($x_t - s_t^*$) G2=B2 – F2

colonne H : calcul de la série des écarts ($e_t = x_t^* - y_t$) H4=G4 – C4

4.3.2. Schéma multiplicatif

Nous indiquons ici les étapes nécessaires pour obtenir la série CVS et la série des résidus. Dans cet exemple, la synthèse des coefficients saisonniers (obtenus via un tableur) a été réalisée par la moyenne.

colonne C : moyennes mobiles de longueur 4 évaluant la tendance

colonne D : quotient entre valeurs observées et tendance$\frac{x_t}{y_{p,t}}$

colonne E : Calcul des 1+$\hat{s}_t$

E4 = (D4 + D8 + D12)/3 E5 = (D5 + D9 + D13)/3

E6 = (D6 + D10 + D14)/3 E7 = (D7 + D11 + D15)/3

⇒ 1[ères] valeurs des 4 coefficients $(1 + \hat{s}_t)$ qu'on reporte sur toute la colonne

Tableau 7. Décomposition de la chronique du tableau précèdent avec le schéma multiplicatif

	A	B	C	D	E	F	G	H	I	J
1	t	x_t	$M_4(t)$	$x_t/M_4(t)$	$1+s_t$	$1+s^*_t$	CVS	$1+e_t$	e_t	Ecart
2	1	3,6			0,69	0,70	5,18			
3	2	7			1,19	1,20	5,84			
4	3	7,6	5,48	1,39	1,36	1,37	5,55	1,01	0,01	0,10
5	4	3,7	5,44	0,68	0,73	0,74	5,03	0,92	-0,08	-0,30
6	5	3,6	5,38	0,67	0,69	0,70	5,18	0,96	-0,04	-0,14
7	6	6,7	5,38	1,25	1,19	1,20	5,59	1,04	0,04	0,26
8	7	7,4	5,41	1,37	1,36	1,37	5,40	1,00	0,00	-0,01
9	8	3,9	5,39	0,72	0,73	0,74	5,30	0,98	-0,02	-0,07
10	9	3,7	5,31	0,70	0,69	0,70	5,32	1,00	0,00	0,01
11	10	6,4	5,30	1,21	1,19	1,20	5,34	1,01	0,01	0,05
12	11	7,1	5,31	1,34	1,36	1,37	5,18	0,98	-0,02	-0,18
13	12	4,1	5,21	0,79	0,73	0,74	5,57	1,07	0,07	0,26
14	13	3,6	5,13	0,70	0,69	0,70	5,18	1,01	0,01	0,04
15	14	5,7	5,08	1,12	1,19	1,20	4,76	0,94	-0,06	-0,38
16	15	7,1			1,36	1,37	5,18			
17	16	3,7			0,73	0,74	5,03			

colonne F : calcul des coefficients saisonniers « normalisés » : $1 + s_t^* = \frac{1+\hat{s}_t}{1+\bar{\hat{s}}}$,

colonne G : calcul de la série CVS x_t^* : =$\frac{x_t}{1+s_t^*}$, G2 = B2 / F2

colonne H : calcul de la série ($1 + e_t$=$\frac{x_t^*}{y_{p,t}}$), H4 = G4 / C4

colonne I : calcul de la série e_t, I4 = H4 – 1

colonne J : calcul de la série des écarts ($y_{p,t}(1 + s_t^*)e_t$), J4 = C4 · F4 · I4,

4.3.3. Comparaison des deux décompositions

La représentation des séries des écarts (colonne H du tableau pour la décomposition en modèle additif et colonne J du tableau pour la décomposition en modèle multiplicatif) permet de comparer les ajustements entre les deux modèles et les observations (figure ci-dessous). On constate que les deux séries des écarts sont presque confondues

