Rapport final

Paul GUILLOTTE & Jules CORBEL

01/02/2019

Contents

Introduction

Dans le cadre de la formation Génie Informatique et Statistique (GIS), l'un des modules qui nous est proposé est la réalisation d'un projet de fin d'études (PFE) en lien avec une entreprise ou un laboratoire. Notre PFE s'effectue en lien avec l'Institut des Retraites Complémentaires des Employés de Maison (IRCEM) et a pour but de construire des modèles statistiques afin de prédire la masse salariale du secteur d'activité des services à la personne pour les années à venir.

L'IRCEM a été créée en 1973. Cet organisme à but non lucratif s'occupe de la protection sociale des employés du secteur du service à la personne. Dans ce but, il doit verser des compléments de salaire aux employés à l'aide à la personne. Il effectue alors régulièrement des prévisions de la masse salariale de ce secteur d'activités, afin d'estimer l'argent qu'il devra verser.

Notre mission pour ce projet est donc de modéliser cette masse salariale et d'effectuer des prévisions pour les années 2018 et 2019, qui seront effectuées en utilisant différentes méthodes de prédiction. La première partie du projet nous a vus nous concentrer sur des méthodes de prédiction univariées, tel que le lissage exponentiel et des modélisations basées sur des processus ARMA. Nous avions également commencé à modéliser la masse salariale à l'aide des variables auxiliaires. Ici, nous présentons nos travaux sur les modèles vectoriels (modèles VAR plus des essais sur les modèles VARMA). L'intégralité du projet s'est effectuée sur le logiciel R.

1 Description des jeux de données

Les données que nous a fourni l'IRCEM sont comprises dans deux jeux de données représentant deux ensembles de variables distincts : l'un contient des variables annuelles et l'autre des variables trimestrielles, dans les deux cas à partir de 1990. Afin d'aider à la prédiction des valeurs de la masse salariale, nous devrons nous appuyer sur plusieurs variables auxiliaires. Pour les données annuelles, nous disposons de 4 variables : le SMIC horaire brut, le PIB, le taux de chômage et le montant de l'Allocation pour la Garde des Enfants à Domicile. La masse salariale annuelle est connue jusqu'en 2017, et on possède les informations sur les autres variables jusqu'à 2019. En ce qui concerne le modèle trimestriel, nous disposons de 3 variables : le SMIC, le PIB et le taux de chômage des femmes. La masse salariale trimestrielle est connue jusqu'au 2e trimestre de 2017. Le PIB trimestriel est lui connu jusqu'au 1er trimestre 2017. Pour les deux autres variables, les informations que nous possédons vont jusqu'au dernier trimestre de 2017. Le faible nombre de données (surtout pour les variables annuelles) pourra cependant être un obstacle. En effet, dans le jeu annuel, il n'y a 28 années, ce qui représente peu de valeurs pour créer des modèles pertinents.

2 Analyse descriptive des séries

2.1 Rappel sur la stationnarité du second ordre

Avant de commencer à analyser les séries, nous rappelons des bases sur des notions dont nous aurons besoin par la suite.

Dans de nombreux modèles de séries temporelles, la série en entrée doit satisfaire une hypothèse de stationnarité. Les conditions de la stationnarité du second ordre sont les suivantes :

$E[y_t] = \mu \forall t = 1...T$

$Var[y_t] = \sigma^2 \neq \infty \forall t = 1...t$

$Cov[y_i, Z_{i-k}] = f(k) \forall i = 1...t, \forall k = 1...t$

Nous nous intéressons dans ce rapport aux différentes séries trimestrielles à notre disposition. Dans un premier temps, nous nous intéressons aux corrélations entre les variables deux à deux afin de nous faire une première idée du lien qu'il existe entre les variables.

2.2 Masse salariale

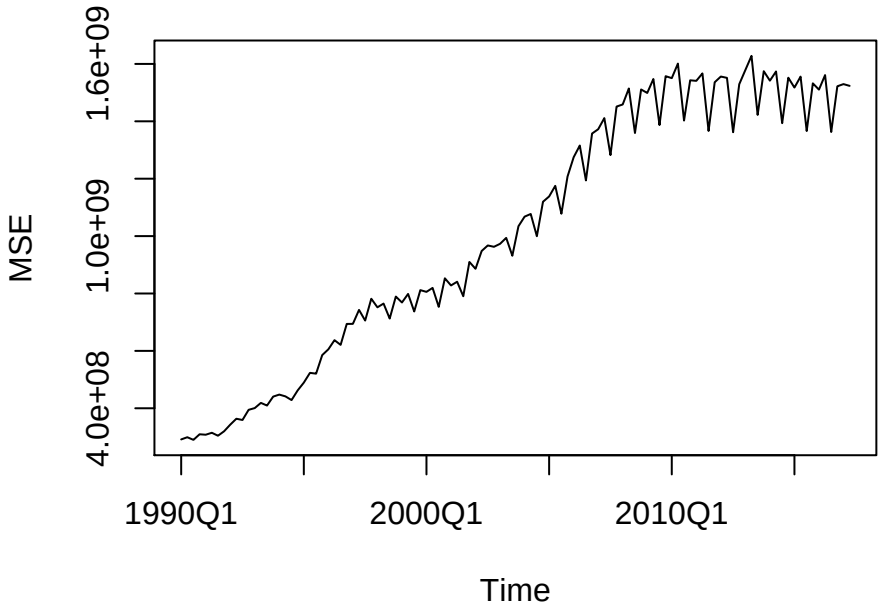


Figure 1: Evolution trimestrielle de la masse salariale

La masse salariale trimestrielle, représentée en Figure 1 possède une composante de tendance de 1990 à 2010. La série tend par la suite à stagner. Nous remarquons également une saisonnalité sur

cette série, qui est de plus en plus marquée à mesure que le temps passe.

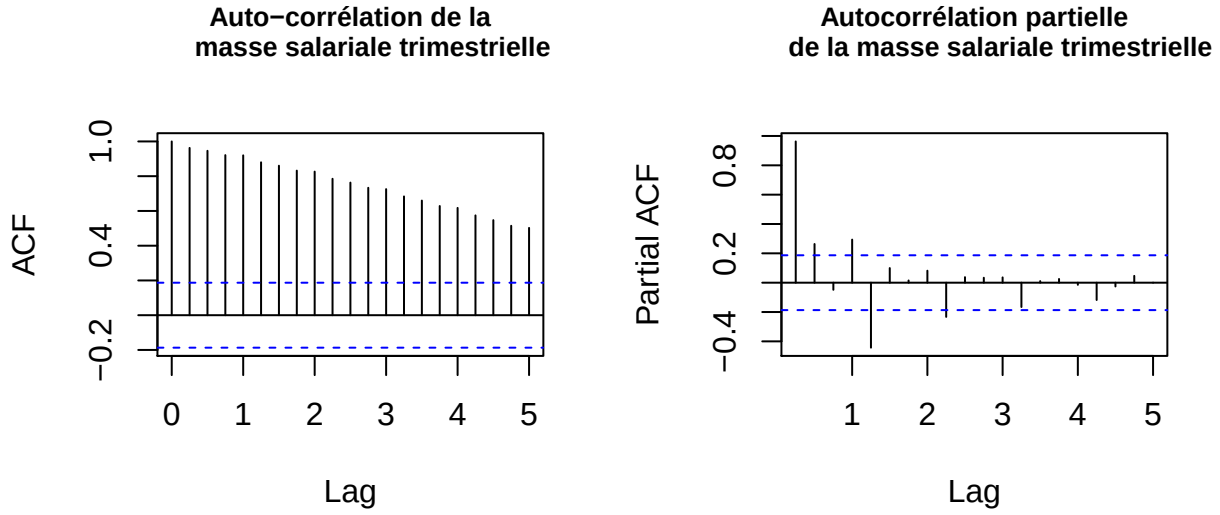


Figure 2: Fonctions d'autocorrélation de la masse salariale trimestrielle

```
## Warning in kpss.test(MSE): p-value smaller than printed p-value

##
##  KPSS Test for Level Stationarity
##
## data:  MSE
## KPSS Level = 3.6772, Truncation lag parameter = 2, p-value = 0.01

## Warning in adf.test(MSE): p-value greater than printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  MSE
## Dickey-Fuller = -0.20821, Lag order = 4, p-value = 0.99
## alternative hypothesis: stationary
```

Comme la série comporte une tendance et une saisonnalité, elle ne correspond pas aux deux premières conditions de la stationnarité du second ordre, soit que la série possède une moyenne et un écart-type constants. Cela est confirmé par la Figure 2, qui nous montre fonction ACF qui décroît régulièrement. Nous effectuons également un test de KPSS (test de stationnarité) servant à vérifier si la série est stationnaire ou non (sous l'hypothèse H_0 la série est stationnaire, et sous l'hypothèse H_1 elle ne l'est pas). La série est dite stationnaire si ses propriétés statistiques (espérance, variance et auto-corrélation) sont fixes au cours du temps. La p-value est de 0.01 ce qui nous confirme que la série n'est pas stationnaire avec un risque de première espèce de 5%. Nous mettons également en place un test de racines unitaires, le test de Dickey Fuller augmenté. Son hypothèse nulle est que la série a été générée par un processus présentant une racine unitaire, et donc que la série n'est pas stationnaire. Ici, avec un risque de première espèce à 5%, on conserve l'hypothèse nulle et on

conclut, à l'aide des deux tests effectués, que la série n'est pas stationnaire.

2.3 PIB

La Figure 3 nous montre l'évolution trimestriel du PIB qui, comme pour la masse salariale possède une tendance. Cependant, elle ne semble pas posséder de saisonnalité. Cette série ne semble donc pas non plus stationnaire. Nous effectuons à nouveau un test de KPSS. La p-value est de 0.01 ce qui nous confirme que la série n'est pas stationnaire avec un risque de première espèce de 5%. Même conclusion au regard du test augmenté de Dickey Fuller.

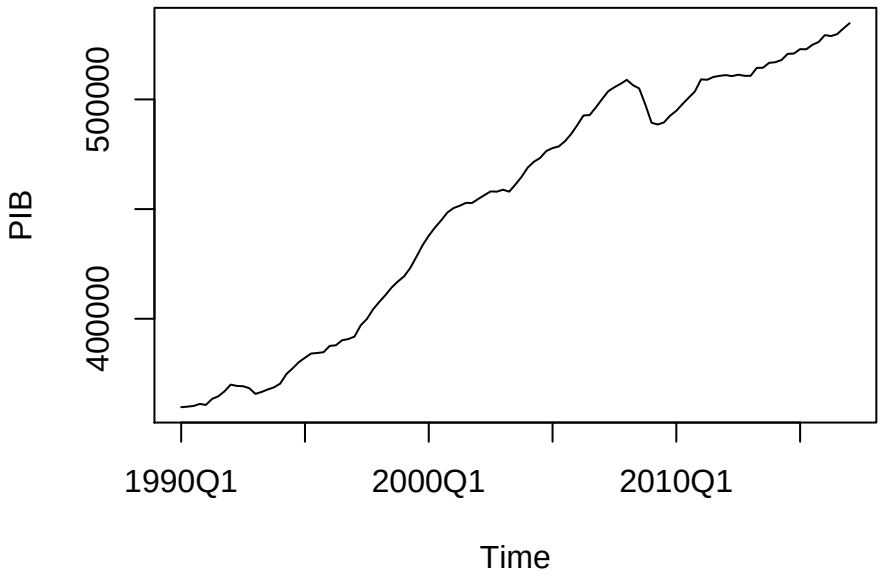


Figure 3: Evolution trimestrielle du PIB

```
## Warning in kpss.test(PIB): p-value smaller than printed p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  PIB
## KPSS Level = 3.6473, Truncation lag parameter = 2, p-value = 0.01
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  PIB
## Dickey-Fuller = -1.3274, Lag order = 4, p-value = 0.8557
## alternative hypothesis: stationary
```

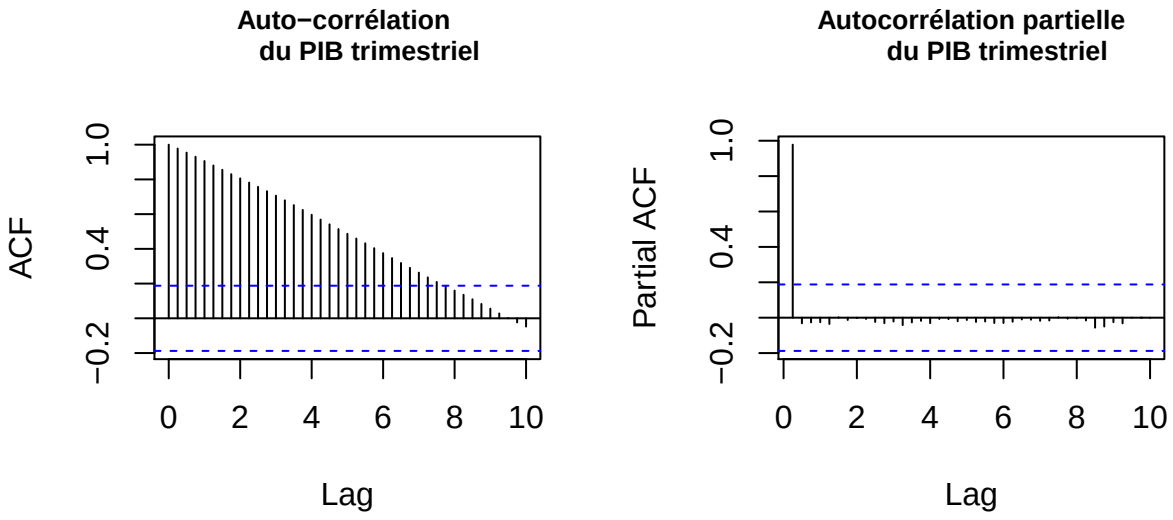


Figure 4: ACF et PACF du PIB trimestriel

2.4 SMIC

Au regard de la Figure 5, on s'aperçoit qu'il y a bien une tendance. Pour la saisonnalité, il est plus difficile de savoir s'il en existe une ou pas, puisque la série semble augmenter seulement à certains temps. Les tests de KPSS et de Dickey Fuller augmenté nous confirment que la série n'est pas stationnaire.

