IAE Nantes
Économie & Management
Pôle Sociétés

Nantes Université

Master 2 ECAP – Économétrie Statistique

Régressions pénalisées sur le taux de chômage américain

Ndéye BAKHOUM
Yohann TESSON

Année 2023/2024

SOMMAIRE

I-Introduction :

Selon le BIT[1], Bureau International du Travail, le chômage peut se définir comme des personnes actives de plus de 15 ans, actuellement sans emploi à la recherche d'un travail rémunéré. On peut calculer son taux, en divisant le nombre de personnes au chômage par le nombre total de personne active. Le taux de chômage est un indicateur très couramment utilisé pour mesurer le bien-être économique d'un territoire ainsi que le bon fonctionnement du marché du travail. Il permet également d'identifier des phases de récession ou d'expansion en termes de conjoncture économique. Enfin il aide à la décision en ce qui concerne les politiques économiques et sociales, en relançant l'emploi, en formant de la main-d'œuvre ou en soutenant les travailleurs en difficultés. En conséquence, plus le taux de chômage est élevé, plus la santé économique sera détériorée et le marché du travail inefficient.

Aux États-Unis[2], le taux de chômage en 2022 s'élevait à 3,57 %. Un taux considéré comme du plein emploi contrairement à celui de la France[3] qui atteignait les 7,2 % en 2022. Cette différence de valeurs se traduit par des conditions et une protection différente de l'emploi au sein de ces deux pays. En effet, en France, le licenciement est plus complexe et coûteux pour les employeurs et les aides aux chômages plus importantes. Tandis que le licenciement aux États-Unis est plus souple, de même pour l'embauche. La période de chômage dure donc moins longtemps aux États-Unis qu'en France. Au cours de son histoire, les États-Unis ont subi quelques pics majeurs du chômage lors d'importantes crises économiques et financières, comme la Grande Dépression où pratiquement un quart de la population active s'est retrouvé au chômage, mais notamment la crise des subprimes et du COVID où le taux avoisinait et dépassait les 10 %.

Le taux de chômage est également en relation avec d'autres indicateurs économiques comme le PIB. En effet selon la loi d'Okun[4], en dessous d'un certain seuil de croissance du PIB, le chômage augmente. C'est donc dans ce contexte que nous allons réaliser une étude sur le taux de chômage américain afin d'identifier des indicateurs macroéconomiques qui expliqueraient sa fluctuation. Pour se faire nous allons utiliser différentes approches économétriques et de Machine Learning afin de sélectionner des variables. Nous utiliserons une base de données issue de la Federal Reserve Bank of St-Louis, composée de 128 variables et 644 observations.

[1]Définition du chômage : https://travail-emploi.gouv.fr/IMG/pdf/Definitions-2.pdf
[2]Taux de chômage USA en 2022 : https://fr.statista.com/statistiques/550404/taux-de-chomage-aux-etats-unis-1990/#:~:text=Cette%20statistique%20repr%C3%A9sente%20le%20taux,%C3%A9tant%20quasiment%20du%20plein%20emploi.
[3]Taux de chômage France 2022 : https://www.insee.fr/fr/statistiques/4805248
[4]Loi d'Okun : https://www.pourleco.com/le-dico-de-l-eco/loi-dokun

Nous allons dans premier temps réaliser une analyse exploratoire et descriptive de la base de données. Nous continuerons ensuite par une sélection de variables à partir d'une approche économétrique puis d'une approche de Machine Learning linéaires et non-linéaires. Le but est de comparer in fine les variables qui ont été retenues parmi les différentes modélisations.

II-Analyses exploratoires et descriptives

1. Contenu de la base de données :

Pour notre présente étude nous utiliserons la base de données FRED-MD publiée par la Federal Reserve Bank of St-Louis[5]. Cette base de données est constituée de 127 variables macroéconomiques, toutes quantitatives, pour 644 observations mensuelles allant de janvier 1970 à août 2023. Huit groupes de variables y sont représentés :

- Production et revenu
- Marché du travail
- Logement
- Consommation, commandes et stocks
- Monnaie et crédit
- Taux d'intérêt et de change
- Prix
- Marché des actions

Nous avons comme variable dépendante, le taux de chômage américain[6], qui représente le nombre de personnes actives sans emploi à la recherche d'un travail rémunéré sur la population totale active des États-Unis. Le calcul de ce taux est illustré par la formule ci-dessous :

Taux de chômage = Nombre de chômeurs / Nombre de personnes actives

2. Analyse de valeurs manquantes

Le premier constat lorsque nous regardons la base de données est qu'il existe de valeurs manquantes. En effet, ceci peut être dû à l'introduction de nouvelles variables dont les mesures sur les années antérieures n'ont pas été effectuées. Cela peut s'expliquer également par des indicateurs

[5]FRED-MD: https://research.stlouisfed.org/econ/mccracken/fred-databases/
[6]Taux de chômage américain : https://data.ca.gov/dataset/civilian-unemployment-rate-for-us-and-california/resource/6b59d10b-ca5d-465f-8ee2-0780fb1044c0?filters=Area%20Name%3AU.S.

en cours de calcul dont les valeurs récentes n'ont pas encore été publiées. Mais aussi des valeurs qui n'ont pas pu être calculées pour différentes causes. Nous allons donc, dans un premier temps, identifier les variables possédant des valeurs manquantes, les supprimer si elles sont trop nombreuses ou imputer une nouvelle valeur. Le tableau ci-dessous nous liste les différentes variables présentant des valeurs manquantes.

Tableau 1 : Tableau des variables présentant des valeurs manquantes

Variable	**Nombre de valeurs manquantes**
ACOGNO	266
UMCSENTx	64
TWEXAFEGSMTHx	36
S.P.div.yield	5
S.P.PE.ratio	4
DTCTHFNM	1
DTCOLNVHFNM	1
COMPAPFFx	1
CP3Mx	1
CONSPI	1
NONREVSL	1
ISRATIOx	1
BUSINVx	1
HWIURATIO	1
HWI	1
CMRMTSPLx	1
Total	**386**

Source : BAKHOUM&TESSON, dossier régressions pénalisées

Ce tableau nous montre qu'il y a 16 variables disposant de valeurs manquantes pour un total de 386. Nous supprimons donc les variables « ACOGNO » et « TWEXAFEGSMTHx » ayant le plus grand nombre de valeurs manquantes. Les garder serait inutiles et nous empêcherait de réaliser les modélisations. Nous conservons, cependant la variables UMCSENTx car ses valeurs manquantes apparaissent de manière régulière au cours du temps. Des valeurs vont donc être imputées pour les autres variables.

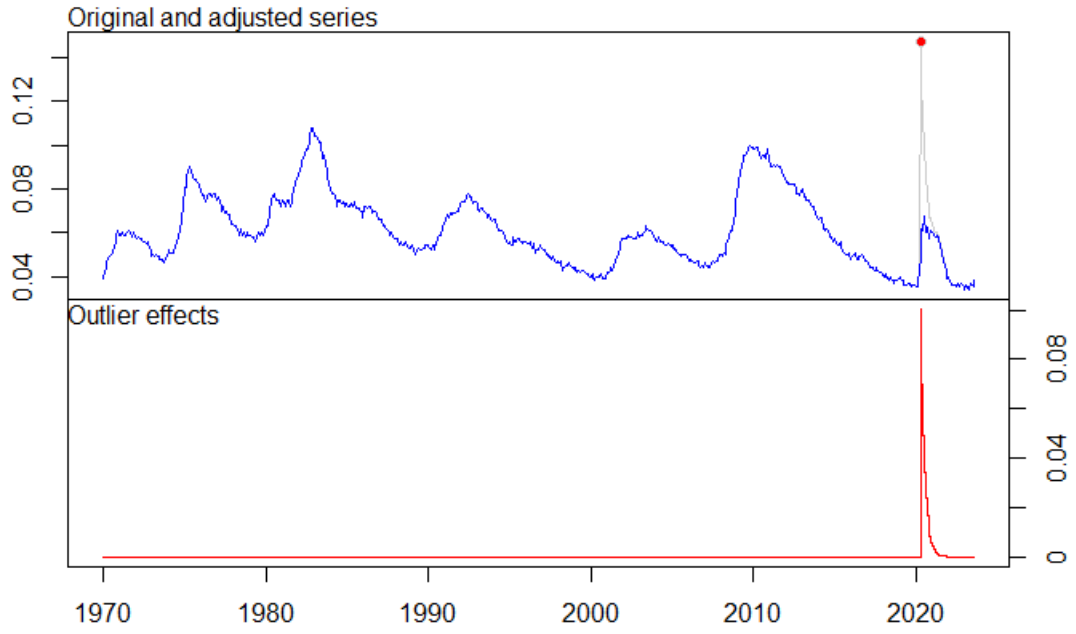
Pour réaliser l'imputation, il existe plusieurs méthodes. On peut soit remplacer une valeur manquante par la moyenne, par la médiane ou donner la valeur de l'observation précédente ou suivante dans le cas des séries temporelle. Dans notre cas nous utilisons un procédé plus efficace à partir de la fonction **missForest** du package **missForest**. A l'aide d'une forêt aléatoire, une valeur (la

plus représentative), en prenant en compte toutes les valeurs de la variable, sera attribué. Après imputation on remarque qu'il n'y a plus aucunes valeurs manquantes.

3. Analyse des outliers : détection et correction

En temps normal, il aurait fallu réaliser une analyse des valeurs atypiques sur toute la base de données, mais pour des problèmes de temps de calcul, nous nous focaliserons uniquement sur la variable à expliquer. Pour ce faire, nous utiliserons le package **tsoutliers** afin d'identifier et de corriger les valeurs atypiques.

Figure 1 : Représentation des valeurs atypiques de la variable à expliquer



Source : BAKHOUM&TESSON, dossier régressions pénalisées

Le graphique ci-dessus, nous représente les valeurs atypiques détectées ainsi que la correction effectuée. Ici, on voit la présence d'un seul outlier avec une valeur de 0,147 au mois d'avril 2020. Ce point représente une importante discontinuité, ce qui explique pourquoi il a été retenu comme valeur atypique. Cet outlier est de type TC, « Temporary Change » qui dispose d'un effet temporaire et qui retourne à son niveau précédent de manière rapide.

Pour expliquer cette valeur, il faut se référer au contexte économique et social de cette époque. En effet cette hausse drastique du chômage de 4,4 % en mars 2020 à 14,7 % en avril intervient durant la pandémie du Covid 19. Durant cette période de nombreuses restrictions ont été mise en place par

l'État afin de limiter la propagation du virus sur le sol américain. Les frontières ont été fermé aux voyages non essentiels avec le reste du monde. Ceci a plongé le pays en récession et favorisé les licenciements pour limiter les coûts des employeurs, dû à une nette baisse de la demande. Le taux de chômage a donc fortement augmenté.

Tableau 2 : Correction de la valeur atypique

Date	Valeur réelle	Valeur corrigée	Différence
01/04/2020	14,7	4,67	10,03
01/05/2020	13,2	6,18	7,02
01/06/2020	11,0	6,08	4,92
01/07/2020	10,2	6,76	3,44
01/08/2020	8,4	5,99	2,41
01/09/2020	7,9	6,21	1,69
01/10/2020	6,9	5,72	1,18
01/11/2020	6,7	5,87	0,83
01/12/2020	6,7	6,12	0,58
01/01/2021	6,4	5,99	0,41
01/03/2021	6,2	5,92	0,28
01/04/2021	6,0	5,80	0,2
01/05/2021	6,0	5,86	0,14
01/06/2021	5,8	5,70	0,1
01/07/2021	5,9	5,83	0,07
01/08/2021	5,4	5,35	0,05
01/09/2021	5,2	5,20	0

Source : BAKHOUM&TESSON, dossier régressions pénalisées

Ce tableau nous montre la correction de la valeur atypique. On peut voir que celle-ci se fait graduellement sur une période d'un an et demi avant de retrouver un niveau normal.

4. Détection de saisonnalité de la variable à expliquer

Nous allons maintenant déterminer s'il existe la présence de saisonnalité dans la série temporelle de la variable dépendante. Pour ce faire, nous allons effectuer différents tests. L'hypothèse nulle pour chacun est l'absence de saisonnalité.

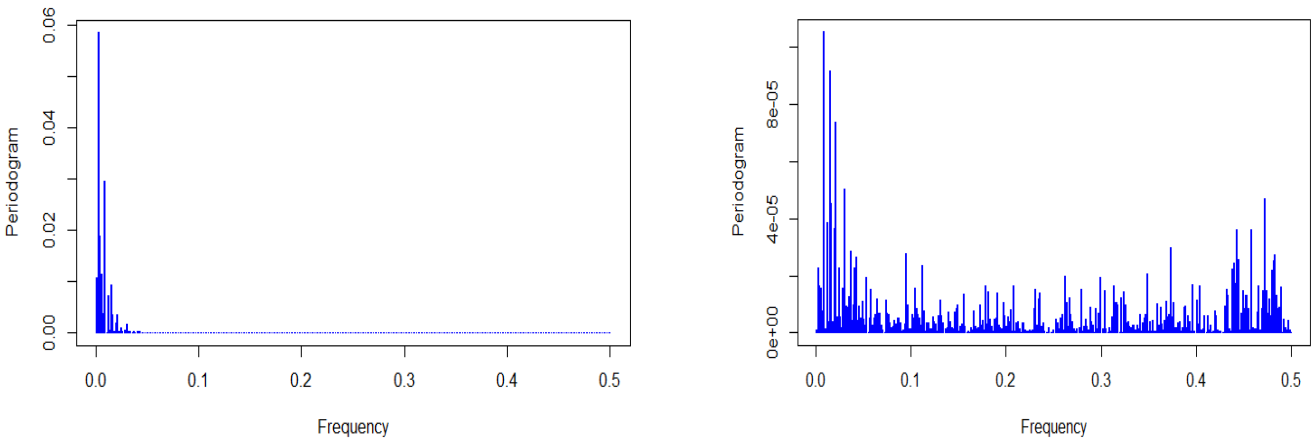
Tableau 3 : Test de saisonnalité

Test	p-value	Saisonnalité
Friedman	0,898	Non
Kruskal-Wallis	0,927	Non
Seasonal dummies	0,979	Non
Welch	0,972	Non
Webel-Ollech	0,909	Non

Source : BAKHOUM&TESSON, dossier régressions pénalisées

Le tableau, ci-dessus, nous montre que toutes les hypothèses nulles des tests sont acceptées puisque les p-value sont supérieur à 5 %. Il n'y a donc pas de saisonnalité dans la série dépendante. Nous allons confirmera cela graphiquement à l'aide de périodogramme qui montre l'évolution de la série en fonction de ses fréquences.

