




Projet de Séries Temporelles

Kim Antunez et Alain Quartier-la-Tente

19/05/2020

Table des matières

1	Partie 1 : Les données	1
1.1	Question 1 : description de la série choisie	1
1.2	Questions 2 et 3 : transformation de la série	1
2	Partie 2 : Modèles ARIMA	2
3	Partie 3 : Prévisions	3
3.1	Question 6, 7 et 8 : construction d'un intervalle de confiance	3
3.2	Question 9 : question ouverte sur la causalité	5
A	Tests supplémentaires sur la qualité des modèles	i
B	Code 	iii
B.1	Fichier 0 - Creation des donnees.R	iii
B.2	Fichier 1 - Stationnarisation.R	iii
B.3	Fichier 2 - Estimation du modele ARIMA.R	vi
B.4	Fichier 3 - Previsions.R	viii

1 Partie 1 : Les données

1.1 Question 1 : description de la série choisie

Pour ce projet, travaillons sur la série d'indice de production industrielle (IPI) dans l'industrie automobile (identifiant : [010537940](#)). Il s'agit d'une série au niveau A64 de la nomenclature d'activités française révision 2 (NAF rév. 2, division 29), poste CL1. L'industrie automobile concerne aussi bien la production des constructeurs de voitures particulières, de véhicules de loisir, de véhicules utilitaires que les équipementiers spécialisés, les carrossiers, les assembleurs ou les prestataires de services d'aménagement de véhicules automobiles. Cette production intègre donc la filière complète, y compris moteurs et organes mécaniques en amont, dès lors qu'ils sont principalement destinés à des véhicules automobiles (à l'exception des parties de moteur).

Il s'agit d'un indice de Laspeyres¹, en base 2015, chaîné avec des pondérations annuelles (les pondérations correspondant aux valeurs ajoutées des branches associées). L'IPI dans l'industrie automobile est calculé à partir de l'enquête mensuelle de branche, par agrégation de séries "élémentaires" estimées en volume², calculées à un niveau plus fin.

Les séries de l'IPI sont corrigées des variations saisonnières et des jours ouvrables (CVS-CJO) à partir de la méthode X13-ARIMA. La désaisonnalisation est réalisée de manière indirecte : elle est effectuée à un niveau fin et les agrégats CVS-CJO sont ensuite calculés directement à partir de ces séries en agréant les séries CVS-CJO. Cette désaisonnalisation est réalisée par sous-périodes pour prendre en compte le fait que la structure économique des séries a beaucoup évolué en 30 ans, et donc qu'il serait peut pertinent d'appliquer un seul modèle de désaisonnalisation sur l'ensemble de la période. Ainsi, les modèles utilisés pour la désaisonnalisation commencent en 2005 et ces modèles sont utilisés pour estimer les séries CVS-CJO à partir de 2012.

Les séries CVS-CJO avant et après 2012 n'étant pas évaluées sur les mêmes modèles, l'idéal serait d'étudier notre série après janvier 2012 pour éviter des ruptures liées à ce changement de modèle. En revanche, cela laisserait une faible profondeur temporelle risquant de fragiliser l'estimation de nos modèles ARIMA. C'est pourquoi nous allons étudier la série d'IPI dans l'industrie automobile entre **janvier 2010 et décembre 2019**³, c'est-à-dire sur **120 observations**.

Nous n'effectuons pas ici de correction de point atypique ou de transformation logarithmique.

1.2 Questions 2 et 3 : transformation de la série

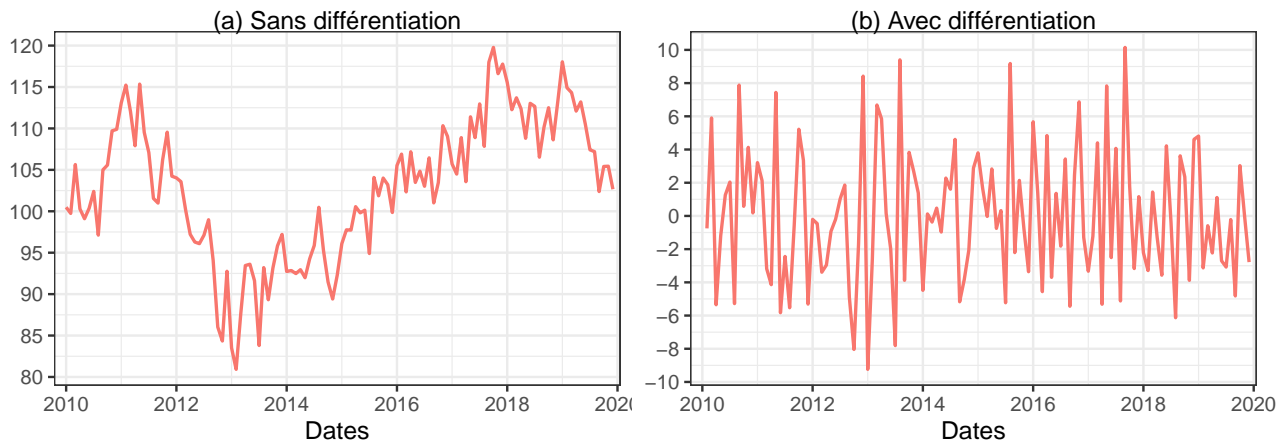


FIGURE 1 – IPI dans l'automobile (CVS-CJO) sans et avec différentiation.

1. Les indices de Laspeyres et de Paasche permettent de synthétiser en un indice unique un certain nombre d'indices. L'indice de Laspeyres le plus célèbre est l'IPC (indice des prix à la consommation).

2. La série d'IPI dans l'industrie automobile ne tient donc pas compte des variations de prix.

3. Lorsque nous avons commencé ce projet, l'IPI était disponible jusqu'en février 2020. En revanche, les derniers points étant souvent sujets à révisions, nous avons préféré ne pas prendre en compte les points de janvier et février 2020.

Le graphique 1-(a) ne montre pas de tendance linéaire déterministe nette sur la période 2010-2020 : on observe plutôt une alternance entre des périodes à tendance croissante (2010-2011, 2013-2018) et à tendance décroissante (2011-2013 et 2018-2020). La série de l'IPI dans l'automobile semble plutôt montrer une tendance stochastique : elle n'est sûrement **pas stationnaire**. Ceci est vérifié en faisant le test Dickey-Fuller augmenté (ADF) avec une constante non nulle et sans tendance : on ne rejette pas l'hypothèse de présence de racine unitaire au seuil de 5 % (tableau 1). Ceci est également confirmé par le test de racine unitaire de Phillips-Perron, non rejeté au seuil de 5 %, et par le test de stationnarité⁴ de Kwiatkowski-Phillips-Schmidt-Shin (KPSS), rejeté au seuil de 5 %. Nous **différencions** donc la série.