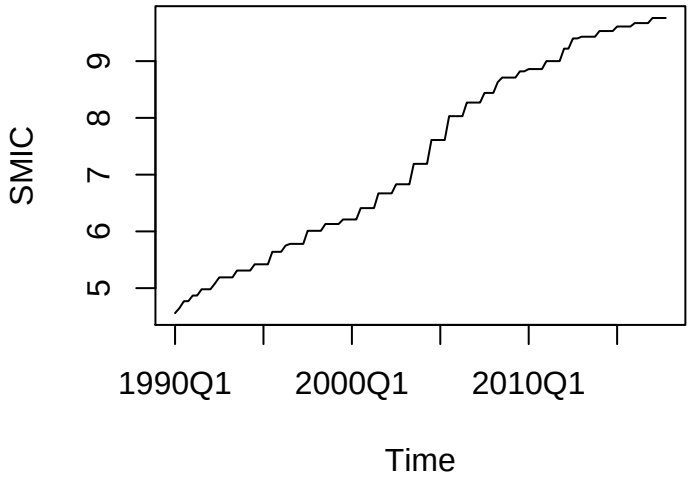


Figure 5: Evolution trimestrielle du SMIC

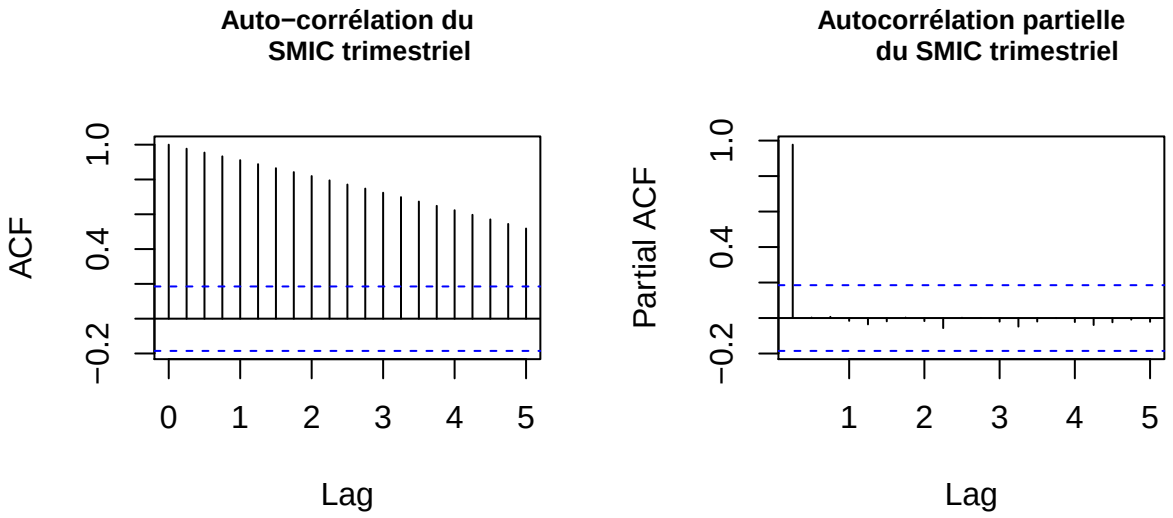


Figure 6: ACF et PACF du SMIC trimestriel

```
## Warning in kpss.test(SMIC): p-value smaller than printed p-value

##
##  KPSS Test for Level Stationarity
##
## data:  SMIC
## KPSS Level = 3.8382, Truncation lag parameter = 2, p-value = 0.01

##
##  Augmented Dickey-Fuller Test
##
## data:  SMIC
## Dickey-Fuller = -1.4174, Lag order = 4, p-value = 0.8184
## alternative hypothesis: stationary
```

2.5 Taux de chômage des femmes

Pour cette dernière série (Figure 7) qui représente le taux de chômage trimestriel des femmes, il ne semble pas y avoir de saisonnalité. On remarque cependant qu'il y a bien une tendance, au regard de la Figure 8. En regardant la série de plus près, on s'aperçoit que la tendance semble être "par morceaux" : d'abord une hausse de 1990 à 1996, puis elle décroît jusqu'en 2002, avant d'augmenter à nouveau jusqu'en 2007, de chuter jusqu'en 2010. Si la série ne possède pas une tendance uniforme sur toute la durée étudiée, elle semble donc bien posséder une tendance par morceaux. Les tests KPSS et de Dickey Fuller augmenté nous confirment que la série n'est pas stationnaire, avec un risque de première espèce de 5%.

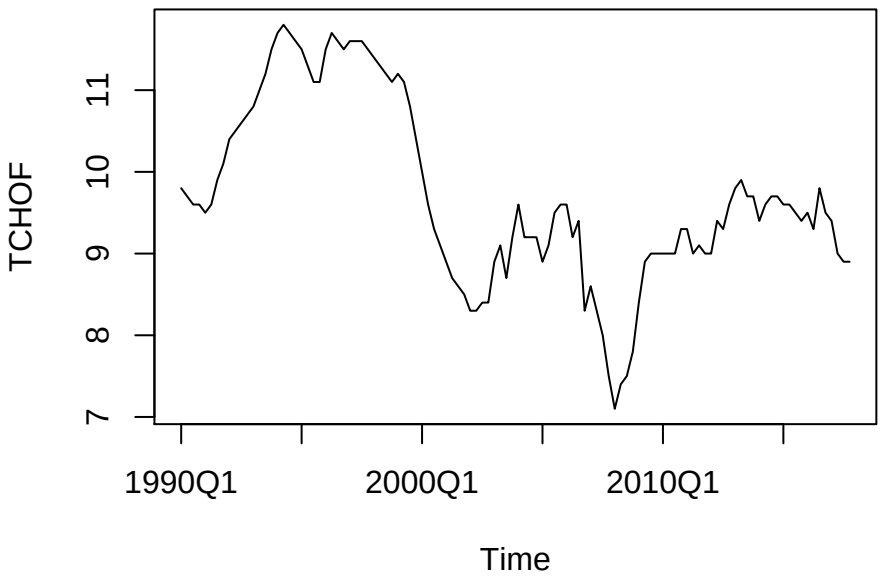


Figure 7: Evolution trimestrielle du taux de chômage des femmes

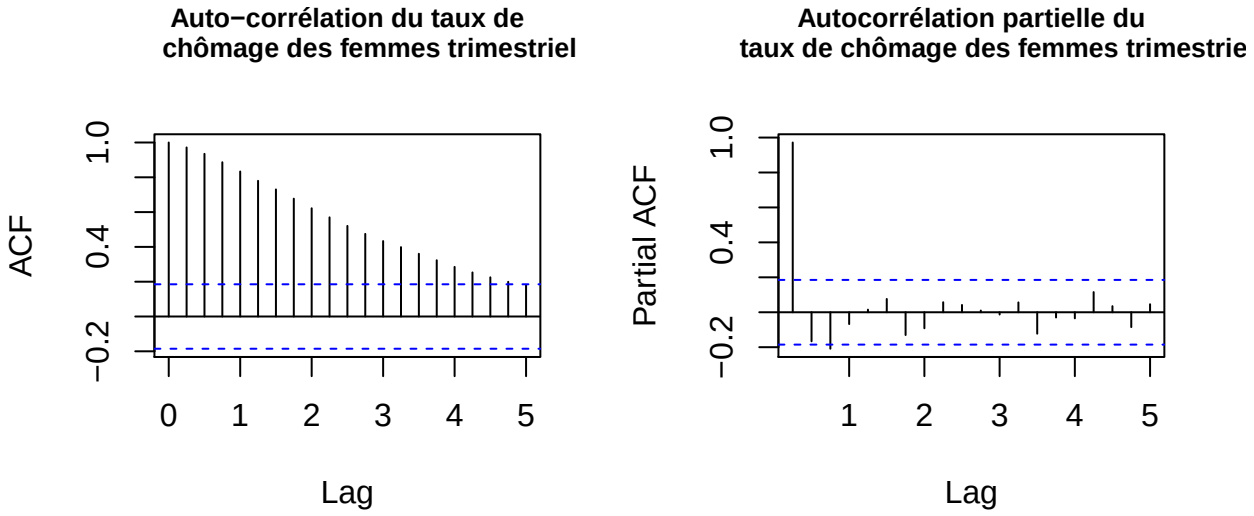


Figure 8: ACF et PACF du taux de chômage des femmes trimestriel

```
## Warning in kpss.test(TCHOF): p-value smaller than printed p-value

##
##  KPSS Test for Level Stationarity
```

```
##
## data:  TCHOF
## KPSS Level = 1.6407, Truncation lag parameter = 2, p-value = 0.01

##
##  Augmented Dickey-Fuller Test
##
## data:  TCHOF
## Dickey-Fuller = -2.5838, Lag order = 4, p-value = 0.3344
## alternative hypothesis: stationary
```

2.6 Calcul des corrélations

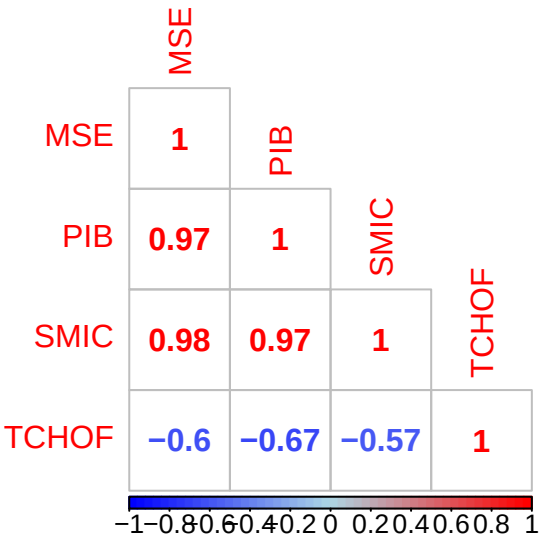


Figure 9: Corrélations entre les variables trimestrielles

```
##                MSE          PIB          SMIC        TCHOF
## MSE   0.000000e+00 3.851955e-69 1.436967e-74 3.321841e-12
## PIB   3.851955e-69 0.000000e+00 1.898200e-71 2.387179e-15
## SMIC  1.436967e-74 1.898200e-71 0.000000e+00 1.377731e-10
## TCHOF 3.321841e-12 2.387179e-15 1.377731e-10 0.000000e+00
```

Nous affichons la matrice des corrélations des différentes variables en Figure 9. On se rend compte que le taux de chômage des femmes est corrélé négativement avec toutes les autres variables. Les variables PIB, masse salariale et SMIC sont extrêmement liées entre elles. En regardant le tableau des p-values associées au test de Student (H0 : La corrélation entre les deux variables est nulle), on s'aperçoit que toutes les variables prises deux à deux présentes une corrélation.

2.7 Découpage des séries

Pour chacune des séries, nous allons créer un échantillon d'apprentissage, qui nous permettra de construire les différents modèles, ainsi qu'un échantillon de test, qui nous permettra de comparer les prédictions des modèles construits avec des vraies valeurs. L'échantillon d'apprentissage sera composé de toutes les valeurs du premier trimestre 1990 jusqu'au 4e trimestre 2015, tandis que celui de test comprendra toutes les valeurs à partir du 1er trimestre 2016.

2.8 Estimation de la valeur manquante du PIB

Contrairement aux autres variables, nous n'avons à notre disposition pour le PIB que les valeurs jusqu'au premier trimestre de 2017. Ceci nous impose de négliger la dernière valeur de toutes les autres séries pour que toutes les variables soient étudiées sur la même période. Pour éviter ce problème, nous décidons d'estimer la variable du PIB pour le 2e trimestre de 2017. Afin de faire cela, nous allons utiliser la valeur estimée par un modèle SARIMA correspondant à la variable PIB, étant donné que c'est celui qui donnait les meilleures prédictions (voir A).

3 Modélisation vectorielle

Dans la première partie du projet, nous avons modélisé les différentes séries séparément, à l'aide de modèles de lissage exponentiel ou des processus ARMA, dont les résultats sont présents en A. Nous avons ensuite effectué d'autres modèles ARMA sur la masse salariale en uilisant les autres variables pour l'expliquer. Ces modèles sont présents en A.2. Cependant, ce type de modélisation ne nous donne pas de résultats plus performants en terme de prédictions.

Pendant la deuxième partie du projet, nous nous sommes donc attachés à utiliser d'autres méthodes de modélisation. Nous nous sommes concentrés sur les modèles de type vectoriels, qui permettent donc de prédire plusieurs séries temporelles simultanément. La plus grande partie de nos travaux portent sur des modèles Vector Auto-Regressive (VAR) plus quelques tests avec l'ajout d'une partie Moving Average(MA) ce qui nous donne des modèles VARMA.

3.1 Définition des modèles

3.1.1 Ecriture

Un modèle VAR s'écrit sous la forme suivante :

$y_t = \sum_{i=1}^{p} A_i y_{t-i} + u_t$

A_i représentent les matrices de coefficients du modèle pour un ordre i et u_t une matrice K-dimensionnelle composée des résidus du modèle (indépendants et identiquement distribués). Enfin, p correspond à l'ordre du modèle, qui est en fait le nombre de valeurs du passé prises en compte pour calculer la valeur présente.

3.1.2 Hypothèses

3.1.2.1 Stabilité du modèle

Pour que le modèle soit valide, nous devons vérifier que l'hypothèse de stabilité est bien respecté. Cette dernière permet d'assurer que les différentes séries du modèle sont stationnaires. Pour vérifier si un processus VAR est stable, nous devons calculer les valeurs propres de la matrice des coefficients suivantes :

$$A = \begin{bmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I & 0 & \cdots & 0 & 0 \\ 0 & I & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I & 0 \end{bmatrix}$$

Si les modules des valeurs propres de A sont inférieures à 1, alors le processus VAR est stable.