Figure 2 : Périodogrammes de la série Unemployment_Rate



Série sans différenciation *Série avec différenciation d'ordre 1*

Source : BAKHOUM&TESSON, dossier régressions pénalisées

On remarque que les deux périodogrammes ne présentent pas de pics réguliers. Ceci prouve donc la non-présence de saisonnalité.

5. Analyse de la stationnarité de la variable à expliquer

Nous allons maintenant vérifier si la série dépendante est stationnaire, c'est à dire que la moyenne et la variance de la variable sont constante au cours du temps. S'il existe une tendance dans la série, ceci peut expliquer une non-stationnarité au niveau de la moyenne. Il faudra alors effectuer une différenciation. Si la variance n'est pas constante au cours du temps, il faudra alors réaliser une transformation par le log ou la racine carré. Pour déceler cela, nous allons utiliser le test de Dickey-Fuller augmenté avec intégration de 4 retards. L'hypothèse nulle de ce test est la présence de racine unitaire, soit la non-stationnarité de la série. Nous obtenons le résultat suivant :

Tableau 4 : Test ADF avec intégration de 4 retards

Statistique	lag order	p-value
-3,1815	4	0,09109

Source : BAKHOUM&TESSON, dossier régressions pénalisées

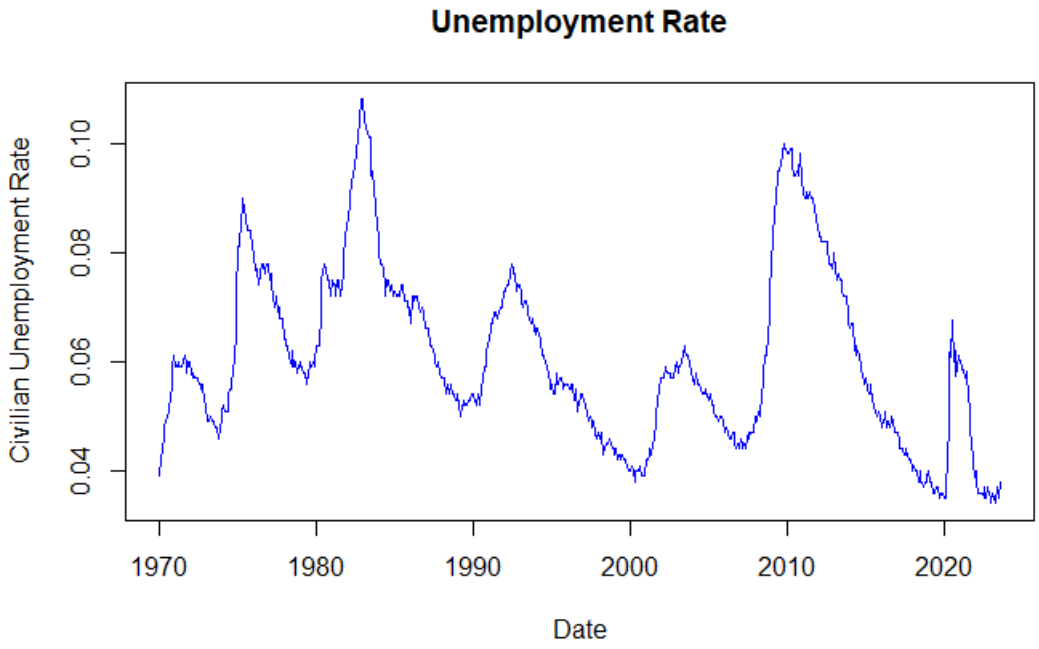
Le test nous donne une p-value de 0,09109. L'hypothèse de présence de racine unitaire est rejetée au seuil de 10 %. La série est donc stationnaire.

Il faut également vérifier la stationnarité des variables explicatives. Pour cela nous avons tout d'abord réalisé les différentes transformations mentionnées par la FRED[7] et réalisé un test ADF avec intégration d'un retard. Nous comptabilisons 6 variables non stationnaires. Pour la suite de l'étude nous considérons toutes les variables stationnaires.

[7]Transformations : file:///C:/Users/Admi/Desktop/M2%20ECAP/r%C3%A9gression%20p%C3%A9nalis%C3%A9e/FRED-MD_description%20(2).pdf

6. Graphique de la variable à expliquer

Figure 3 : Graphique de la variable à expliquer en niveau



Source : BAKHOUM&TESSON, dossier régressions pénalisées

Le graphique ci-dessus nous montre une représentation de l'évolution du taux de chômage de 1970 à 2023. On peut remarquer la présence de quelques pics que l'on peut associer aux différentes crises. On peut observer une forte hausse du taux en 1973 et 1979 correspondants à la période des crises pétrolières, mais également en 2008 et 2020 avec la crise des subprimes et du COVID-19.

7. Statistiques descriptives de la variable à expliquer

Voici quelques statistiques descriptives du taux de chômage américain :

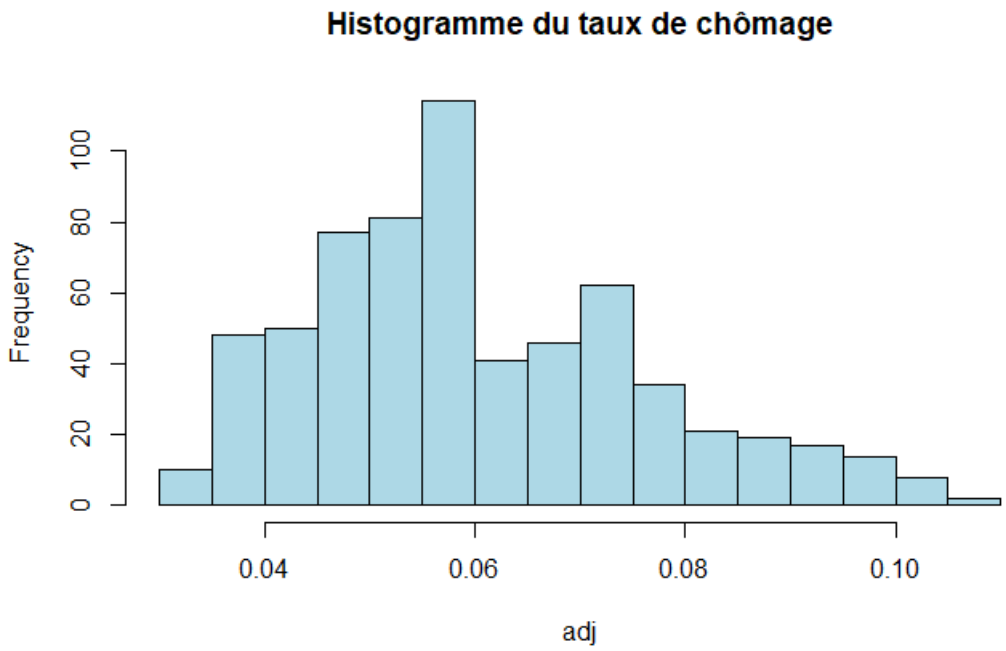
Tableau 5 : Statistiques descriptives de la variables taux de chômage

Min	Q1	Médiane	Moyenne	Q3	Max	Ecart-type	Kurtosis	Skewness
0,034	0,049	0,058	0,0609	0,072	0,108	0,016	-0,133	0,646

Source : BAKHOUM&TESSON, dossier régressions pénalisées

D'après le tableau, on peut voir que les valeurs fluctuent entre 0,034 et 0,108. Le minimum et le maximum sont atteint respectivement en janvier 2023 et novembre 1982. Mais il faut rappeler que la valeur d'avril 2020 (0,147) a été corrigé. La médiane de la série est de 0,058 inférieur à la moyenne qui est de 0,0609. Ceci explique une plus grande concentration vers les petites valeurs. Le coefficient d'aplatissement, le kurtosis, dispose d'une valeur négative signifiant une distribution plus plate que la normale et donc une apparition de valeurs extrêmes moins fréquentes qu'une distribution gaussienne. De même, le coefficient d'asymétrie, le skewness, est positif. Cela implique que la queue de la distribution à droite est plus étirée et donc que les valeurs se concentrent plus vers la gauche. Le test de shapiro-wilk sur la normalité de la série nous indique une p-value inférieur à 0,05 (1,636e-12). Le taux de chômage ne suit donc pas une loi normale. Le graphique ci-dessous nous montre la distribution de la variable dépendante qui illustre bien les valeurs du kurtosis et du skewness.

Figure 4 : Histogramme de la série dépendante



Source : BAKHOUM&TESSON, dossier régressions pénalisées

8. Classifications

8.1 Analyse en composantes principales :

Nous allons maintenant réaliser une analyse en composante principale afin de synthétiser l'information de la base de données sur plusieurs axes et avoir un aperçu visuel des différentes corrélations qu'il pourrait y avoir entre les variables.

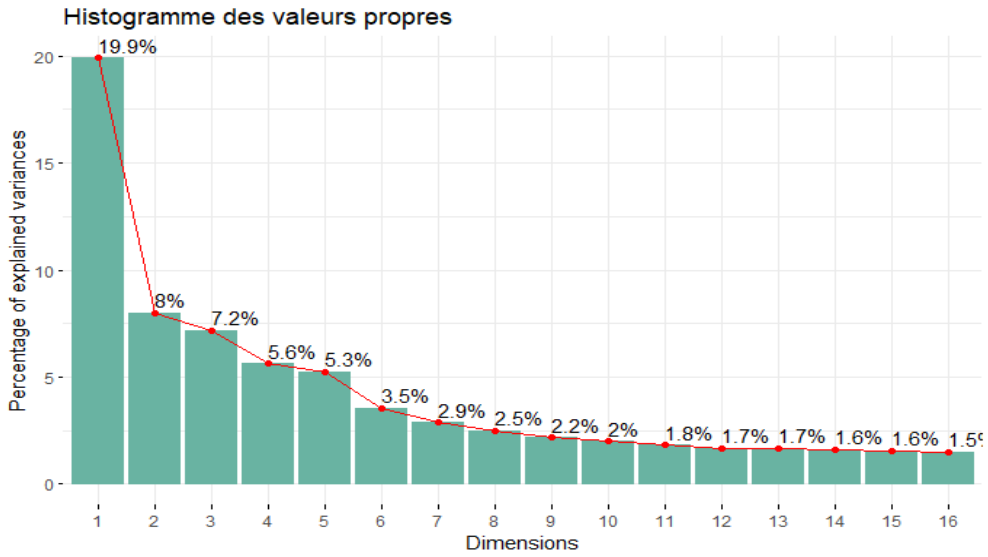
Tableau 6 : Valeurs propres des premières composantes

	Valeur propre	Pourcentage de la variance	Pourcentage cumulée de la variance
Comp 1	24,88	19,9	19,9
Comp 2	10,01	8	27,91
Comp 3	8,99	7,19	35,1
Comp 4	7,03	5,62	40,72
Comp 5	6,56	5,25	45,97
Comp 6	4,42	3,54	49,51

Source : BAKHOUM&TESSON, dossier régressions pénalisées

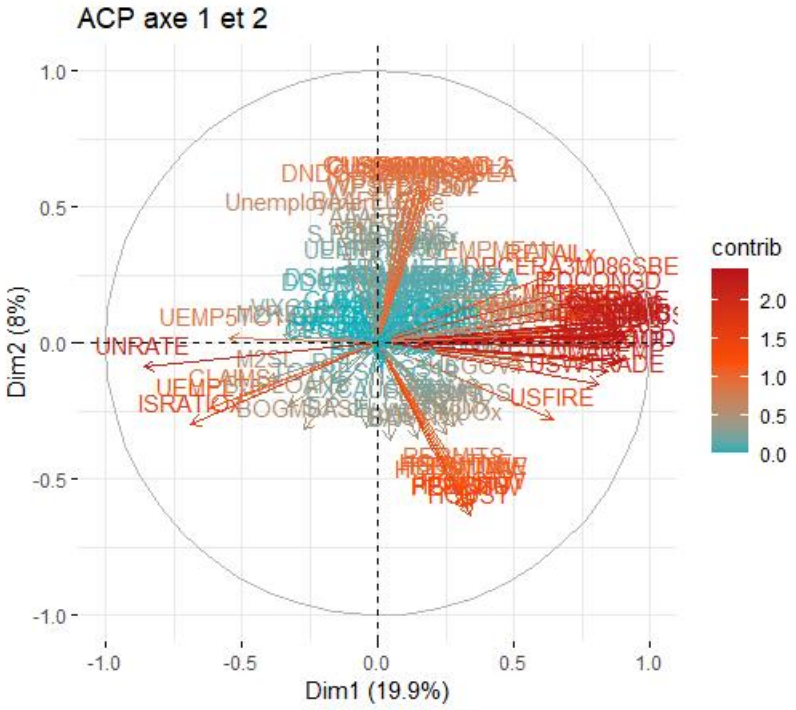
Ce tableau nous présente les valeurs propres sur les 6 premières composantes avec leur variance. On remarque que sur la première dimension, l'inertie est de 19,9 %. C'est une valeur très faible, l'information est mal représentée sur ce premier axe. De même avec en combinant les axes 2 et 3, l'inertie totale s'élève à 35,1 %. Ce qui est médiocre. Il faut attendre d'atteindre la 43ème composante pour avoir une variance cumulée supérieure à 90 %. Ceci peut s'expliquer par un grand nombre de variables fortement corrélés entre elles créant ainsi de la redondance dans l'information. Il faudrait donc diminuer le nombre de variable afin d'éviter que l'information ne se dilue trop. Pour l'analyse nous nous focaliserons que sur les deux premières dimensions car les valeurs propres sont supérieures à 1, jusqu'à la 27ème composantes et la technique du coude (voir histogramme ci-dessous) nous donne un grand nombre de variables latentes.

