IUT de Paris - Rives de Seine
Université Paris Cité

Séries chronologiques - Partie 4

Résidus, valeurs atypiques, prévision

BUT Science des Données, deuxième année

Objectifs de l'analyse des résidus

- Juger de l'ajustement d'un modèle.
- Comparer différents modèles entre eux.
- Evaluer la capacité prédictive d'un modèle.

L'analyse des résidus va reposer sur l'utilisation de **critères d'erreur**.

L'erreur de prévision (ou d'ajustement)

Les critères d'erreur se basent tous sur l'**erreur de prévision** (ou d'ajustement) du modèle.

Rappel : L'erreur d'ajustement représente l'écart entre valeurs observées et valeurs estimées et se calcule de la façon suivante, quel que soit le modèle (additif ou multiplicatif, parmétrique ou non paramétrique) :

$$\hat{e}_i = y_i - \hat{y}_i$$

où $(y_i)_{i=1,\dots,n}$ désigne la série initiale et $(\hat{y}_i)_{i=1,\dots,n}$ la série ajustée.

Exemple

On reprend les ventes trimestrielles d'un grand magasin parisien.

La série $(y_i)_{i=1,\dots,10}$ des ventes (en milliers d'euros) est donnée du premier trimestre 1995 au deuxième trimestre 1997 dans le tableau ci-dessous :

t_i	1	2	3	4	5	6	7	8	9	10
y_i	662	742	683	842	717	792	742	875	767	805

Si l'on estime la tendance à l'aide par $\hat{f}_i = MMC(4)_i$ puis les coefficients saisonniers centrés $\hat{s}_i$, on trouve les résidus $\hat{e}_i = y_i - \hat{f}_i - \hat{s}_i$:

t_i	3	4	5	6	7	8
$\hat{e}_i$	-5.34375	5.03125	-0.15625	-0.15625	5.03125	-5.34375

Exemple

On a vu que ces résidus n'étaient pas centrés: $\bar{\hat{e}} = -0.15625$. On modifie donc la tendance et les résidus:

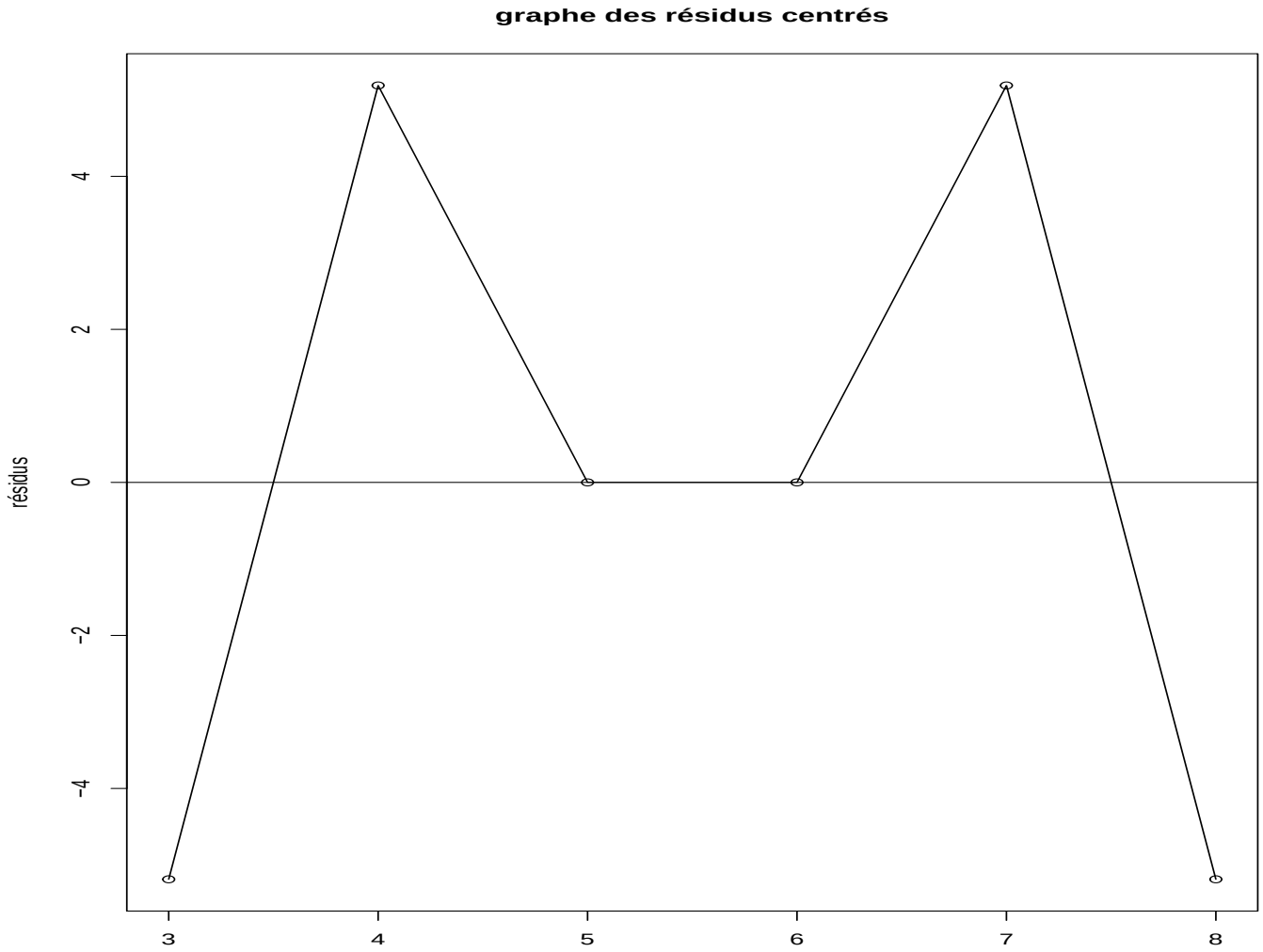
$$\tilde{f}_i = \hat{f}_i + \bar{\hat{e}}\,, \quad \tilde{e}_i = \hat{e}_i - \bar{\hat{e}}\,.$$

Les résidus $\tilde{e}_i$ sont donc centrés, et valent

t_i	3	4	5	6	7	8
$\tilde{e}_i$	-5.1875	5.1875	0	0	5.1875	-5.1875

Dans la suite, on supposera toujours les résidus centrés, et on les notera encore $\hat{e}_i$.

Exemple



Deux outils: graphe et variance des résidus

Le graphe des résidus permet de repérer les observations qui sont mal prédites par le modèle.