3.1.2.2 Hypothèses sur les résidus

Pour que le modèle soit valide, certaines conditions sur les résidus doivent également être validées. Il s'agit des suivantes :

- Homoscédasticité
- Normalité
- Absence d'auto-corrélations et de corrélations croisées

3.2 Transformation des séries

Nous allons maintenant transformer les séries pour les rendre stationnaires, afin de pouvoir appliquer les modèles VAR ensuite. Afin de stationnariser les séries, nous utiliserons la fonction *decompose* qui permet de découper la série en trois : la tendance, la saisonnalité et les résidus, afin de pouvoir ensuite travailler avec les résidus. Nous ne stationnariserons que les échantillons d'apprentissage.

3.2.1 Masse salariale

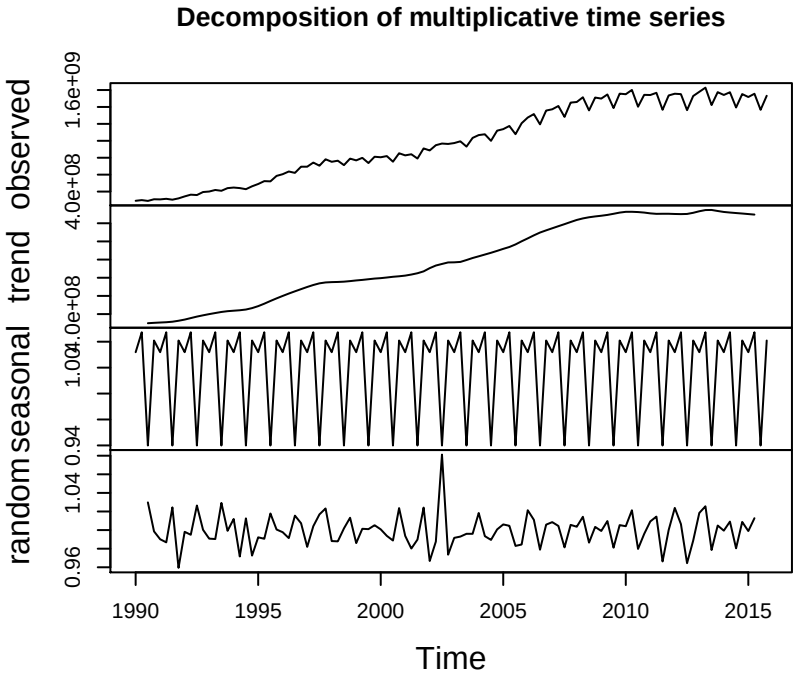


Figure 10: Décomposition de la masse salariale

Nous nous intéressons aux ACF, PACF (visibles en Figure 27) et test de KPSS afin de vérifier si les résidus obtenus à l'aide de la fonction *decompose* sont stationnaires. Bien que l'ACF et la PACF nous mettent en garde d'une possible non stationnarité de la série, la p-value des tests de KPSS et Dickey Fuller augmenté nous amène à confirmer que notre série est désormais stationnarisée (avec un seuil de confiance à 5% pour les deux tests effectués).

```
## Warning in kpss.test(MSESta): p-value greater than printed p-value
```

```
##
##  KPSS Test for Level Stationarity
```

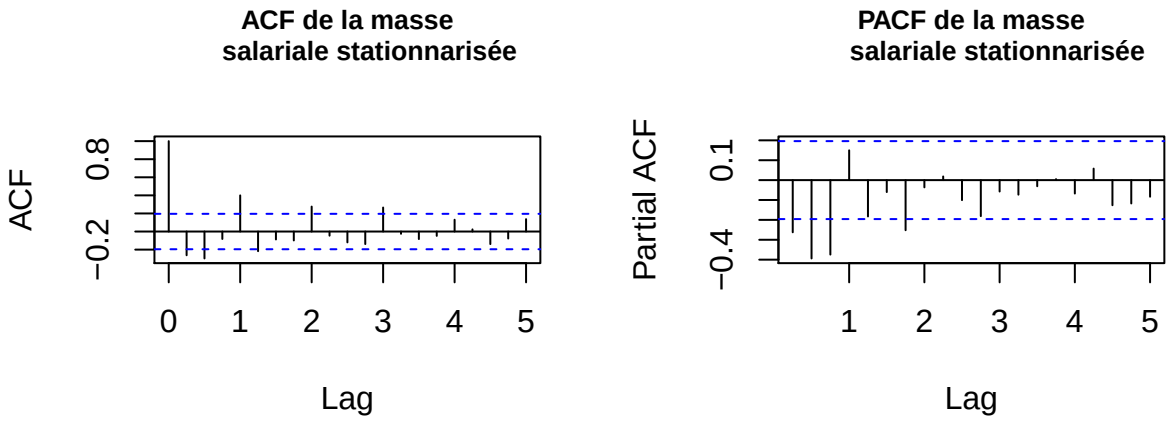


Figure 11: Fonctions d'autocorrélation de la masse salariale stationnarisée

```
##
## data:  MSESta
## KPSS Level = 0.017376, Truncation lag parameter = 2, p-value = 0.1

## Warning in adf.test(MSESta): p-value smaller than printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  MSESta
## Dickey-Fuller = -6.3219, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

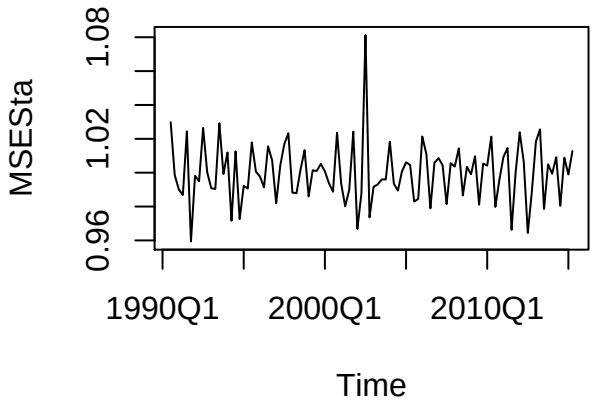


Figure 12: Masse salariale trimestrielle stationnarisée

3.2.2 PIB

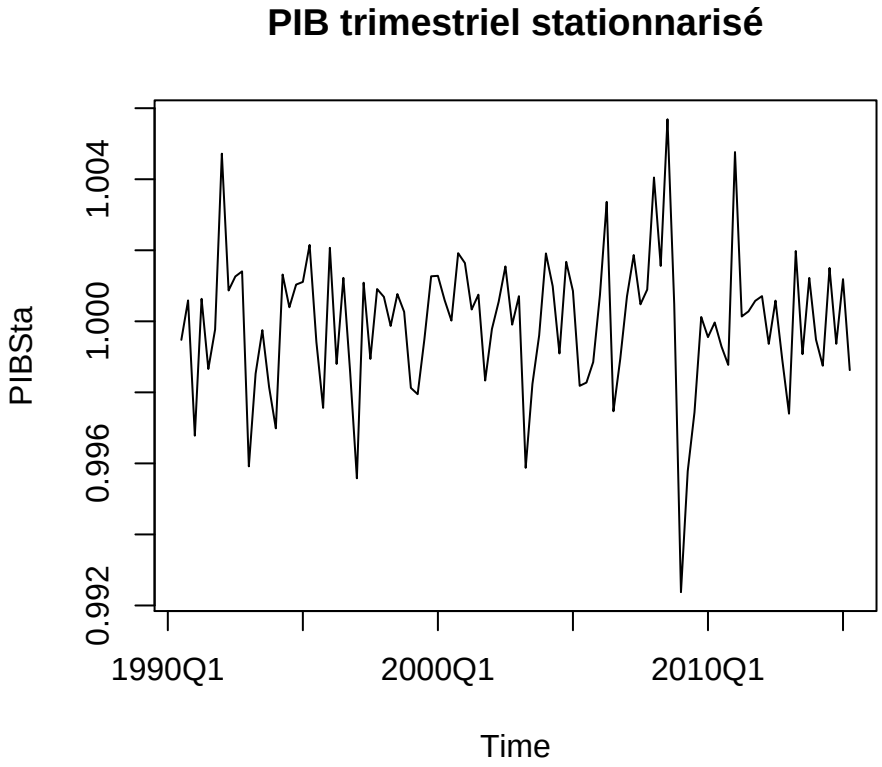


Figure 13: PIB trimestriel stationnarisé

```
## Warning in kpss.test(PIBSta): p-value greater than printed p-value
##
##  KPSS Test for Level Stationarity
##
## data:  PIBSta
## KPSS Level = 0.027524, Truncation lag parameter = 2, p-value = 0.1
## Warning in adf.test(PIBSta): p-value smaller than printed p-value
##
##  Augmented Dickey-Fuller Test
##
## data:  PIBSta
## Dickey-Fuller = -5.0084, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

Nous nous intéressons aux ACF, PACF, test de KPSS et test de Dickey Fuller augmenté (Figure 28) afin de vérifier si les résidus obtenus à l'aide de la fonction *decompose* sont stationnaires. Au regard de ces différentes informations, nous pouvons conclure à la stationnarité des résidus.

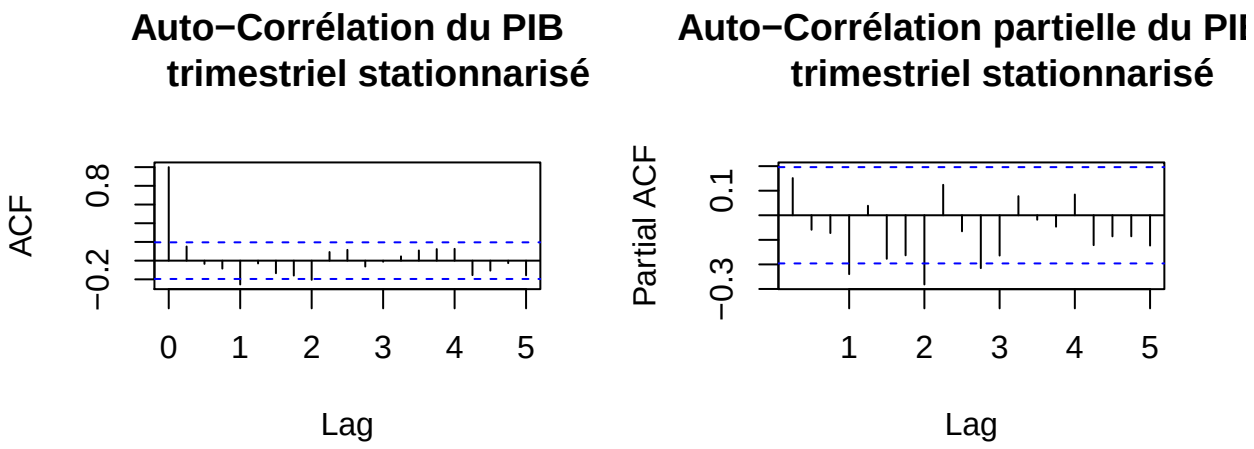


Figure 14: Fonctions d'autocorrélation du PIB stationnarisé

3.2.3 SMIC

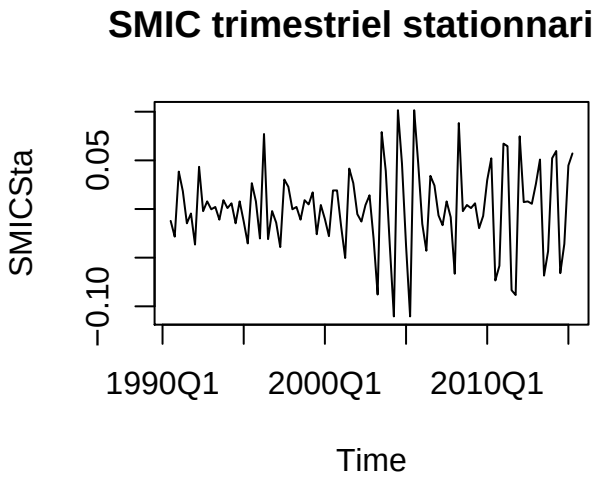


Figure 15: SMIC trimestriel stationnarisé

Comme pour la masse salariale, les ACF et PACF de la Figure 16 semblent montrer que la série résiduelle pourrait ne pas être stationnaire. Cependant le test de KPSS ainsi que le test de Dickey Fuller augmenté nous permettent de conclure à la stationnarité des résidus.

```
## Warning in kpss.test(SMICSta): p-value greater than printed p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  SMICSta
```