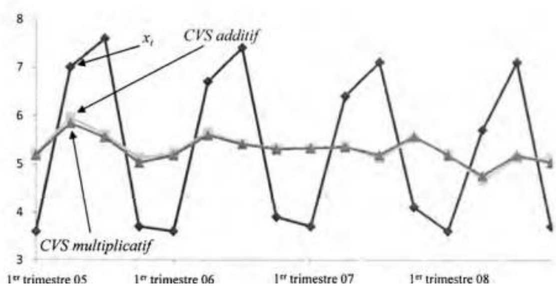


Figure 17. Séries CVS

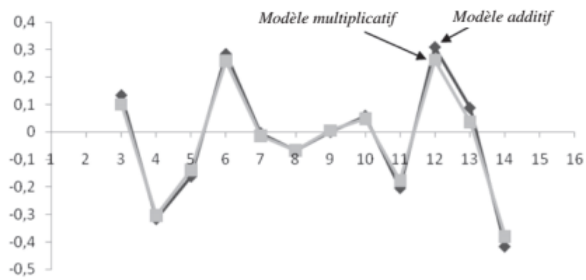


Figure 18. Écarts entre les modèles et les observations

5. PREVISION

Pour faire de la prévision, deux approches seront vues dans cet exposé :

Prévision par modélisation de ses composantes Prévision par lissage exponentiel.

5.1. Prévision par modélisation de ses trois composantes

Quand une série comporte une tendance et une périodicité, il faut d'abord enlever la saisonnalité avant d'estimer la tendance. Selon que le modèle est additif ou multiplicatif, les méthodes d'estimation ne sont pas les mêmes.

Pour enlever la saisonnalité, on utilisera des méthodes de lissage (filtrage linéaire par moyennes mobiles, médianes mobiles), qui permettent d'éliminer la composante saisonnière.

Quand on n'a plus que la tendance entachée d'erreur, on utilise des techniques de régression pour modéliser la tendance (linéaire, non linéaire, robuste, ...)

Quand la série ne comporte, à des erreurs près, que les composantes tendancielle et saisonnière, la modélisation de la tendance et l'estimation des coefficients saisonniers permet de faire des prévisions.

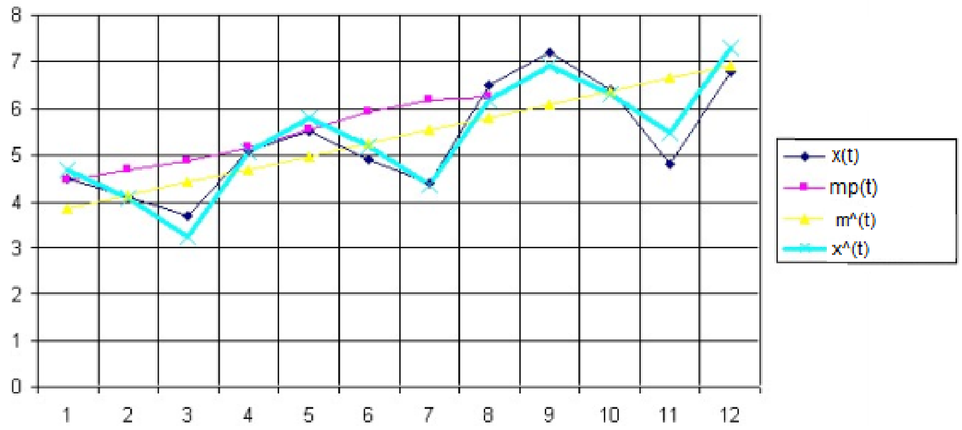
<u>**Modèle additif**</u>	<u>**Modèle multiplicatif**</u>
$\hat{x}_t = \hat{m}_t + \hat{s}_t$	$\hat{x}_t = \hat{m}_t \times (1 + \hat{s}_t)$

avec

$\hat{m}_t$ =l'estimation de la tendance (obtenue via la régression linéaire des moyennes mobiles), m_t est de la forme $\hat{m}_t = a + bt$

$\hat{s}_t$ =l'estimation de la composante saisonnière, $\hat{s}_t$ est périodique et ils sont tous connus

Voici (en trait bleu) la représentation graphique du modèle que l'on a construit pour décrire une série chronologique en tenant compte des variations saisonnières.



Pour faire de la prédiction, il faut

- calculer $\hat{m}_t$ et $\hat{s}_t$ au temps t désirés
- les combiner en utilisant le modèle de décomposition

<u>Remarque :</u> Le chapeau sur les séries permet de bien différencier les données théoriques (x_t) des données estimées($\hat{x}_t$).

