IAE Nantes
Économie & Management
Pôle Sociétés

Nantes Université

Master 2 Econométrie appliquée

Régressions pénalisées et sélection de variables

# Analyse du taux d'emploi canadien de janvier 1981 à juillet 2024

BELANE TANGA Lauriane
TRILLAUD Valorys

Année universitaire 2024-2025

# Résumé

L'objectif de ce projet est d'explorer une base de données étendue contenant diverses variables potentiellement associées au taux d'emploi au Canada entre janvier 1981 et juillet 2024. Nous avons appliqué plusieurs techniques de sélection de variables, incluant l'approche SIS (Sure Independence Screening) et un modèle Random Forest. Des décalages temporels ont également été intégrés pour mieux saisir la relation dynamique entre les variables explicatives et la variable cible au fil du temps. Les résultats montrent que des méthodes telles que GETS, Elastic Net grid search et random search permettent de sélectionner un plus grand nombre de variables, tandis que Lasso, SCAD et Elastic Net fixed favorisent une approche plus parcimonieuse. Deux variables clés, le taux de chômage et le taux d'emploi retardé de deux mois, se révèlent particulièrement déterminantes pour expliquer les variations du taux d'emploi.



**Mots clés :** taux d'emploi au Canada, techniques de sélection de variables, SIS, GETS, Elastic Net grid search , random search , Lasso, SCAD et Elastic Net fixed.



# Abstract

The objective of this project is to explore an extensive database containing various variables potentially associated with the employment rate in Canada between January 1981 and July 2024. We applied several variable selection techniques, including the SIS (Sure Independence Screening) approach and a Random Forest model. Temporal lags were also integrated to better capture the dynamic relationship between the explanatory variables and the target variable over time. The results show that methods such as GETS, Elastic Net grid search, and random search allow for the selection of a larger number of variables, while Lasso, SCAD, and Elastic Net fixed favor a more parsimonious approach. Two key variables, the unemployment rate and the two-month lagged employment rate, prove to be particularly significant in explaining variations in the employment rate.



**Keywords:** employment rate in Canada, variable selection techniques, SIS, GETS, Elastic Net grid search, random search, Lasso, SCAD, and Elastic Net fixed.

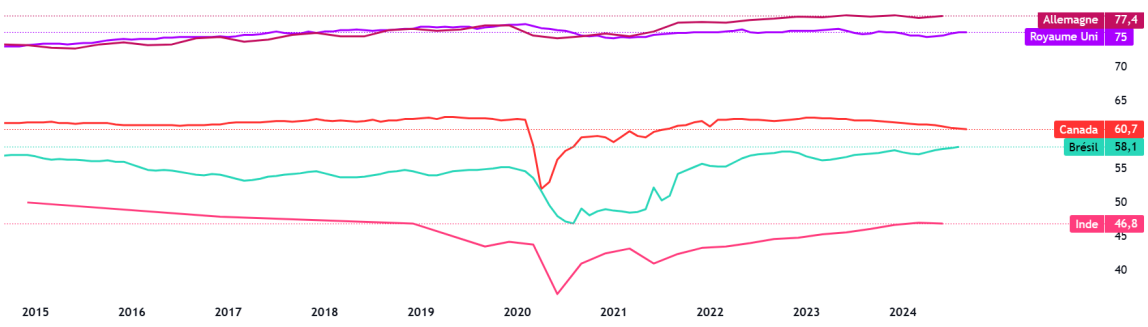
# Sommaire

# 1. Introduction

Le taux d'emploi est un indicateur clé pour évaluer la santé du marché du travail et l'activité économique d'un pays, en l'occurrence ici le Canada. Calculé en divisant le nombre de personnes actives occupées par la population en âge de travailler, généralement définie comme celle âgée de 15 à 64 ans, il constitue l'un des indicateurs économiques les plus pertinents pour apprécier la capacité d'un pays à offrir des opportunités d'emploi stables à ses citoyens. En mesurant la proportion de la population en âge de travailler qui est effectivement employée, le taux d'emploi fournit un aperçu précis du dynamisme du marché du travail et de l'inclusion économique des actifs.

Comprendre les déterminants du taux d'emploi permet de cerner les facteurs ayant un impact direct sur l'emploi et de guider des politiques économiques efficaces tout en formulant des prévisions plus précises. Le Canada, comme la plupart des pays du G20, a connu une évolution du taux d'emploi marquée par des fluctuations significatives dues à divers facteurs économiques, sociaux et politiques. Pour mieux situer le Canada dans ce contexte mondial, il est utile de comparer son taux d'emploi à celui de plusieurs pays du G20.