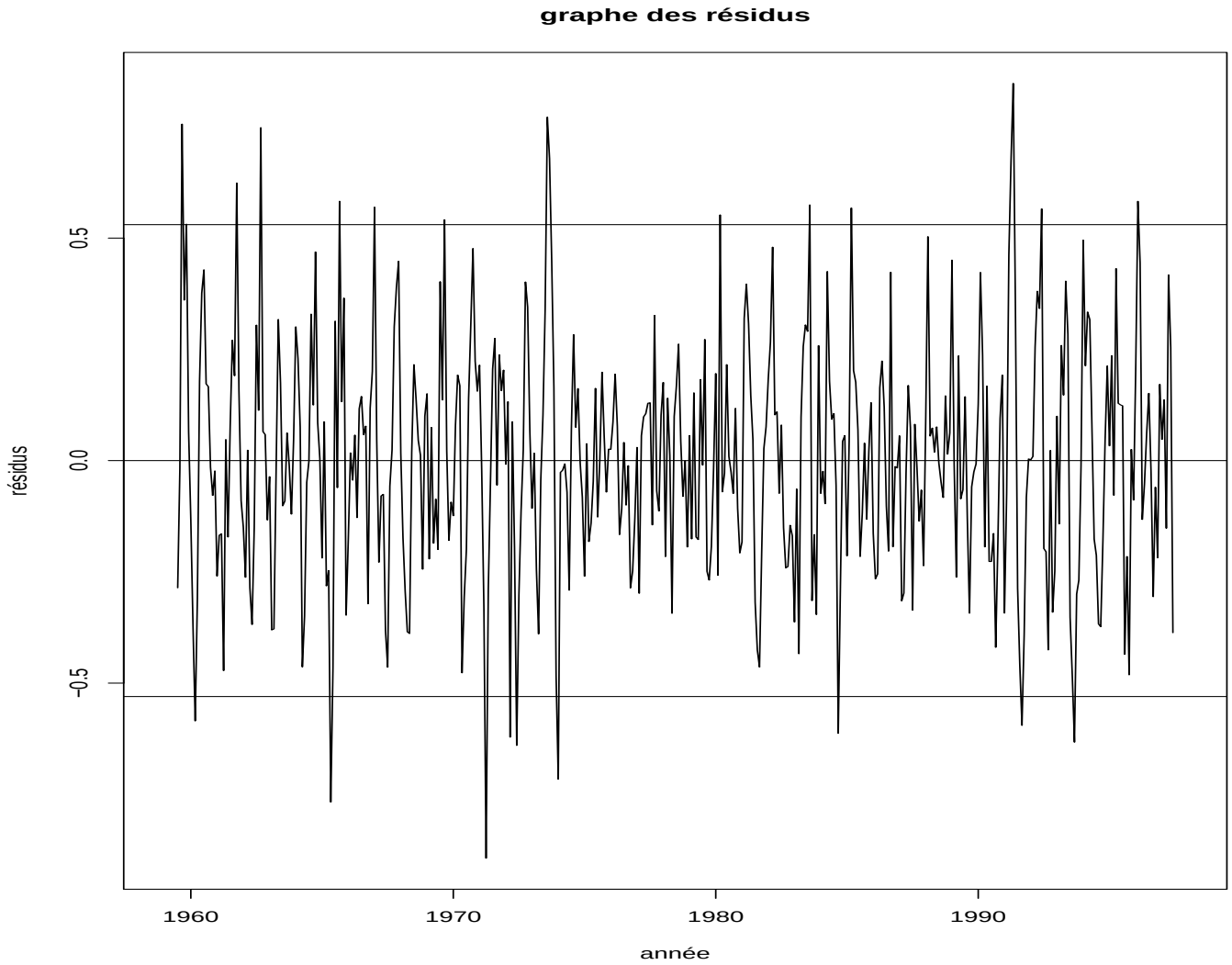
Un critère global est l'erreur quadratique moyenne d'ajustement, c'est à dire la variance des résidus :

$$EQM(\hat{e}) = \mathrm{Var}(\hat{e})\,.$$

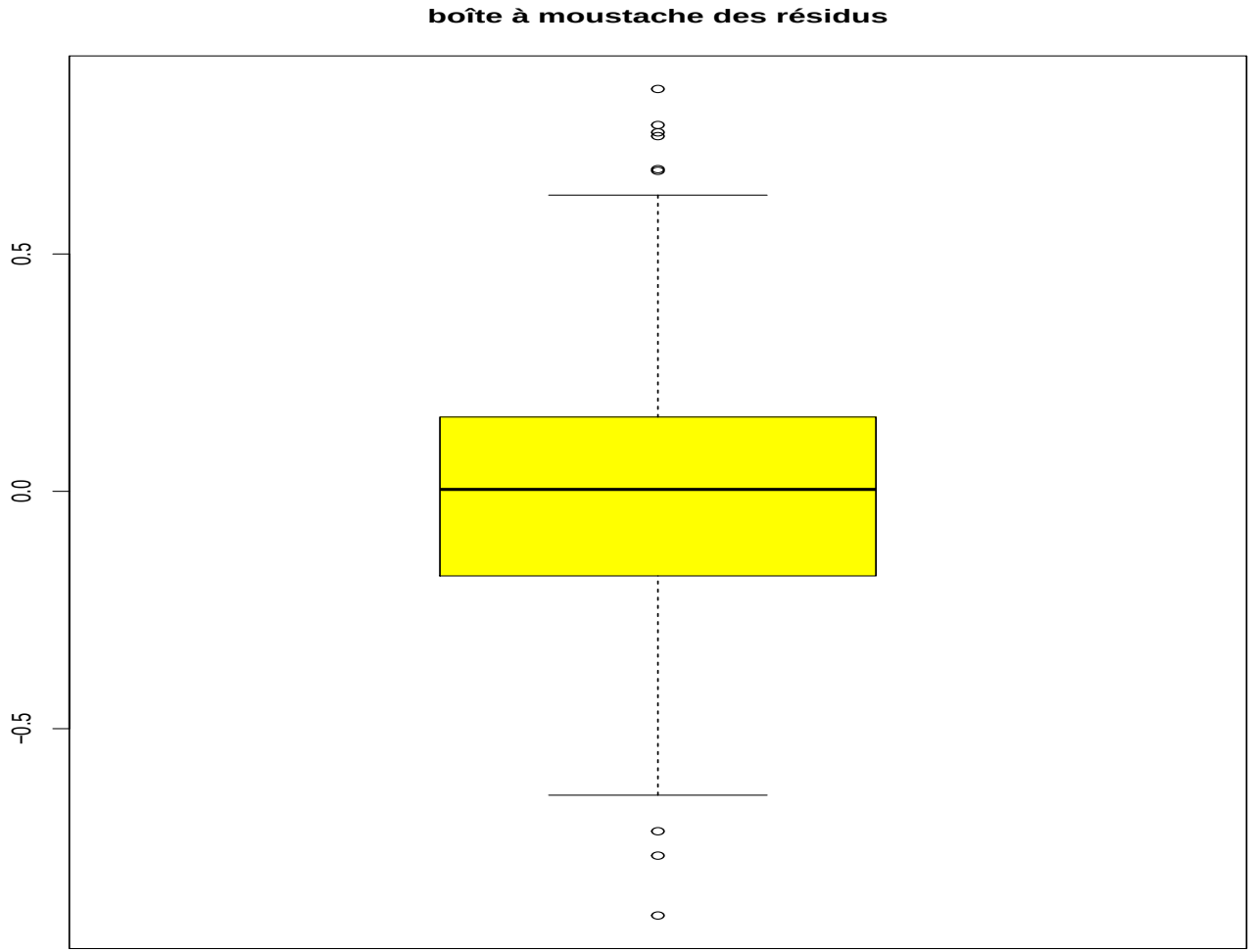
On peut combiner ces deux outils pour repérer les résidus qui s'écartent de plus de deux écarts-type, en traçant les droites horizontales d'ordonnées $+2\sigma(\hat{e})$ et $-2\sigma(\hat{e})$.