Figure 5 : Histogramme de valeurs propres



Source : BAKHOUM&TESSON, dossier régressions pénalisées

Figure 6 : ACP sur les deux premières dimensions



Source : BAKHOUM&TESSON, dossier régressions pénalisées

A première vue, il semble difficile d'analyser le cercle de corrélations, mais nous arrivons à distinguer des groupes qui se forment. Plus une variable est proche du bord du cercle et proche de l'un des axes, plus il contribue à l'inertie de la composante en question. Afin de déterminer les différentes variables latentes nous allons nous appuyer sur le tableau des contributions. Le long de l'axe 1, nous avons un regroupement de variables, en rouge sur la droite, qui correspondent pour la plupart à des variables en référence à la production et revenu et au marché du travail.

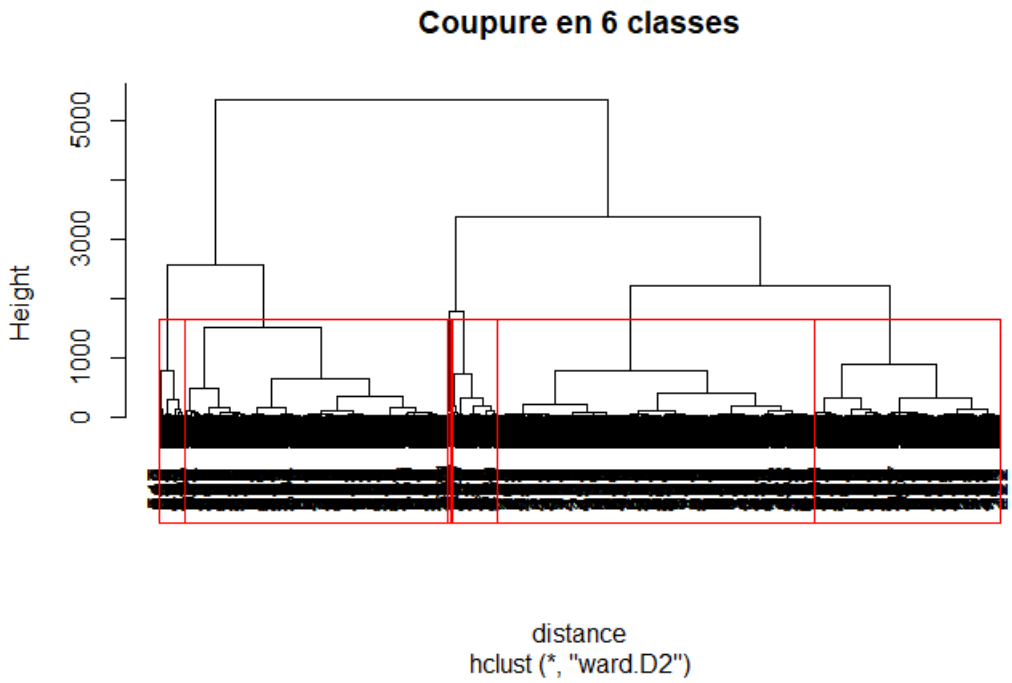
Pour l'axe 2, nous avons deux regroupements en bas et en haut en orange. Ces variables correspondent plus aux variables liées au logement, à l'immobilier. Mais nous avons également des variables liées au prix.

Cependant avec le nombre de variables, il est difficile de faire une analyse claire des composantes principales. On constate de plus, qu'il y a de nombreuses variables qui sont corrélées entre elles et qui diffèrent très peu en termes de signification, notamment avec les variables du logement.

8.2 Classification ascendante hiérarchique

Clustérisons maintenant réaliser une classification ascendante hiérarchique afin de créer des clusters entre les variables. Pour se faire, il nous faut déterminer tout d'abord le nombre de groupe optimal pour réaliser la classification. Nous utiliserons donc la fonction **pamk** du package **fpc**. Cette fonction nous recommande comme nombre représentatif de cluster 6. Nous obtenons donc la classification ascendante hiérarchique suivante :

Figure 7 : Classification ascendante hiérarchique



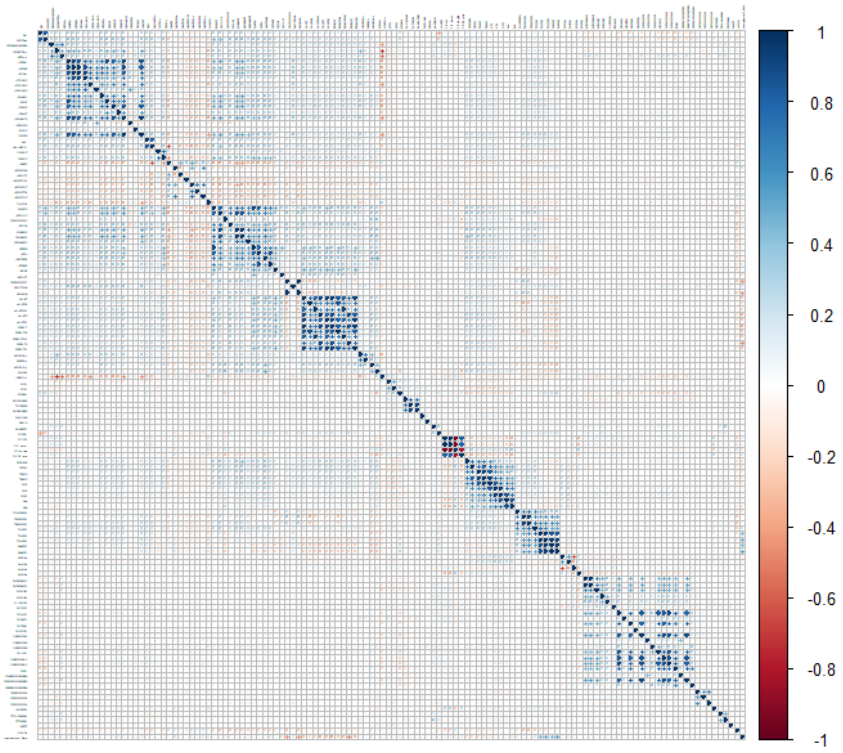
Source : BAKHOUM&TESSON, dossier régressions pénalisées

Cependant cette méthode n'aboutit à rien de concluant puisque les groupes sont déséquilibrés et le nombre important de variables nous empêche d'interpréter convenablement ces groupes.

9. Analyse des corrélations :

Vu que la plupart des variables ne suivent pas une loi normale, nous allons utiliser la méthode de corrélation de Spearman pour déterminer les différentes relations entre les variables.

Figure 8 : Corrélogramme de Spearman entre toutes les variables



Source : BAKHOUM&TESSON, dossier régressions pénalisées

Ce graphique ci-dessus, nous montre que la plupart des variables appartenant à une même catégorie, sont corrélées positivement entre elles (en bleu). En regardant le long de la diagonale représentée par la corrélation des variables par elles-même, on constate la formation de groupe. Il n'y a pas vraiment de variables anti corrélées importantes. Les corrélations sont montrées plus en détail en annexe.

III-Sélection de variables

Dans cette partie nous allons modéliser le taux de chômage américain par différentes approches. L'objectif étant de diminuer la dimension des données, c'est-à-dire limiter la complexité des modèles afin de synthétiser au mieux l'information fournie par les différentes variables. Nous utiliserons dans un premier temps une approche économétrique avec GETS avant d'appliquer des méthodes paramétriques, les régressions pénalisées. Enfin nous procéderons par une approche non linéaire de sélection de variables avec les forêts aléatoires.

Pour supprimer le biais d'échelle, nos données sont standardisées. Elles ont donc été centrées et réduites par leur moyenne et leur écart-type. Les différentes variables disposent donc d'une

moyenne égale à 0 et d'un écart-type égale à 1. Ceci met donc toutes les variables à la même échelle, les rendant indépendantes de leur unité respective. Il est donc plus simple de les comparer.

Pour la suite de notre étude nous intégrerons à notre base de données un retard pour chaque variable explicative et 4 retards pour la variable dépendante. Pour réaliser chaque méthode, il nous a fallu également supprimer la variable « NONBORRES ».

1.Approche économétrique : GETS

L'approche économétrique GETS (General-to-Specific) est une méthodologie qui vise à créer des modèles économétriques de manière itérative en commençant par des spécifications générales et en les réduisant progressivement à des spécifications plus spécifiques en utilisant des critères statistiques appropriés.

On a en premier lieu, les spécifications générales qui incluent toutes nos variables explicatives. L'approche GETS va ensuite réduire progressivement le modèle en éliminant des variables explicatives et des retards de manière séquentielle. Cela se fait en utilisant des critères de sélection tels que le critère d'information d'Akaike (AIC) ou le critère d'information bayésien (BIC). Ces critères mesurent la qualité du modèle en termes de précision et de parcimonie. A chaque étape de la réduction, des tests de spécification sont effectués pour vérifier si les hypothèses économétriques sont satisfaites. Ces tests peuvent inclure des tests d'hétéroscédasticité, d'autocorrélation ou de non-linéarité. Si un test échoue, cela peut indiquer que la spécification du modèle doit être ajustée. Le processus de réduction s'arrête lorsque le modèle atteint une spécification qui satisfait les critères de sélection de modèle et les tests de spécifications, ce qui signifie qu'il offre un équilibre entre précision et parcimonie.

Nous n'avons pas pu réaliser la méthode GETS pour cause de problème de singularité. En effet vu le nombre de variables incluses dans la base de données, il existe de fortes relations linéaires entre les variables. La matrice n'est pas de rang plein et donc non inversible. Il est suggéré pour palier à ce problème de réaliser une sélection de variables avant d'effectuer la modélisation. Nous verrons cela un peu plus loin dans notre étude.

2.Régressions pénalisées :

2.1 Ridge :

La régression Ridge consiste à ajouter des contraintes sur les estimateurs en réduisant leur amplitude et par conséquent leur sur-influence sur les prédictions. L'objectif est donc non pas de faire une simple régression linéaire classique mais une régression pénalisée qui va permettre de limiter l'impact des variables semblant être les moins pertinentes. Ce n'est pas une méthode de sélection de variables mais une méthode d'estimation de rétrécissement. Les estimateurs des variables les moins pertinentes ont donc des valeurs proches de 0 et ne sont pas annulés. L'avantage de cette méthode est que les effets de variables explicatives très corrélées se combinent pour se renforcer mutuellement. En termes d'inconvénient, vu que l'on garde toutes les variables, il n'y a donc pas moyen de savoir quelles variables sont les plus importantes. Nous avons ci-dessous la formule mathématique de la régression Ridge.

$$\hat{\beta}^{ridge} = \underset{\beta_0,\ldots,\beta_p}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

On part d'un modèle MCO (Moindres carrés ordinaires) où l'on rajoute une contrainte quadratique. L'estimateur n'est donc pas BLUE puisque qu'il est biaisé. Son espérance n'est pas égale à la vraie valeur du paramètre. Cependant la variance y est plus faible.

Tableau 7 : Résultats de la méthode Ridge

Nombre de variables	Valeur du lambda
252	0,001

Source : BAKHOUM&TESSON, dossier régressions pénalisées

Les résultats obtenus sont cohérents puisque nous obtenons 252 variables (nombre égale aux variables explicatives). En effet, la régression Ridge ne permet pas de sélectionner des variables. Le lambda quant à lui est égale à 0,001 et est estimé par cross-validation. Ce qui correspond à une contrainte plutôt faible. Les variables les moins pertinentes sont donc moins fortement pénalisées et proches des coefficients réalisés par MCO.

2.2 LASSO

La méthode du Lasso, contrairement à celle du Ridge, permet de faire une sélection de variables en éliminant celles qui sont inutiles. C'est le principe de parcimonie. Elle rend nul certains

coefficients de l'estimation par une méthode de shrinkage. Le calcul des estimateurs se fait numériquement par des algorithmes d'optimisation convexe. Cependant, il existe un inconvénient avec cette méthode. En effet, en présence de variables explicatives corrélées, la régression Lasso en choisit une arbitrairement et met les autres à 0. C'est le biais de sélection. La formule du Lasso est indiquée ci-dessous.

$$\hat{\beta}^{lasso} = \underset{\beta_0,\ldots,\beta_p}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Comme la méthode du Ridge, nous utilisons la régression des moindres carrés ordinaires pour laquelle on ajoute une norme de régularisation en valeurs absolues.

Tableau 8 : Résultats de la méthode Lasso

Nombre de variables	Valeur du lambda
5	0,002535364

Source : BAKHOUM&TESSON, dossier régressions pénalisées

Cette régression nous donne 5 variables. En effet les autres ont été supprimées car jugées inutiles. Le modèle est donc plus parcimonieux que celui du Ridge. De plus, le lambda est plus important avec une valeur de 0,002. La pénalité est donc plus forte que le modèle précédent. Par la présence du biais de sélection, nous ne pouvons pas choisir les variables les plus intéressantes à garder pour l'interprétation parmi le groupement de variables corrélées. Il existe pour cela d'autres méthodes qui limite cela.