5.2. Introduction au lissage exponentiel

Le lissage exponentiel est une généralisation de la notion de moyenne mobile. Il a pour unique but de prévoir, en fonction des dernières observations faites, les observations à venir, et ceci à moindre frais. En effet, contrairement aux modèles décrits précédemment, ce type de lissage permet de faire des estimations de valeurs futures en tenant compte des dernières valeurs prises, sans avoir à refaire tous les calculs d'estimation des paramètres (coefficients saisonniers et tendance).

Etant donnée une série d'observations $\{x_1, x_2,..., x_T\}$, on s'intéresse aux prévisions qu'on peut donner à la date T pour les dates futures. De façon générale, la prévision faite à une date t pour l'horizon h, c.à.d. pour la date t+h, sera notée $\hat{x}_{t+h}$ ou $\hat{x}_t(h)$.

La méthode de lissage exponentiel simple procède par filtrage de la série de données avec les particularités suivantes:

- le filtre utilisé fait intervenir tout le passé (il est donc décentré à gauche, contrairement aux moyennes mobiles centrées) ;
- les poids attribués aux observations décroissent de façon exponentielle en fonction de l'ancienneté de ces observations.

Les méthodes de lissage exponentiel, développées dans les années 60, sont des méthodes d'extrapolation qui donnent un poids prépondérant aux valeurs récentes.

Le lissage exponentiel simple ne s'applique qu'aux séries sans tendance ni saisonnalité. Toutes ces méthodes consistent à ajuster à une chronique de série temporelle une estimation locale de ce que va être sa valeur future. Selon les variantes:

• une constante pour le lissage exponentiel simple

• une droite pour le lissage exponentiel double ou de Holt

• une droite pour le lissage exponentiel de Holt-Winters pour des séries avec saisonnalités

5.3. Lissage exponentiel simple (LES ou SES, Single Exponential Smoothing)

Cette méthode de prévision s'applique à Le lissage exponentiel simple permet d'effectuer des prévisions pour des séries chronologiques dont la tendance est constante et sans saisonnalité. On suppose la grandeur observée caractérisée par des variations irrégulières autour de la moyenne :

<u>Modèle considéré :</u> $x_t = a + e_t$

Pour des chroniques sans variations saisonnières et à tendance localement constante.

<u>Idée :</u> Etant donné un réel β tel que 0 < β < 1, comme la tendance est constante, on cherche une prévision $\hat{x}_T(h)$ sous la forme de la constante qui s'ajuste le mieux au sens des moindres carrés pondérés au voisinage de T , c'est-à-dire la solution du problème de minimisation

$$\min_a \sum_{j=0}^{T-1} \beta^j \, (x_{t-j} - a)^2$$

<u>Définition :</u> Pour $\beta \epsilon [0,1]$, La prévision de la série à l'horizon h, $\hat{x}_T(h)$, fournie par la méthode de lissage exponentiel simple est donnée par $\hat{x}_T(h) = (1-\beta) \sum_{j=0}^{T-1} \beta^j x_{t-j}$ où β est la **constante de lissage**.

<u>Remarques :</u>

0. l'appellation exponentiel vient de la décroissance exponentielle des poids dans cette dernière formule.

1. Dans le cas où β est indépendant de h, on notera simplement $\hat{x}_T$au lieu de $\hat{x}_T(h)$. Les prévisions sont dans ce cas identiques pour tout h.

2. Cette méthode de prévision prend en compte tout le passé d'une série temporelle, mais en accordant de moins en moins d'importance aux observations les plus éloignées de l'instant T (puisque β^j décroît avec j).

3. Si β est proche de 0, la prévision est souple, c'est-à-dire fortement influencée par les observations les plus récentes (β^j devenant négligeable pour les grandes valeurs de j). Dans le cas extrême où β=0, la prévision est alors égale à la dernière valeur observée.

4. si β est proche de 1, l'influence des observations passées est d'autant plus importante et remonte loin dans le passé. On dit dans ce cas que **la prévision est rigide** en ce sens qu'elle est peu sensible aux fluctuations exceptionnelles (aussi appelées fluctuations conjoncturelles). Dans l'autre cas extrême où β=1, alors toutes les prévisions sont identiques (et donc à la valeur choisie pour l'initialisation).

A partir de la définition ci-dessus, on obtient directement la formule de mise à jour qui permet de calculer directement (à partir de la prévision $\hat{x}_{T-1}$ à la date T-1) une nouvelle prévision $\hat{x}_T$ lorsqu'une nouvelle observation x_T est effectuée.