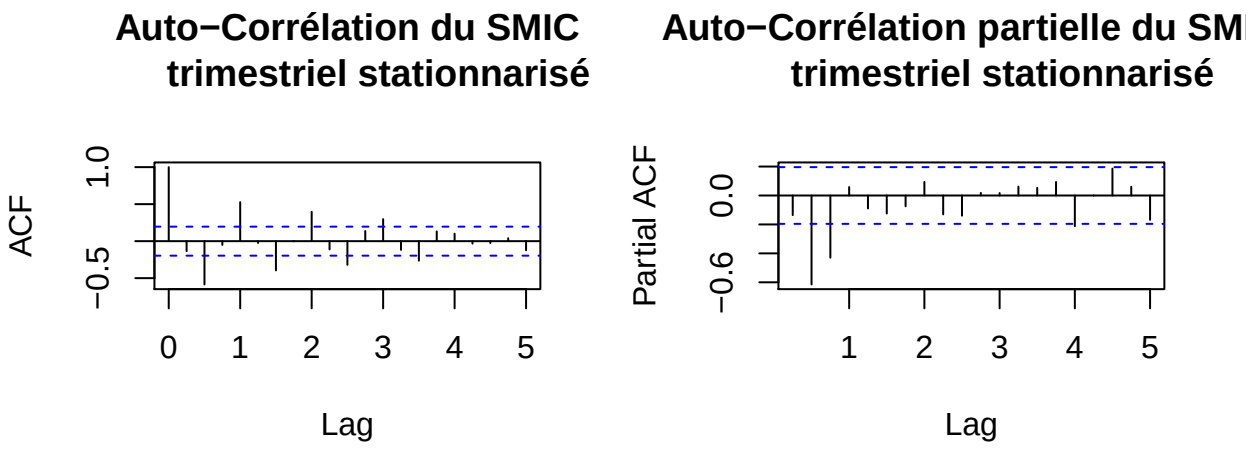


Figure 16: Fonctions d'autocorrélation du SMIC stationnarisé

```
## KPSS Level = 0.043771, Truncation lag parameter = 2, p-value = 0.1

## Warning in adf.test(SMICSta): p-value smaller than printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  SMICSta
## Dickey-Fuller = -6.357, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

3.2.4 Taux de chômage des femmes

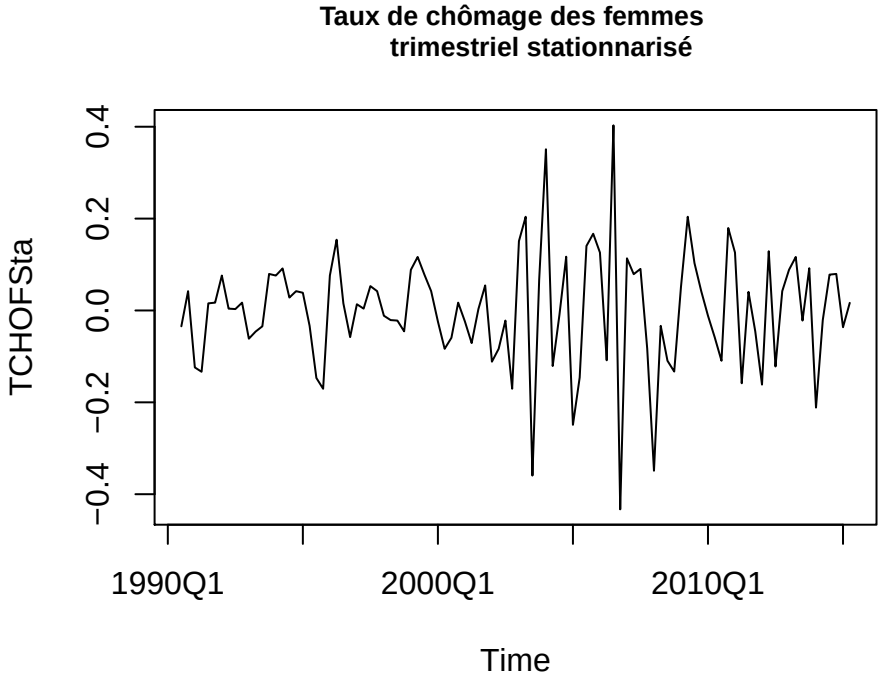


Figure 17: Taux de chômage trimestriel stationnarisé

En ce qui concerne le taux de chômage des femmes, en regardant l'ACF, PACF, le test de KPSS et le test de Dickey Fuller augmenté présents en Figure 18, on peut conclure que la série résiduelle est stationnaire.

```
## Warning in kpss.test(TCHOFSta): p-value greater than printed p-value

##
##  KPSS Test for Level Stationarity
##
## data:  TCHOFSta
## KPSS Level = 0.022077, Truncation lag parameter = 2, p-value = 0.1

## Warning in adf.test(TCHOFSta): p-value smaller than printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  TCHOFSta
## Dickey-Fuller = -6.6221, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

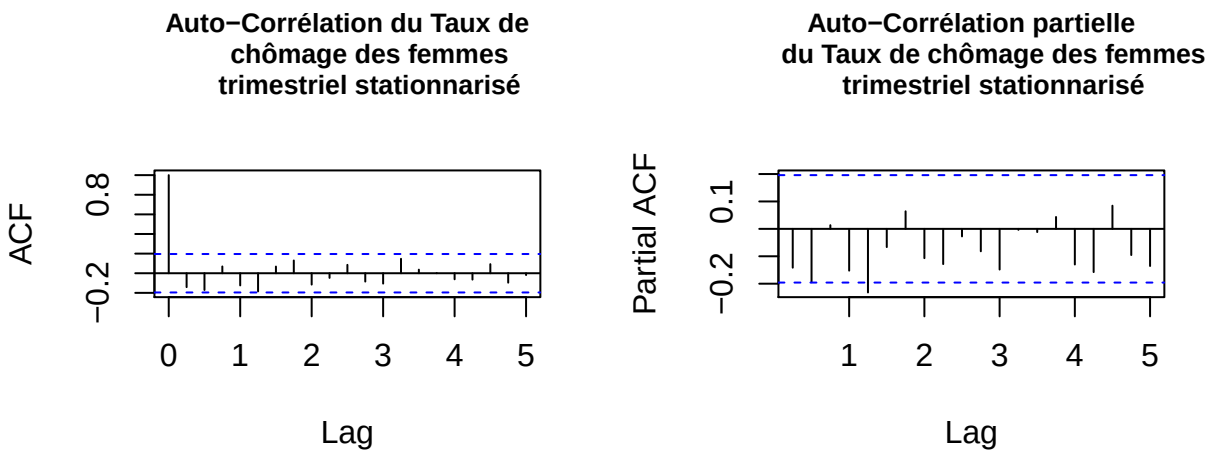


Figure 18: Fonctions d'autocorrélation du taux de chômage stationnarisé

3.3 Corrélation entre les variables stationnarisées

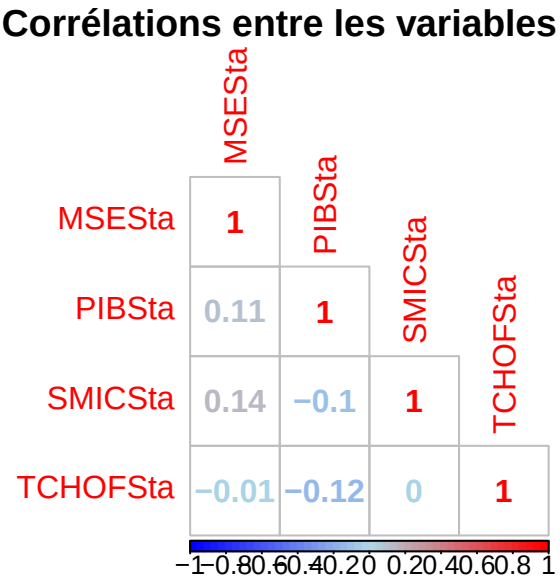


Figure 19: Corrélations entre les variables trimestrielles stationnarisées

```
##              MSE       PIB       SMIC      TCHOF
## MSE   0.0000000 0.2808591 0.1511096 0.9397487
## PIB   0.2808591 0.0000000 0.3027994 0.2329023
## SMIC  0.1511096 0.3027994 0.0000000 0.9786726
## TCHOF 0.9397487 0.2329023 0.9786726 0.0000000
```

On s'aperçoit que la transformation de nos séries a permis de supprimer les corrélations entre elles. En effet, la matrice des corrélations présentes en Figure 19 nous montre que la corrélation la plus élevée vaut 0.14 ce qui reste très faible. De plus, en regardant le tableau des p-values, l'hypothèse nulle de non significativité du coefficient de corrélation n'est rejetée pour aucun couple de variables

(avec un seuil de 5%).

Maintenant que toutes les séries ont été stationnarisées, elles peuvent être utilisées pour construire un modèle VAR.

3.4 Mise en place de modèles VAR avec le package vars

Le package vars a été construit par Bernhard Pfaff. Il permet de construire des modèles vectoriels (VAR), et dispose également des différentes fonctions de diagnostics permettant de vérifier que les hypothèses du modèle sont bien remplies. La version utilisée est la version 1.5-3 et date du 6 Août 2018.

3.4.1 Calcul de l'ordre p

Afin de mettre en place une modélisation VAR, nous devons dans un premier temps nous intéresser à l'ordre p du modèle VAR. L'ordre p correspond à l'ordre de l'opérateur de retard, c'est-à-dire le nombre de valeurs du passé qui ont un impact sur la valeur à un instant t. Dans le package **vars**, la fonction *VARselect* permet de déterminer l'ordre des modèles VAR à selectionner en fonction de 4 critères (AIC, HQ, SC et FPE).

Pour les critères suivants, p correspond à l'ordre du modèle VAR, T le nombre d'observations utilisées pour la phase d'apprentissage, K le nombre de variables et $\tilde{\Sigma}_u(p) = \frac{1}{T}\Sigma_{t=1}^{T}\hat{u}_t\hat{u}_t'$ (la matrice de covariance des résidus du modèle).

Dans cette partie, nous développerons le fonctionement de la méthodologie en l'appliquant uniquement au modèle complet, soit celui prenant en compte les variables PIB, SMIC et taux de chômage des femmes.

Le critère AIC (Aikaike information criterion) se calcule, dans ce package, de la manière suivante : $AIC(p) = \ln\det(\tilde{\Sigma}_u(p)) + \frac{2}{T}pK^2$. L'objectif est de minimiser ce critère. Cela suppose donc que le déterminant de la matrice $\tilde{\Sigma}_u(p)$ soit strictement positif. Ce critère est asymptotiquement effiace : si le nombre d'observations tend vers l'infini, sa variance est aussi faible que possible.

Le critère HQ (Hannan-Quinn criterion) se calcule, dans ce package, de la manière suivante : $HQ(p) = \ln\det(\tilde{\Sigma}_u(p)) + \frac{2\ln(\ln(T))}{T}pK^2$. L'objectif est de minimiser ce critère. Encore une fois, cela suppose que le déterminant de la matrice $\tilde{\Sigma}_u(p)$ soit strictement positif.

Le critère SC (Schwarz criterion) se calcule dans ce package de la manière suivante : $SC(p) = \ln\det(\tilde{\Sigma}_u(p)) + \frac{\ln(T)}{T}pK^2$. L'objectif est de minimiser ce critère. Ce critère est un autre nom du BIC.

On s'aperçoit que les différents critères à notre disposition, visibles sur les graphiques de la Figure 20, nous donnent des ordres à choisir différents. Ainsi, le meilleur AIC correspond à un modèle d'ordre 10, le meilleur HQ à un modèle d'ordre 3 et le meilleur SC à un modèle d'ordre 2. L'ordre de l'AIC étant trop grand (car trop de coefficients à estimer par rapport au nombre d'observations à notre disposition), nous ne souhaitons pas conserver cet ordre. De plus, on se rend compte que l'AIC du modèle avec un ordre 10 est similaire à celle d'un modèle avec un ordre 4. Les modèles HQ et SC sont meilleurs avec respectivement un ordre 3 et 2. Nous allons donc, dans la suite de l'analyse, essayer les trois modèles définit par les différents critères : ici, nous allons donc nous intéresser aux modèles d'ordre 2, 3 et 4.

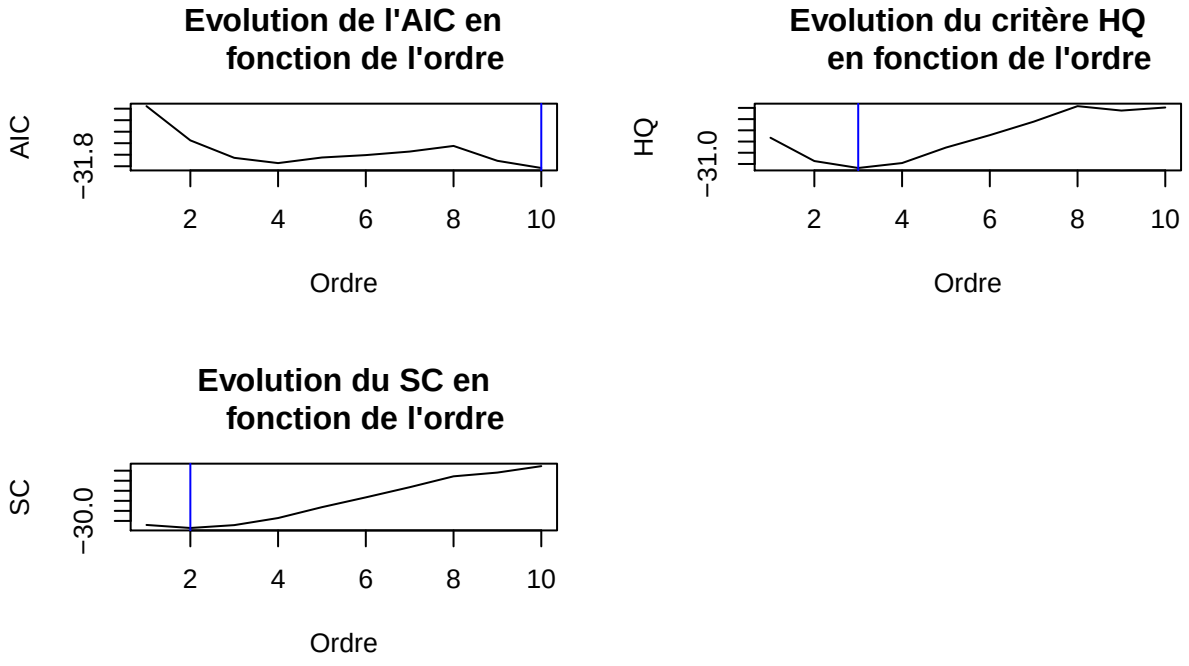


Figure 20: Critères associés au modèle complet

3.4.2 Estimation du modèle VAR(p)

Dans la partie précédente, nous avons sélectionné le meilleur ordre pour notre modèle VAR. Il s'agit maintenant d'estimer différents modèles afin de pouvoir prédire la MSE. L'exemple que nous avons pris est pour le modèle complet, avec les trois ordres déterminés précédemment (2, 3 et 4).