Pour évaluer le nombre d'observations mal prédites, on peut aussi tracer la boîte à moustache des résidus.

Exemple du taux de CO2



Exemple du taux de CO2



Comparaison de modèles : exemple du taux de CO2

L'erreur quadratique d'ajustement peut-être utilisée pour comparer la qualité prédictive de deux modèles.

Si l'on reprend l'exemple du taux de CO2, on peut ajuster un polynôme (pour représenter la tendance globale non saisonnière) + un polynôme trigonométrique (pour représenter la tendance saisonnière), par la méthode des moindre carrés.
On cherchera donc à minimiser

$$\varphi(a_0, a_1, \ldots, a_p, b_1, b_2) = \sum_{i=1}^{n} \Big(y_i - a_0 - a_1 t_i - \cdots - a_p t_i^p - b_1 \sin(\theta t_i) - b_2 \cos(\theta t_i) \Big)^2 .$$

où θ est connu et représente la période ($\theta = (2\pi)/12$ pour une série qui a tendance à se répéter toutes les 12 mesures).
Pour un polynôme de degré 3, on obtient les coefficients:
$\hat{a}_0 = 316.1, \hat{a}_1 = 0.03311, \hat{a}_2 = 2.726x10^{-4}, \hat{a}_3 = -2.61x10^{-7}$,
$\hat{b}_1 = 2.185, \hat{b}_2 = -1.719$ et le coefficient $R^2 = 0.9975$.

Comparaison de modèles : exemple du taux de CO2

On peut comparer ce modèle paramétrique (ajustement d'un polynôme de degré 3 + polynôme trigonométrique) et le modèle non paramétrique additif (tendance par moyenne mobile + calcul des coefficients saisonniers).

Notons $\hat{e}_i^{(p)}$ les résidus du modèle paramétrique, et $\hat{e}_i^{(np)}$ les résidus centrés de la décomposition du modèle additif.

On obtient $\sigma(\hat{e}^{(p)}) = 0.747$ et $\sigma(\hat{e}^{(np)}) = 0.265$. L'erreur d'ajustement du modèle non paramétrique est plus petite que celle du modèle paramétrique.

Valeurs atypiques, un exemple

On reprend l'exemple de l'indice trimestriel de production industrielle, entre 1962 et 1969, base 100 en 1962 (source : INSEE).

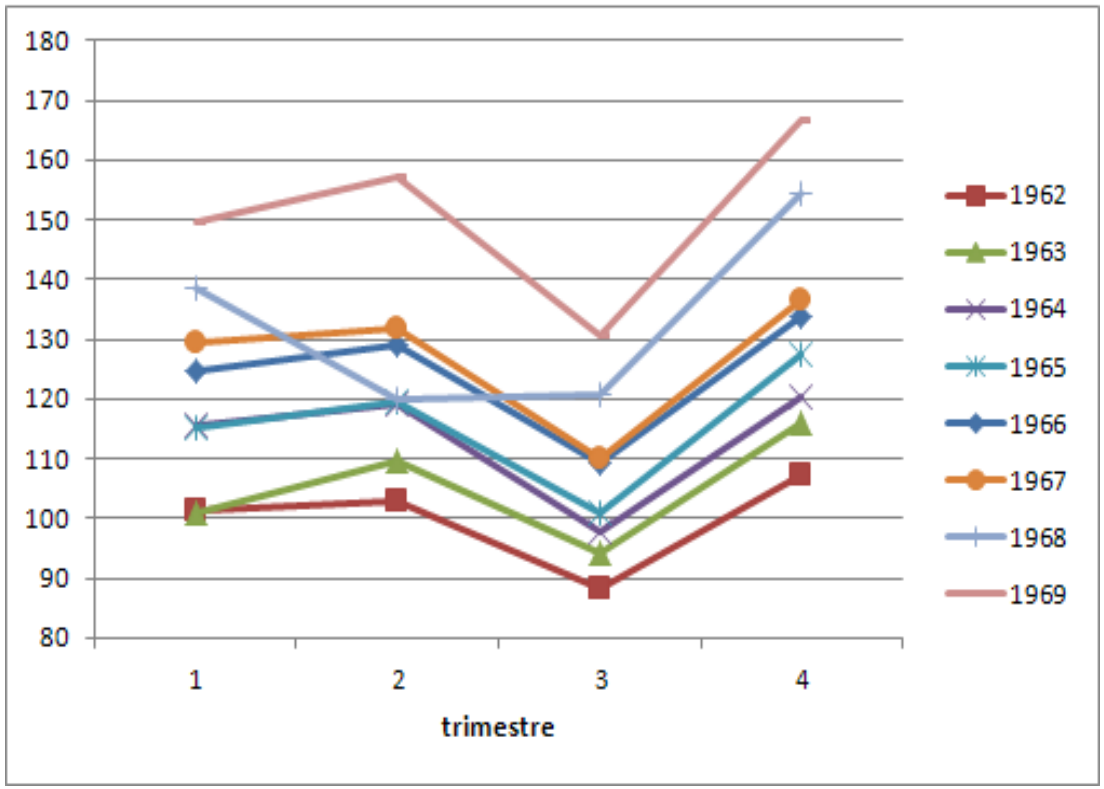
Année	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
1962	101.3	102.9	88.4	107.3
1963	101	109.8	94.1	116.1
1964	115.6	119.2	97.7	120.3
1965	115.1	119.5	101.1	127.4
1966	124.8	129	109.3	133.6
1967	129.4	131.8	110.2	136.4
1968	138.5	120.1	120.8	154.4
1969	149.5	157.1	130.8	166.5

Analyse graphique

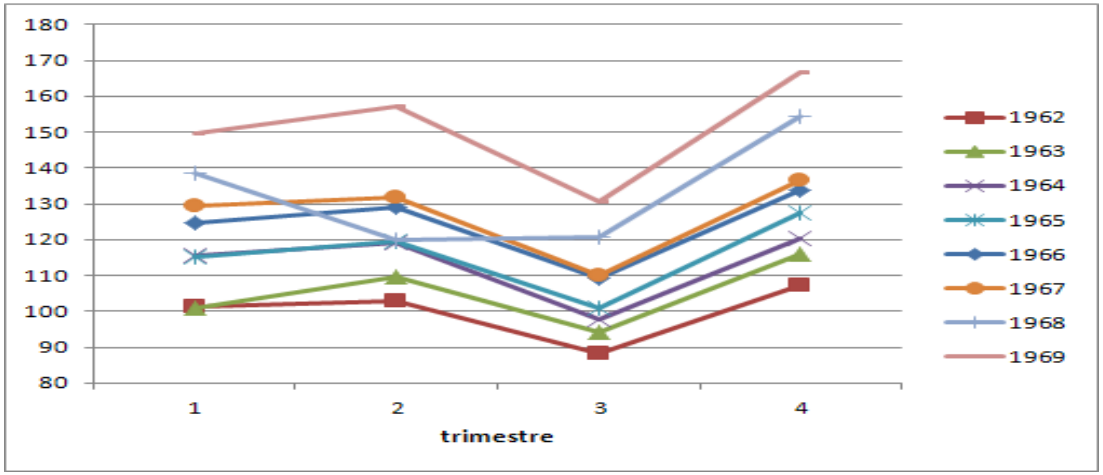
Quels graphiques peut on produire pour:

- Repérer les données atypiques.
- Faire apparaître la saisonnalité.