Proposition (formule de mise à jour) :

$$\hat{x}_T(h) = \beta\hat{x}_{T-1}(h) + (1-\beta)x_T = \hat{x}_{T-1}(h) + +(1-\beta)(x_T - \hat{x}_{T-1}(h))$$

Remarques :

1. Cette équation permet donc de mettre à jour les prévisions à l'horizon h à partir de la dernière prévision de manière extrêmement simple. L'initialisation de la récurrence est en général faite en prenant $\hat{x}_1(h) = x_1$ ou $\hat{x}_1(h) = \bar{x}$.

Si T est assez grand, ce choix a en fait peu d'importance. Le choix de l'initialisation importe peu en réalité puisque cette initialisation est rapidement "oubliée". Cet oubli est d'autant plus rapide que la constante de lissage est proche de 1.

2. La première égalité fait apparaître $\hat{x}_T(h)$ comme le barycentre entre $\hat{x}_{T-1}(h)$, la valeur prédite à l'horizon h à partir des T−1 premières observations, et x_T la dernière observation.

3. La seconde égalité fait intervenir $(x_T - \hat{x}_{T-1}(h))$ la dernière erreur de prévision.

5.4. Le lissage exponentiel double de Holt (LED)

Holt (1957) a étendu le lissage exponentiel simple au cas du lissage exponentielle linéaire. L'idée est d'ajuster une droite au lieu d'une constante dans l'approximation locale de la série.

Modèle considéré : $x_t = b + at + e_t$

Pour des chroniques sans variations saisonnières et à tendance localement tendance localement linéaire

Idée : Etant donné un réel β tel que 0 < β < 1, comme la tendance est constante, on cherche une prévision $\hat{x}_T(h)$ sous la forme de la droite qui s'ajuste le mieux au sens des moindres carrés pondérés au voisinage de T, c'est-à-dire la solution du problème de minimisation

$$\min_a \sum_{j=0}^{T-1} \beta^j \left(x_{t-j} - (b + ah)\right)^2$$

<u>Définition :</u> Pour $\beta\epsilon[0,1]$, La prévision de la série à l'horizon h, $\hat{x}_T(h)$, fournie par la méthode de lissage exponentiel double est donnée par $\hat{x}_T(h) = \hat{b}_T + \hat{a}_T h$

où β est la **constante de lissage**.

$$\begin{cases} \hat{a}_T = \frac{1-\beta}{\beta}(S_1(T) - S_2(T)) \\ \hat{b}_T = 2S_1(T) - S_2(T) \end{cases} \text{avec} \begin{cases} S_1(T) = (1-\beta)\sum_{j=1}^{j=T-1} \beta^j x_{t-j} \\ S_2(T) = (1-\beta)\sum_{j=1}^{j=T-1} \beta^j S_1(t-j) \end{cases}$$

<u>Remarque :</u>

Les expressions de $\hat{a}_T$ et $\hat{b}_T$ dépendent de $S_1(T)$ et $S_2(T)$ qui sont respectivement le lissage exponentiel simple de la série initiale et le lissage exponentiel simple de la série lissée. On a donc effectué deux lissages consécutifs d'où le nom de lissage exponentiel double.

A partir de la définition ci-dessus, on obtient directement la formule de mise à jour et permet de calculer directement (à partir de la prévision $\hat{x}_{T-1}$ à la date T-1) une nouvelle prévision $\hat{x}_T$ lorsqu'une nouvelle observation x_T est effectuée.

<u>Proposition (formule de mise à jour) :</u>

$$\begin{cases} \hat{a}_T = \hat{a}_{T-1} + (1-\beta)^2(x_T - \hat{x}_{T-1}(1)) \\ \hat{b}_T = \hat{b}_{T-1} + \hat{a}_{T-1} + (1-\beta^2)(x_T - \hat{x}_{T-1}(1)) \end{cases}$$

<u>Remarque :</u>

1. Pour l'initialisation, on prend $\begin{cases} \hat{a}_2 = x_2 - x_1 \\ \hat{b}_2 = x_2 \end{cases}$

2. On peut généraliser cette technique de lissage pour traiter des séries sans saisonnalité présentant une tendance polynômiale de degré supérieur à 2. Les résultats font intervenir, dans ce cas, les opérateurs de lissage d'ordre p $S_p(T)$, p ∈ ℕ, itérées d'ordre p de$S_1(T)$