Un modèle VAR s'écrit sous la forme suivante :

$y_t = \sum_{i=1}^{p} A_i y_{t-i} + u_t$

A_i représentent les matrices de coefficients du modèle pour un ordre i, t le décalage de la série et u_t une matrice K-dimensionnelle composée des résidus du modèle (indépendants et identiquement distribués).

3.4.2.1 Ordre 4

Dans le package **vars**, la fonction utilisée pour construire des modèles VAR est VAR, qui prend en entrée la série temporelle multivariée, l'ordre du processus et le type de régresseurs à inclure. Dans notre cas, *type* vaut *const* car la série est stationnarisée et donc centrée en une constante μ. Ci-dessous, le modèle d'ordre 4.

Les coefficients du modèle sont les suivants. Les erreurs standards associées sont présentes en B.

```
##                MSE        PIB         SMIC         TCHOF
## MSE   -0.49571073  0.8961454  0.066963585 -0.0127991267
## PIB   -0.03134947  0.1217021 -0.005125468  0.0005912438
```

```
## SMIC   0.05931412  1.1813374 -0.496926461 -0.0096012159
## TCHOF -0.88275343 -5.5627534  0.660087437 -0.3220844544
```

```
##                MSE          PIB         SMIC        TCHOF
## MSE   -0.44837936  -0.85710828  0.059065357 -0.016289594
## PIB   -0.03290389  -0.03056789 -0.006426264 -0.001447922
## SMIC  -0.12186822  -0.59050790 -0.674266226 -0.004993847
## TCHOF  0.40185192 -17.96140240  0.821446464 -0.310004540
```

```
##                MSE          PIB         SMIC         TCHOF
## MSE   -0.27254709  -0.67990488  0.04532636  0.0043284001
## PIB   -0.02833629  -0.06107288 -0.01176086 -0.0004898846
## SMIC  -0.27835989  -0.37298995 -0.41721629 -0.0094899806
## TCHOF  0.25697997 -16.67936660  0.63877023 -0.1024043275
```

```
##                MSE        PIB         SMIC        TCHOF
## MSE    0.19486895 -0.4557750  0.09080679 -0.001506479
## PIB   -0.02602058 -0.2762433 -0.00804238  0.001727524
## SMIC   0.14617955  1.3033993  0.08614964 -0.022040339
## TCHOF -1.24236337 -2.7851730  1.22452627 -0.152158590
```

Les indicateurs de qualité du modèle sont présents ci-dessous.

```
##        AIC(n)         HQ(n)         SC(n)        FPE(n)
## -3.174520e+01 -3.098355e+01 -2.985646e+01  1.664221e-14
```

Enfin, l'erreur quadratique moyenne de ce modèle pour les données prédites est la suivante :

```
## [1] 7.711856e+14
```

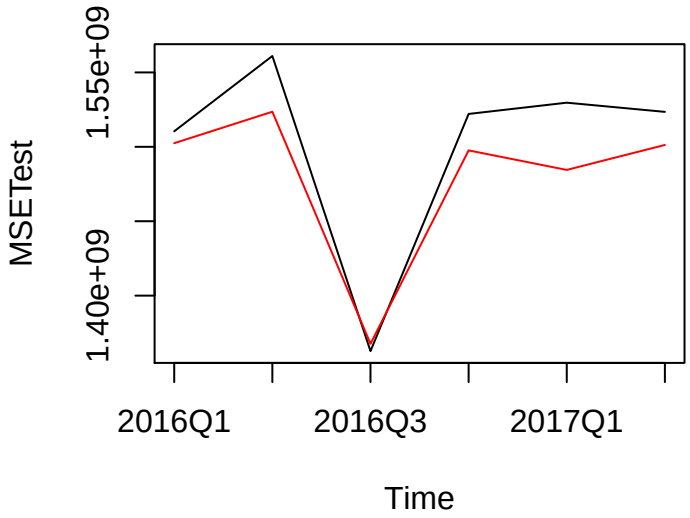


Figure 21: Comparaison entre les vraies valeur et le modèle complet pour un ordre 4

La représentation graphique de la Figure 21 nous montre des prédictions proches des vraies valeurs, à part pour le premier trimestre 2017.

3.4.2.2 Ordre 3

On s'intéresse ensuite au modèle d'ordre 3, soit celui avec le meilleur critère HQ.

Les coefficients du modèle associés à un retard de 1 sont les suivants :

```
##                MSE        PIB           SMIC        TCHOF
## MSE   -0.55534062  0.5875606  0.0078931593 -0.007157107
## PIB   -0.01712717  0.1810606 -0.0001595774  0.001149490
## SMIC  -0.01904054  0.4355559 -0.5342311296 -0.015680593
## TCHOF -0.40492349 -6.7105659  0.1543612225 -0.278488213
```

Les indicateurs de qualité du modèle sont présents ci-dessous.

```
##        AIC(n)         HQ(n)         SC(n)        FPE(n)
## -3.165371e+01 -3.107127e+01 -3.020938e+01  1.805099e-14
```

L'erreur quadratique moyenne de ce modèle pour les données prédites est la suivante :

```
## [1] 9.682886e+14
```

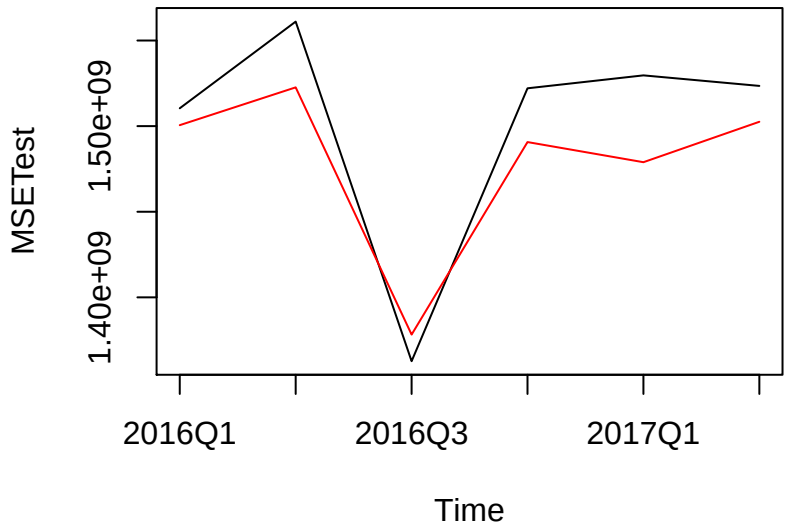


Figure 22: Comparaison entre les vraies valeur et le modèle complet pour un ordre 3

Graphiquement, la Figure 22 associée à l'ordre 3 nous donne des prédictions semblant inférieures à celles du modèle d'ordre 4. L'apport d'un lag supplémentaire dans la construction du retard a donc bien une importance dans la qualité des prédictions.

3.4.2.3 Ordre 2

Enfin, nous construisons le modèle avec le meilleur BIC, soit celui d'ordre 2.

Les coefficients du modèle associés à un retard de 1 sont les suivants :

```
##                MSE        PIB          SMIC        TCHOF
## MSE   -0.39572755  0.8295593  0.030196701 -0.009163581
## PIB   -0.01295909  0.1873301  0.006086945  0.001036053
## SMIC   0.12910889  0.9413863 -0.239926026 -0.022141074
## TCHOF -0.51603938 -6.5905696  0.103373369 -0.217258398
```

Les indicateurs de qualité du modèle sont présents ci-dessous.

```
##        AIC(n)         HQ(n)         SC(n)        FPE(n)
## -3.135082e+01 -3.094759e+01 -3.035090e+01  2.430391e-14
```

L'erreur quadratique moyenne de ce modèle pour les données prédites est la suivante :

```
## [1] 9.987686e+14
```

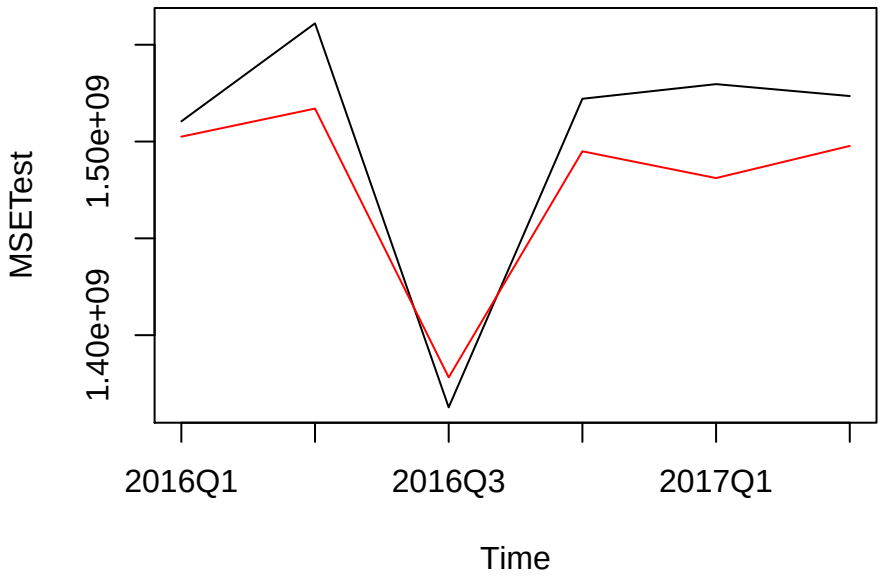


Figure 23: Comparaison entre les vraies valeur et le modèle complet pour un ordre 2

Lorsque l'on compare les erreurs quadratiques moyennes, nous remarquons que la plus faible est celle associée à un modèle d'ordre 4, soit celui avec le meilleur AIC. Il nous faut maintenant vérifier que les hypothèses associées à ce modèle soient bien vérifiées.

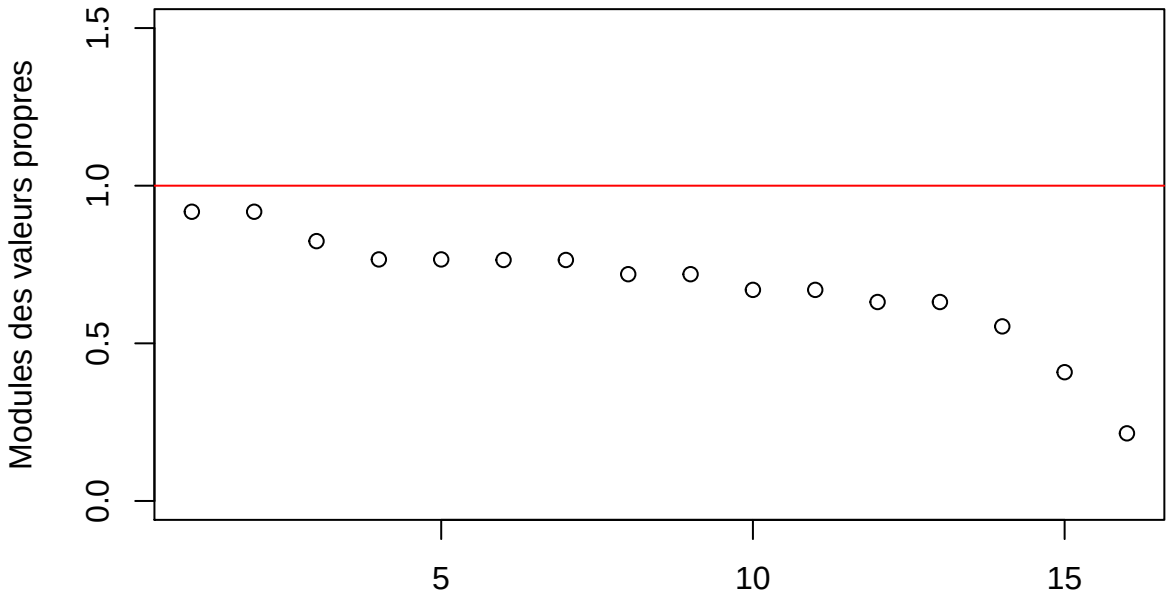
3.4.3 Verification de la stabilité

Pour vérifier si le processus VAR est stable, c'est-à-dire qu'il génère des séries stationnaires, nous devons calculer les valeurs propres de la matrice des coefficients :

$$A = \begin{bmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I & 0 & \cdots & 0 & 0 \\ 0 & I & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I & 0 \end{bmatrix}$$

Si les modules des valeurs propres de A sont inférieures à 1, alors le processus VAR est stable. Nous allons donc vérifier que le processus VAR(4) créé précédemment avec toutes les variables à notre disposition est bien stable.

```
##  [1] 0.9174340 0.9174340 0.8241534 0.7661190 0.7661190 0.7644083 0.7644083
##  [8] 0.7192911 0.7192911 0.6695288 0.6695288 0.6310697 0.6310697 0.5537349
## [15] 0.4084401 0.2145192
```



On s'aperçoit que tous les modules sont inférieurs à 1, le processus VAR(4) est donc stable.

3.4.4 Test ARCH (homoscédasticité des résidus)

Le test multivarié de ARCH-LM permet de tester l'homoscédasticité des résidus. La statistique de test est la suivante : $VARCH_{LM}(q) = \frac{1}{2}TK(K+1)R_m^2$, où $R_m^2 = 1 - \frac{2}{K(K+1)}tr(\hat{\Omega}\hat{\Omega}_O^{-1})$, et $\hat{\Omega}$ est la matrice de covariance de la régression suivante : $vech(\hat{u}_t\hat{u}_t{}^T) = \beta_0 + B_1 vech(\hat{u}_{t-1}\hat{u}_{t-1}^T) + ... + B_q vech(\hat{u}_{t-q}\hat{u}_{t-q}^T) + v_t$. La dimension de β_O est $\frac{1}{2}K(K+1)$ et celle des matrices des coefficients B_i est $\frac{1}{2}K(K+1)$Ö$\frac{1}{2}K(K+1)$. La statistique de test suit une loi de $\chi^2(qK^2(K+1)^2/4)$, donc dans notre cas $\chi^2(16q * 25/4)$. L'hypothèse nulle de ce test est $H_0 : B_0 = B_1 = ... = B_q = 0$ (homoscédasticité).