Figure 1- Evolution du taux d'emploi de cinq pays




source: Comparer les pays par Taux d'emploi: Pick & Chart — TradingView

À l'été 2024, le taux d'emploi au Canada est estimé à environ 60,7%, reflétant une reprise progressive après les perturbations causées par la pandémie de COVID-19 et les défis économiques récents. En comparaison, l'Allemagne et le Royaume-Uni affichent des taux d'emploi plus élevés, respectivement autour de 77,4 % et 75%. Ces chiffres

indiquent non seulement la solidité des marchés du travail dans ces pays, mais aussi leur capacité à maintenir des niveaux d'emploi stables même en période de crise. D'autre part, des économies émergentes comme le Brésil et l'Inde montrent des taux d'emploi plus variables, se chiffrant autour de 58 % et 46,8 %, respectivement. Ces statistiques mettent en lumière les défis structurels auxquels ces pays sont confrontés, notamment en matière de transition vers une économie plus formelle et de création d'emplois durables.

Ainsi, ces comparaisons soulèvent une question cruciale : quels sont les facteurs qui peuvent influencer le taux d'emploi au Canada ? Dans cette étude, nous cherchons à identifier le modèle le plus performant pour expliquer le taux d'emploi canadien en procédant à des sélections de variables. Il est important de noter que le taux d'emploi que nous avons utilisé dans cette analyse a déjà été stationnarisé avec une différence première en logarithme, ce qui nous permet de nous focaliser sur l'évolution du taux d'emploi entre 1981 et 2024.

Nous débuterons par une analyse exploratoire et descriptive des données, incluant le traitement des valeurs manquantes et des valeurs aberrantes. Par la suite nous procéderons à la sélection de variables en utilisant une méthode économétrique et des régressions pénalisées. Ces techniques sont appliquées sur la base de données complète et sur une version réduite grâce à la méthode SIS. Enfin, nous utilisons une troisième approche, la Random Forest, pour compléter notre analyse et comparer les résultats obtenus afin d'identifier les variables les plus influentes.

# 2. Analyses exploratoire et descriptive

## 2.1. Analyse des valeurs manquantes

Pour analyser le taux d'emploi canadien entre 1981 et 2024, nous avons utilisé une base de données préexistante, la "Large Canadian Database for Macroeconomic Analysis"[1]. Cette base regroupe de nombreux indicateurs économiques, avec un total de 411 variables couvrant divers aspects de l'économie. Elle inclut des indicateurs tels que

[1] Stevanovic, Dalibor; Surprenant, Stéphane; Leroux, Maxime; Fortin-Gagnon, Olivier, 2021, « Large Canadian Database for Macroeconomic Analysis (LCDMA) – Vintages », https://doi.org/10.5683/SP3/59JYPU, Scholars Portal Dataverse, V1, UNF:6:ncxLZ5kO683egg7+EUOcSw== [fileUNF]

le PIB total, les importations et exportations totales, ainsi que l'indice des prix à la consommation. Bien que la base fournisse les données pour l'ensemble du Canada et pour chacune des provinces du Canada, notre étant centrée sur le pays dans son ensemble, nous avons choisi d'exclure les données concernant les provinces, réduisant ainsi le nombre de variables à 120 variables.