5.5. Le lissage exponentiel de Holt-Winter

Cette approche est une généralisation du lissage double, qui permet entre autre de proposer les modèles suivants:

• tendance linéaire locale

• tendance linéaire locale + saisonnalité (modèle additif), présenté dans ce cours

• tendance linéaire locale * saisonnalité (modèle multiplicatif), non-présenté dans ce cours

Elle est une variante du lissage exponentiel double, et s'applique au même type de série. Dans ce cas, 2 voire 3 paramètres de lissage entrent en jeu.

5.5.1. La méthode non saisonnière de HW

<u>Modèle considéré :</u> $x_t = b + at + e_t$

Pour des chroniques sans variations saisonnières et à tendance localement tendance localement linéaire

Définition (formules de mise à jour) : Pour $\alpha, \beta \epsilon [0,1]$, La prévision de la série x_t à l'horizon h, $\hat{x}_T(h)$, fournie par la méthode de lissage exponentiel de Holt-Winters de paramètres $\alpha\ et\ \beta$ est donnée par $\hat{x}_T(h) = \hat{b}_T + \hat{a}_T h$

où α, β est la **constante de lissage**.

$$\begin{cases} \hat{a}_T = \beta\left(\hat{b}_T - \hat{b}_{T-1}\right) + (1-\beta)\hat{a}_{T-1} \\ \hat{b}_T = \alpha\, x_t + (1-\alpha)(\hat{b}_{T-1} + \hat{a}_{T-1}) \end{cases}$$

Remarques :

1. L'initialisation peut se faire comme dans le cas du lissage exponentiel double

2. L'avantage de cette approche est d'avoir une plus grande flexibilité mais la contrepartie est de devoir régler deux paramètres. Si α et β sont proches de 1 tous les deux, la prévision est lisse (fort poids du passé).

5.5.2. La méthode saisonnière additive de HW

Modèle considéré : $x_t = b + at + s_t + e_t$

Pour des chroniques sans variations saisonnières et à tendance localement tendance localement linéaire, modèle additif. On note P la période de s_t.

Définition (formules de mise à jour) : Pour $\alpha, \beta\ et\ \gamma\ \epsilon [0,1]$, La prévision de la série à l'horizon h, $\hat{x}_T(h)$, fournie par la méthode de lissage exponentiel de Holt-Winters de paramètres $\alpha, \beta\ et\ \gamma$ est donnée par

$$\hat{x}_T(h) = f(x) = \begin{cases} \hat{a}_T h + \hat{b}_T + \hat{S}_{T+h-P}, & si\ 1 \le h \le P \\ \hat{a}_T h + \hat{b}_T + \hat{S}_{T+h-2P}, & si\ P+1 \le h \le 2P \\ \qquad \dots \\ \qquad \dots \end{cases}$$

où $\alpha, \beta\ et\ \gamma$ sont les **paramètres de lissage** (à déterminer).

$$\begin{cases} \hat{a}_T = \beta\left(\hat{b}_T - \hat{b}_{T-1}\right) + (1-\beta)\hat{a}_{T-1} \\ \hat{b}_T = \alpha\,(x_t - \hat{S}_{T-P}) + (1-\alpha)\left(\hat{b}_{T-1} + \hat{a}_{T-1}\right) \\ \hat{S}_T = \gamma\left(x_t - \hat{b}_T\right) + (1-\gamma)\hat{S}_{T-P} \end{cases}$$

Remarques :

1. La première formule de mise à jour s'interprète comme une moyenne pondérée de la différence des niveaux estimés aux instants T et T-1 et la pente estimée à l'instant T-1.

La deuxième comme une moyenne pondérée de l'observation x_t (à laquelle on a retranché la composante saisonnière estimée à l'étape précédente) et l'estimation de la tendance faite à l'instant T-1.

La troisième comme une moyenne pondérée de l'observation x_t (à laquelle on a retranché le niveau calculé à l'instant T) et de la composante saisonnière calculée à l'instant T−P.

2. Le choix des constantes de lissage dans la pratique peut s'effectuer de la même manière que précédemment, c'est-à-dire en minimisant la somme des carrés des erreurs de prévision.