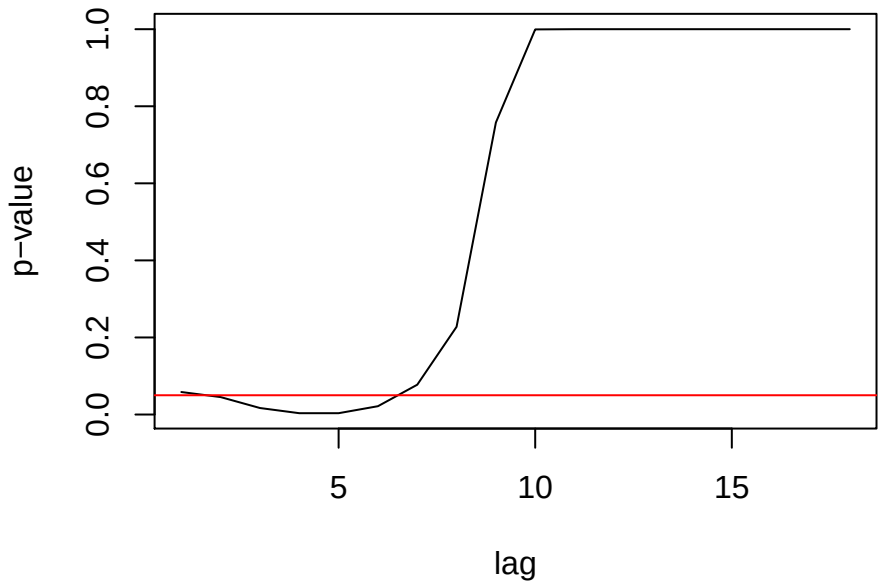


Figure 24: Evolution de la p-value en fonction du lag

On s'aperçoit au regard de la figure 25, avec un seuil de confiance de 5% (ligne rouge), qu'on rejette l'hypothèse nulle d'homoscédasticité pour un retard faible (inférieur à 7). Cependant, en augmentant le nombre de valeurs prises en compte pour calculer la nouvelle, on se rend compte qu'on conserve l'hypothèse d'homoscédasticité. On observe également que la valeur de la p-value converge vers 1 au fur et à mesure qu'on augmente le retard. Ainsi, en prenant l'ensemble des résidus, nous conservons l'hypothèse d'homoscédasticité.

3.4.5 Test normalité (normalité des résidus)

Le test de Jarque-Bera pour séries multivariées permet de tester la normalité des résidus. Il utilise les résidus standardisés à l'aide d'une décomposition de Cholesky de la matrice de variance-covariance des résidus centrés. Il est important noter que l'ordre dans lequel les variables sont stockées dans la matrice a une importance sur les résultats. La statistique de test est la suivante : $JB_{mv} = s_3^2 + s_4^2$,

où s_3^2 et s_4^2 se calculent de la sorte : $s_3^2 = Tb_1^T b_1/6$ et $s_4^2 = T(b_2 - 3_K)^T(b_2 - 3_K)/24$, avec b_1 et b_2 qui sont respectivement les vecteurs des moments non-centrés d'ordre trois et quatre des résidus standardisés. La statistique de test suit une loi de $\chi^2(2K)$. Ce test compare en fait le coefficient kurtosis K (l'aplatissement de la fonction de densité) et le coefficient skewness S (asymétrie de la fonction de densité) d'une loi normale à ceux des résidus testés. L'hypothèse nulle est donc $H0 : S = 0$ et $K = 3$.

```
##
##  JB-Test (multivariate)
##
## data:  Residuals of VAR object modele
## Chi-squared = 71.388, df = 8, p-value = 2.599e-12
```

Ici, on rejette l'hypothèse H0, avec un seuil de confiance de 5%. Les résidus obtenus ne suivent pas une loi normale.

3.4.6 Test Portmanteau (corrélations des résidus)

Le test de Portmanteau multivarié permet de tester l'auto-corrélation (au sein d'une même série) et la corrélation croisée (entre les différentes séries) des résidus.

3.4.6.1 Verification de la Matrice C_0

La statistique de Portmanteau est $Q_h = T\sum_{j=1}^{h} tr(\hat{C}_j^T \hat{C}_0^{-1} \hat{C}_j \hat{C}_0^{-1})$, et elle suit une loi de $\chi^2(K^2h - n^*)$, n^* étant le nombre de coefficients à estimer. Pour qu'elle existe, il faut donc vérifier que la matrice $\hat{C}_0$ est inversible pour que la statistique puisse être définie. Les matrices $\hat{C}_i$ s'écrivent $\hat{C}_i = \frac{1}{T}\sum_{t=i+1}^{T} \hat{u}_t \hat{u}_{t-i}^T$, donc $\hat{C}_0$ s'écrit $\hat{C}_0 = \frac{1}{T}\sum_{t=1}^{T} \hat{u}_t \hat{u}_t^T$. Nous allons donc vérifier qu'elle est inversible pour le modèle complet que nous avons mis en place. $tr()$ correspond à la trace de la matrice, soit la somme des éléments diagonaux de la matrice.

```
##              MSESta        PIBSta       SMICSta      TCHOFSta
## [1,]  1.777427e-04  1.302790e-06  7.666911e-06 -9.660844e-05
## [2,]  1.302790e-06  3.036813e-06 -7.311009e-06 -2.784179e-05
## [3,]  7.666911e-06 -7.311009e-06  7.893080e-04  2.705167e-04
## [4,] -9.660844e-05 -2.784179e-05  2.705167e-04  1.041230e-02
```

```
## Déterminant de la matrice
##               4.173756e-15
```

Le déterminant de la matrice n'étant pas nul, la matrice $\hat{C}_0$ est donc inversible.

3.4.6.2 Application du test

Les 3 premières p-values ne peuvent être calculées à cause de la valeur des degrés de liberté. En effet, comme nous l'avons expliqué plus haut, la statistique de test suit une loi de $\chi^2(K^2h - n^*)$. Or, avec un retard compris entre 1 et 3, les degrés de liberté sont négatifs et il n'est donc pas possible d'appliquer le test. L'hypothèse nulle de ce test est l'absence de corrélations croisées et d'auto-corrélations.

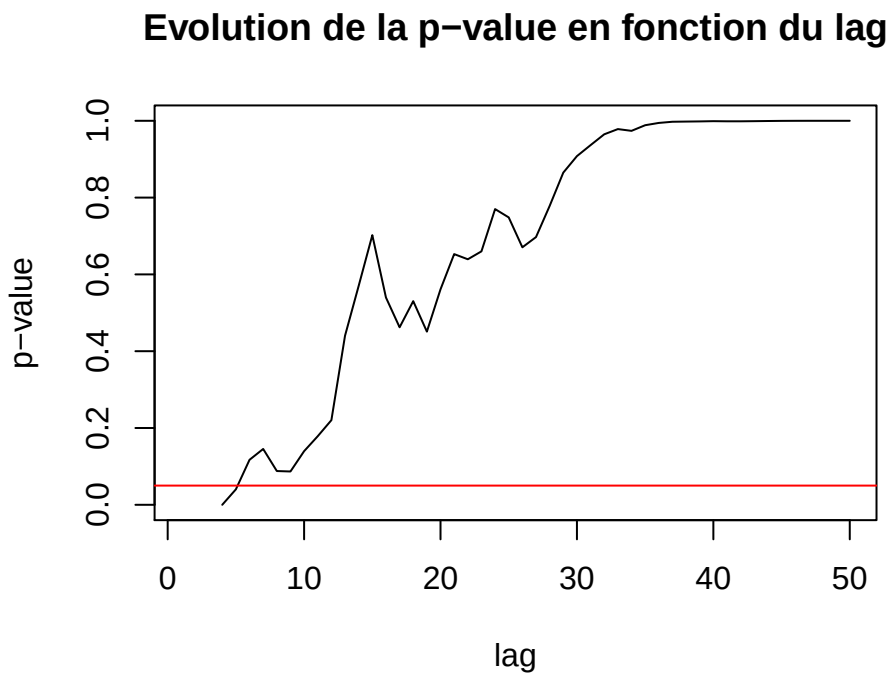


Figure 25: Evolution de la p-value en fonction du lag

Au regarde la figure 21 Comme pour le test ARCH, on rejette l'hypothèse nulle, avec un seuil de confiance à 5% (ligne rouge) pour un retard faible (5 ou moins). Cependant, pour un retard grand (supérieur à 5), on conserve l'hypothèse nulle d'absence d'auto-corrélations et de corrélations croisées. On observe également que la p-value converge vers 1 à mesure qu'on augmente le retard. Ainsi, en prenant en compte l'ensemble des résidus, on conserve l'hypothèse d'absence d'auto-corrélations et de corrélations croisées.

3.4.7 Prévisions

Maintenant que nous avons estimé l'ordre des différents modèle VAR, et que nous avons explicité l'estimation des modèles, nous cherchons désormais à trouver celui dont les prédictions sont les plus proches de la réalité.

Après avoir comparé tous les modèles possibles (7 : 3 modèles avec deux variables, 3 modèles avec trois variables et un modèle avec les quatre variables), nous nous apercevons que le meilleur en terme de prédictions est le modèle prenant en compte le SMIC (en plus de la masse salariale).

```
#SMIC
VARselect(cbind(MSESta, PIBSta), lag.max=10)
```

```
## $selection
## AIC(n)  HQ(n)  SC(n) FPE(n)
##      4      3      2      4
##
```

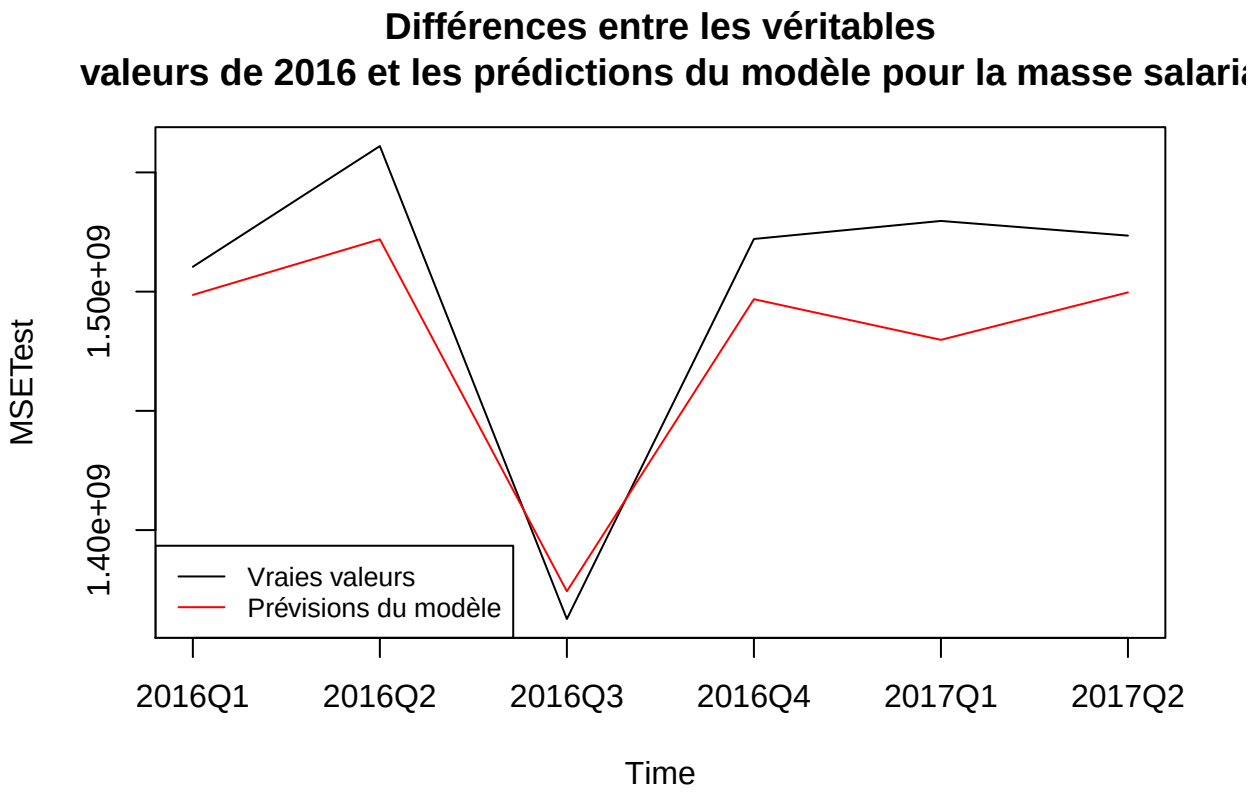
```
## $criteria
##                       1             2             3             4
## AIC(n) -2.058779e+01 -2.074116e+01 -2.084002e+01 -2.085328e+01
## HQ(n)  -2.052059e+01 -2.062915e+01 -2.068321e+01 -2.065166e+01
## SC(n)  -2.042114e+01 -2.046340e+01 -2.045116e+01 -2.035331e+01
## FPE(n)  1.145135e-09  9.824929e-10  8.903636e-10  8.792655e-10
##                       5             6             7             8
## AIC(n) -2.079925e+01 -2.075571e+01 -2.077752e+01 -2.079075e+01
## HQ(n)  -2.055283e+01 -2.046449e+01 -2.044150e+01 -2.040992e+01
## SC(n)  -2.018818e+01 -2.003354e+01 -1.994425e+01 -1.984637e+01
## FPE(n)  9.291143e-10  9.720286e-10  9.531524e-10  9.433686e-10
##                       9            10
## AIC(n) -2.075250e+01 -2.067957e+01
## HQ(n)  -2.032687e+01 -2.020914e+01
## SC(n)  -1.969703e+01 -1.951299e+01
## FPE(n)  9.837885e-10  1.063124e-09
```

```
VAR(cbind(MSESta, PIBSta), p=4, type="const")
```

```
##
## VAR Estimation Results:
## ========================
##
## Estimated coefficients for equation MSESta:
## ===========================================
## Call:
## MSESta = MSESta.l1 + PIBSta.l1 + MSESta.l2 + PIBSta.l2 + MSESta.l3 + PIBSta.l3 + MSESta.l4 +
##
##  MSESta.l1  PIBSta.l1  MSESta.l2  PIBSta.l2  MSESta.l3  PIBSta.l3
## -0.4918839  0.5281328 -0.4589271 -0.5838285 -0.2994483 -0.3154436
##  MSESta.l4  PIBSta.l4      const
##  0.1528415 -0.3365771  2.8046318
##
##
## Estimated coefficients for equation PIBSta:
## ===========================================
## Call:
## PIBSta = MSESta.l1 + PIBSta.l1 + MSESta.l2 + PIBSta.l2 + MSESta.l3 + PIBSta.l3 + MSESta.l4 +
##
##   MSESta.l1   PIBSta.l1   MSESta.l2   PIBSta.l2   MSESta.l3   PIBSta.l3
## -0.02828762  0.14135650 -0.03296562 -0.05553778 -0.02897191 -0.03074332
##   MSESta.l4   PIBSta.l4       const
## -0.02590789 -0.25928322  1.32028238
```

```
VARSMICSta <- forecast(VAR(cbind(MSESta, PIBSta), p=4, type="const"))
plot(MSETest, xlim=c(2016,2017.25), main="Différences entre les véritables
     valeurs de 2016 et les prédictions du modèle pour la masse salariale", xaxt="n")
axis(side=1, at=seq(2016,2017.25,0.25), labels=c("2016Q1", "2016Q2", "2016Q3", "2016Q4", "2017
lines(window(VARSMICSta$forecast$MSESta$mean*MSETrendTest*MSESeasonalTest, start=2016, end=c(20
```

```
legend('bottomleft', legend = c('Vraies valeurs', 'Prévisions du modèle'),
       col=c('black', 'red'), lty=1, cex=0.8)
```