2.3 Elastic-net

Pour pallier le problème de parcimonie de la méthode Ridge et au biais de sélection de celle de Lasso, nous utilisons la régression Elastic-net. Elle combine les pénalisations des deux modèles précédents. Elle permet de moins discriminer les variables trop corrélées qui étaient néanmoins importante dans la conception du modèle final, ce qui va réduire le biais de sélection et va appliquer un coefficient nul pour les variables les moins pertinentes. Il y a également un effet de regroupement dans cette régression, c'est-à-dire que les variables fortement corrélées entre elles auront des coefficients proches.

$$\hat{\beta}^{en} = \underset{\beta_0,\ldots,\beta_p}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} \left(\alpha \beta_j^2 + (1-\alpha)|\beta_j| \right) \right\}$$

Comme pour les deux régressions précédentes, ce modèle suit une régression des moindres carrés ordinaires en introduisant une pénalité qui est un compromis entre la méthode Ridge et Lasso. On retrouve une contrainte quadratique et une norme en valeur absolue. Le lambda dépend donc de la pondération du paramètre alpha. Comme pour la régression du Lasso, des méthodes numériques sont utilisées pour calculer les estimateurs.

Tableau 9 : Résultats de la méthode Elastic-net

Alpha	Nombre de variables	Valeur du lambda
0,75	12	0,001747528

Source : BAKHOUM&TESSON, dossier régressions pénalisées

Les résultats obtenus nous donnent comme valeur du paramètre alpha égale à 0,75, qui est un compromis entre les pénalités Lasso et Ridge. Ceci nous donne 12 variables explicatives, ce qui est plus que la régression Lasso. Ce résultat est cohérent puisque la régression Elastic-Net pénalise moins les variables corrélées. Il est donc normal d'avoir plus de variables. Nous avons en lambda une valeur de 0,00175 qui est moins importante que la régression Lasso et plus forte que la régression Ridge.

2.4 SCAD

Pour résoudre les problèmes de pénalités non convexes, nous pouvons utiliser la méthode SCAD. Elle propose une fonction de pénalité non-concave. La formule mathématique des estimateurs SCAD est donné ci-dessous :

$$\hat{\beta}_{scad} = \underset{\beta_0,\ldots,\beta_p}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \sum_{j=1}^{p} p_\lambda^{scad}(|\beta_j|) \right\}$$

où $p_\lambda^{scad}(\beta)$ est la fonction de pénalité définie sur $[0,\infty)$, avec $a > 2$, par

$$p_\lambda^{scad}(\beta) = \begin{cases} \lambda|\beta| & \text{si } |\beta| \leq \lambda \\ \frac{(2a\lambda|\beta|)-\beta^2-\lambda^2)}{2(a-1)} & \text{si } \lambda < |\beta| \leq a\lambda \\ \frac{\lambda^2(a^2+1)}{2} & \text{si } |\beta| > a\lambda \end{cases}$$

L'avantage de cette pénalité est de varier en fonction de la valeur absolue de l'estimateur beta. En effet, pour des petites valeurs de beta, où la valeur absolue de beta est inférieure à lambda, la pénalité sera linéaire en beta. Lorsque la valeur absolue de beta est égale au paramètre lambda, la pénalité SCAD coïncide avec celle du Lasso. Pour des valeurs moyennes où la valeur absolue de beta est comprise entre lambda et le produite de $\boldsymbol{\alpha}$ fois lambda, la pénalité devient quadratique. Enfin pour des valeurs élevées où la valeur absolue de beta est supérieure au produit a fois lambda, la pénalité est constante par rapport à beta. Les valeurs les plus élevées ne sont donc plus pénalisées. Cette méthode permet donc de pénaliser plus sévèrement les petits coefficients donnant un modèle plus parcimonieux que le Ridge et moins les grands coefficients pour réduire le biais de sélection.

Tableau 10 : Résultats de la méthode SCAD

Nombre de variables	Valeur de lambda
2	0,01931655

Source : BAKHOUM&TESSON, dossier régressions pénalisées

Les résultats du modèle nous montrent un modèle plus parcimonieux que le Ridge, l'Elastic-Net et le Lasso avec seulement 2 variables retenues. La valeur du lambda est 0,019. Comparé aux autres modèles, c'est celui qui pénalise le plus.

2.5 Adaptive Lasso

Pour résoudre le problème de sélection de variables du Lasso, nous pouvons opter pour la régression Adaptive Lasso. En effet cette approche utilise la même pénalité que le lasso à la différence que l'aLASSO introduit dans son modèle un paramètre de pondération généré par un premier modèle qui peut être un Ridge, Lasso ou bien Elastic-Net.

$$\hat{\beta}^{alasso} = \underset{\beta_0,\ldots,\beta_p}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} \widehat{\omega}_j |\beta_j| \right\}$$

$\widehat{\omega}_j = 1/|\widehat{\beta}_j^*|^{\gamma}$, avec $\gamma > 0$

$\widehat{\beta}^*$: estimation du paramètre initial en 1ère étape

Dans notre cas, nous avons utilisé les pondérations issues modèle Ridge dans la première étape avant d'estimer le modèle aLASSO. Le lambda optimal de notre modèle est estimé ici à 0.2485499 qui correspond au lambda minimum. Nous avons constaté que le modèle aLASSO n'avait sélectionnés que deux variables.

Tableau 11 : Résultats de la méthode aLASSO

Nombre de variables	Valeur de lambda
2	0,2485499

Source : BAKHOUM&TESSON, dossier régressions pénalisées

A présent nous allons présenter le tableau récapulatif de l'estimation de nos hyperparamètres pour les différentes méthodes

Tableau 12 : Estimation des hyperparamètres

	Ridge	Lasso	Elastic-net	SCAD	aLASSO
Lambda	0,001	0,002535364	0,001747528	0,01931655	0,2485499
alpha			0,75		

Source : BAKHOUM&TESSON, dossier régressions pénalisées

3.Approche de réduction de dimension : SIS

L'utilisation de grande quantité de données peu entrainée parfois des problèmes de multicolinarité ou de surajustement. De ce fait des méthodes de réduction de dimension ont été initié afin de pallier ces problèmes. Dans le cas de notre étude nous choisit la méthode SIS (Sure Independent Screening) qui est une méthode de présélection de variables basé sur un apprentissage des corrélations. En effet on fait le classement des prédicateurs en fonction de la mesure d'utilité entre la cible et chaque prédicateur.

Nous avons ainsi estimé la méthode SIS en appliquant une validation croisée et un maximum de 100 prédicteurs. Une fois le modèle estimé nous avons appliqué en amont des méthodes de régressions pénalisées (Ridge, Lasso, EN, aLASSO et SCAD) mais aussi la méthode GETS.

Tableau 13 : Sélection des variables selon l'approche SIS

	Sans SIS	**Avec SIS**
Ridge	252	66
Lasso	5	10
Elastic-Net	12	18
aLASSO	2	5
SCAD	2	4
GETS	.	20

Source : BAKHOUM&TESSON, dossier régressions pénalisées

La sélection de variable nous a permis de réaliser l'approche GETS. Il n'y a pas de problème de singularité et il n'y a pas non plus de présence d'autocorrélations. En regardant le tableau ci-dessus on remarque que la sélection de variables par SIS permet une fois les différentes modélisations effectuées de retenir plus de variables.

4.Approche non linéaire : Random Forest

Le Random Forest est une méthode d'apprentissage ensembliste qui a été mis en place par Breiman et Cutler. L'algorithme qui sous-tend le Random Forest est une extension de la méthode Bagging, car elle repose sur le bagging et le feature randomness pour créer une forêt d'arbres de décisions non corrélées.

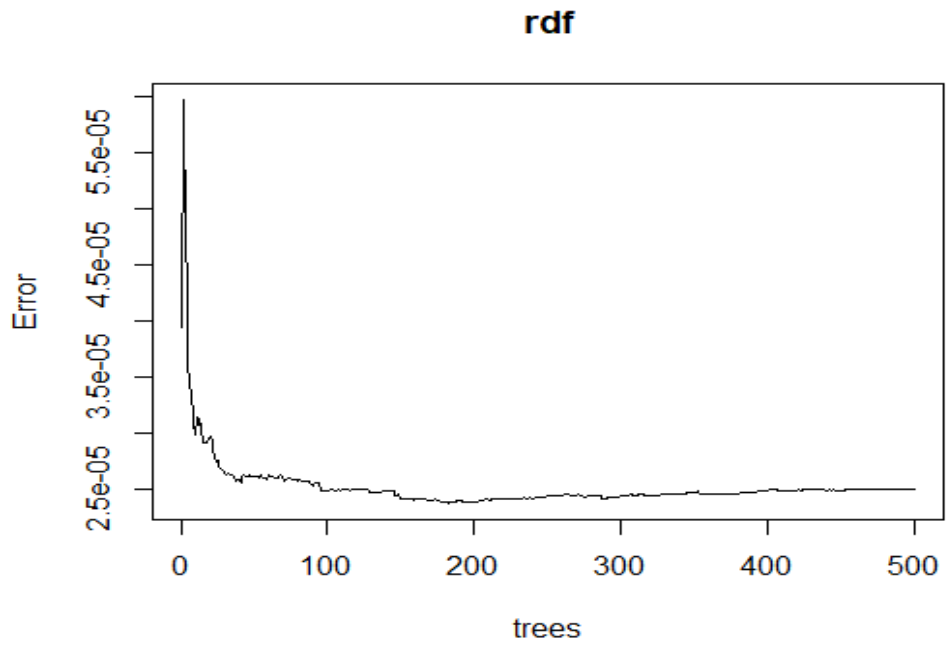
Les principaux hyperparamètres à savoir le nombre d'arbres, la taille des nœuds et le nombre de fonction d'échantillonnées n'ont pas à être défini avant l'entrainement dans l'algorithme. Le classificateur du Random Forest peut être utilisé pour résoudre soit des problèmes de classification (variable cible est numérique) ou de régression (variable cible est qualitative).

L'algorithme du Random Forest est composé par un ensemble d'arbres de décisions dont chaque arbre comprend un ensemble de données tiré d'un échantillon Bootstrap. Une partie de ces données sont mis de côté en tant que données de test appelés échantillon Out-Of-Bag (OOB). La prédiction est déterminée suivant le type d'arbres. Pour les arbres de classification, on utilise la catégorie ayant la plus grande fréquence au niveau de la feuille considérée (échantillon d'apprentissage) et pour les arbres de régression, on prend la moyenne des valeurs de la variable cible pour les observations (échantillon d'apprentissage) atteignant une feuille donnée. Enfin l'échantillon OOB est utilisé pour effectuer une validation croisée afin de finaliser la prédiction.

L'avantage sur l'utilisation des Random Forest est qu'il permet la réduction du risque de surajustement. Mais aussi d'évaluer facilement l'importance ou la contribution des variables au modèle. Pour déterminer cette importance est mesuré par le **MDA** (Mean Decrease Accuracy) ou le **MDI** (Mean Decrease in node Impurity).

Nous allons ainsi appliquer la méthode des Random Forest sur notre jeu donné afin de faire une sélection de variables en passant par la contribution ou l'importance des variables au modèle. Nous avons d'abord estimé le modèle **rdf** avec un nombre d'arbres égale à 500 puis représenter le graphique de cette fonction.

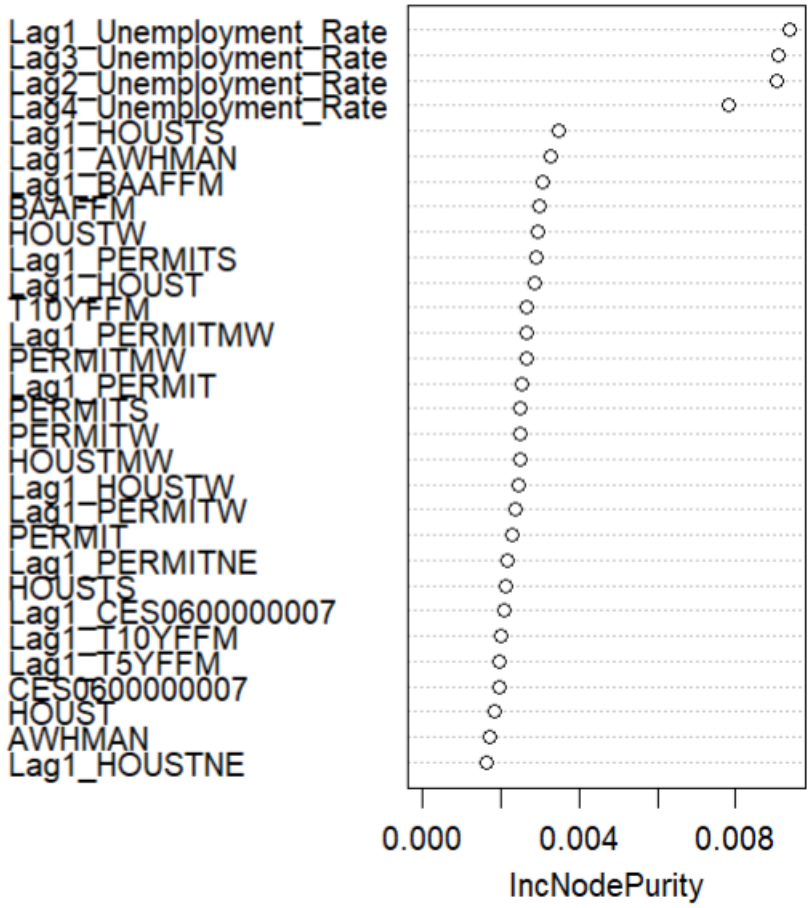
Figure 9 : Nombres d'arbres optimal



Source: BAKHOUM&TESSON, dossier régressions pénalisées

Nous constatons que nous obtenons quasiment la même erreur de prévision aussi bien avec 300 arbres qu'avec 500. Donc il serait plus optimal de retenir 300 arbres pour la suite. Nous notons également sur l'échantillon d'apprentissage que nos variables explicatives expliquent 91,31% du chômage aux États-Unis. Une fois le modèle estimé nous pouvons à présent déterminer l'importance des variables au modèle et faire ensuite la sélection. Nous utiliserons par défaut un **mtry** égale à 5.