5.6. Mise en œuvre de lissages exponentiels

Dans cette partie nous détaillerons seulement la mise en œuvre dans le cas du LES, les autres lissages reposant sur la même démarche.

Initialisation

$\hat{x}_1 = x_1$ ou $\hat{x}_1 = \bar{x}$. La valeur choisie pour aura d'autant moins d'influence sur que T sera grand.

<u>Choix de la constante de lissage</u>

Ce choix peut se faire selon des critères subjectifs de « rigidité » ou de « souplesse » de la prévision. Si on veut une prévision rigide, on choisira β ∈ [0.7; 0.99]. Si on veut une prévision souple, on choisira β ∈ [0.01; 0.3]. Mais une méthode plus objective consiste à choisir α minimisant les résidus suivant un critère qui mesure la grandeur globale des résidus, tels que les suivants :

Erreur Quadratique Moyenne de prévision $$EQM = \frac{1}{T-1}\sum_{i=1}^{i=N}(x_{t+1} - \hat{x}_t)^2$$

Erreur Absolue Moyenne de prévision $$EAM = \frac{1}{T-1}\sum_{i=1}^{i=N}|x_{t+1} - \hat{x}_t|$$

Erreur Moyenne de prévision $$EM = \frac{1}{T-1}\sum_{i=1}^{i=N}(x_{t+1} - \hat{x}_t)$$

<u>Exemple :</u> nous reprenons les données de ventes en France d'essence aviation

Tableau 8. Ventes en France d'essence aviation (en milliers de tonnes, source : Comité Professionnel du Pétrole)

Trimestre / Année	1er trimestre	2e trimestre	3e trimestre	4e trimestre	Moyenne annuelle
2005	3,6	7,0	7,6	3,7	5,5
2006	3,6	6,7	7,4	3,9	5,4
2007	3,7	6,4	7,1	4,1	5,3
2008	3,6	5,7	7,1	3,7	5
Moyenne trimestrielle	3,7	6,5	7,6	3,9	5,3

Tableau 9. Chronique du tableau précédent et série obtenue par LES avec α=0,4

C4 =0,4*B3+0,6*C3

	A	B	C
1	t	x_t	*LES* α = 0,4
2	1	130	
3	2	121	130,00
4	3	120	126,40
5	4	127	123,84
6	5	122	125,10
7	6	129	123,86
8	7	124	125,92
9	8	120	125,15
10	9	139	123,09
11	10	136	129,45
12	11	135	132,07
13	12	134	133,24
14	13	136	133,55
15	14	133	134,53
16	15	138	133,92
17	16	133	135,55
18	17		134,53
19	18		134,53