TABLE 1 – Tests de racine unitaire et de stationnarité sur la série d'IPI dans l'automobile.

Test	Statistique	p-valeur
Dickey-Fuller augmenté ^a	-1,678	0,434
Phillips-Perron	-2,578	0,336
KPSS	0,892	0,010 **

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

^a Le test ADF a été fait en rajoutant 2 retards. De cette façon les résidus utilisés dans ce test sont bien indépendants et le test ADF est bien interprétable

D'après le graphique 1-(b), la série différenciée semble **stationnaire**. Cette hypothèse est confirmée par le test de Dickey-Fuller augmenté, effectué avec une constante nulle et sans tendance, le test de Phillips-Perron et le test KPSS (tableau 2).

TABLE 2 – Tests de racine unitaire et de stationnarité sur la série différenciée d'IPI dans l'automobile.

Test	Statistique	p-valeur
Dickey-Fuller augmenté ^a	-10,261	0,010 **
Phillips-Perron	-15,132	0,010 **
KPSS	0,074	0,100 .

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

^a Le test ADF a été fait en rajoutant 1 retard. De cette façon les résidus utilisés dans ce test sont bien indépendants et le test ADF est bien interprétable

2 Partie 2 : Modèles ARIMA

Afin de déterminer les ordres maximaux, p_{max} et q_{max} , du modèle $ARMA(p, q)$ suivi par la série différenciée de l'IPI dans l'automobile, nous analysons les autocorrélogrammes et les autocorrélogrammes partiels (graphique 2). À partir de retard 2 (inclus), aucun autocorrélogramme est significatif à 5 % : on en déduit que $p_{max} = 1$. À partir de retard 2 (inclus), aucun autocorrélogramme partiel est significatif à 5 % : on en déduit que $q_{max} = 1$. Ainsi, pour savoir quel(s) modèle(s) retenir, nous allons tester tous les modèles $ARMA(p, q)$ tels que $p \leq 1$ et $q \leq 1$.

Quatre modèles ARMA ont donc été testés⁵ afin de s'assurer de l'indépendance des résidus (tableau 4) et, si c'est bien le cas, de la significativité des coefficients associés aux ordres maximaux des parties AR et MA des modèles (tableau 5) :

- $ARMA(0, 0)$: les résidus de ce modèle ne sont pas indépendants ➡ **modèle non retenu**
- $ARMA(1, 0)$: les résidus de ce modèle ne sont pas indépendants ➡ **modèle non retenu**

4. Ici, l'hypothèse alternative est la non-stationnarité de la série.

5. Ils ont été estimés sans constante.

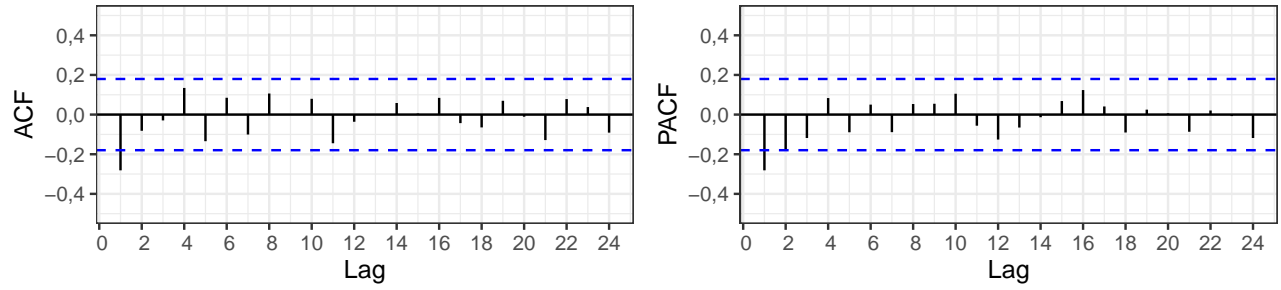


FIGURE 2 – Autocorrélogrammes (ACF) et autocorrélogrammes partiels (PACF) pour la série différenciée de l'IPI dans l'automobile.

- $ARMA(0,1)$: les résidus de ce modèle sont indépendants et le coefficient associé au MA(1) est significatif **👉 modèle retenu**
- $ARMA(1,1)$: les résidus de ce modèle sont indépendants (tableau mais le coefficient associé au AR(1) n'est significatif **👉 modèle non retenu**

Finalement, seul le modèle $ARMA(0,1)$ est valide sur la série différenciée. Sur la série non différenciée de l'IPI automobile, on retient donc le modèle **ARIMA(0,1,1)** défini mathématiquement par :

$$\Delta X_t = \varepsilon_t - 0,38 \varepsilon_{t-1} \quad (0,09)$$

ε_t est bien un bruit blanc : les $(\varepsilon_t)_t$ sont indépendants (tableau 4), homoscedastiques (tableau 6) et suivent aussi une loi normale (tableau 7).

Parmi l'ensemble des modèles testés, l'ARIMA(0,1,1) est aussi le modèle qui minimise les critères d'information (tableau 3).

TABLE 3 – Critères d'information des modèles ARIMA sur l'IPI de l'automobile.

	ARIMA(0,1,0)	ARIMA(1,1,0)	ARIMA(0,1,1)	ARIMA(1,1,1)
AIC	672,439	664,677	660,932	662,345
BIC	675,219	670,235	666,490	670,683

3 Partie 3 : Prévisions

3.1 Question 6, 7 et 8 : construction d'un intervalle de confiance

On cherche désormais à faire une prévision de X_t à l'horizon $T+2$. Notons θ_1 le coefficient associé à la partie MA de notre modèle $ARMA(0,1,1)$, qu'on estime par $\hat{\theta}_1 \simeq -0,38$ (tableau 5) en estimant le modèle entre janvier 2010 et décembre 2019. On a donc :

$$\Delta X_T = \varepsilon_T + \theta_1 \varepsilon_{T-1} \iff X_T = X_{T-1} + \varepsilon_T + \theta_1 \varepsilon_{T-1} \quad \text{où } \varepsilon_t \stackrel{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$$

En considérant θ_1 connu, les prévisions de X_{T+1} et X_{T+2} réalisées à l'instant T , notées $\hat{X}_{T+1|T}$ et $\hat{X}_{T+2|T}$, vérifient l'équation :

$$\begin{cases} \hat{X}_{T+1|T} = X_T + \theta_1 \varepsilon_T \\ \hat{X}_{T+2|T} = \hat{X}_{T+1|T} = X_T + \theta_1 \varepsilon_T \end{cases}$$