Figure 10 : Variables importantes selon le modèle Random Forest



Source : BAKHOUM&TESSON, dossier régressions pénalisées

Le graphique ci-dessus décrit l'importance ou la contribution des variables au modèle en fonction du Mean Decrease in node Impurity (MDI) qui mesure la contribution de chaque variable à la réduction de l'impureté lors de la construction des arbres de décision. Plus la diminution est élevée et plus la variable est importante. Nous retrouvons ainsi les variables retardées du taux de chômage, « Lag1_Unemployment_Rate », « Lag3_Unemployment_Rate », « Lag2_Unemployment_Rate » et « Lag4_Unemployment_Rate » comme variables les plus importantes.
Pour la sélection des variables nous avons considéré la fonction VSURF compris dans le package VSURF, une variable ont été sélectionnée à des fins d'interprétation. C'est la variable « Lag1_VIXCLSx ».

Après avoir effectué le filtrage avec la méthode SIS, nous avons réestimer le modèle Random Forest. Nous trouvons ainsi que le nombre d'arbres optimal est évalué à 300. Pour l'importance des variables, nous retrouvons trois variables retardées du taux de chômage. La sélection des variables a retenu 1 variables pour l'interprétation, celle du « Lag1_Unemployment_Rate ».

Tableau 14 : Variables retenues avec l'approche du Random Forest

	Variables retenues
Random Forest sans filtrage	1
Random Forest avec filtrage	1

Source : BAKHOUM&TESSON, dossier régressions pénalisées

5.Comparaison des méthodes de régression

Dans cette partie, nous avons les variables sélectionnées par les différentes méthodes

Tableau 15 : Comparaisons des variables sélectionnées suivant différentes méthodes

Variables	**Gets**	**Ridge**	**Lasso**	**Elastic-net**	**aLasso**	**SCAD**	**S_ridge**	**S_gets**	**S_lasso**	**S_EN**	**S_SCAD**	**S_aLasso**
UEMP15OV		6.824718e-03					2.9596e-02	0.0306026837	0.0136052322	0.01753960 78	8.788254e-03	0.0136052322
UEMP27OV		-2.990696e-03					3.017055e-03			0.00019107 41		
PAYEMS		-2.186478e-03					-1.665235e-01		-0.12209 16918	-0.08890238 54	-2.554027e-01	-0.12209169 18
USGOOD		-7.492997e-03					6.445907e-02					
USCONS		3.639534e-03					-8.532449e-03					
MANEMP		1.158453e-03					-6.504803e-02	-0.019458849				
DMANEMP		2.254930e-03					2.737576e-02					
SRVPRD		-1.799568e-03					-1.131022e-01	-0.2531347 03	-0.09310 50198	-0.11841401 25		-0.0931050198
USWTRADE		-7.617708e-04		-1.638286e-04			1.100320e-02	0.0215953 58				
USFIRE		-7.261219e-05					8.078879e-03					
USGOVT		-1.558827e-04					2.364754e-02	0.0237160 43				
CES0600000007		-6.425018e-03					1.482892e-02					
AWHMAN		4.435060e-03		-4.151067e-04			-1.130624e-02		-0.01711 59473	-0.02024659 05		-0.0171159473
HOUST		-3.143257e-04					-2.904763e-02	-0.028130824				
HOUSTNE		2.449786e-03					1.248952e-02			0.00031239 71	2.184325e-04	

HOUSTMW		-9.656732e-04					-5.919555e-06					
HOUSTS		-1.710499e-03					-8.220271e-03					
HOUSTW		-3.395644e-03					4.253740e-04					
PERMIT		5.043310e-03					6.433400e-03					
PERMITNE		1.959730e-04					4.606663e-03					
PERMITMW		-3.743711e-06					1.451747e-02					
PERMITS		3.918902e-03					-7.573597e-03		-0.0009356592	-0.01105018 53		-0.0009356592
PERMITW		-7.077625e-04					-2.244832e-02					
AMDMUOx		-2.156660e-03					9.144227e-03	0.01028054 1				
M2REAL		1.477426e-02					-3.304292e-03					
BAA		1.703808e-03					-5.029087e-03					
T5YFFM		-2.648965e-03					-5.698769e-02	-0.051281610				
T10YFFM		5.683484e-03					4.835416e-02					
AAAFFM		-4.374607e-03					-2.284673e-02					
BAAFFM		3.537379e-03					6.767538e-02	0.08442338 2				
Lag1_HWI		1.559013e-03					9.963077e-03	0.00967956 1				
Lag1_CLF16OV		2.471249e-02					1.338858e-03					
Lag1_UEMP5TO14		7.699077e-04					2.646828e-02	0.02971634 9		0.01746936 51		
Lag1_UEMP15T26		1.958679e-03					-5.316664e-03		-0.0008017481	-0.00301155 58		-0.0008017481
Lag1_CLAIMSx		4.610553e-04		1.021516e-03			3.021576e-02	0.02928049 9	0.03130 91114	0.04730891 31		0.03130911 14
Lag1_PAYEMS		-9.928992e-03		-2.659321e-04			-9.240275e-02	-0.08237987 9		-0.02386091 39		
Lag1_CES1021000 0001		5.697707e-04					-6.549318e-03					
Lag1_USCONS		-9.632467e-04					1.162330e-02					
Lag1_MANEMP		1.522704e-03					1.366458e-02					
Lag1_NDMANEMP		-1.938327e-04					-7.356700e-03					

Lag1_USTPU		-5.760621e-04					4.691963e-02					
Lag1_USTRADE		1.317342e-04					-4.475335e-02			-0.00102816 38		
Lag1_USFIRE		7.731449e-04					1.519076e-02	0.022761494				
Lag1_USGOVT		1.783777e-03					1.353288e-02	0.013124326				
Lag1_AWOTMAN		1.043059e-03					3.646040e-03					
Lag1_AWHMAN		-4.315540e-03					-1.842002e-02			-0.00491502 48		
Lag1_HOUST		1.612181e-03					8.042768e-03					
Lag1_HOUSTNE		1.326543e-03					2.538601e-03					
Lag1_HOUSTMW		3.204442e-03					1.885058e-02	0.013285167				
Lag1_HOUSTS		-4.258195e-03					8.571683e-03					
Lag1_HOUSTW		3.095219e-05					1.100604e-02					
Lag1_PERMIT		3.551169e-03		-4.653099e-05			-1.374898e-02					
Lag1_PERMITNE		-5.899083e-03					-5.810795e-03					
Lag1_PERMITMW		-3.410475e-03					-2.157954e-02					
Lag1_PERMITS		-3.042958e-03					-3.335363e-04		-0.00692 50771	-0.00353590 64		-0.00692507 71
Lag1_ANDENOx		1.117541e-04					-1.141834e-03					
Lag1_M2SL		-4.193497e-04					1.960177e-03			0.001374226 3		
Lag1_T1YFFM		5.952703e-04					3.718952e-03					
Lag1_T5YFFM		-1.215639e-03					5.292059e-02	0.089594760				
Lag1_T10YFFM		5.229751e-03					1.749273e-03					
Lag1_AAAFFM		-5.914436e-03					-8.671886e-02	-0.122377637				
Lag1_INVEST		-3.584358e-04					1.320782e-03					
Lag1_VIXCLSx		1.052139e-04					1.994680e-03					
Lag1_Unemployment_Rate		6.156779e-01	0.9951604315	9.422807e-01	0.7133985	9.988846e-01	6.365140e-01	0.649574755	0.9008072724	0.720095401	9.804687e-01	0.9008072724
Lag2_Unemployment_Rate		3.336613e-01	0.0006650252	5.470325e-02			3.226385e-01	0.323564329	0.0594845434	0.219977371 4		0.0594845434
Lag3_Unemployment_Rate		3.357524e-02					-6.358867e-04			0.018240754 3		

CE160V		-7.706972e-02	-0.0019967168	-2.784821e-03								
UNRATE		2.138631e-01	0.2696041165	2.678246e-01	2.748898e-01							
UEMP5TO14		3.352368e-03	0.0007436274	2.956785e-03								
Lag1_UNRATE		5.955508e-02		1.208845e-02								
Lag1_UEMPLT5		-3.954970e-03		1.635975e-04								

Source : BAKHOUM&TESSON, dossier régressions pénalisées

D'après le tableau ci-dessus, nous pouvons constater que seule la variable retardée d'une période du taux de chômage est présente dans tous les modèles. Elle a également été sélectionnée par le random forest après filtrage. Elle dispose pour chacun des modèles, d'un coefficient positif. On peut donc interpréter que l'augmentation d'une unité du taux de chômage retardé d'une période augmente le taux de chômage à l'instant t.

On peut observer de même que pour les modèles sans filtrage, la variable la plus retenue est la variable « UNRATE » qui correspond au taux de chômage civil. Cette variable a une signification semblable à la variable dépendante. On aurait pu la supprimer de notre base pour effectuer l'analyse.

Enfin la variable la plus sélectionnée après filtrage, est la variable « UEMP150V ». Cette variable correspond au taux de chômage supérieur à 15 semaines. On voit que la signification est également proche de celle de la variable dépendante.

En soit on constate que ce sont les mêmes variables, en général, qui ressortent des différents modèles.

IV Conclusion :

Le but de cette étude était de déterminer des indicateurs macroéconomiques qui influençait la fluctuation du taux de chômage américain sur le période de janvier 1970 à Aout 2023. Pour ce faire nous avons choisi un jeu de donnée issues de différents groupes à savoir la production, la consommation, le logement, le marché du travail, les taux intérêts/changes, le prix et le marché des actions. Nous avons récolté au total 127 variables sur lesquelles un pré-tr**a**itement (valeurs manquantes, outliers, saisonnalité, stationnarité) a été effectuées au préalable.

Nous avons tout ensuite réalisé des classifications telle l'analyse en composantes principales (ACP) ou la classification ascendante hiérarchique qui nous ont permis de déterminer des groupes qui se formaient entre les variables. On a constaté la présence de nombreuses corrélations, surtout des variables qui appartenaient à une même catégorie. Le nombre de variables étant important avec une base de données constituée de 127 variables explicatives, nous avons conclue qu'il était difficile de réaliser une interprétation claire avec les classifications.

Par la suite dans le but de trouver des variables qui serait plus pertinentes sur nos variables pour expliquer le chômage américain, nous avons opté sur différentes méthodes de sélection de variables. Parmi celles-ci nous avons l'approche économétrique GETS mais aussi des méthodes de régressions pénalisées (Ridge, Lasso, EN, SCAD et aLASSO). Une approche de réduction de dimension SIS et approche non linéaire (Random Forest) ont été également appliquée à nos données. Toutes ces méthodes ont sélectionné au moins une variable pour expliquer le chômage à l'exception du GETS. Il aura fallu l'utilisation de la méthode à réduction de dimension, SIS pour qu'il en sélectionne 20. A la suite de l'application de ces méthodes, nous avons en appliquant différentes régressions pénalisées nous avons constaté que le paramètre Lambda varie entre les modèles et elle est forte au niveau de l'Adaptative Lasso.

Par l'utilisation de toutes ces méthodes, la conclusion fut qu'on retrouvait la plupart du temps les mêmes variables explicatives. Nous avons également constaté que les fluctuations des taux de chômage retardés impactaient la valeur en t du taux de chômage, particulièrement en t-1 et t-2.

Bibliographie :

Définition du chômage :
https://travail-emploi.gouv.fr/IMG/pdf/Definitions-2.pdf

« *Taux de chômage aux Etats-Unis entre 1990 et 2022* », site Statista, publié par Sheelah Delestre, 25 oct. 2023 :
https://fr.statista.com/statistiques/550404/taux-de-chomage-aux-etats-unis-1990/#:~:text=Cette%20statistique%20repr%C3%A9sente%20le%20taux,%C3%A9tant%20quasiment%20du%20plein%20emploi.

« *L'essentiel sur...le chômage* » site de L'INSEE, paru le 11/08/2023 :
https://www.insee.fr/fr/statistiques/4805248

Michael W. McCracken, Economic research Federal Reserve Bank of St.Louis:
https://research.stlouisfed.org/econ/mccracken/fred-databases/

Civilian Unemployment Rate for US and California, site California open data portal:
https://data.ca.gov/dataset/civilian-unemployment-rate-for-us-and-california/resource/6b59d10b-ca5d-465f-8ee2-0780fb1044c0?filters=Area%20Name%3AU.S.