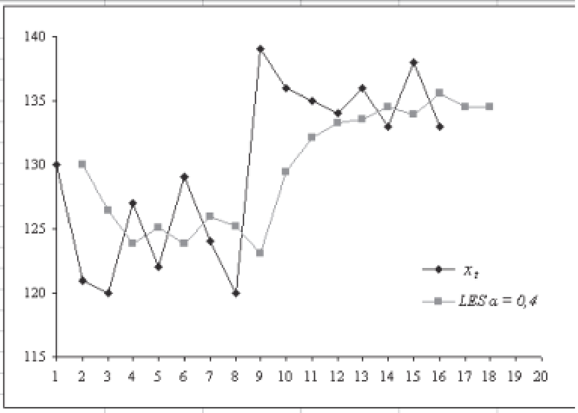


Tableau 10. Présentation des calculs du LES avec les critères calculés

		α = 0,4				α = 0,5			
t	x_t	*LES*	e_t	*ABS* (e_t)	$(e_t)^2$	*LES*	e_t	*ABS* (e_t)	$(e_t)^2$
1	130								
2	121	130,00	−9,00	9,00	81,00	130,00	−9,00	9,00	81,00
3	120	126,40	−6,40	6,40	40,96	125,50	−5,50	5,50	30,25
4	127	123,84	3,16	3,16	9,99	122,75	4,25	4,25	18,06
5	122	125,10	−3,10	3,10	9,63	124,88	−2,88	2,88	8,27
6	129	123,86	5,14	5,14	26,39	123,44	5,56	5,56	30,94
7	124	125,92	−1,92	1,92	3,68	126,22	−2,22	2,22	4,92
8	120	125,15	−5,15	5,15	26,53	125,11	−5,11	5,11	26,11
9	139	123,09	15,91	15,91	253,12	122,55	16,45	16,45	270,45
10	136	129,45	6,55	6,55	42,85	130,78	5,22	5,22	27,28
11	135	132,07	2,93	2,93	8,57	133,39	1,61	1,61	2,60
12	134	133,24	0,76	0,76	0,57	134,19	−0,19	0,19	0,04
13	136	133,55	2,45	2,45	6,02	134,10	1,90	1,90	3,62
14	133	134,53	−1,53	1,53	2,33	135,05	−2,05	2,05	4,20
15	138	133,92	4,08	4,08	16,67	134,02	3,98	3,98	15,81
16	133	135,55	−2,55	2,55	6,50	136,01	−3,01	3,01	9,07
17		134,53				134,51			
			EM = 0,64	*EAM* = 2,27	*EQM* = 6,42		*EM* = 0,12	*EAM* = 2,23	*EQM* = 6,55

Tableau 11. Valeurs des critères calculés sur le dernier tiers de la série

Valeur de α	*EM*	*EQM*	*EAM*
0,1	4,548	25,311	4,548
0,2	2,931	14,068	3,101
0,3	1,545	8,151	2,495
0,4	0,643	6,421	2,274
0,5	0,125	6,547	2,227
0,6	−0,148	7,361	2,449
0,7	−0,280	8,436	2,648
0,8	−0,339	9,670	2,833
0,9	−0,369	11,095	3,012

5.7. Critique des méthodes de lissage exponentiel

L'avantage des méthodes vues dans ce chapitre pour la prévision, est de fournir une prévision "bon marché" (peu coûteuse en moyens) et parfois très satisfaisante comme dans l'exemple précédent.

Les inconvénients les plus flagrants sont de deux ordres :

1. rien ne garantit l'optimalité de la méthode sur une série de donnée : les méthodes de lissage exponentiel sont parfois loin d'être les mieux adaptées (encore faut-il s'en apercevoir).

2. elles sont incapables de fournir des intervalles de prévision, c'est-à-dire un intervalle contenant la prévision avec une probabilité donnée. Et pour cause, aucun cadre probabiliste n'a été défini pour le moment.

Pour pallier ces insuffisances, on est amené à réaliser des prévisions au moyen de modèles probabilistes. Il est à noter que les méthodes de lissage exponentiel correspondent (à l'exception de la version multiplicative de Holt-Winters) à des modèles probabilistes particuliers. On peut donc voir les méthodes probabilistes comme des techniques plus générales permettant de justifier l'emploi des méthodes élémentaires et d'en élargir le champ d'application.

6. COMPLEMENTS

Dans ce chapitre, nous énonçons des notions non abordés dans ce cours, et invitons le lecteur intéressé à se documenter dans les références laissées en bibliographie.

6.1. Autre indices descriptifs

Indices de dépendance

Auto-covariances empiriques.

Autocorrélations empiriques.

6.2. Autres méthodes de décomposition

La méthode Census

Décomposition via LOESS (locally estimated scatterplot smoothing)

6.3. Prévisions via les modèles autorégressifs ARMA

Modèle ARMA(p,q)

Modèle ARMA saisonnier (modèle SARMA)

Modèle Modèles ARIMA(p,d,q) et SARIMA

Modèle ARMAX

6.4. Autres

Hétéroscédasticité conditionnelle

Etude du bruit lié à la décomposition

7. BIBLIOGRAPHIE

Aragon, Y., Séries temporelles avec R, Collection Pratique R, EDP sciences, ISBN : 978-2-7598-1779-5, 2011

Box, G.E. P., Jenkins G. M., Reinsel G.C.. *Time series analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, fourth édition, 2008

Chatfield, C.. *The Analysis of Time Series, an Introduction*. Fourth edition. Chapman & Hall, 1989

Goldfarb, B.,Pardoux, C.. Introduction à la méthode statistique,Dunod, 2011

Hamilton, J.D. *Time series analysis*. Princeton University Press, Princeton, NJ, 1994

Lagnoux, A.. *Statistique pour les Sciences Humaines II*. Disponible à l'adresse https://perso.math.univ-toulouse.fr/lagnoux/enseignements/.