Les erreurs de prévision sont égales à :

$$\begin{cases} \hat{\varepsilon}_{T+1|T} = X_{T+1} - \hat{X}_{T+1|T} = \varepsilon_{T+1} + (\theta_1 - \theta_1) \varepsilon_T & = \varepsilon_{T+1} \\ \hat{\varepsilon}_{T+2|T} = X_{T+2} - \hat{X}_{T+2|T} = \varepsilon_{T+2} + (1 + \theta_1) \varepsilon_{T+1} + (\theta_1 - \theta_1) \varepsilon_T & = \varepsilon_{T+2} + (1 + \theta_1) \varepsilon_{T+1} \end{cases}$$

Les ε_t étant i.i.d., $\widehat{\varepsilon}_{T+h|T} \stackrel{(H_0)}{\sim} \mathcal{N}(0, \sigma_h^2)$ avec $\sigma_1^2 = \sigma^2$ et $\sigma_h^2 = \sigma^2(1 + (1 + \theta_1)^2)$. De plus, $\text{Cov}(\widehat{\varepsilon}_{T+1|T}, \widehat{\varepsilon}_{T+2|T}) = \sigma^2(1 + \theta_1)$. Donc :

$$\begin{pmatrix} \widehat{\varepsilon}_{T+1|T} \\ \widehat{\varepsilon}_{T+2|T} \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \underbrace{\sigma^2 \begin{pmatrix} 1 & 1 + \theta_1 \\ 1 + \theta_1 & 1 + (1 + \theta_1)^2 \end{pmatrix}}_{\Sigma} \right)$$

D'où :

$$(\widehat{\varepsilon}_{T+1|T} \quad \widehat{\varepsilon}_{T+2|T}) \Sigma^{-1} \begin{pmatrix} \widehat{\varepsilon}_{T+1|T} \\ \widehat{\varepsilon}_{T+2|T} \end{pmatrix} \sim \chi^2(2) \quad \text{avec} \quad \Sigma^{-1} = \frac{1}{\sigma^2} \begin{pmatrix} 1 + (1 + \theta_1)^2 & -(1 + \theta_1) \\ -(1 + \theta_1) & 1 \end{pmatrix}$$

En notant $q_{1-\alpha}$ le quantile $1 - \alpha$ d'une loi $\chi^2(2)$, une région de confiance de niveau α pour (X_{T+1}, X_{T+2}) est :

$$R_\alpha = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} : (1 + (1 + \theta_1)^2)(x - \widehat{X}_{T+1|T})^2 - 2(1 + \theta_1)(x - \widehat{X}_{T+1|T})(y - \widehat{X}_{T+2|T}) + (y - \widehat{X}_{T+2|T})^2 \leq \sigma^2 q_{1-\alpha} \right\} \quad (1)$$

Le problème est que σ_h et θ_1 sont ici inconnus. On estime donc θ_1 par $\widehat{\theta}_1$, qui est l'estimation que l'on fait à partir de nos données et σ par $\widehat{\sigma} = \frac{1}{T-2} \sum_{t=2}^T \widehat{\varepsilon}_t^2$. En remplaçant σ_h et θ_1 par leurs valeurs estimées, la région de confiance définis dans l'équation (1) reste valide mais **asymptotiquement uniquement**.

[Question 6] En somme, la région de confiance pour (X_{T+1}, X_{T+2}) est une ellipse, dont le centre est $(X_T + \widehat{\theta}_1 \varepsilon_T, X_T + \widehat{\theta}_1 \varepsilon_T)$ et caractérisé par l'équation :

$$R_{1-\alpha} = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} : (1 + (1 + \widehat{\theta}_1)^2)(X_T + \widehat{\theta}_1 \varepsilon_T - x)^2 - 2(1 + \widehat{\theta}_1)(X_T + \widehat{\theta}_1 \varepsilon_T - x)(X_T + \widehat{\theta}_1 \varepsilon_T - y) + (X_T + \widehat{\theta}_1 \varepsilon_T - y)^2 \leq \widehat{\sigma}^2 q_{1-\alpha} \right\} \quad (2)$$

L'application numérique ($\alpha = 0,05$, $\widehat{\theta}_1 \simeq -0,38$, $X_T \simeq 102,66$, $\varepsilon_T \simeq -2,63$, $\widehat{\sigma}^2 \simeq 14,73$, $q_{0,95} \simeq 5,99$) donne :

$$R_{95\%} = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} : 0,016x^2 - 0,014 \times x \times y + 0,011y^2 - 1,798x - 0,885y + 139,044 = 1 \right\} \quad (3)$$

[Question 7] Pour obtenir cette région de confiance il faut :

- que le modèle suivit par notre série entre janvier 2010 et février 2020 soit bien un modèle ARIMA(0,1,1). Le modèle doit être en théorie parfaitement identifié (σ et θ_1 connus ou a minima que leurs estimateurs convergent vers leurs valeurs) ;
- que les résidus de notre modèle ARIMA soient **indépendants, homoscedastiques et suivent une loi normale** : ce qui a bien été vérifié dans la partie précédente ;
- que T soit grand (dans notre cas $T = 120$).

[Question 8] Le graphique 3 présente la région de confiance au seuil 95 %, les intervalles de confiance associés aux deux prévision (lorsqu'on les calcule de manière indépendante pour X_{T+1} et X_{T+2}), ainsi que les dernières valeurs publiées de l'IPI automobile de janvier et de février 2020. On retrouve ce que l'on a montré par l'équation (2) : la même valeur est prédite pour X_{T+1} et X_{T+2} . Prédire les mêmes valeurs pour les deux dates paraît économiquement peu cohérent, mais cela reflète la dynamique du modèle ARIMA(0,1,1) :

- Puisqu'il y a aucun ordre autorégressif, ΔX_t ne dépend pas des valeurs passées prises par $(\Delta X_{t'})_{t' \leq t-1}$.
- Puisque l'ordre MA est égal à 1, il n'y a aucune influence du bruit à l'horizon supérieur ou égal à 2 : sans aucune information supplémentaire, la seule prévision possible pour ΔX_{t+h} , $h \geq 2$, est une prévision nulle, et donc pour X_{t+h} la seule prévision possible est $\widehat{X}_{t+1|t}$. Cette incertitude se traduit par une région de confiance très large ⁶.