« *Loi d'Okun* », site Pour l'Eco paru, le 12 juin 2023
https://www.pourleco.com/le-dico-de-l-eco/loi-dokun

Qu'est-ce que l'algorithme de forêt aléatoire (Random Forest)
https://www.ibm.com/fr-fr/topics/random-forest

Annexes :

Figure 11 : Valeurs manquantes

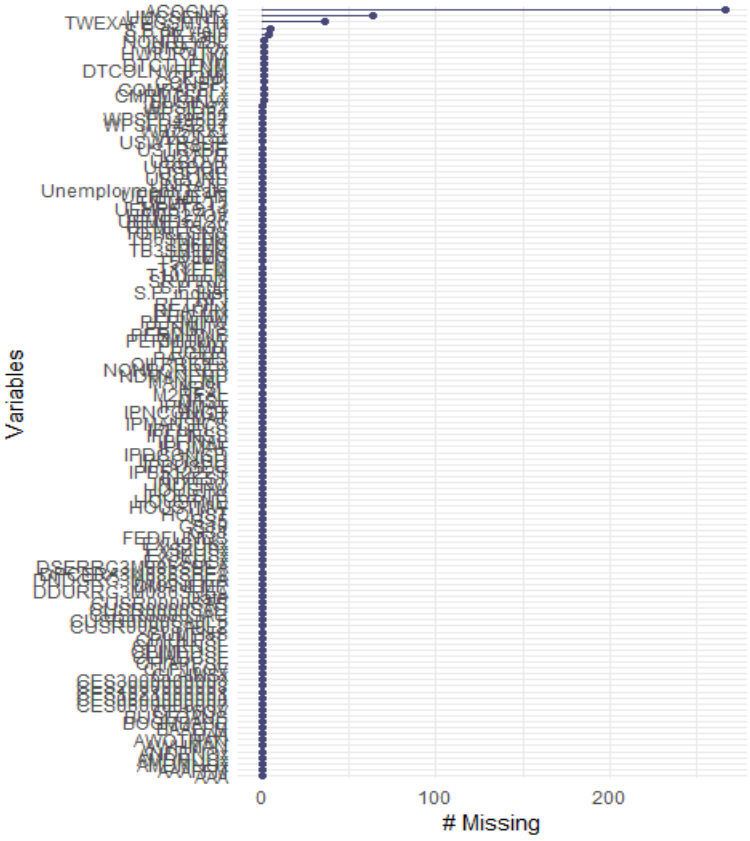


Figure 12 : Nombre de groupe optimal pour la classification

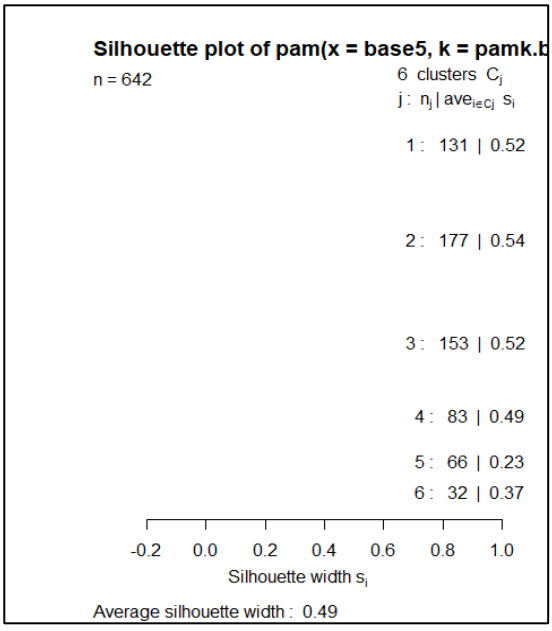


Figure 13 : Cor_matrice1

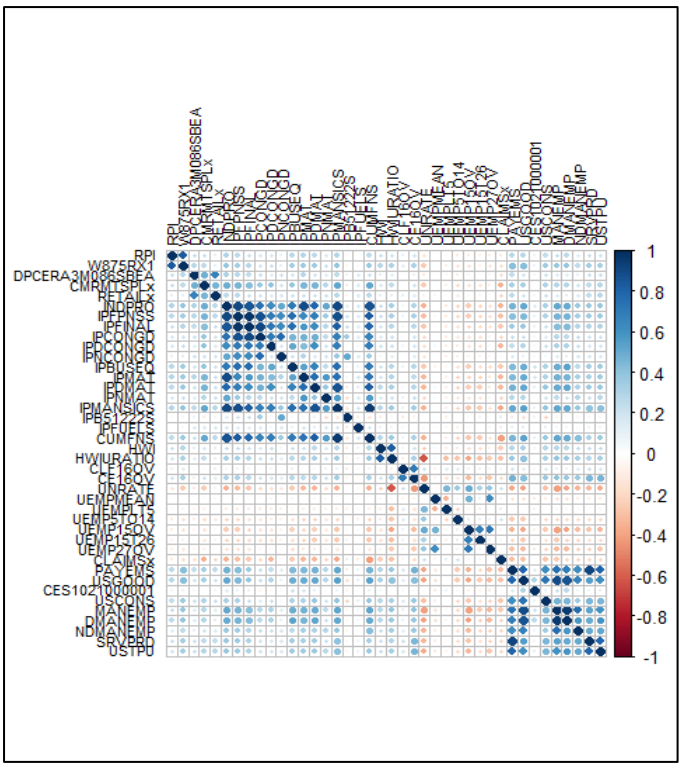


Figure 14 : Cor_matrice2

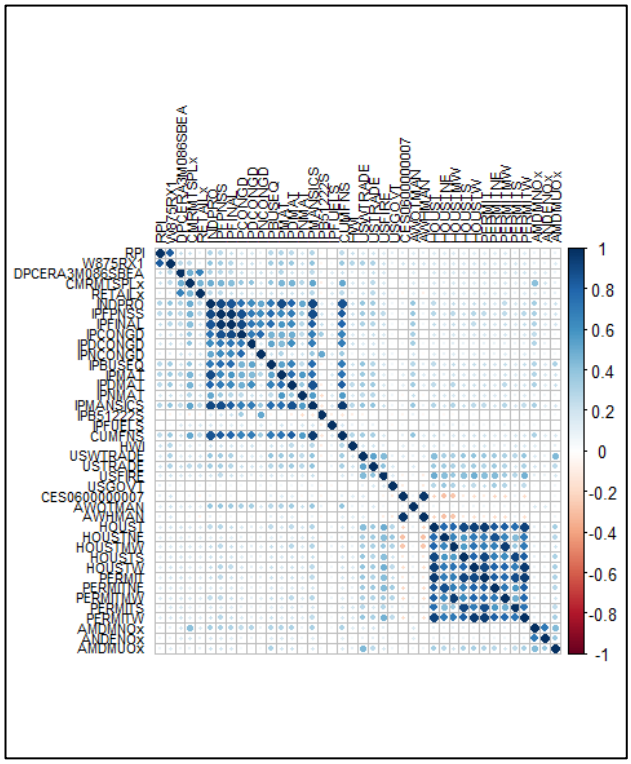


Figure 15 : Cor_matrice3

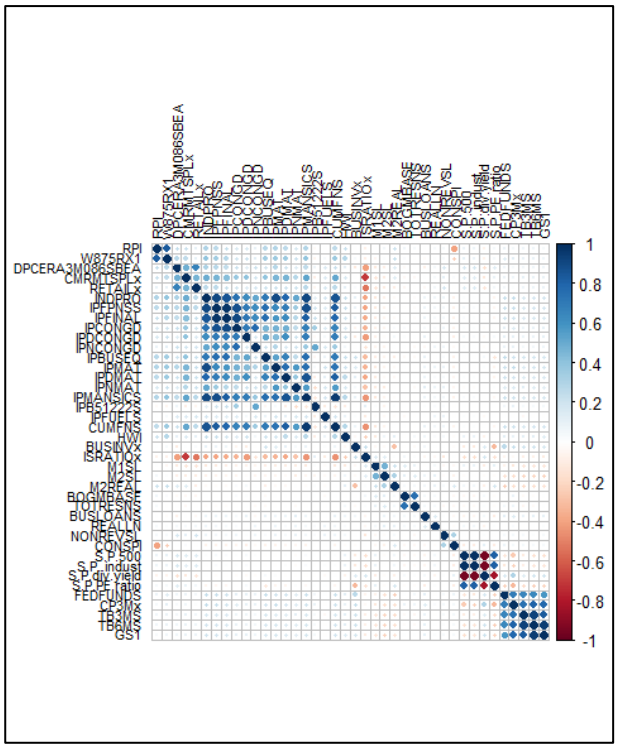


Figure 16 : Cor_matrice4

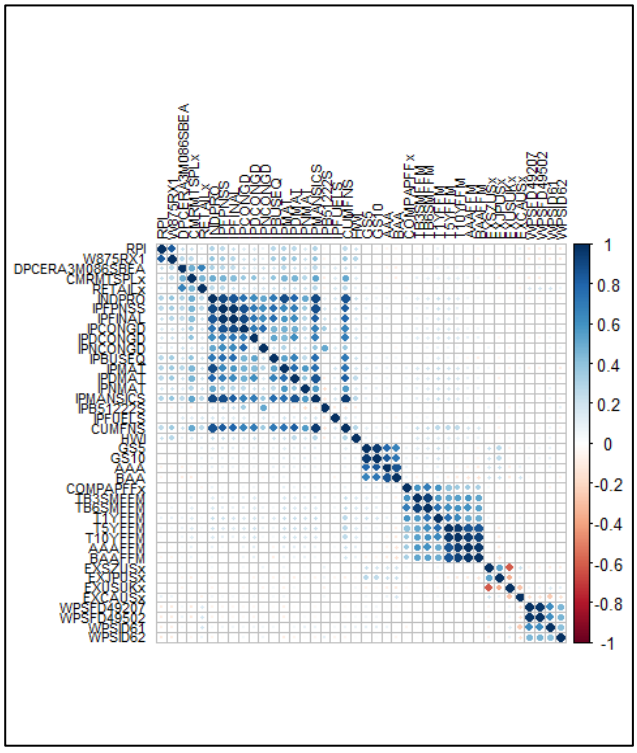