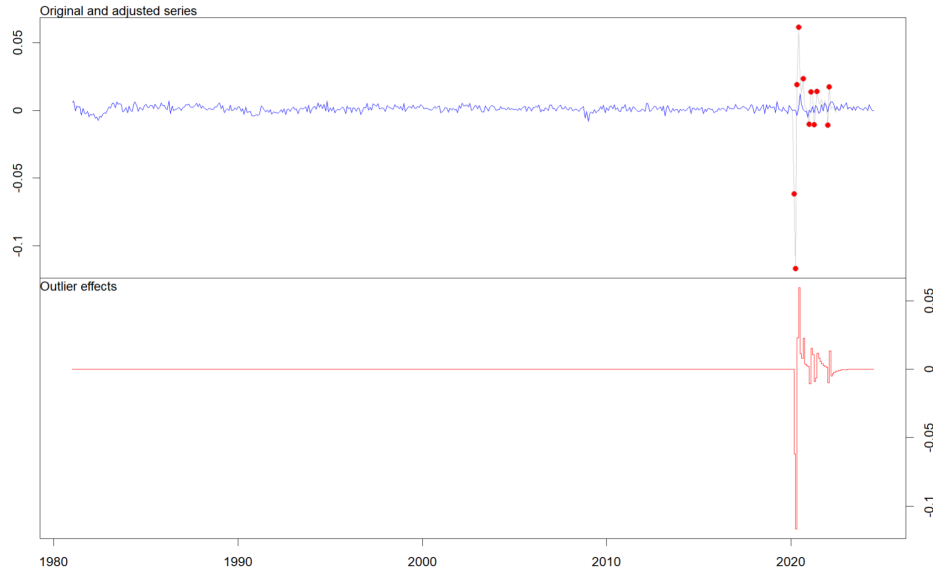
Avec ce sous-ensemble de variables, nous vérifions ensuite la présence de valeurs manquantes. En effet, les valeurs manquantes peuvent fausser les résultats en introduisant une perte d'information ou des biais. Cette vérification a permis de révéler cinq variables présentant chacune 46 valeurs manquantes (Annexe n°1) : les crédits totaux (CRED_T_discontinued), les crédits aux ménages (CRED_HOUS_discontinued), les crédits hypothécaire (CRED_MORT_ discontinued), les crédits à la consommation (CRED_CONS_discontinued) et les crédits aux entreprises (CRED_BUS_discontinued). Une analyse approfondie a montré que les valeurs manquantes pour chacune de ces variables correspondent aux mois d'octobre 2020 à juillet 2024.

Remplacer les valeurs manquantes sur une période de près de quatre années consécutives pourrait rendre difficile l'imputation de données prenant en compte des changements structurels, remplacer les données alors pourrait introduire des biais. Par conséquent, nous avons décidé de retirer ces cinq variables pour éviter les approximations biaisées et préserver la qualité de l'analyse. Finalement, notre base de données contient 115 variables à laquelle on supprime la date soit 114 variables et 523 observations.

## 2.2. Analyse des outliers : détection et correction

Nous nous concentrons maintenant sur la détection des valeurs aberrantes dans la série temporelle. Pour identifier d'éventuelles valeurs aberrantes et les corriger si nécessaire, nous avons utilisé la bibliothèque *outliers* et la fonction *ts*. Cette méthode nous permet d'identifier les valeurs atypiques présentes dans nos données.

Figure 2- Points atypiques de la variable dépendante




Le graphique ci-dessus met en évidence les valeurs atypiques. Un aperçu des valeurs et des individu est présenté dans le tableau suivant, et les autres valeurs peuvent être consultées dans l'annexe n°2 :

Table 1- les six valeurs aberrantes les plus importantes

| | Type | Individu | Time |
|---|---|---|---|
| 1 | TC | 471 | 2020:03 |
| 2 | AO | 472 | 2020:04 |
| 3 | TC | 473 | 2020:05 |
| 4 | AO | 474 | 2020:06 |
| 5 | AO | 477 | 2020:09 |
| 6 | AO | 481 | 2021:01 |

Les valeurs aberrantes détectées dans l'analyse de la variable dépendante taux d'emploi canadien, sont fortement concentrées en 2020, année marquée par la crise de la COVID-19. Cette crise a entraîné des perturbations soudaines dans l'emploi, avec des changements temporaires (TC) reflétant des fluctuations transitoires du taux d'emploi, probablement causées par les périodes de confinement et les réouvertures partielles. Ces variations traduisent une chute brutale suivie d'une reprise temporaire, indiquant

l'impact temporaire des mesures sanitaires. En parallèle, les outliers additifs (AO) capturent des anomalies ponctuelles correspondant à des chocs uniques dans le taux d'emploi, comme les mois de confinement strict, qui ont entraîné des baisses abruptes sans effets durables. Ces valeurs aberrantes soulignent ainsi l'impact significatif et à la fois temporaire et ponctuel de la pandémie sur le taux d'emploi canadien. Les valeurs aberrantes relevées jusqu'en février 2022 coïncident avec la levée progressive des restrictions liées au Covid-19, marquée par le dernier confinement en janvier au Nouveau-Brunswick, suivi de l'abandon des passeports vaccinaux fin février et début mars.

Nous corrigeons ainsi ces valeurs atypiques pour la suite de notre étude afin d'assurer une analyse plus précise de l'évolution à long terme.

## 2.3. Analyse de la stationnarité du taux emploi

Après avoir corrigé les valeurs atypiques dans notre analyse, nous avons vérifié la stationnarité de la série ajustée. Il est à noter que toutes les variables de notre base de données ont déjà été soumises à un processus de différenciation afin de les rendre stationnaires. Cependant, nous allons tout de même effectuer le test de stationnarité pour confirmer cette hypothèse.