6. On prévoit une évolution mensuelle entre décembre 2019 et janvier 2020 comprise entre -6,8 % et +8,7 %, ce qui est très grand compte tenu de la volatilité de la série (l'écart-type de la série en évolution est de 4,1 et sa moyenne de 0,1).

La forme allongée et orientée à 45 degrés de la région de confiance reflète une certaine cohérence entre les prévisions de X_{T+1} et X_{T+2} (que l'on a pas quand on construit des intervalles de confiance de manière indépendante). En effet, plus $\hat{X}_{T+1|T}$ est grand, plus les valeurs « plausibles » de $\hat{X}_{T+2|T}$ sont élevées. Toutefois, le grand axe de l'ellipse n'est pas aligné à la première bissectrice : si $\hat{X}_{T+1|T}$ atteint sa plus grande valeur possible alors $\hat{X}_{T+2|T} < \hat{X}_{T+1|T}$ (il y a dans ce cas un contrecoup économique).

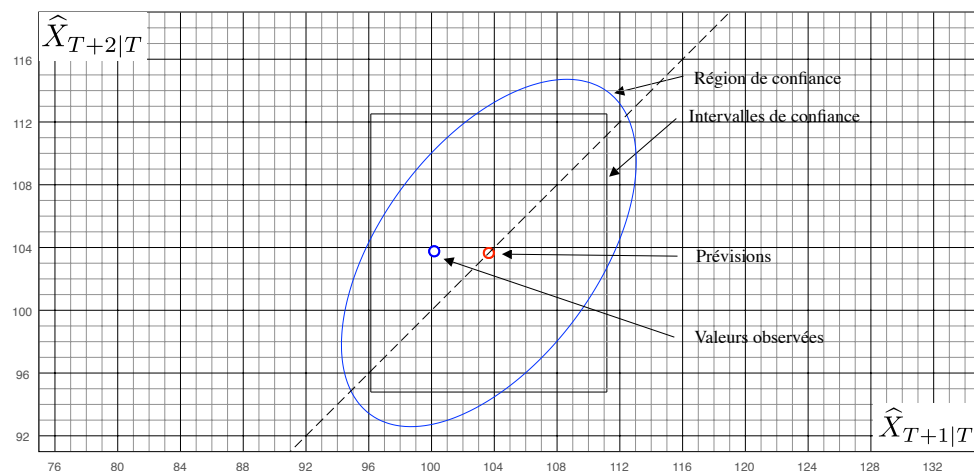


FIGURE 3 – Région de confiance pour la prévision de l'IPI automobile CVS-CJO pour janvier et février 2020 par un modèle ARIMA(0,1,1)

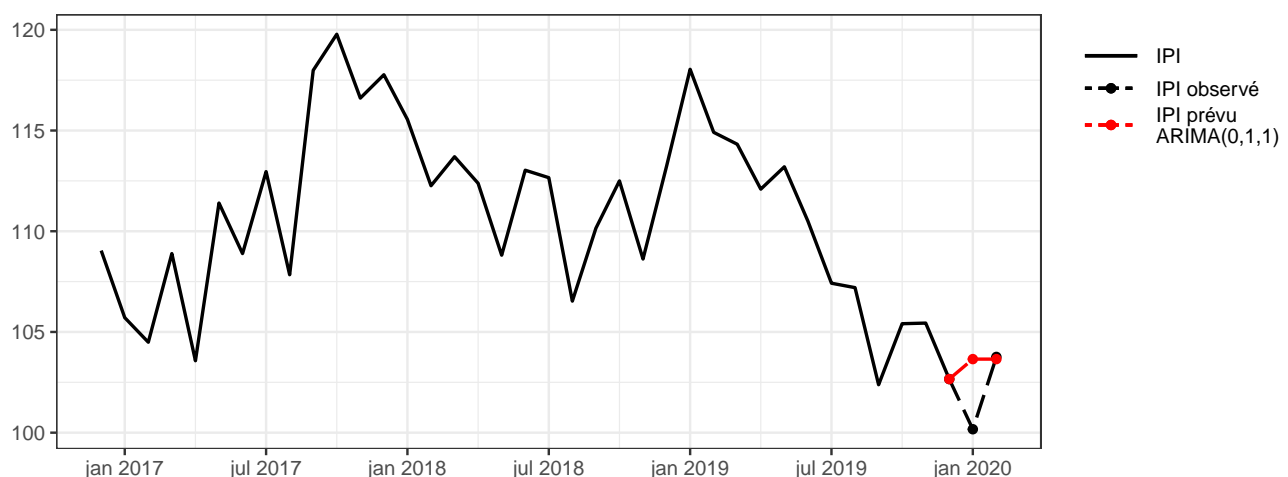


FIGURE 4 – Prévisions de l'IPI automobile CVS-CJO pour janvier et février 2020 par un modèle ARIMA(0,1,1).

3.2 Question 9 : question ouverte sur la causalité

Soit Y_t une série stationnaire disponible de $t = 1$ à T telle que Y_{T+1} est disponible plus rapidement que X_{T+1} .

Y_t cause⁷ instantanément X_t au sens de Granger (C_{X-Y} ⁸) si et seulement si pour tout t , la valeur de Y_{t+1} permet d'améliorer la prévision de X_{t+1} . Ainsi, la causalité instantanée au sens de Granger est une condition suffisante pour que Y_{T+1} permette d'améliorer la prévision de X_{T+1} . Mathématiquement :

7. Cette notion de causalité n'est pas la même que celle utilisée en économétrie classique (une influence directe de Y_t sur X_t), cela signifie simplement ici que Y_t est utile pour prévoir X_t .

8. La relation de causalité instantanée de Granger est symétrique : $C_{X-Y} \iff C_{Y-X}$.

$$\underbrace{\forall t \quad \widehat{X}_{t+1|\{Y_u, X_u, u \leq t\} \cup \{Y_{t+1}\}} \neq \widehat{X}_{t+1|\{Y_u, X_u, u \leq t\}}}_{C_{X-Y} : Y_t \text{ cause instantanément au sens de Granger } X_t} \xRightarrow{\text{en particulier}} \widehat{X}_{T+1|\{Y_{1...T}, X_{1...T}\} \cup \{Y_{T+1}\}} \neq \widehat{X}_{T+1|\{Y_{1...T}, X_{1...T}\}}$$

Cette définition de causalité instantanée est fondée sur les corrélations entre les erreurs. En effet, si $(Y'_t, X'_t)'$ est bien **stationnaire** alors on a montré dans le cours que :

$$C_{X-Y} \iff Cov(\epsilon_{1t}, \epsilon_{2t}) = 0 \quad \text{avec} \quad \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix} = \begin{pmatrix} Y_t - \widehat{Y}_{t|\{Y_u, X_u, u \leq t-1\}} \\ X_t - \widehat{X}_{t|\{Y_u, X_u, u \leq t-1\}} \end{pmatrix}$$