Graphique des courbes superposées

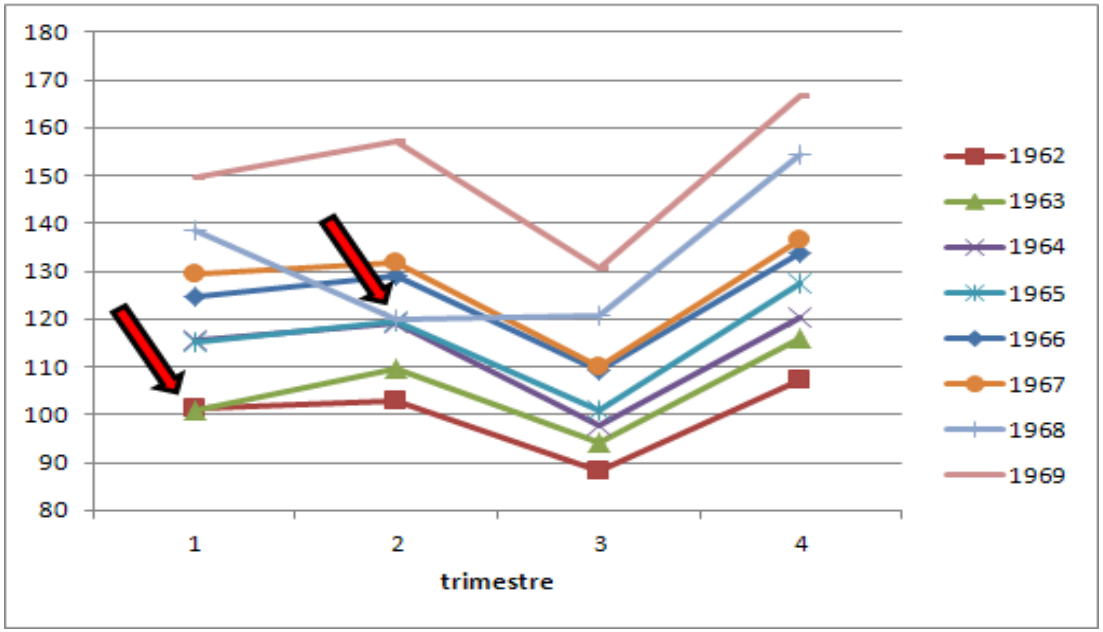


Graphique des courbes superposées



- Confirmation de variations saisonnières importantes : saisonnalité de période $p = 4$.
- Amplitude croissante des variations saisonnières : choix possible d'un modèle multiplicatif (voir aussi la méthode de la bande pour les données log-transformées).

Graphique des courbes superposées



Mise en évidence de deux variations accidentelles exceptionnelles :

- T1-1963 : rigueur exceptionnelle de l'hiver 1962-1963.
- T2-1968 : grève des mineurs en mars 1968.

Synthèse de la démarche

- Analyse descriptive des données : analyse graphique, données manquantes, données atypiques, . . .
  L'examen des graphiques superposés par période sur les données brutes permet de déceler des atypicités.

  Une bonne connaissance du phénomène étudié est nécessaire pour identifier la cause des variations accidentelles d'ampleur exceptionnelle (grève, accident climatique, . . .).
- Modélisation et estimation : plusieurs méthodes possibles (paramétriques, non-paramétriques).
- Analyse des résidus: qualité de l'ajustement, détections de valeurs atypiques.
- Prévision.

Prévision

La plupart du temps, l'objectif principal de l'étude des séries chronologiques est la prévision des valeurs futures.

On va décrire ici 3 méthodes différentes:

- la première méthode repose sur des notions que nous avons vues en détail, et reste donc dans le cadre de ce cours.
- les deux dernières débordent du cadre de ce cours, et seront vues plus en détail en TP.

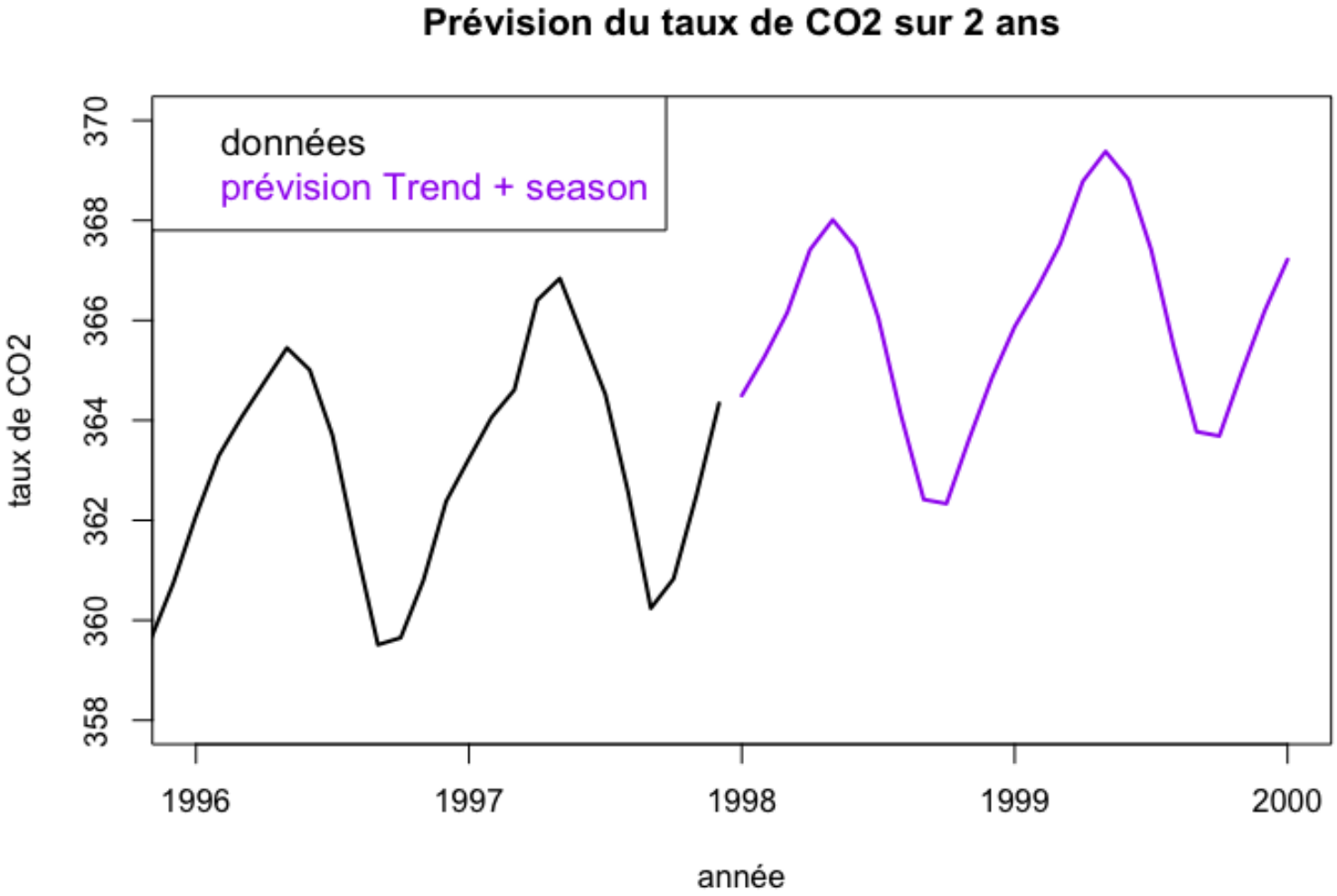
Prévision par estimation de tendance + coefficients saisonniers