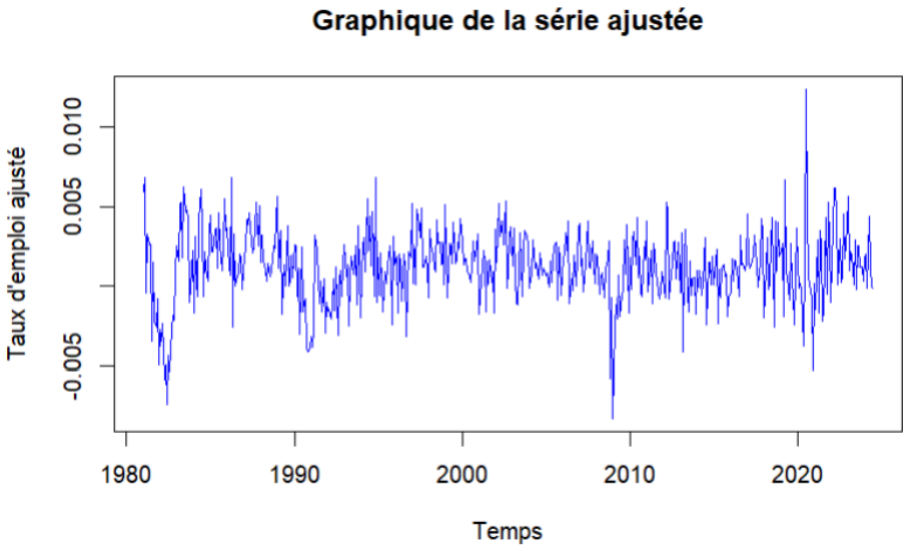
Nous utilisons le test de Dickey-Fuller augmenté (ADF) pour évaluer la stationnarité de notre série. Ce test est appliqué à la fois aux résidus du modèle AR(1) et à la série corrigée. Les résultats montrent que les résidus du modèle AR(1) présentent une statistique Dickey-Fuller de -8,33 avec un p-value inférieur à 0,05, permettant ainsi de rejeter l'hypothèse nulle de non-stationnarité au seuil de 5%. Cela indique que les résidus sont stationnaires et ne présentent pas de tendance significative. Par ailleurs, le coefficient AR(1) est de 0,29, ce qui traduit une composante autorégressive modeste dans la série, indiquant une dépendance limitée à ses valeurs passées immédiates. De même, le test ADF sur la série corrigée (adj) révèle une statistique de -5,76 avec un p-value également inférieur à 0,05, confirmant que la série corrigée est stationnaire.

Table 2- Test de stationnarité ADF

| Test | Hypothèse nulle | p.value | Statistique Dickey-Fuller |
|---|---|---|---|
| Résidu du modèle AR(1) | Non-stationnaire | 0.01 | -8.3327 |
| Série corrigée(adj) | Non-stationnaire | 0.01 | -5.7581 |

La graphique ci-dessous offre une vue d'ensemble de l'évolution temporelle de la variable dépendante, confirmée comme stationnaire. Cette représentation permet d'observer clairement les fluctuations autour de la moyenne et de valider les résultats des tests de stationnarité. Elle illustre les variations du taux d'emploi ajusté au Canada de 1980 à 2024, mettant en évidence une série de cycles économiques, avec des périodes de stabilité entrecoupées de crises économiques majeures. Les pics les plus marqués correspondent à des événements économiques mondiaux, tels que la crise de 2008 et la pandémie de COVID-19, qui ont entraîné des fluctuations soudaines du taux d'emploi au Canada.