Figure 17 : Cor_matrice5

Figure 18 : Cor_matrice6

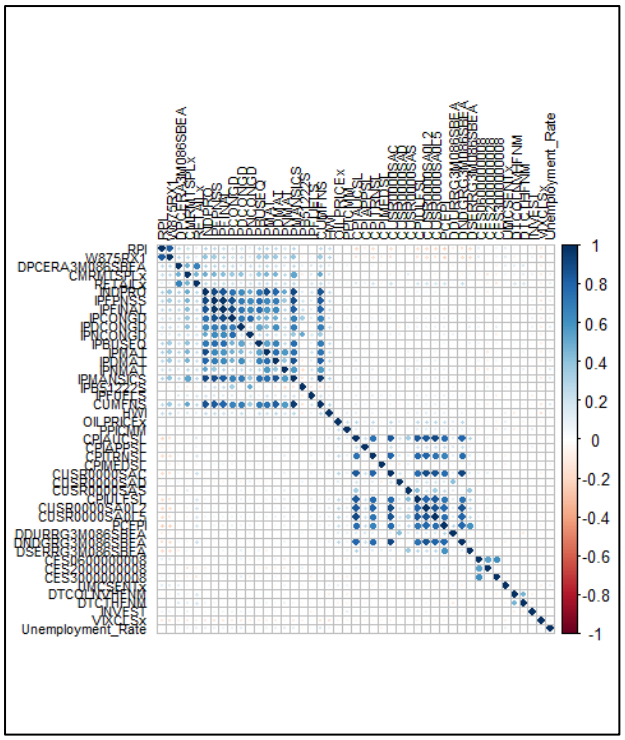


Figure 19 : Cor_matrice7

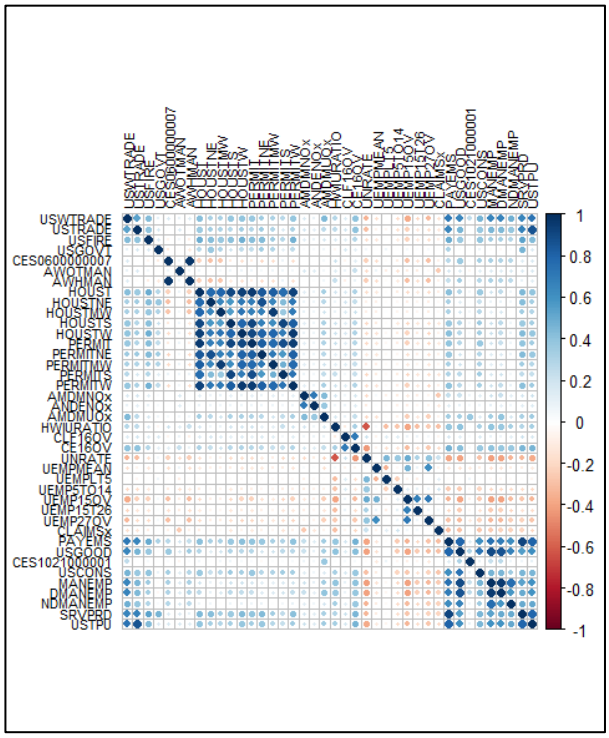


Figure 20 : Cor_matrice8

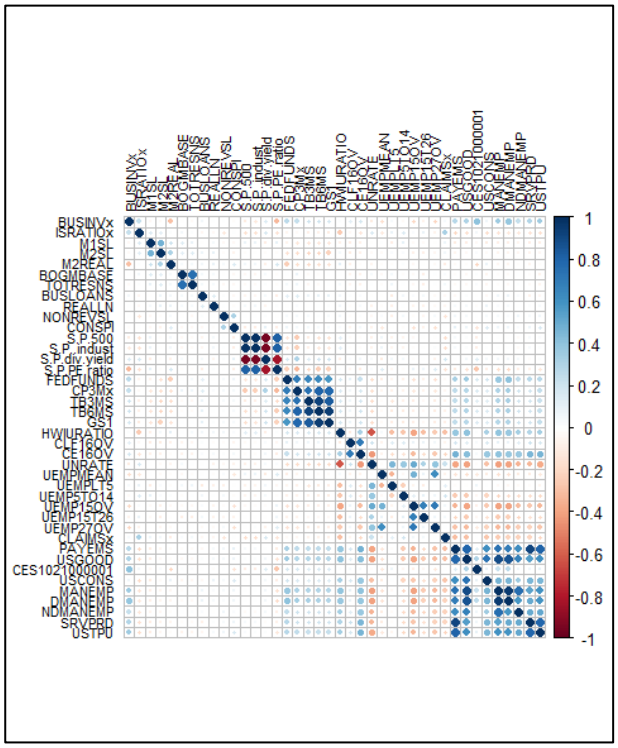


Figure 21 : Cor_matrice9

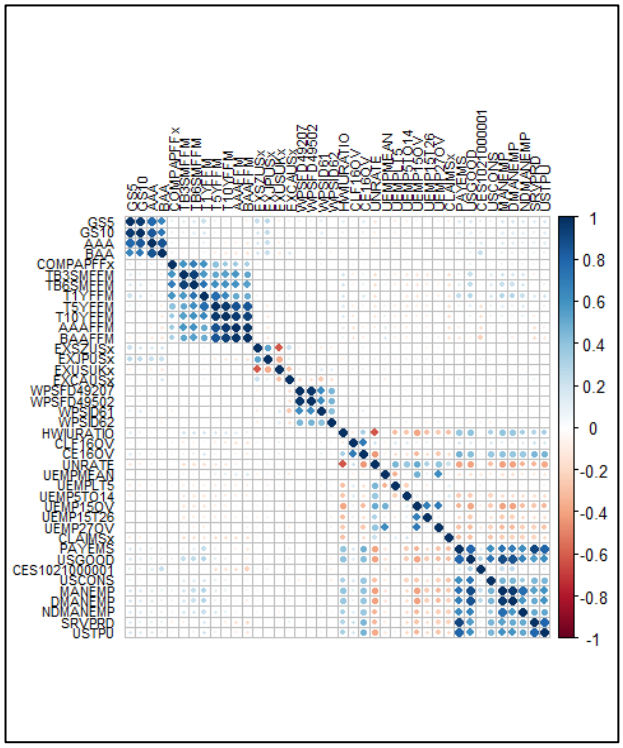


Figure 22 : Cor_matrice10

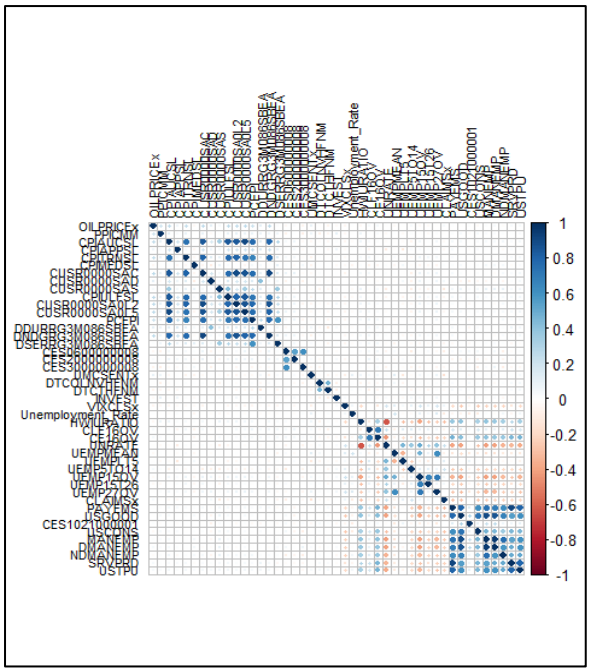


Figure 23 : Cor_matrice11

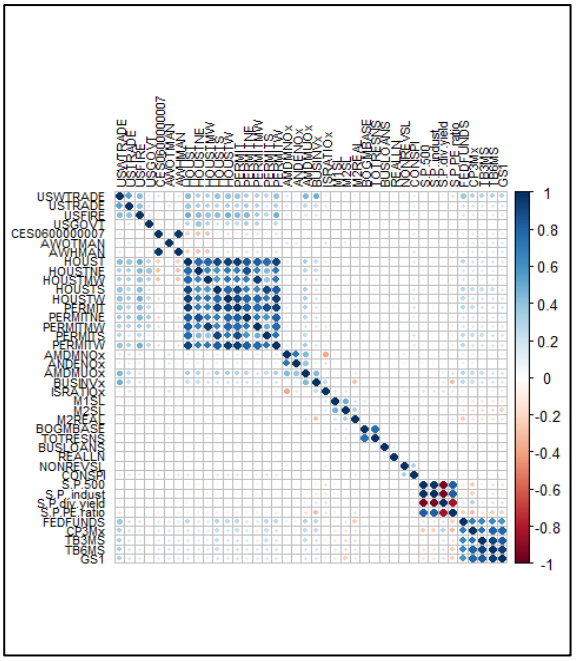


Figure 24 : Cor_matrice12

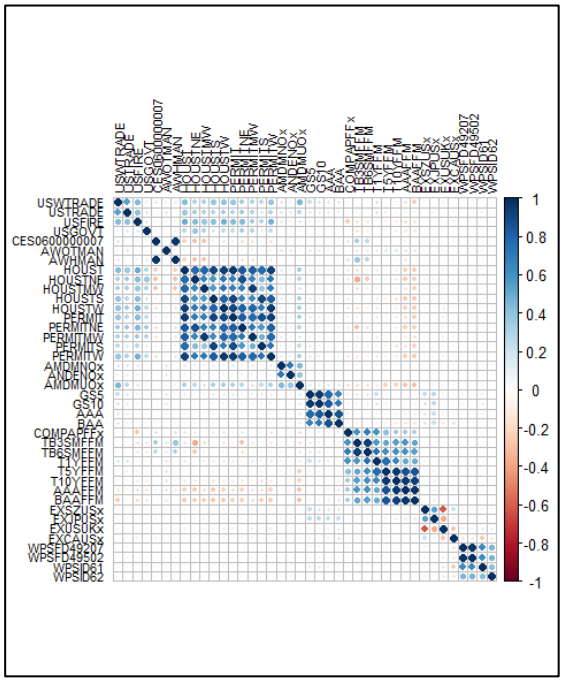


Figure 25 : Cor_matrice13

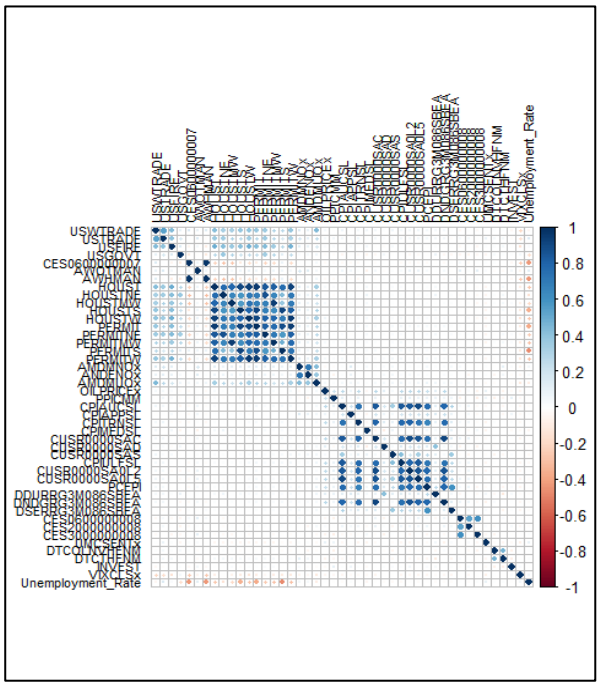


Figure 26 : Cor_matrice14

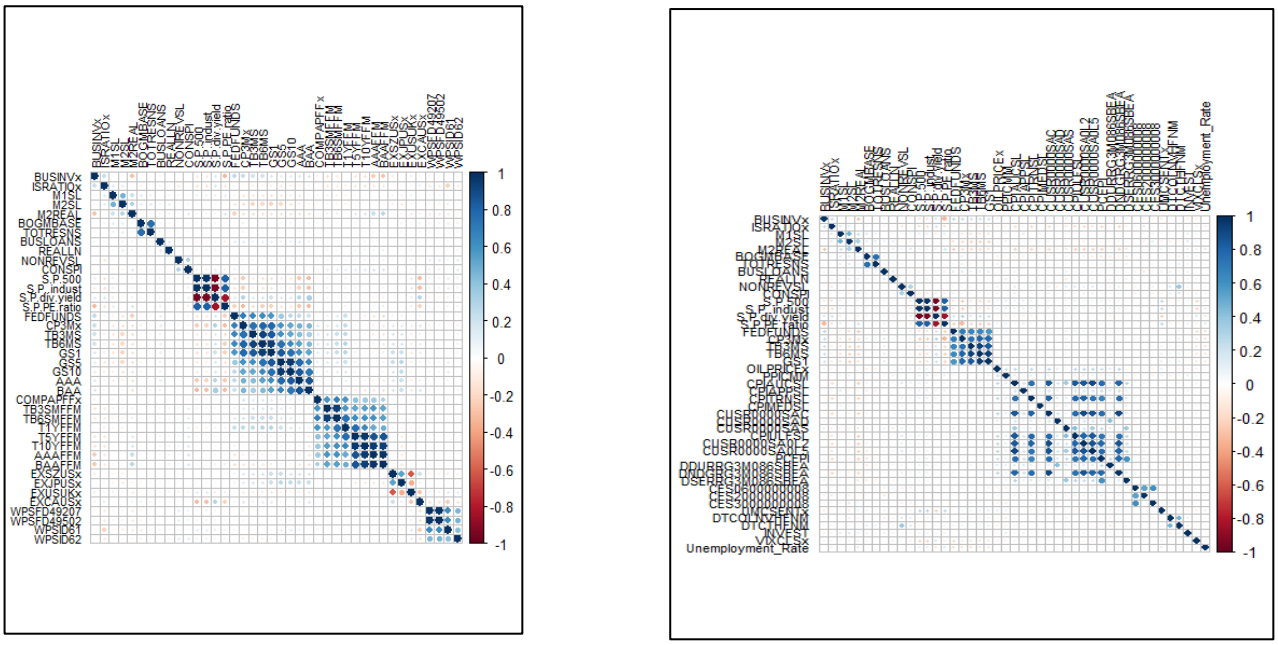