La première méthode, pour un modèle additif, est constituée de trois étapes:

- On calcule des coefficients saisonniers $\hat{s}_i$ (cf. Partie 3).
- On en déduit la série corrigée des variation saisonnières: $CVS_i = y_i - \hat{s}_i$.
- On ajuste une tendance paramétrique $\tilde{f}$ sur la série CVS_i (par exemple par moindres carrés). La plupart du temps, un ajustement polynomial suffit, mais selon la forme de la CVS ou de la tendance, il peut être judicieux d'envisager d'autres ajustements (exponentiel, logarithmique,...).

On peut alors prévoir la valeur future au temps $k > n$ par $\hat{y}_k = \tilde{f}_k + \hat{s}_k$.

Exemple du taux de CO2



Prévision par la méthode de Holt-Winters

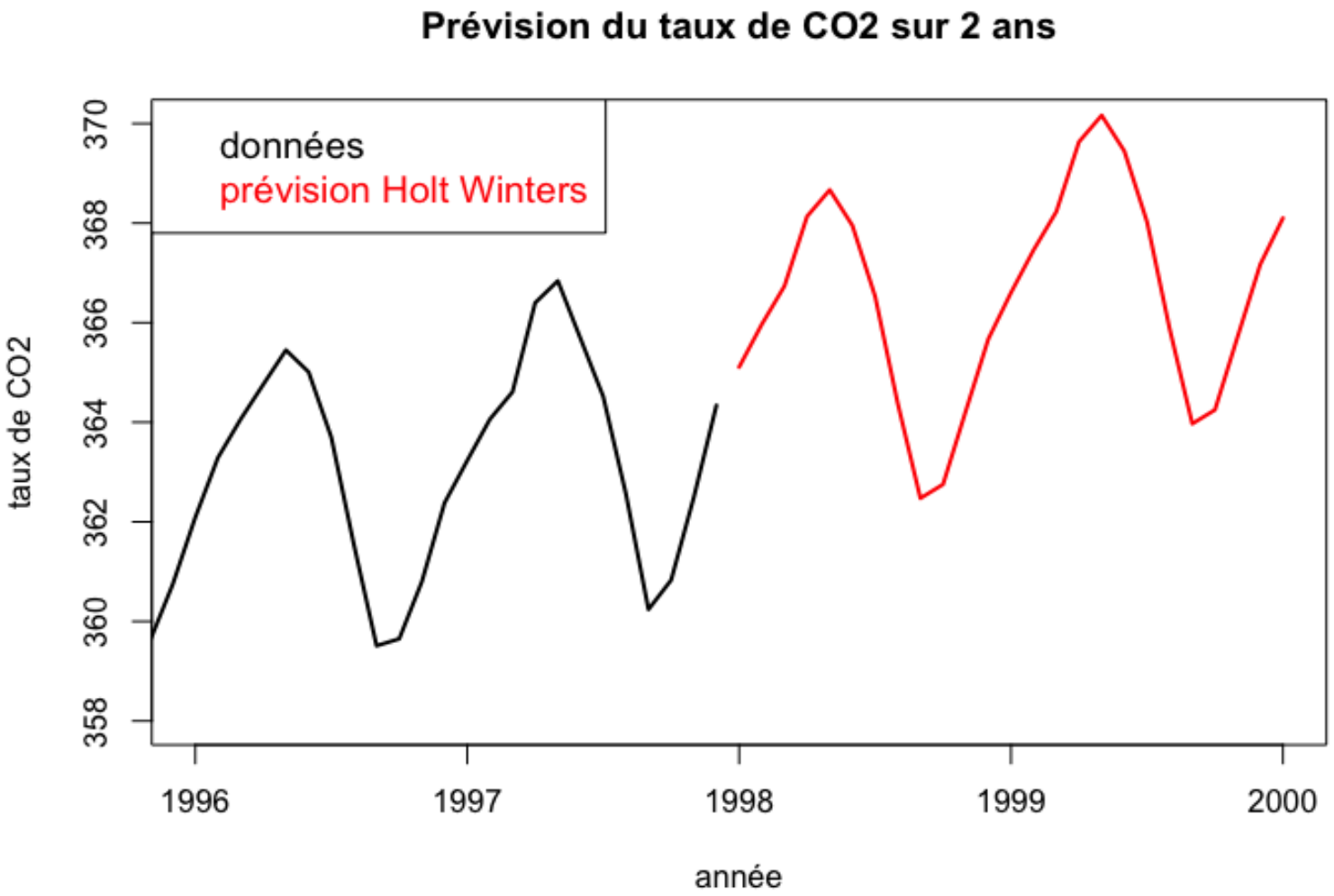
Le principe est le suivant:
on suppose qu'on peut prévoir y_i à l'aide des observations antérieures en affectant un poids qui décroit exponentiellement avec la distance à i; et de même pour le comportement saisonnier, avec les observations espacées de p: $y_{i-p}, y_{i-2p}, \ldots$.

On détermine le meilleure procédure de ce type en ajustant les paramètres par moindre carrés.

Une propriété importante de cette procédure est qu'elle est récursive: lorsqu'une nouvelle observation arrive on peut mettre à jour les paramètres sans refaire toute l'optimisation (coût de calcul très faible).

On peut donc prévoir $\hat{y}_{n+1}$, puis $\hat{y}_{n+2}, \ldots$ en mettant à jour l'algorithme au fur et à mesure.