Figure 3- Evolution de la série Taux d'emploi ajusté




## 2.4. Analyse descriptive

Pour analyser le taux d'emploi, nous avons examiné ses principales statistiques

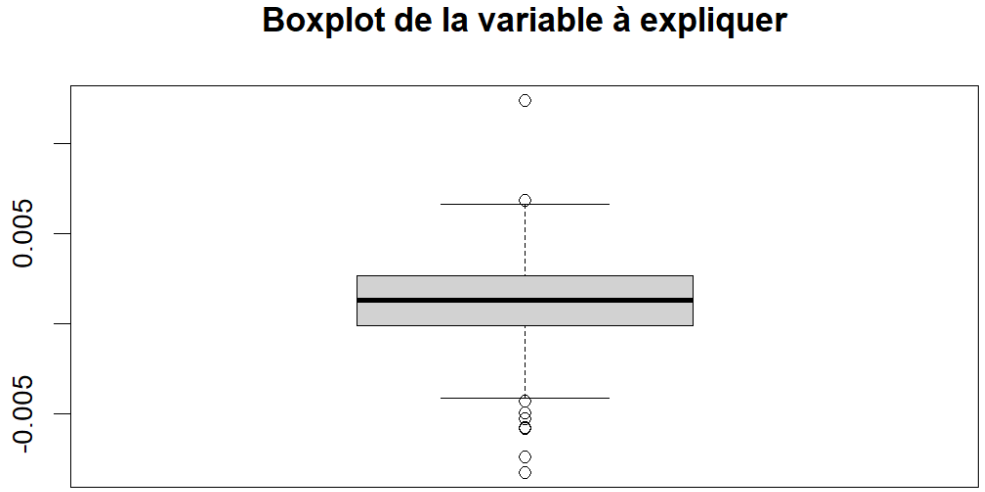
descriptives, afin de mieux comprendre la structure des données.

Tableau 3- Statistiques Descriptives

| Moyenne | Médiane | Variance | Ecart-type | skewness | Kurtosis |
|---|---|---|---|---|---|
| 0.0012 | 0.0013 | 5.3919e-06 | 0.0023 | -0.1897 | 4.5977 |

La moyenne de 0.0012 indique une légère tendance à la hausse, avec une concentration générale des valeurs autour de cette moyenne. Les valeurs extrêmes, allant de -0.0083 à 0.0124 (voir annexe n°3), révèlent une certaine amplitude de variation, mais la faible variance (5.39e-06) et l'écart-type (0.0023) indiquent une faible dispersion. La distribution présente une légère asymétrie vers la gauche identifié par la valeur de la skewness (-0.1896), ce qui est confirmé par le boxplot (figure 4) montrant une légère asymétrie et quelques valeurs atypiques situées au-dessus de la boîte principale. Ces valeurs extrêmes suggèrent des fluctuations ponctuelles, probablement dues à des événements spécifiques ayant influencé le taux d'emploi. Par ailleurs, la kurtosis élevée (4.6) indique des queues épaisses, reflétant une fréquence plus élevée de valeurs extrêmes par rapport à une distribution normale. En somme, le taux d'emploi présente en moyenne une légère croissance, une répartition relativement homogène autour de la moyenne, mais ponctuée de valeurs extrêmes, potentiellement associées à des fluctuations économiques importantes.

Figure 4- Boxplot de la variable à expliquer



## 2.5. Classification

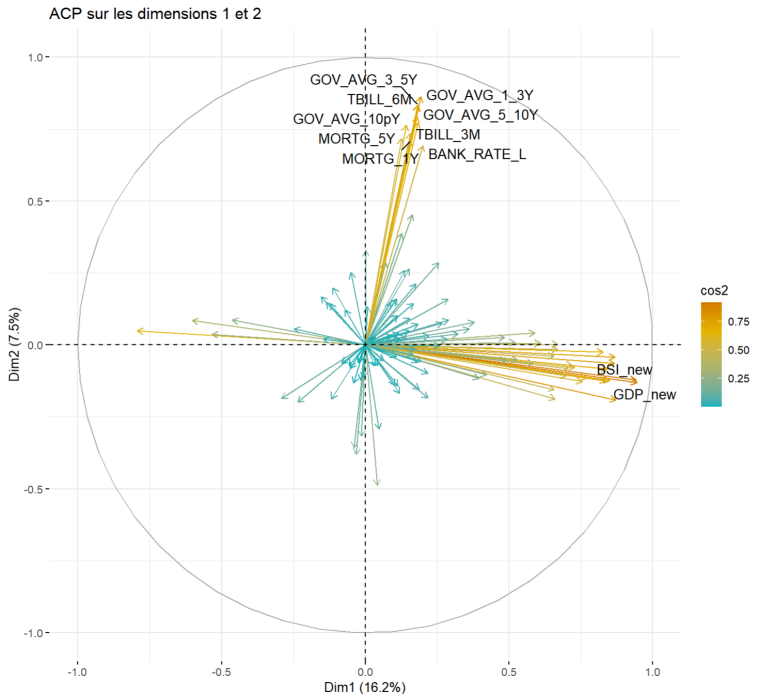
### 2.5.1. Analyse en composantes principales

Étant donné le grand nombre de variables présentes dans notre base de données, il est pertinent d'effectuer une Analyse en Composante Principale (ACP). L'ACP permet d'identifier les impacts que certaines groupes de variables peuvent avoir sur le taux d'emploi. En effet, cette méthode permet de regrouper les variables, facilitant ainsi leur interprétation.