Figure 27 : Cor_matrice15

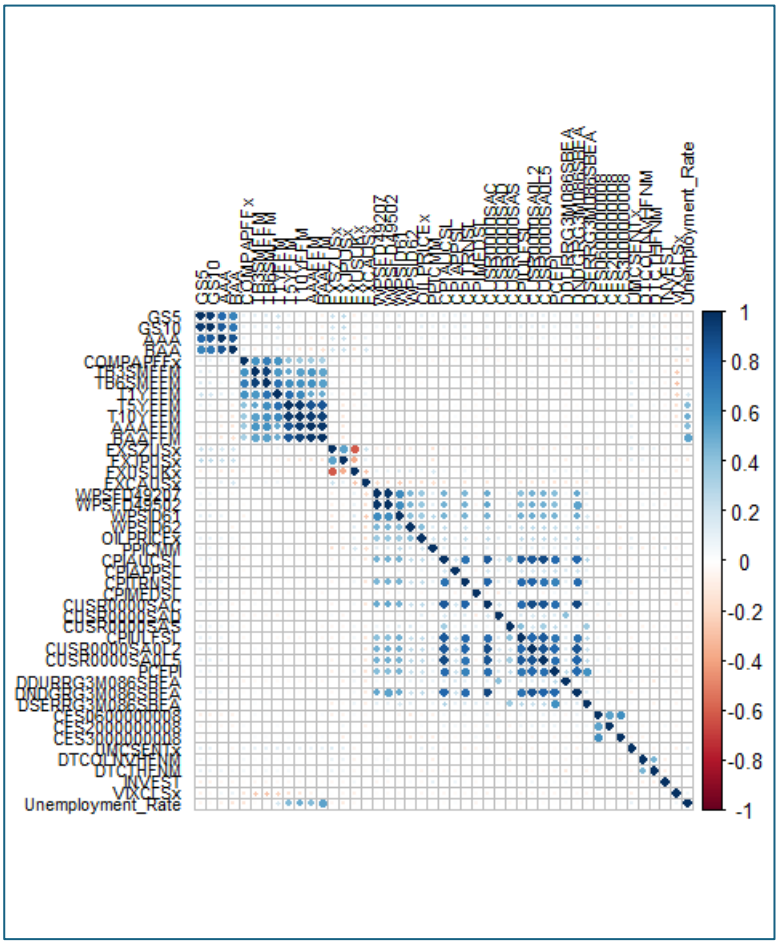


Figure 28 : Représentation du paramètre lambda avec le modèle Ridge

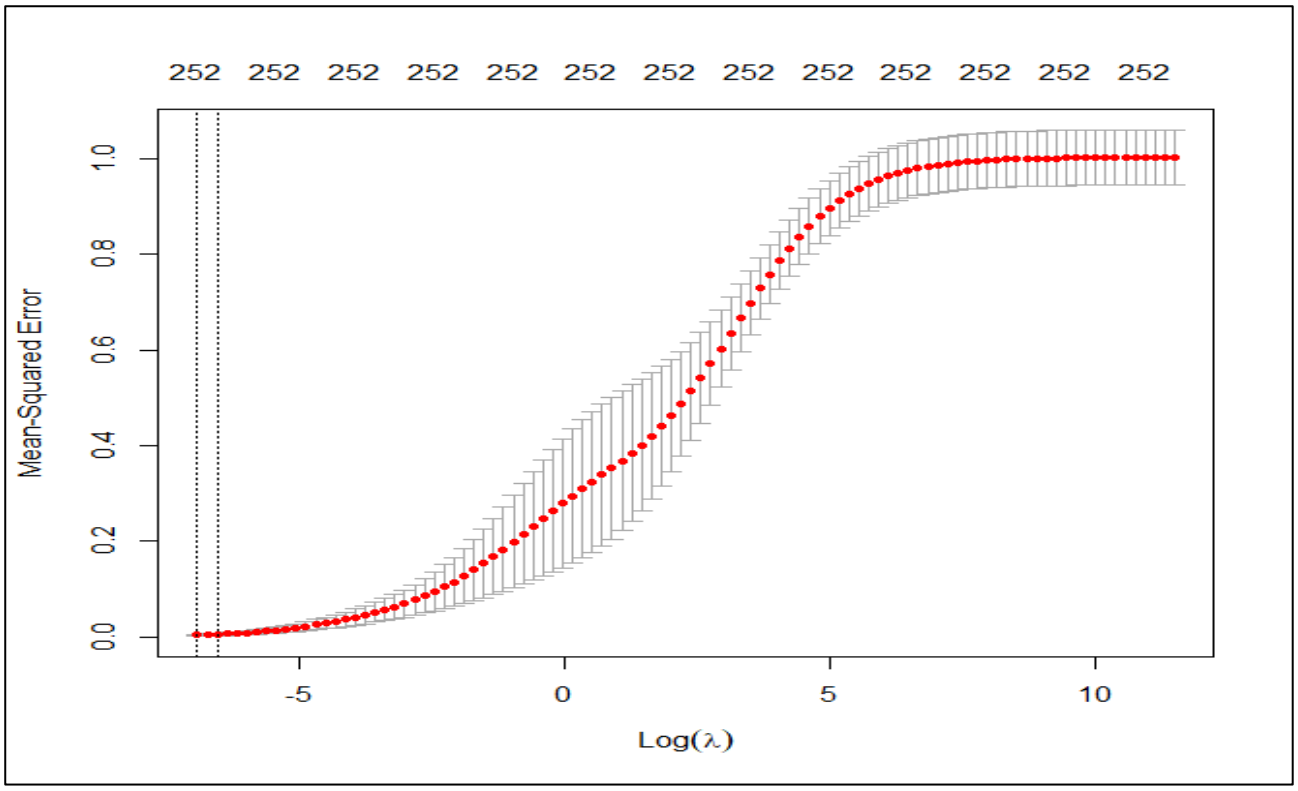


Figure 29 : Représentation du paramètre lambda avec le modèle Lasso

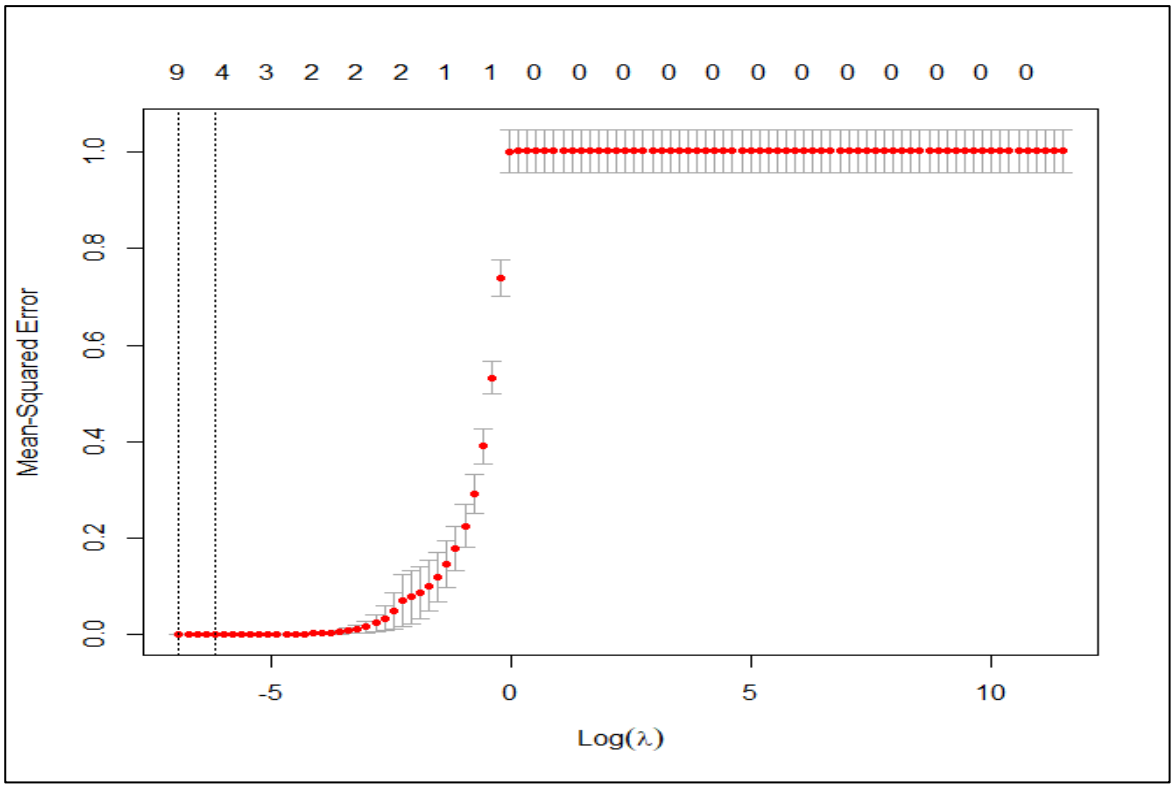


Figure 30 : Représentation du paramètre lambda avec le modèle Elastic-net

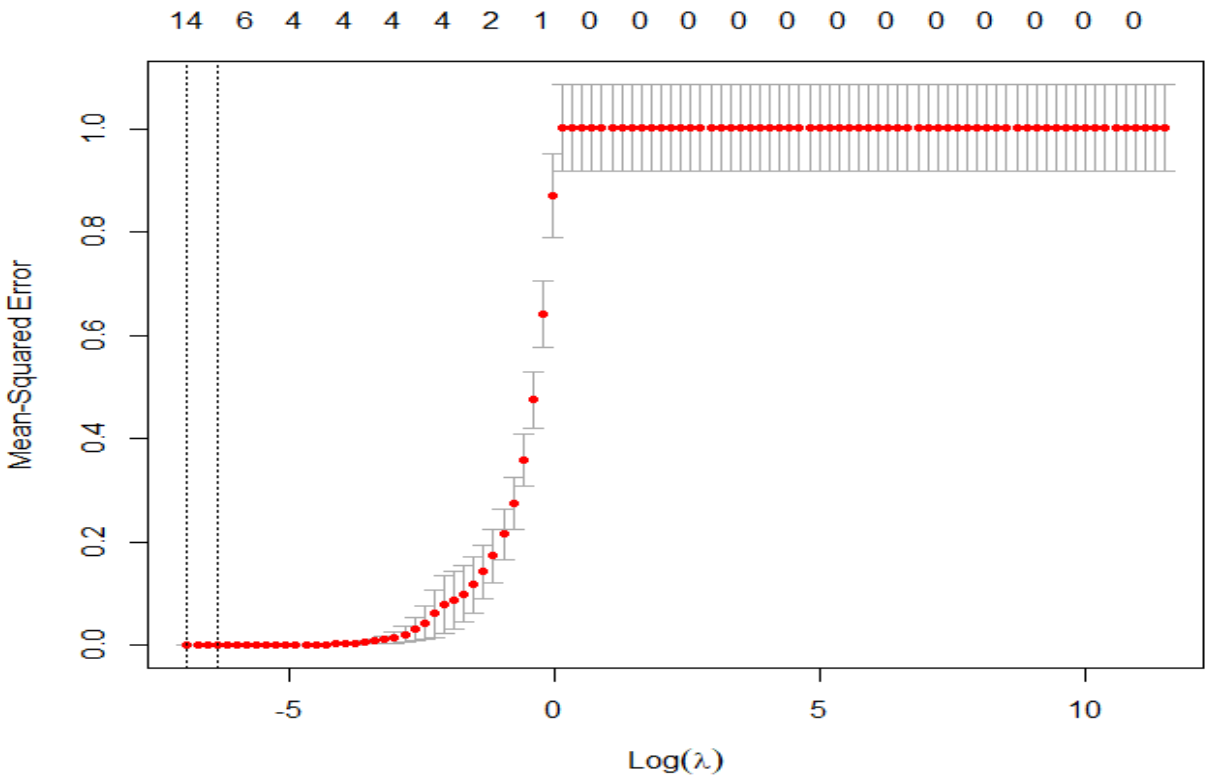


Figure 31 : Représentation du paramètre lambda avec le modèle SCAD

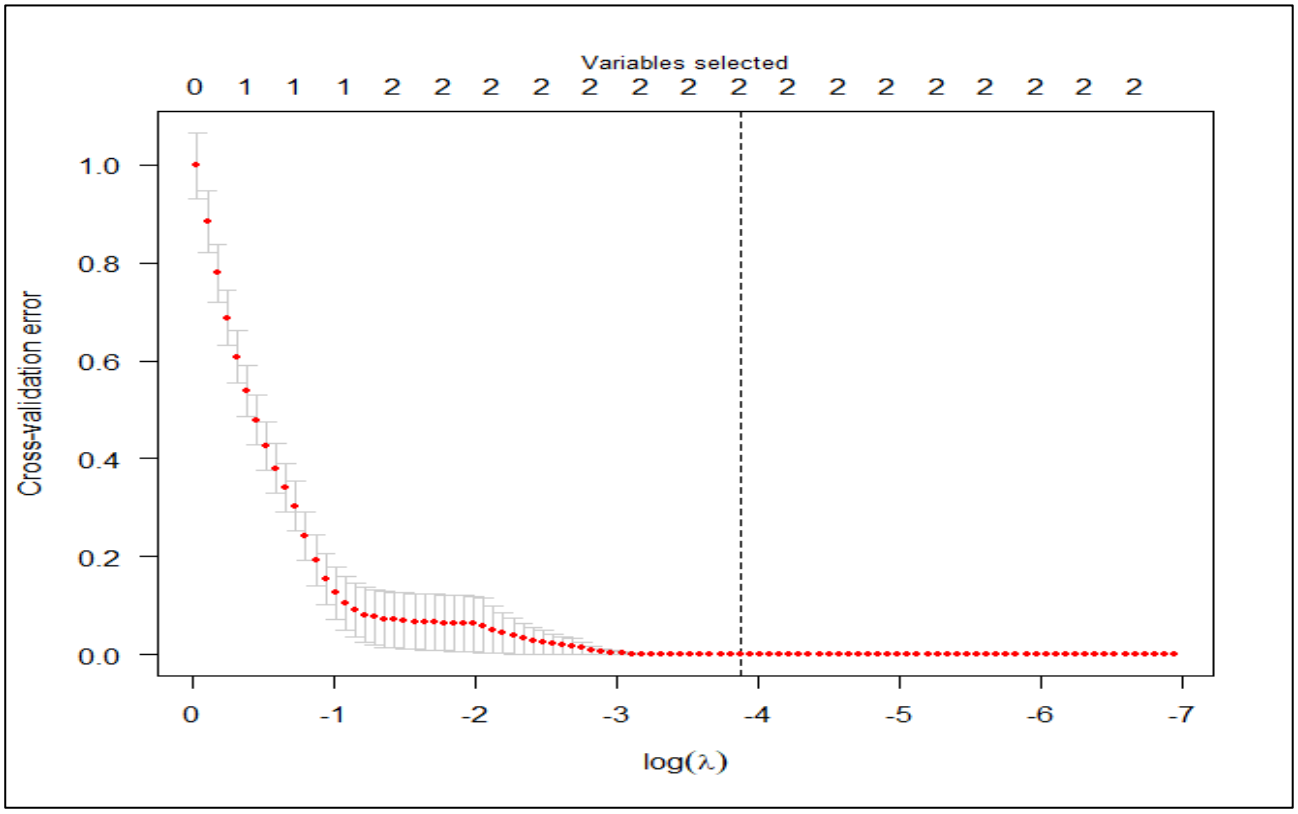


Figure 32 : Représentation du paramètre lambda avec le modèle aLASSO

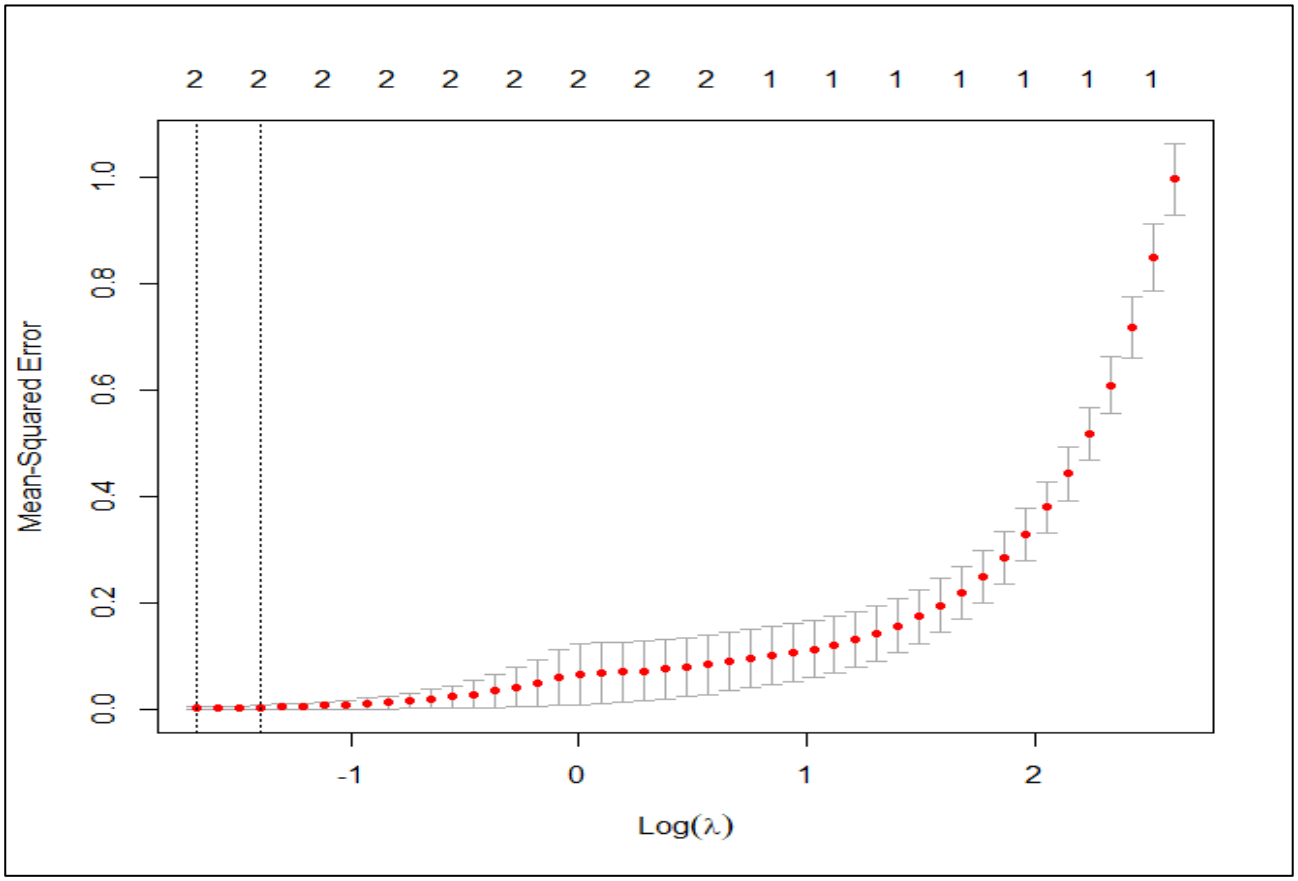


Table des Matières :