En prenant bien le soin de choisir X_t et Y_t de manière à ce que $(Y'_t, X'_t)'$ soit stationnaire, vérifier que Y_{T+1} permet d'améliorer la prévision de X_{T+1} revient alors à tester la condition suivante après avoir estimé les modèles correspondants à l'évolution de X_t et Y_t :

$$Cov(Y_{T+1} - \widehat{Y}_{T+1|\{Y_{1...T}, X_{1...T}\}}, X_{T+1} - \widehat{X}_{T+1|\{Y_{1...T}, X_{1...T}\}}) = 0$$

En notant X_t notre série de l'IPI de l'industrie automobile stationnarisée, il faudrait par exemple prendre pour Y_t des données susceptibles d'être disponibles avant X_t et feraient de bons candidats pour prévoir l'évolution de l'IPI dans l'industrie automobile :

- un IPI correspondant à une composante de la division « industrie automobile » (29)⁹ qu'il faudrait différencier autant de fois que nécessaire pour qu'elle soit bien stationnaire. Certaines composantes peuvent en effet disponibles être avant leur agrégat.
- une série stationnaire issue d'une autre enquête qui donnerait de l'information sur la production de l'IPI de l'industrie automobile. Cela peut par exemple être le cas des résultats des enquêtes de conjoncture, de l'Insee ou de la Banque de France, auprès des entreprises de ce secteur. Dans ces enquêtes on demande en effet l'opinion des chefs d'entreprise sur l'évolution passée et future de leur production : ces informations sont qualitatives et sont donc connues bien avant les informations quantitatives demandées par l'enquête mensuelle de branche.

Il peut également exister des cas où Y_{T+1} permet d'améliorer la prévision de X_{T+1} mais sans que Y_t cause instantanément X_t .

Prenons, par exemple, pour X_T la série stationnarisée de l'indice de production industrielle dans la cokéfaction-raffinage. Cette série est très bruitée car en France la production de ce poste est concentrée en une dizaine de raffineries. Ainsi, il est très difficile d'avoir une prévision précise de cette série qui sera très sensible, par exemple, à l'arrêt d'une raffinerie pendant quelques jours. En revanche, si on dispose d'un indicateur Y_t du nombre de jours de fermeture dans le mois, cette série peut, dans certains cas, permettre d'améliorer de manière conséquente la prévision de X_t . C'est par exemple le cas pendant les périodes de grèves : l'analyse de l'évolution de la production pendant les grèves passées permet d'estimer l'effet moyen d'un jour de grève, et cela permet donc d'estimer l'évolution future de la production pendant les mouvements sociaux futurs.

Si en $T+1$ il y a des grèves dans les raffineries, Y_{T+1} est connu bien avant X_{T+1} et permet d'améliorer la prévision de X_{T+1} .

Puisque les fermetures de raffineries sont rares (elles sont ouvertes tous les jours de la semaine), pour la majorité des périodes t , Y_t ne permet pas d'améliorer la prévision de X_t : elle ne cause donc a priori pas instantanément X_t au sens de Granger.

9. Par exemple une des classes parmi la construction de véhicules automobiles (29.1), la fabrication de carrosseries et remorques (29.2) ou les équipements automobiles (29.3).

A Tests supplémentaires sur la qualité des modèles

TABLE 4 – Tests de Ljung-Box sur les résidus (tests d'autocorrélation) des modèles ARIMA sur l'IPI de l'automobile.

Retards	ARIMA(0,1,0)			ARIMA(1,1,0)			ARIMA(0,1,1)		ARIMA(1,1,1)	
	Statistique	p-valeur		Statistique	p-valeur		Statistique	p-valeur	Statistique	p-valeur
1	9,652	0,002	**							
2	10,473	0,005	**	4,757	0,029	*	0,858	0,354		
3	10,580	0,014	*	4,796	0,091	.	0,925	0,630	0,106	0,744
4	12,843	0,012	*	6,191	0,103		1,985	0,576	1,544	0,462
5	15,122	0,010	**	7,239	0,124		3,024	0,554	2,606	0,456
6	16,041	0,014	*	7,327	0,197		3,165	0,675	2,907	0,573
7	17,341	0,015	*	7,800	0,253		3,566	0,735	3,389	0,640
8	18,808	0,016	*	8,936	0,257		4,884	0,674	4,676	0,586
9	18,809	0,027	*	9,306	0,317		5,129	0,744	4,781	0,687
10	19,645	0,033	*	9,610	0,383		5,338	0,804	5,037	0,754
11	22,433	0,021	*	12,895	0,230		8,684	0,562	8,103	0,524
12	22,609	0,031	*	13,822	0,243		9,649	0,562	8,787	0,552
13	22,610	0,047	*	13,838	0,311		9,649	0,647	8,787	0,642
14	23,078	0,059	.	14,496	0,340		10,455	0,656	9,497	0,660
15	23,082	0,082	.	14,860	0,388		11,003	0,686	9,871	0,704
16	24,077	0,088	.	15,947	0,386		12,232	0,661	11,074	0,680
17	24,339	0,111		16,263	0,435		12,411	0,715	11,227	0,736
18	24,935	0,127		16,911	0,460		12,993	0,737	11,795	0,758
19	25,633	0,141		17,430	0,494		13,173	0,781	11,966	0,802
20	25,649	0,178		17,579	0,551		13,432	0,816	12,208	0,836
21	28,080	0,138		20,078	0,453		15,782	0,730	14,643	0,745
22	28,992	0,145		20,694	0,478		16,064	0,766	14,878	0,783
23	29,209	0,173		20,920	0,526		16,146	0,809	14,914	0,827
24	30,464	0,170		21,907	0,526		17,136	0,803	16,104	0,811

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

L'hypothèse (H0) d'indépendance des résidus n'est pas rejetée à 5 % pour le modèle retenu ARIMA(0,1,1).

TABLE 5 – Estimation des coefficients associés aux modèles ARIMA sur l'IPI de l'automobile.

	AR(1)				MA(1)			
	Coefficient	Écart-type	p-valeur		Coefficient	Écart-type	p-valeur	
ARIMA(0,1,0)								
ARIMA(1,1,0)	-0,280	0,088	0,001	**				
ARIMA(0,1,1)					-0,377	0,091	0,000	***
ARIMA(1,1,1)	0,165	0,214	0,439		-0,515	0,183	0,005	**

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

TABLE 6 – Tests de Ljung-Box sur le carré des résidus (tests d’homoscédasticité) des modèles ARIMA sur l’IPI de l’automobile.