Nous nous intéressons donc à l'erreur quadratique moyenne de cette prévision.

```
EQM(MSETest, window(VARSMICSta$forecast$MSESta$mean*MSETrendTest*MSESeasonalTest, start=2016, e
```

```
## [1] 9.153141e+14
```

A Annexe 1 : Modélisation univariée des séries

A.1 Modélisation individuelle

Une fois que nous avons analysé le comportement des différentes séries temporelles à notre disposition, nous souhaitons les modéliser afin de prédire les valeurs futures de ces différentes séries. En effet, si nous voulons prédire la MSE pour des valeurs futures, nous aurons également besoin des valeurs associées pour les variables explicatives, qui ne seront peut-être pas à notre disposition. Nous avons utilisé à la fois des modèles basés sur un lissage exponentiel et des processus ARMA.

A.1.1 Comparaison des différents modèles

Afin de comparer les modèles construits pour chaque série avec les différentes méthodes, nous calculons l'Erreur Quadratique Moyenne (EQM), soit les moyennes des différences au carré entre les valeurs de test et les valeurs prédites par le modèle.

A.1.2 Lissage exponentiel

A.1.2.1 Définition

Le lissage exponentiel permet de prédire les valeurs d'une série temporelle en lissant successivement les données à partir d'une valeur initiale. Plus les observations sont éloignées dans le passé, moins leur poids est important lors du calcul. Pour une série stationnaire, la formule de calcul d'une valeur est la suivante : $s_t = \alpha y_t + (1 - \alpha)s_{t-1}$, le paramètre α étant le facteur de lissage. Le nom de cette méthode est un lissage exponentiel **simple**. Afin de modéliser les séries possédant une tendance, nous introduisons un paramètre β permettant de la prendre en compte, la méthode étant appelée lissage exponentiel **double**. Enfin, Holt et Winters ont également modifié la méthode pour qu'elle puisse modéliser les séries comportant une saisonnalité en introduisant un paramètre γ. Ils ont donné leur nom à cette méthode, qui est donc un lissage exponentiel de **Holt-Winters**.

Dans notre cas, nous ne calculons pas nous-mêmes α, β et γ Ces paramètres sont déterminés automatiquement par la fonction *ets* du package **forecast** de façon à optimiser la qualité de la prédiction. Cette fonction permet également de choisir la méthode à utiliser, grâce à l'argument model. Afin de mesurer la qualité de notre modèle, nous avons choisi d'utiliser l'**AICc** (Akaike Information Criterion with correction). Le choix de l'AICc par rapport à l'AIC s'explique par le faible nombre de données que nous possédons par rapport au nombre de paramètres à estimer. C'est ce critère qui nous servira par la suite afin de comparer nos différents modèles.

Prenons l'exemple de la MSE. Nous avons vu dans la partie 2.2 que la série possédait une tendance linéaire ainsi qu'une saisonnalité multiplicative. L'argument model de la fonction *ets** prendra donc la valeur "ZAM", (erreur sélectionnée automatiquement, tendance linéaire, saisonnalité multiplicative). On peut également remarquer que lorsque tous les paramètres sont automatiquement sélectionnés (valeur "ZZZ"), les paramètres retenus sont les mêmes que ceux que nous avions rentré.

```
## ETS(M,A,M)
##
## Call:
##  ets(y = MSETrain, model = "ZAM")
##
##   Smoothing parameters:
##     alpha = 0.7675
##     beta  = 0.1111
##     gamma = 0.2325
##
##   Initial states:
##     l = 279219343.2211
##     b = 12053621.0848
##     s = 1.0092 0.9655 1.0159 1.0094
##
```

```
##   sigma:  0.0279
##
##       AIC     AICc      BIC
## 4031.536 4033.451 4055.336
```

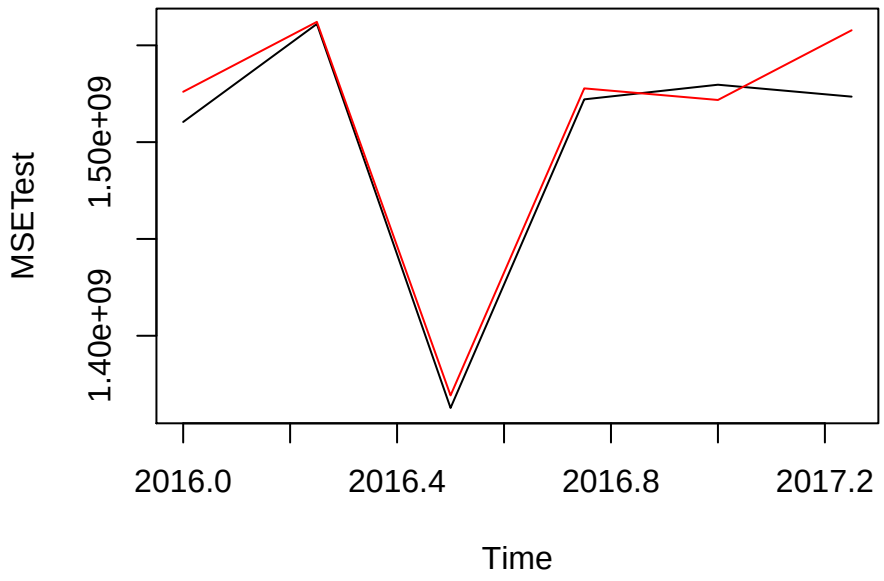


Figure 26: Comparaison entre la prédiction du lissage exponentiel et les valeurs réelles pour la masse salariale trimestrielle

```
## [1] 2.592281e+14
```

On obtient donc un AICc de 4033.451 pour le modèle ainsi qu'une erreur quadratique moyenne de $2.6 * 10^{14}$. Le graphique obtenu en figure 26 nous montrent que le modèle obtenu nous donne des prédictions très proches de la réalité.

A.1.2.2 Résultats obtenus

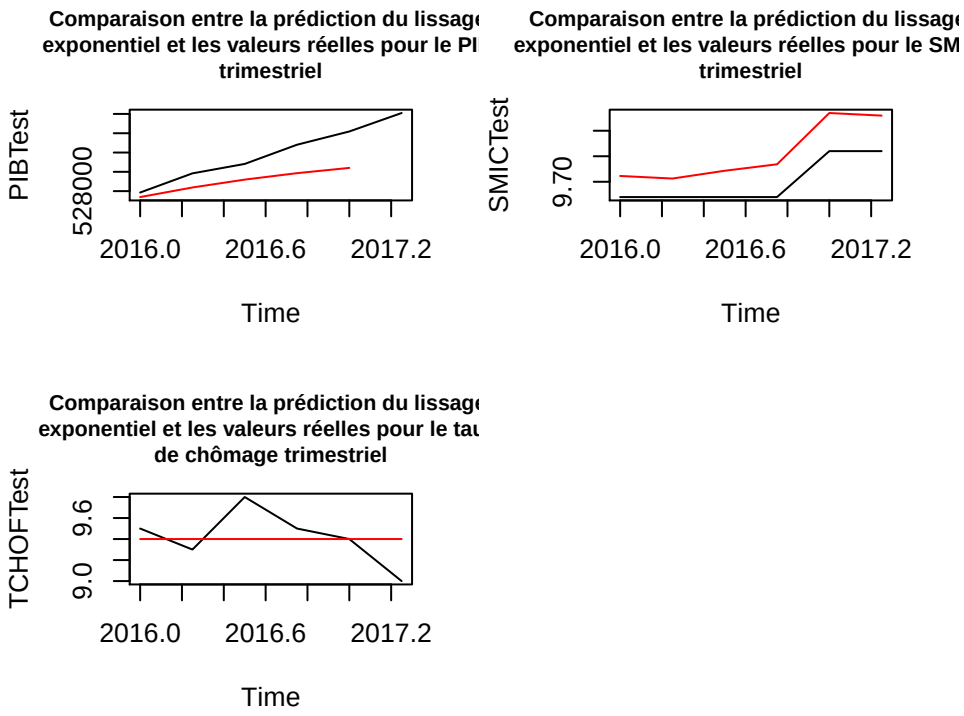


Figure 27: Résultats obtenus pour le lissage exponentiel

Les graphiques de la figure 27 nous montrent des résultats mitigés. Pour le PIB et le SMIC, les prédictions suivent la forme de la série mais en sont éloignées. Pour le taux de chômage des femmes, la méthode de lissage utilisée est un lissage exponentiel simple, ce qui nous donne donc des prédictions constantes soit de mauvaise qualité.

Nous résumons dans le tableau suivant les résultats obtenus pour chaque série estimée par un lissage exponentiel.

Variable	Tendance	Saisonnalité	Argument model	AIC
MSE	linéaire	multiplicative	ZAM	4033.45
PIB	linéaire	absente	ZAN	2053.15
SMIC	linéaire	additive	ZAA	-84.96
TCHOF	absente	absente	ZNN	204.37

A.1.3 Modèles ARMA

A.1.3.1 Définition

Les modèles **ARMA(p,q)** sont une autre famille de modèles permettant d'estimer une série temporelle. Il est divisé en deux parties : une partie autorégressive **AR** auquel est associé un ordre p qui donne le nombre de valeurs passées qui vont être utiles dans la prédiction, et une partie moyennes mobiles **MA** qui permet de de prendre en compte les q innovations de la série dans le

futur.

L'une des propriétés des processus ARMA est qu'ils sont utilisés pour modéliser des séries stationnaires, donc par extension des séries qui ne possèdent ni tendance ni saisonnalité. Afin de modéliser des séries non stationnaires, on généralise les processus ARMA en processus **ARIMA(p,d,q)**, d représentant l'ordre de différenciation de la série. Les séries saisonnières sont elles modélisées par des processus $SARIMA(p,d,q)(P,D,Q)_s$ qui modélisent des séries avec une saisonnalité de période s.

Comme pour le lissage exponentiel, nous ne calculons pas nous-mêmes les ordres des processus. Pour cela, la fonction *auto.arima* du package **forecast** nous a été très utile. Elle permet en effet de trouver les ordres du processus qui optimisent un critère défini à l'avance et de calculer un modèle avec ces coefficients. Nous avons choisi d'optimiser l'**AICc**(Akaike Information Criterion with correction), pour les raisons évoquées dans la partie A.1.2.1

```
## Series: MSETrain
## ARIMA(0,1,1)(0,1,1)[4]
##
## Coefficients:
##           ma1     sma1
##       -0.1986  -0.3857
## s.e.   0.1014   0.0929
##
## sigma^2 estimated as 7.396e+14:  log likelihood=-1834.54
## AIC=3675.09   AICc=3675.34   BIC=3682.87
```

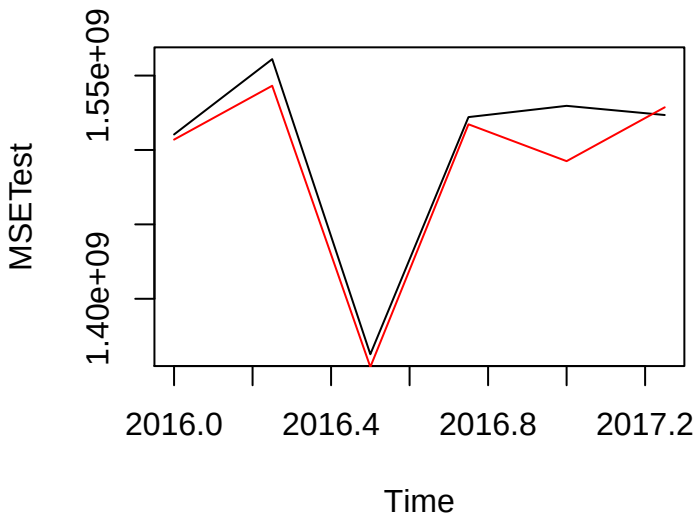


Figure 28: Comparaison entre le modèle SARIMA et les données de validation pour la masse salariale trimestrielle

Pour la MSE, on obtient par exemple un modèle $SARIMA(0,1,1)(0,1,1)_4$ ainsi qu'un AICc de

3896.01. La figure 28 nous donne des prédictions d'assez bonne qualité mais qui semblent moins bonnes que celles obtenues par lissage exponentiel.