Exemple du taux de CO2



Prévision à l'aide de modèles ARMA

La troisième méthode est constituée de trois étapes:

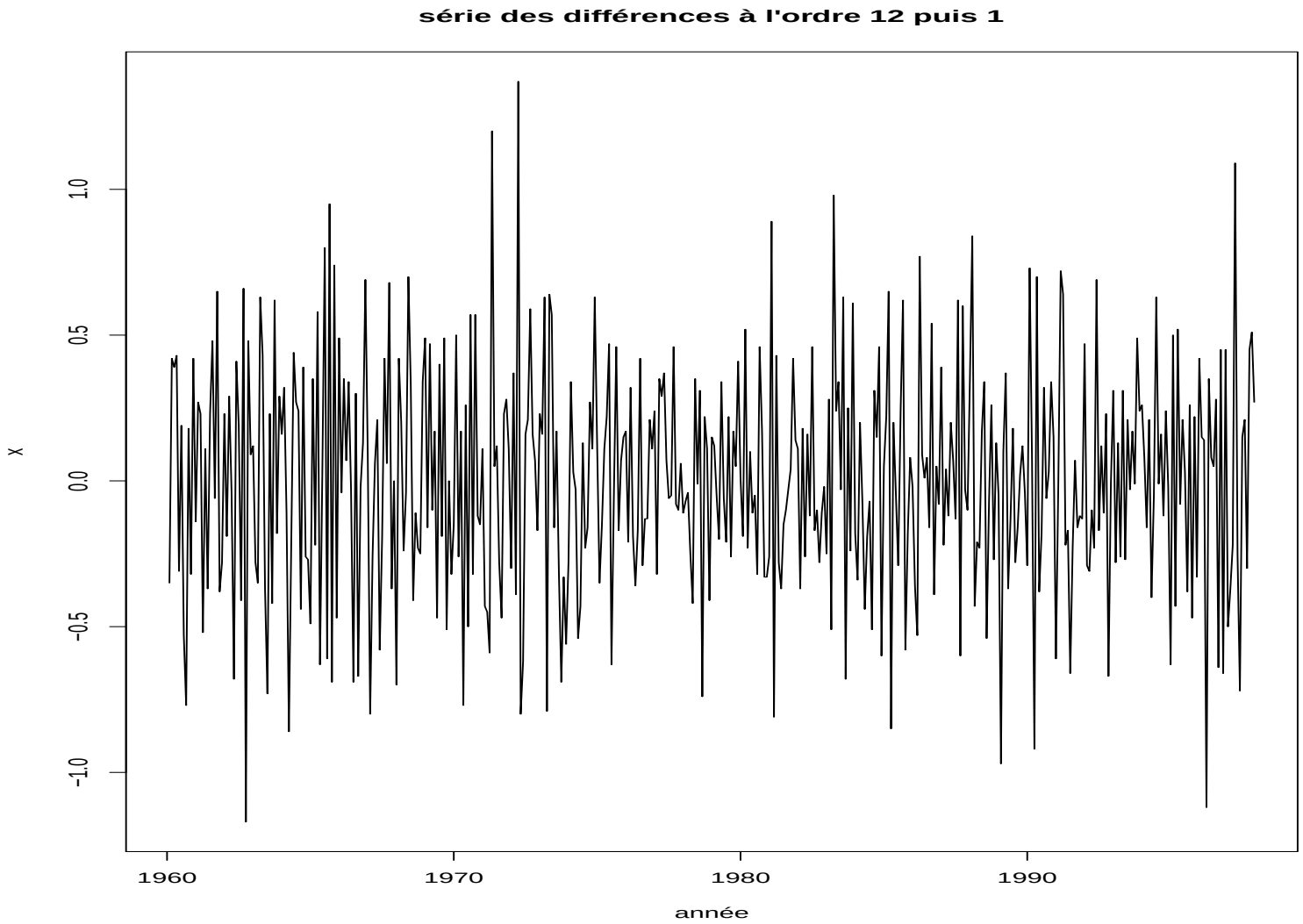
- On constitue la série des différences (à l'ordre p puis à l'ordre 1). $z_i = y_i - y_{i-p}$ puis $x_i = z_i - z_{i-1}$.
- On vérifie que la série des x_i présente bien les caractéristiques d'une suite stationnaire (sans tendance ni saisonnalité). Si ce n'est pas le cas, on différencie encore (ordre p ou ordre 1).
- On ajuste un modèle ARMA sur la série des différences; ce sont des modèles qui expliquent la variable aléatoire X_i par le passé proche et le passé à l'ordre p. Le modèle le plus simple de ce type est

$$X_i = aX_{i-1} + bX_{i-p} + \varepsilon_i \,,$$

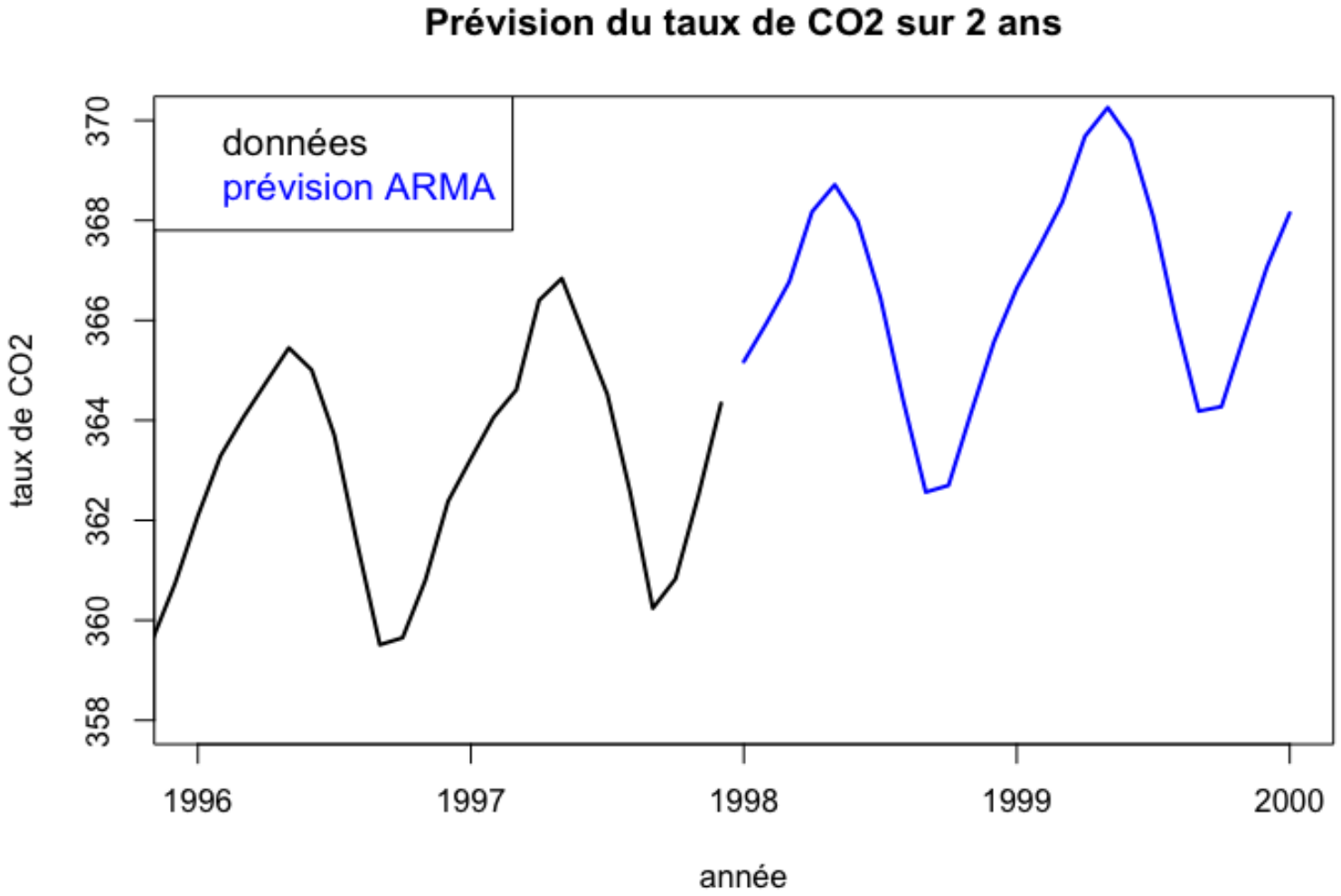
où les variables ε_i sont iid et ε_i est indépendante de toutes les $X_k, k < i$.