Retards	ARIMA(0,1,0)		ARIMA(1,1,0)		ARIMA(0,1,1)		ARIMA(1,1,1)	
	Statistique	p-valeur	Statistique	p-valeur	Statistique	p-valeur	Statistique	p-valeur
1	2,832	0,092						
2	2,843	0,241	2,917	0,088	2,032	0,154		
3	2,860	0,414	3,844	0,146	3,569	0,168	3,140	0,076
4	3,227	0,521	5,425	0,143	4,164	0,244	3,575	0,167
5	3,233	0,664	5,448	0,244	4,183	0,382	3,576	0,311
6	3,262	0,775	7,479	0,187	6,916	0,227	5,513	0,239
7	3,263	0,860	7,836	0,250	7,515	0,276	6,127	0,294
8	3,270	0,916	8,294	0,307	7,781	0,352	6,290	0,392
9	3,772	0,926	8,613	0,376	7,787	0,455	6,314	0,504
10	3,797	0,956	8,727	0,463	7,838	0,551	6,605	0,580
11	5,286	0,917	9,343	0,500	8,071	0,622	6,897	0,648
12	5,482	0,940	10,431	0,492	9,136	0,609	7,987	0,630
13	5,619	0,959	11,058	0,524	9,537	0,656	8,373	0,680
14	7,532	0,912	12,221	0,510	10,714	0,635	9,807	0,633
15	7,727	0,934	12,252	0,586	10,784	0,703	9,807	0,710
16	7,760	0,956	12,256	0,660	10,802	0,766	9,823	0,775
17	7,866	0,969	12,268	0,725	10,968	0,811	9,890	0,827
18	11,403	0,876	12,791	0,750	11,573	0,825	11,468	0,780
19	11,440	0,908	12,791	0,804	11,788	0,858	11,709	0,817
20	12,066	0,914	13,115	0,833	12,156	0,879	12,086	0,843
21	12,213	0,934	14,050	0,828	13,228	0,867	12,984	0,839
22	13,751	0,910	15,070	0,819	14,231	0,859	14,364	0,812
23	14,907	0,898	15,242	0,852	14,284	0,891	14,389	0,852
24	15,944	0,890	16,450	0,835	16,650	0,826	16,688	0,780

Signif. codes : 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

L’hypothèse (H0) d’indépendance des résidus n’est pas rejetée à 5 % pour le modèle retenu ARIMA(0,1,1).

TABLE 7 – Tests de Jarque-Bera de normalité des résidus des modèles ARIMA sur l’IPI de l’automobile.

	Statistique	p-valeur
ARIMA(0,1,0)	2,381	0,304
ARIMA(1,1,0)	2,414	0,299
ARIMA(0,1,1)	2,363	0,307
ARIMA(1,1,1)	2,241	0,326

Signif. codes : 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Le test de Jarque-Bera suppose que les résidus soient indépendants et homoscédastiques.

L’hypothèse (H0) de normalité des résidus n’est pas rejetée à 5 % pour l’ensemble des modèles et en particulier pour le modèle retenu ARIMA(0,1,1).

B Code

L'ensemble du code a été écrit avec l'encodage UTF-8.

B.1 Fichier 0 - Creation des donnees.R

Code utilisé pour télécharger les données : non utile pour la suite puisque les données sont jointes au projet.

```
# Codes pour télécharger les séries :  
# il n'est pas nécessaire de le relancer puisqu'elles  
# sont toutes dans le dossier data/  
  
# devtools::install_github("aqlt/AQLTools")  
library(AQLTools)  
library(zoo)  
  
# CL1 = automobile  
  
ipi_cl1 <- AQLTools::lectureBDM("010537940")  
ipi_cl1_brut <- AQLTools::lectureBDM("010537939")  
data_b <- ts.union(ipi_cl1, ipi_cl1_brut)  
data_2010 <- window(data_b,  
                     start = c(2010, 1),  
                     end = c(2019,12))  
  
saveRDS(data_b,  
         file = "data/donnees_completes.RDS")  
saveRDS(data_2010,  
         file = "data/donnees.RDS")  
  
# Exporter en CSV : non utile pour lancer les programmes mais demandé par les consignes  
write.csv(data.frame(date = format(as.yearmon(time(data_2010))), "%m/%Y"),  
          data_2010),  
          row.names = FALSE,  
          file = "data/donnees.csv")  
  
# Pour tracer le graphique avec ggplot2  
#AQLTools::graph_ts(window(data,start = 2005))
```

B.2 Fichier 1 - Stationnarisation.R

```
library(urca)  
library(fUnitRoots)  
# devtools::install_github("aqlt/AQLTools")  
library(AQLTools) # utilisé pour tracer les séries  
library(patchwork) # pour mettre à côté deux graphiques ggplot2  
  
data <- readRDS(file = "data/donnees.RDS")  
# data <- ts(read.csv("data/donnees.csv"),[-1],  
#           start = 2010, frequency = 12)  
  
x <- data[, "ipi_cl1"]  
p1 <- AQLTools::graph_ts(window(x,  
                               start = c(2009,10),
```

```

        end = c(2020,2),
        extend = TRUE), x_lab = "Dates", y_lab = NULL,
        titre = "IPI-CL1 (sans traitement)", n_xlabel = 6)

p1

summary(lm(x ~ time(x)))

# Même si on observe une tendance dans la régression de la série
# par rapport au temps, étant donné la rupture de tendance, nous
# considérons, comme dans les TD, qu'il y a ici pas de tendance déterministe et
# une moyenne non nulle.
# => On fait le test ADF AVEC constante et SANS tendance
# Pour que le test soit valide il faut rajouter des retards :
# On fait donc le test jusqu'à ce que les résidus du modèles de "ADF" soient bons :
# que les résidus soient indépendants (on ne veut plus d'endogénéité dû aux variables omises)

# Cette fonction permet de faire les tests d'indépendance des résidus du modèle ADF
# en fonction du lag
lb_test <- function(x, lag_max = 24, fitdf = 0){
  t(sapply(seq_len(lag_max),function(l){
    if(l <= fitdf){
      b <- list(statistic = NA, p.value = NA)
    }else{
      b <- Box.test(x,"Ljung-Box",lag = l,
                     fitdf = fitdf)
    }
    data.frame(lag = l,
               b$statistic,
               b$p.value)
  })
)

# Cette fonction a le même objectif que la fonction précédente
# (tests d'indépendance en fonction du lag)
# mais correspond à celle en corrigé des TD.
Qtests <- function(series, k = 24, fitdf=0) {
  pvals <- apply(matrix(1:k), 1, FUN=function(l) {
    pval <- if (l<=fitdf) NA else Box.test(series, lag=l, type="Ljung-Box", fitdf=fitdf)$p.value
    return(c("lag"=l,"pval"=pval))
  })
  return(t(pvals))
}

# tests ADF jusqu'à ce que les résidus ne soient pas autocorrélés
adfTest_valid <- function(series, kmax,type){
  k <- 0
  noautocorr <- 0
  while (noautocorr==0){
    cat(paste0("ADF with ",k, " lags: residuals OK? "))
    adf <- adfTest(series,lags=k,type=type)
    pvals <- Qtests(adf@test$lm$residuals,24,fitdf=length(adf@test$lm$coefficients))[,2]
  }
}