L'analyse des résultats de l'ACP dans le cadre de l'étude du taux d'emploi canadien, de janvier 1981 à juillet 2024, met en lumière des relations entre les indicateurs économiques et sociaux au Canada. La première dimension, qui représente une inertie de 16,2%, indique que 16,2% des informations contenues dans les données sont expliquées par les variables de la dimension 1. Cette dimension (figure n°5) est fortement influencée par des variables économiques telles que le PIB total (GDP_new) et le PIB des entreprises (BSI_new). Cela suggère que les fluctuations économiques générales, comme la croissance du PIB, sont étroitement liées aux tendances d'emploi, indiquant que des périodes de croissance économique pourraient être associées à une augmentation des emplois.

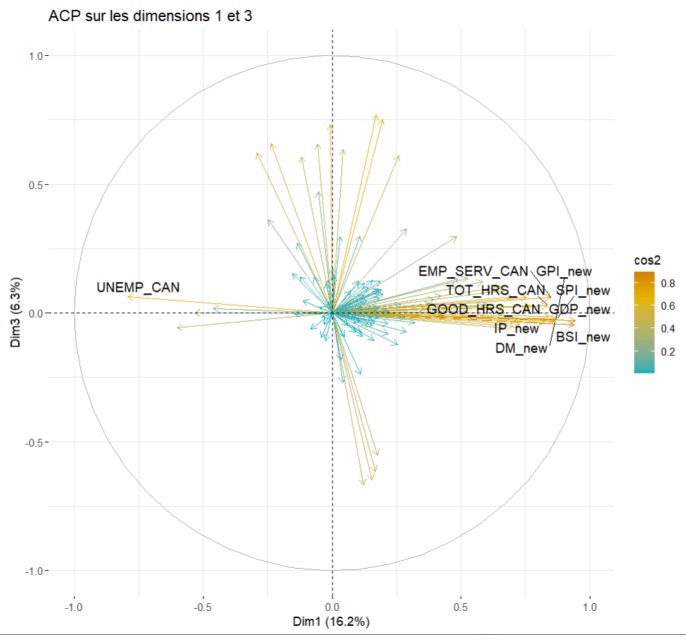
D'autre part, les dimensions 2 et 3 offrent des perspectives complémentaires sur des aspects plus spécifiques, tels que la politique gouvernementale et le marché du travail. Cependant, elles ne présentent respectivement que 7,5% et 6,3% de l'inertie totale. Dans la dimension 2 (figure n°5), la contribution de variables comme les taux moyens des obligations gouvernementales que ce soit de court terme de 1 à 3 ans (GOV_AVG_1_3Y) et à long terme de plus de 10 ans (GOV_AVG_10pY), ainsi que des bons du trésors (TBILL), indique que des décisions politiques, telles que les taux d'intérêt, influencent le taux d'emploi.

Figure 5- Cercle de corrélation de l'ACP sur les dimensions 1 et 2




De plus, la dimension 3 (figure n°6), qui inclut le taux de chômage (UNEMP_CAN), souligne l'importance de ce dernier dans l'analyse des tendances d'emploi. En intégrant ces dimensions dans l'étude du taux d'emploi canadien, nous obtenons une compréhension plus globales des facteurs économiques et politiques qui façonnent le marché du travail, permettant ainsi d'identifier des leviers potentiels pour améliorer l'emploi dans le pays.

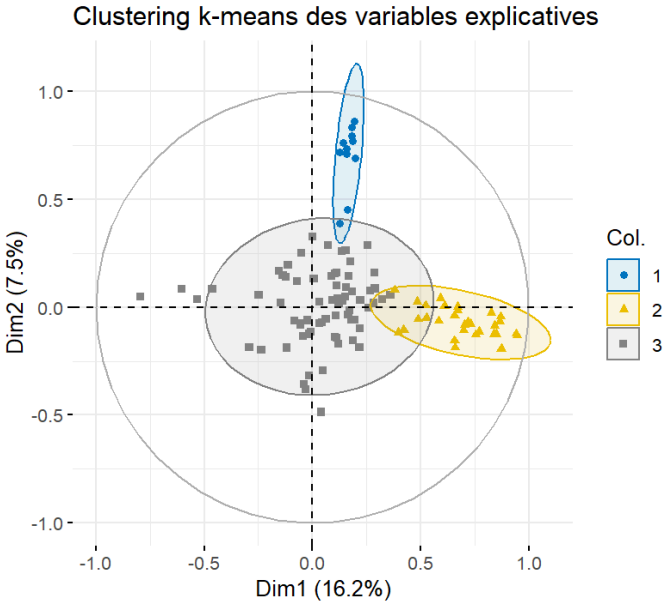
Figure 6- Cercle de corrélation de l'ACP sur les dimensions 2 et 3