On peut alors prévoir la valeur future au temps $k > n$ à l'aide de l'équation du modèle ARMA.

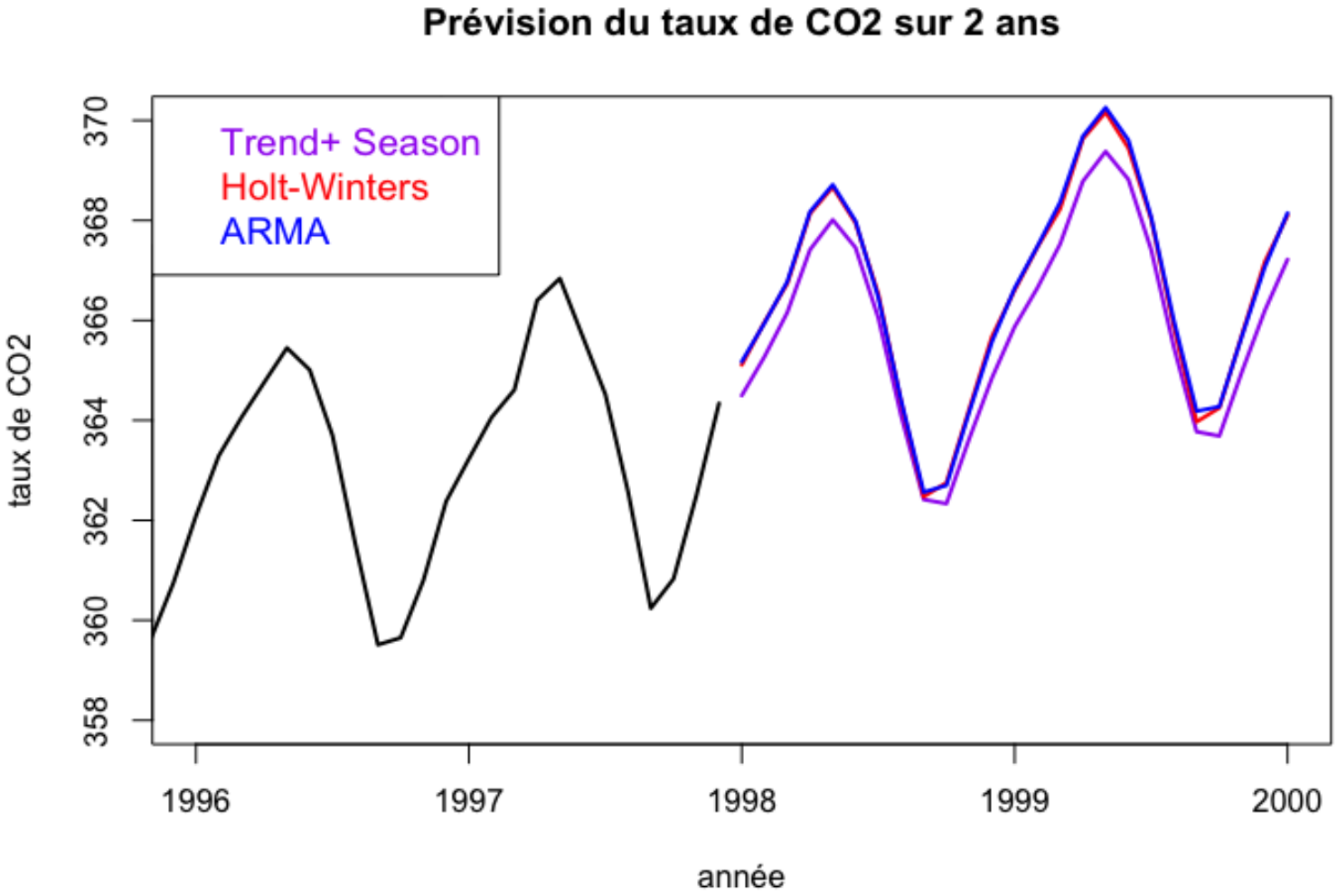
Exemple du taux de CO2



Exemple du taux de CO2



Taux de CO2: comparaison des prévisions



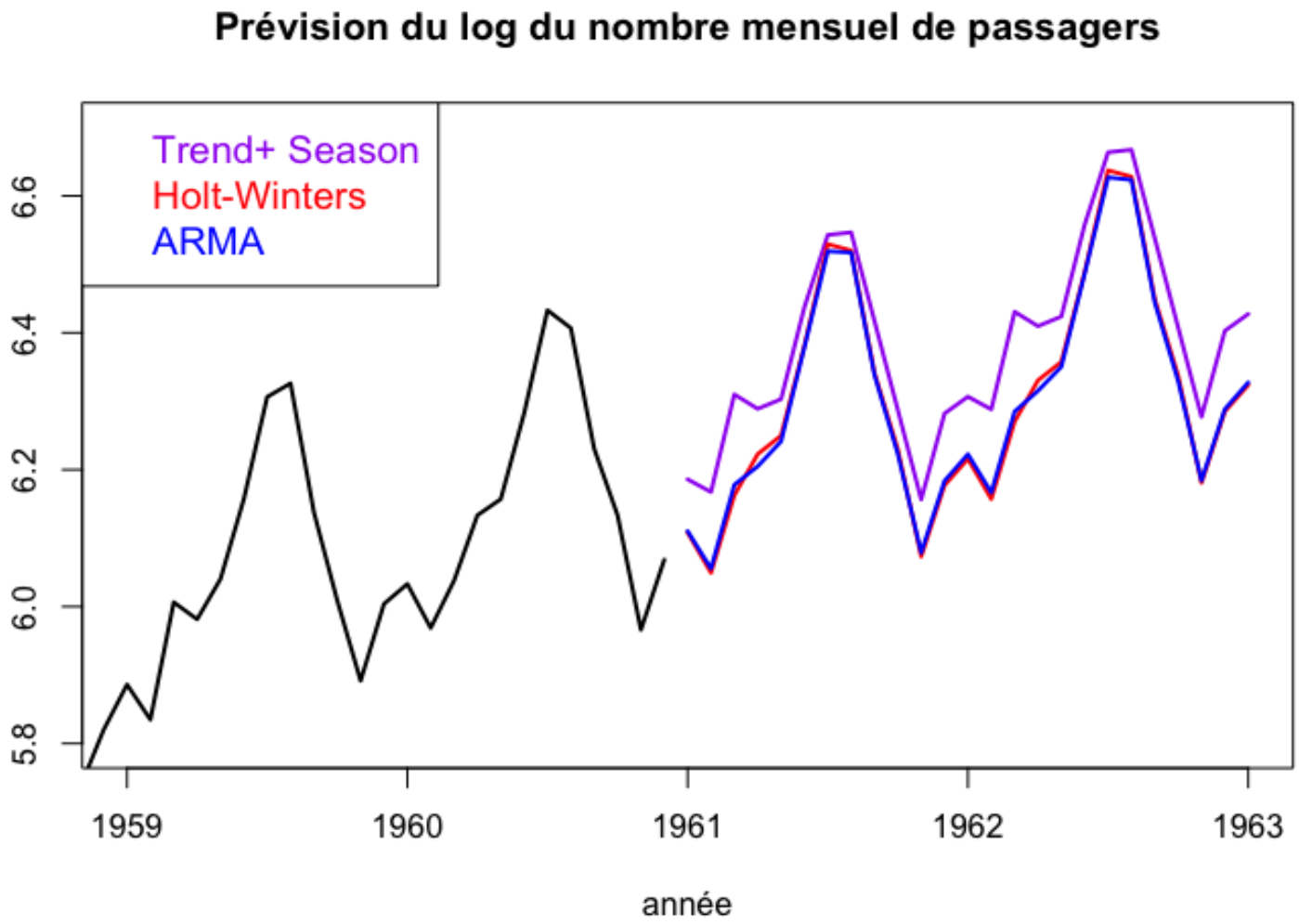
Exemple du trafic aérien

On rappelle que cette série relève du modèle multiplicatif.

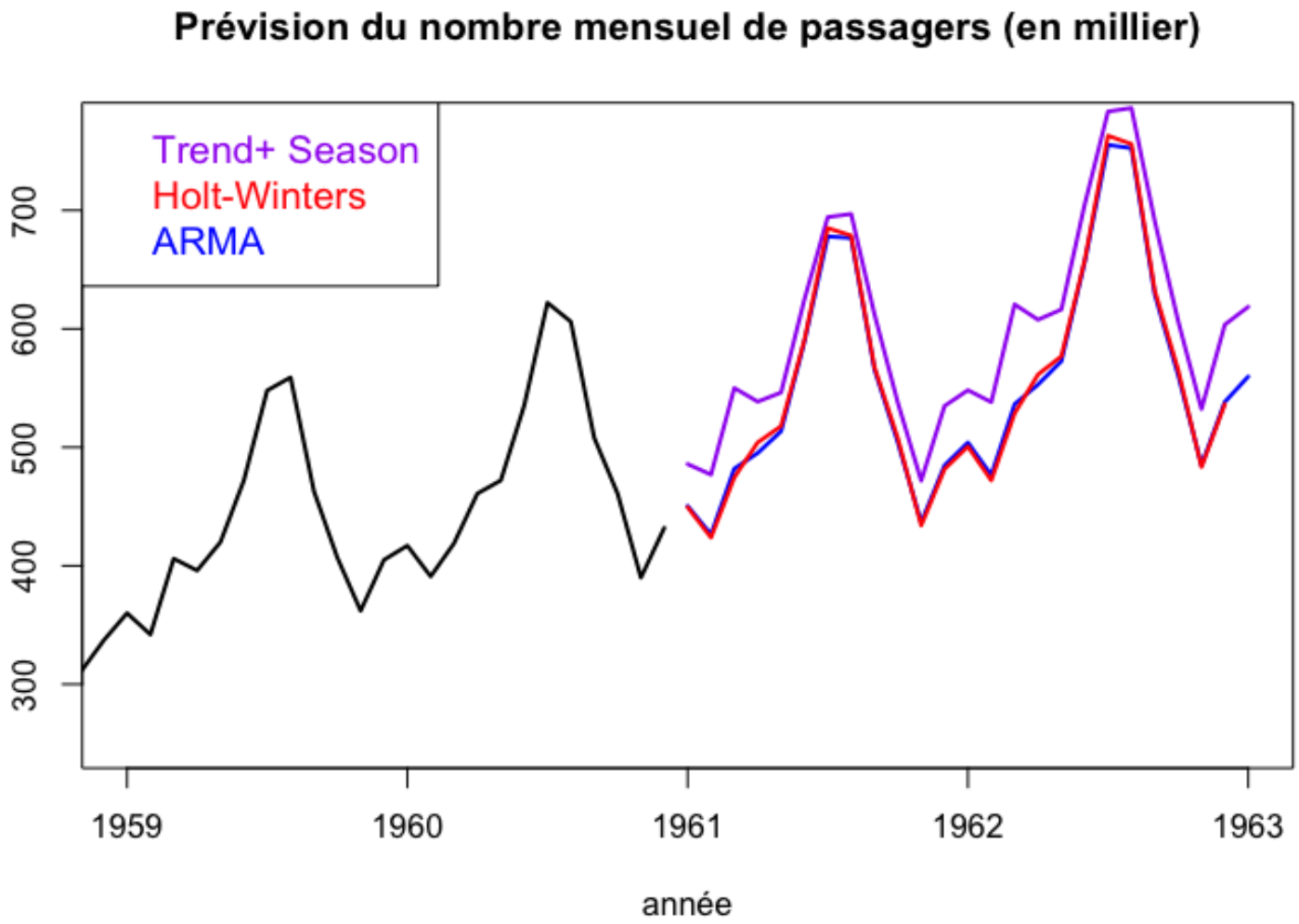
On peut se ramener à un modèle additif en prenant le logarithme des observations, puis mettre en oeuvre les trois méthodes de prévision décrites précédemment.

Pour obtenir des prévisions sur la série initiale, on appliquera simplement la transformation exponentielle aux prévisions du modèle additif.

Exemple du trafic aérien



Exemple du trafic aérien



Quelques remarques sur la prévision

- La première méthode est descriptive: on répète le motif que l'on a repéré sur les observations. Les prévisions pour $k > n$ ne dépendent que des observations $(y_i)_{1 \leq i \leq n}$.
- Les deux dernières méthodes sont dynamiques: pour $k > n$, $\hat{y}_k$ dépend de la série des observations $(y_i)_{1 \leq i \leq n}$ mais aussi des prévisions précédentes $(\hat{y}_j)_{n < j < k}$.
- L'algorithme de Holt-Winters est itératif, et se met à jour à chaque nouvelle prévision.
- En utilisant la méthode ARMA, on peut obtenir des intervalles de prévisions (l'analogue des intervalles de confiance pour la prévision).

Quelques remarques sur la prévision

On a présenté 3 méthodes de prévision fondées uniquement sur la série $(y_i)_{1 \leq i \leq n}$.

Mais pour affiner la prévision il peut être utile d'ajouter des variables dites "exogènes"; ces variables peuvent être d'autres séries temporelles, ou des dates d'évènements particuliers.

Pour le taux de CO2, on peut penser à la série des températures (à l'endroit où est relevé le taux de CO2), ainsi qu'aux occurrences (et leurs intensités) des éruptions des volcans proches.

Pour la production industrielle, on peut penser à la série des prix du pétrole, ansi qu'aux occurrences des grèves ou des crises économiques.