```

```

        if (sum(pvals<0.05,na.rm=T) == 0) {
            noautocorr <- 1; cat("OK \n")
        } else cat("nope \n")
        k <- k + 1
    }
    return(adf)
}

adfTest_valid(x, kmax = 20, type = "c") #juste constante et pas de tendance
# On trouve un lag de 2
adf <- adfTest(x, type = "c",lags = 2) # juste constante et pas de tendance
adf # on ne rejette pas à 5 % : série non stationnaire avec une racine unitaire

# vérification tests d'indépendance des résidus du modèle ADF
# en fonction du lag (aussi vérifié dans adfTest_valid mais pour bien vérifier)
lb_test(adf@test$lm$residuals, fitdf=length(adf@test$lm$coefficients))
# les p-valeurs sont bien toutes supérieures à 0,05 :
# OK indépendance des résidus (le test est valide)

# PP et kpss donnent des résultats similaires
PP.test(x) # on ne rejette pas à 5 % : série non stationnaire avec une racine unitaire
tseries::kpss.test(x) # on rejette à 5 % : série non stationnaire

# On différencie la série pour la stationnariser :
x_st <- diff(x, 1)
# On trace la série différenciée.
AQLTools::graph_ts(window(x_st,
                        start = c(2009,10),
                        end = c(2020,2),
                        extend = TRUE), x_lab = "Dates", y_lab = NULL,
                    titre = "IPI-CL1 différenciée", n_xlabel = 12)
summary(lm(x_st ~ time(x_st)))
# Série qui paraît stationnaire, sans tendance ni constante,
# confirmée par la régression en fonction du temps.

# On le vérifie avec un test ADF
adfTest_valid(x_st, kmax = 24, type = "nc") # test sans tendance ni constante
# Il faut donc utiliser un retard

adf <- adfTest(x_st, type = "nc",lags = 1)# test dans tendance ni constante
adf # on rejette à 5 % : pas de racine unitaire (série stationnaire)

# vérification tests d'indépendance
lb_test(adf@test$lm$residuals, fitdf=length(adf@test$lm$coefficients))
# les p-valeurs sont bien toutes supérieures à 0,05 :
# OK indépendance des résidus (le test est valide)

PP.test(x_st) # vérifié avec test de Phillips-Perron. On rejette à 5 % :
# pas de racine unitaire (série stationnaire)
tseries::kpss.test(x_st) # vérifié avec KPSS. On ne rejette pas à 5 % : série stationnaire

```

```

series_a_tracer <- ts.union(x, x_st)
p2 <- AQLTools::graph_ts(window(x_st,
                             start = c(2009,10),
                             end = c(2020,2),
                             extend = TRUE), x_lab = "Dates", y_lab = NULL,
                             titre = "IPI-CL1 (série différenciée)", n_xlabel = 6)

p1 + p2

saveRDS(x_st, file = "data/x_st.RDS")

```

B.3 Fichier 2 - Estimation du modele ARIMA.R

```

library(forecast)
library(patchwork) # pour mettre à coté deux graphiques ggplot2

data <- readRDS(file = "data/donnees.RDS")
# data <- ts(read.csv("data/donnees.csv"),[-1],
#           start = 2010, frequency = 12)

x <- data[,"ipi_cl1"]
x_st <- readRDS(file = "data/x_st.RDS")

# graphique des ACF
acf(x_st) # q_max = 1
# graphique des PACF
pacf(x_st) # p_max = 1

# Fonctions identiques du package forecast mais où on enlève lag = 0
# Permet d'éviter les confusions pour l'acf
Acf(x_st) # q_max = 1
Pacf(x_st) # p_max = 1

# Deux graphiques côte à côte
ggAcf(x_st) + labs(title = "ACF") +
  ggPacf(x_st) + labs(title = "PACF")

# On va donc tester tous les modèles pour q <= 1, p <= 1

# Grâce à la fonction evaluation_model, on repère les modèles possibles qui vérifient deux conditions :
# 1) tests d'indépendance des résidus de Ljung-Box
# 2) coefficients associés au qmax ET pmax sont bien significatifs
evaluation_model <- function(order, x, lags = 24, include.mean = TRUE){
  # ici on utilise Arima plutôt que arima pour la fonction accuracy
  model <- forecast::Arima(x, order = order,
                           include.mean = include.mean)
  residus <- residuals(model)
  # test d'indépendance
  lbtest <- t(sapply(1:lags,function(l){
    if(l <= sum(model$arima[1:2])){
      b <- list(statistic = NA, p.value = NA)
    }else{
      b <- Box.test(residus,"Ljung-Box",lag = l,