### 2.5.2. Clustering

A la suite de l'ACP, nous avons effectué une analyse de clustering pour mieux comprendre les relations entre les variables influençant le taux d'emploi canadien. Cette analyse a permis de regrouper les variables en trois clusters distincts, chacun représentant des dimensions spécifiques des dynamiques économiques et sociales. Ces clusters sont illustrés dans le graphique suivant :

Figure 7- Représentation des clusters




Le cluster 1 qui comprend 12 variables se concentre sur des mesures financières, telles que le taux d'intérêt de la banque (BANK_RATE_L) et le taux moyenne des obligations gouvernementales de 1 à 3 ans (GOV_AVG_1_3Y), soulignant l'importance des politiques monétaires dans l'impact indirect sur le marché de l'emploi. Les variations des taux d'intérêt peuvent influencer les décisions d'investissement des entreprises, affectant ainsi l'emploi.

Le cluster 2 comprend lui 30 variables, il regroupe des indicateurs économiques majeurs tels que le PIB total (GDP_new), le PIB des entreprises (BSI_new), et les biens du PIB (GPI_new), ainsi que des variables liées à l'emploi, comme le taux d'emploi dans les services (EMP_SERV_CAN) et le taux d'emploi dans la construction (EMP_CONS_CAN). Cela suggère que ces facteurs partagent une relation étroite,

indiquant l'amélioration de ces indicateurs économiques pourrait être corrélée à une augmentation des niveaux d'emploi.

Enfin, le cluster 3 avec 71 variables, englobe un large éventail d'indicateurs, notamment des mesures de chômage comme le taux de chômage (UNEMP_CAN) et divers indicateurs économiques et sociaux comprenant l'indice des prix à la consommation (CPI_ALL_CAN). Ce cluster met en évidence les interactions complexes entre l'emploi et des facteurs socio-économiques variés, suggérant que des défis structurels peuvent influencer le marché du travail.

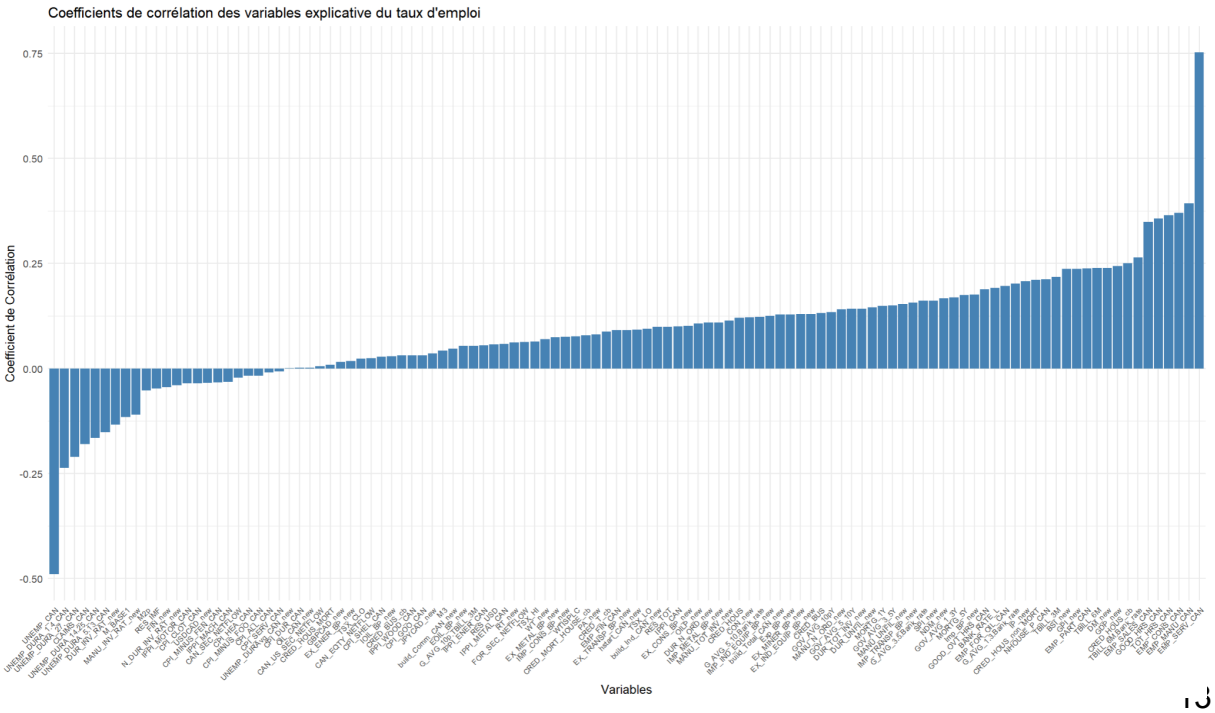
Cette segmentation par clusters, réalisée après l'ACP, nous permet d'obtenir une compréhension plus approfondie des différents facteurs qui interagissent pour influencer le taux d'emploi au Canada, offrant ainsi des pistes pour l'élaboration de politiques publiques et de stratégies économiques adaptées.