A.1.3.2 Résultats obtenus

```
## Series: PIBTrain
## ARIMA(2,1,0) with drift
##
## Coefficients:
##          ar1     ar2      drift
##       0.4701  0.1596  1585.8365
## s.e.  0.0966  0.0966   456.3024
##
## sigma^2 estimated as 3153980:  log likelihood=-915.48
## AIC=1838.95   AICc=1839.36   BIC=1849.49
```

```
## Series: SMICTrain
## ARIMA(1,0,0)(1,1,0)[4] with drift
##
## Coefficients:
##          ar1     sar1   drift
##       0.8645  -0.3643  0.0486
## s.e.  0.0514   0.0961  0.0080
##
## sigma^2 estimated as 0.003998:  log likelihood=134.95
## AIC=-261.89   AICc=-261.47   BIC=-251.47
```

```
## Series: TCHOFTrain
## ARIMA(1,1,1)
##
## Coefficients:
##          ar1      ma1
##       0.6974  -0.5005
## s.e.  0.1669   0.1935
##
## sigma^2 estimated as 0.06173:  log likelihood=-1.76
## AIC=9.52   AICc=9.76   BIC=17.42
```

Comme pour le lissage exponentiel, nous résumons les résultats obtenus dans un tableau pour plus de lisibilité. Le PIB et le taux de chômage des femmes ne comportent pas de partie saisonnière car comme vu dans les parties 2.3 et 2.5 on ne constate pas de saisonnalité dans l'analyse descriptive de la série. On peut également voir sur la figure 29 que les prédictions du PIB semblent de bien meilleure qualité

Variable	Ordre du processus	AICc
MSE	(0,1,1)(0,1,1)	3675.34
PIB	(2,1,0)	1839.36
SMIC	(1,0,0)(1,1,0)	-261.47
TCHOF	(0,1,1)	9.76

Variable	Ordre du processus	AICc

A.1.4 Comparaison des différents modèles

Une fois que nous avons construit les deux types de modèles pour chacune des variables, nous souhaitons les comparer pour savoir quel modèle est le plus efficace pour prédire chacune des variables. Pour cela, les EQM, calculant l'erreur de prédiction, de chacun des modèles sont synthétisées dans le tableau suivant. L'AICc ne peut pas être utilsié ici car les méthodes à comparer sont différentes. Il n'est donc pas sûr que la méthode utilisée pour calculer la vraisemblance soit la même.

```
##            lissage         ARMA
## MSE   2.592281e+14 3.060118e+14
## PIB   4.654031e+06 1.382004e+05
## SMIC  3.371267e-03 8.609164e-03
## TCHOF 5.833300e-02 6.223076e-02
```

Nous nous rendons compte que le lissage a une EQM plus faible pour la masse salariale (notre variable d'intérêt), ainsi que pour le SMIC et le taux de chômage des femmes. En ce qui concerne le PIB, le modèle ARIMA est plus performant. Cette analyse va nous servir par la suite, comme expliqué au début de la partie A.1

A.2 Modélisation ARMA avec variables exogènes

A.2.1 Définition

Maintenant que nous avons modélisé chaque série individuellement, nous souhaitons savoir s'il est possible d'améliorer la qualité de prédiction de la série MSE trimestrielle à l'aide des autres variables à notre disposition. Pour ce faire, nous allons construire des modèles SARIMA prenant en compte des variables exogènes.

$Y_t = \beta_0 + \beta_1 X_{1t} + ... + \beta_k X_{kt} + \epsilon_t$

Y_t est la variable à modéliser. X_i pour $i = 1, ..., k$ correspond à la $i$ème variable exogène. β_i pour i allant de $i = 0, ..., k$ correspond aux coefficients d'une régression linéaire. Enfin, le résidu ϵ_t suit un processus de type ARMA.

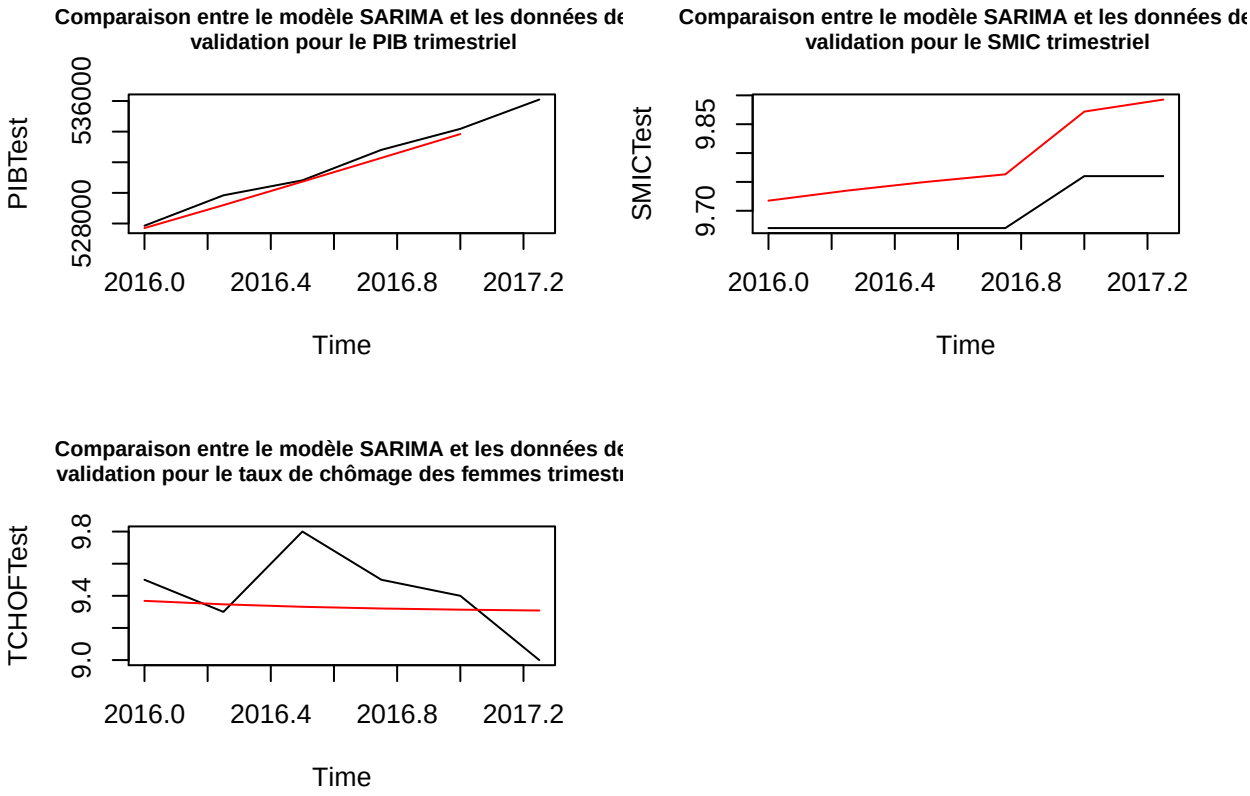


Figure 29: Résultats obtenus avec des modèles ARMA

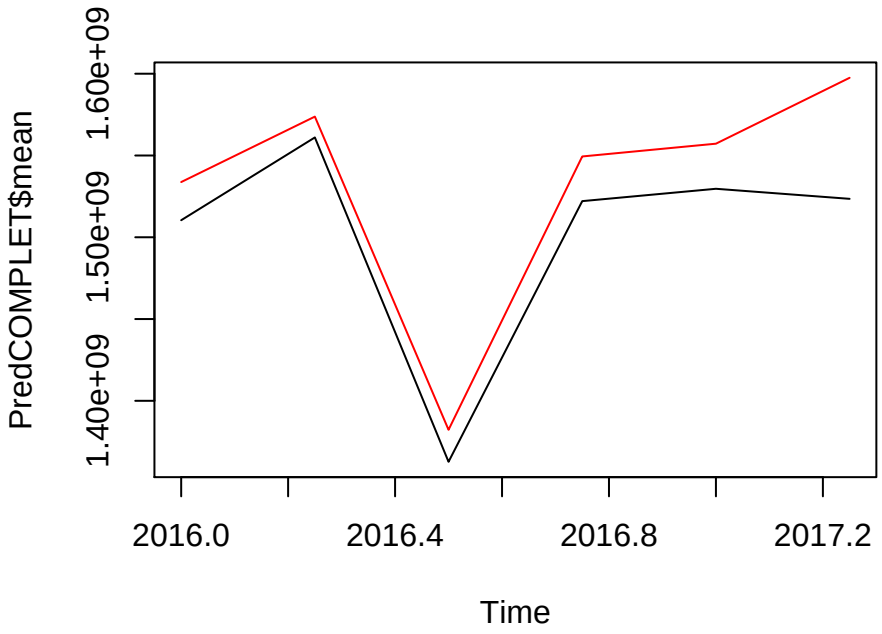


Figure 30: Masse salariale expliquée par le PIB, le SMIC et le taux de chômage vs Vraies valeurs

```
## [1] 1.345414e+15
```

Ici, les résidus suivent un SARIMA(0,0,0)(0,1,0)[4], et le modèle possède 3 variables exogènes : le coefficient correspondant à la variable PIB est $\beta_1 = 706.0045$, celui correspondant à la variable SMIC est $\beta_2 = 209266766$ et enfin celui correspondant au taux de chômage est $\beta_3 = -1644706$

A.2.2 Résultats obtenus

Nous allons désormais nous intéresser à la construction des différents modèles prenant en compte 1 variable exogène (3 modèles), 2 variables exogènes (3 modèles) et 3 variables exogènes (1 modèle). La qualité de ces 7 modèles est representée dans le tableau ci-dessous.

Variable	Ordre du processus	AICc
Aucune	(0,1,1)(0,1,1)	3675.34
PIB	(1,0,0)(2,1,0)	3715.99
SMIC	(0,1,0)(0,1,0)	3688.92
TCHOF	(0,0,0)(1,1,0)	3809.82
PIB & SMIC	(0,0,0)(0,1,0)	3807.37
PIB & TCHOF	(1,0,0)(0,1,0)	3721.36
SMIC & TCHOF	(0,1,0)(0,1,0)	3690.57
COMPLET	(0,0,0)(0,1,0)	3809.5

On se rend compte qu'aucun modèle ARIMA ne prenant en compte des variables exogènes n'a une qualité meilleure que celui ne prenant en compte aucune variable exogène (en comparant les AIC corrigés). Si on omet le modèle sans variable exogène, le meilleur modèle prenant en compte au moins une variable exogène est celui prenant en compte le SMIC (dans la figure 31).

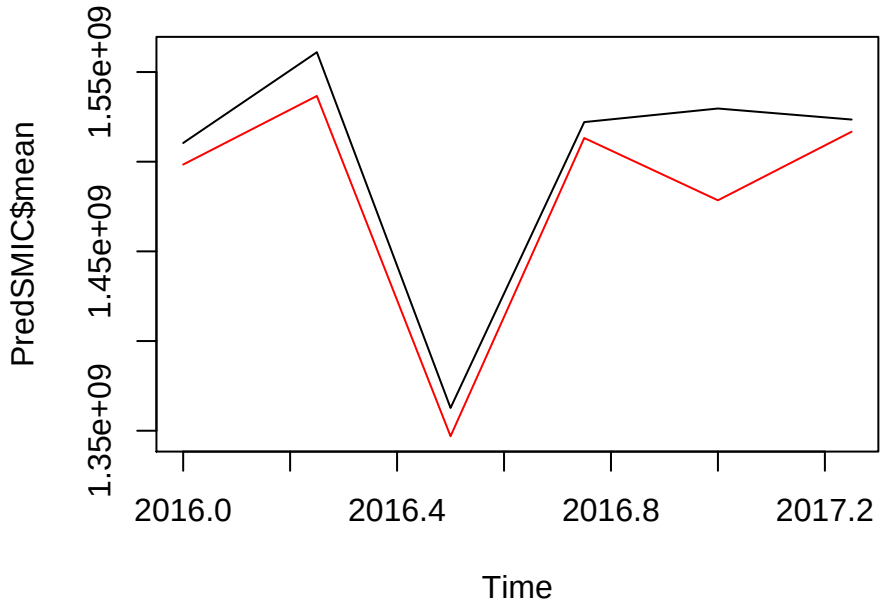


Figure 31: SARIMA expliqué par le SMIC vs Vraies valeurs

```
## [1] 6.219316e+14
```

B Annexe 3 : Erreurs standard associées aux coefficients du modèle VAR d'ordre 4

```
##                MSE          PIB         SMIC        TCHOF
## MSE   -0.44837936  -0.85710828  0.059065357 -0.016289594
## PIB   -0.03290389  -0.03056789 -0.006426264 -0.001447922
## SMIC  -0.12186822  -0.59050790 -0.674266226 -0.004993847
## TCHOF  0.40185192 -17.96140240  0.821446464 -0.310004540
```

```
##                MSE          PIB         SMIC         TCHOF
## MSE   -0.27254709  -0.67990488  0.04532636  0.0043284001
## PIB   -0.02833629  -0.06107288 -0.01176086 -0.0004898846
## SMIC  -0.27835989  -0.37298995 -0.41721629 -0.0094899806
## TCHOF  0.25697997 -16.67936660  0.63877023 -0.1024043275
```

```
##                MSE        PIB        SMIC        TCHOF
```

```
## MSE    0.19486895 -0.4557750  0.09080679 -0.001506479
## PIB   -0.02602058 -0.2762433 -0.00804238  0.001727524
## SMIC   0.14617955  1.3033993  0.08614964 -0.022040339
## TCHOF -1.24236337 -2.7851730  1.22452627 -0.152158590
```