```

```

        fitdf = sum(model$arma[1:2])
    )
}
data.frame(lag = 1,
            b$statistic,
            b$p.value
)
)))
# test d'homoscédasticité
lb2test <- t(sapply(1:lags,function(l){
  if(l <= sum(model$arma[1:2])){
    b <- list(statistic = NA, p.value = NA)
  }else{
    b <- Box.test(residus^2,"Ljung-Box",lag = 1,
                  fitdf = sum(model$arma[1:2])
    )
  }
  data.frame(lag = 1,
              b$statistic,
              b$p.value
            )
}))
# test de normalité
jbtest <- tseries::jarque.bera.test(residus)
# test significativité des coefficients
ttest <- tryCatch(lmtest::coefest(model), error = function(e) 0)
qualite <- c(AIC(model), BIC(model), accuracy(model))
names(qualite) <- c("AIC", "BIC", colnames(accuracy(model)))
list(model = model,
      ttest = ttest,
      lbtest = lbtest, lb2test = lb2test,
      jbtest = jbtest,
      qualite = qualite)
}

models_possibles <- expand.grid(p = 0:1, d = 0, q = 0:1)
models_evalues <- apply(models_possibles,1, evaluation_model, x = x_st,
                        include.mean = FALSE)
names(models_evalues) <- sprintf("ARIMA(%i,%i,%i)", models_possibles[, "p"],
                                models_possibles[, "d"], models_possibles[, "q"])
saveRDS(models_evalues, file = "data/models_evalues.RDS")
## Pour éviter de tout écrire à la main :
#cat(paste(sprintf("models_evalues$`%s`",names(models_evalues)),collapse = "\n"))

models_evalues$`ARIMA(0,0,0)`
# Il n'y a pas indépendance des résidus : modèle non valide
models_evalues$`ARIMA(1,0,0)`
# Il n'y a pas indépendance des résidus : modèle non valide
models_evalues$`ARIMA(0,0,1)`
# Il y a indépendance des résidus et coefficient MA(1) significatif :
# modèle valide
models_evalues$`ARIMA(1,0,1)`
# coef AR1 non significatif : modèle non valide

```

```

# Bilan : seul modèle valide : ARIMA(0,1,1)

# On regarde par curiosité les critères d'informations
qualite_modeles <- sapply(models_evalues, function(x) x$qualite)
round(qualite_modeles,1)
# C'est également le modèle ARIMA(0,0,1) qui présente les meilleurs
# AIC et BIC (les plus petits)

ordres_retenus <- c(0,1,1) #sur la série initiale : d = 1 (on l'a différenciée une fois) et q=1 (MA(1))

saveRDS(ordres_retenus, file = "data/ordres_retenus.RDS")

model_estime <- arima(x, order = ordres_retenus)
model_estime
lmtest::coefest(model_estime) # coefficients significatifs
residus <- residuals(model_estime)

# On fait les tests d'indépendance des résidus d'un modèle ARIMA en fonction du lag
# (déjà vérifié dans evaluation_model mais pour bien vérifier)
lbttest <- t(sapply(1:24,function(l){
  if(l <= sum(model_estime$arima[1:2])){
    b <- list(statistic = NA, p.value = NA)
  }else{
    b <- Box.test(residus,"Ljung-Box",lag = l,
                  fitdf = sum(model_estime$arima[1:2]))
  }
  data.frame(lag = l,
             b$statistic,
             b$p.value
  )
}))
lbttest # résidus biens valides (p-valeur > 5 %)

ggAcf(residus) + labs(title = "ACF") +
ggPacf(residus) + labs(title = "PACF")
#Modèle bien valide : on remarque que aucun ordre de lag reste significativement non nul dans ACF/PACF

tseries::jarque.bera.test(residus) # on ne rejette pas à 5 %. Résidus normaux :
# on peut bien faire les ic

#Remarquons que le même modèle serait déterminé automatiquement avec
# la fonction auto.arima
m <- auto.arima(x)
m

```

B.4 Fichier 3 - Previsions.R

```

library(forecast)
library(patchwork) # pour mettre à côté deux graphiques ggplot2
library(conics) # pour tracer une ellipse

data <- readRDS(file = "data/donnees.RDS")

```

```

# data <- ts(read.csv("data/donnees.csv"),[-1],
#           start = 2010, frequency = 12)

data_complet <- readRDS(file = "data/donnees_completes.RDS")
x <- data[, "ipi_cl1"]
x_complet <- data_complet[, "ipi_cl1"]
ordres_retenus <- readRDS(file = "data/ordres_retenus.RDS")
model_estime <- Arima(x, order = ordres_retenus, include.constant = FALSE)
# Réalise les prévisions sur 2 périodes du modèle retenu
prev <- forecast(model_estime, h = 2)
prev
# On les représente sur un graphique.
plot(prev)
# Attention, ce sont les intervalles de confiance à chaque période
# (et non la région de confiance calculée ci-dessous) qui sont représentés automatiquement.

#Retrouver les IC du graphique :
res <- residuals(model_estime)
sum((res - mean(res))^2) / (length(res) - 2) # sigma2
prev$mean[1]+sqrt(model_estime$sigma2)*qnorm(1-0.05/2)
prev$mean[2]+sqrt(model_estime$sigma2*(1+(1+model_estime$coef[1])^2))*qnorm(1-0.05/2)

# Plutôt que de tracer les IC on veut une région de confiance :
sigma2 <- model_estime$sigma2
theta <- coef(model_estime)

sigma_m1 = matrix(c(1+(1+theta)^2, -(1+theta),
                    -(1+theta),1),ncol = 2)/sigma2
# Pour vérifier qu'on a bien fait l'inversion de la matrice :
# matlib::Inverse(sigma2 * matrix(c(1, (1+theta),
#                                   (1+theta),1+(1+theta)^2),ncol = 2))

alpha = 0.05

sigma_sur_quantile <- sigma_m1/(qchisq(1-alpha, 2))
prevs <- prev$mean

a = sigma_sur_quantile[1,1]
b = sigma_sur_quantile[1,2]
d = sigma_sur_quantile[2,2]
x_p = prevs[1]
y_p = prevs[2]
# coefficients de l'ellipse : a_1 à a_6
# a_1 * x^2 + a_2 * x * y + a_3 * y^2 + a_4 * x + a_5 * y + a_6 = 1
a_1 <- a
a_2 <- 2*b
a_3 <- d
a_4 <- -2*(a*x_p+b*y_p)
a_5 <- -2*(b*x_p+d*y_p)
a_6 <- a*x_p*x_p+2*b*x_p*y_p+d*y_p*y_p

ellipse_eq <- c(a_1, a_2, a_3, a_4,
                a_5, a_6)

```



```

eq <- paste(round(ellipse_eq, 3), c("x^2", "* x * y", "y^2", "x", "y", 1),
            collapse = " + ")
eq <- paste(gsub("+ -", "- ", eq, fixed = TRUE), "= 1")
eq <- gsub(" 1 ", " ", eq, fixed = TRUE)
eq
#equation latex :
cat(gsub(".", "",
          gsub("*", "\\times", eq, fixed = TRUE),
          fixed = TRUE))

# Pour tracer l'ellipse :
# il faut une equation sous la forme :
# a_1 * x^2 + a_2 * x * y + a_3 * y^2 + a_4 * x + a_5 * y + a_6 = 0

# On trace la prévision avec la région de confiance (ellipse) autour
ellipse <- conicPlot(ellipse_eq - c(0,0,0,0,0,1))
ellipse
points(prevs[1], prevs[2])

```