## 2.6. Corrélation

### 2.6.1. Entre la variable à expliquer et les variables explicatives

Pour compléter notre analyse des données, nous allons examiner le taux de corrélation entre nos variables explicatives et le taux d'emploi. Cela nous permettra de mieux comprendre les relations qui existent ces différentes variables.

Figure 8- Corrélations entre le taux d'emploi et les variables explicatives



La figure n°8 illustre le coefficient de corrélation de spearman. Nous observons que deux variables présentent une forte corrélation avec le taux d'emploi : le taux de chômage (UNEMP_CAN), avec un coefficient de presque -0,5 et le taux d'emploi dans le secteur des services (EMP_SERV_CAN), avec un coefficient de 0,75. Cette dernière valeur indique une relation positive significative, suggérant que lorsque le taux d'emploi dans les services augmente, le taux d'emploi global tend également à augmenter. Alors que lorsque le taux de chômage augmente, le taux d'emploi diminue.

De plus, certaines variables présentent des coefficients de corrélation intermédiaires, situés entre 0,25 et 0,5. C'est le cas du temps total travaillé (TOT_HRS_CAN) et du taux d'emploi dans la construction (EMP_CONS_CAN). Ces résultats suggèrent qu'il existe des liens modérés entre ces variables et le taux d'emploi, ce qui mérite une attention particulière lors d'analyses plus approfondies.

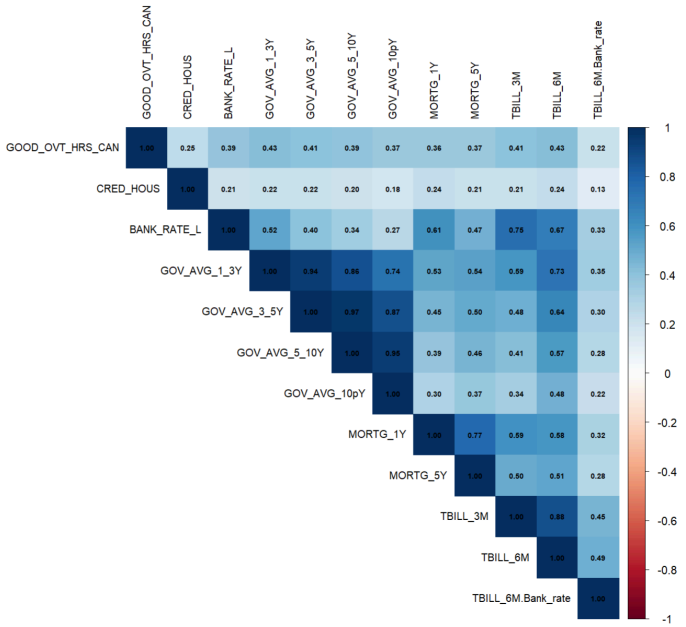
En revanche, les autres variables que nous avons examinées montrent une corrélation faible, avec des coefficients variant entre -0,25 et 0,25. Cela indique qu'elles n'ont qu'un impact marginal sur le taux d'emploi. Dans l'ensemble, cette première analyse des corrélations nous donne des indications précieuses sur les facteurs influençant le taux d'emploi. Cependant, il serait intéressant d'approfondir ces relations pour mieux cerner les effets directs et indirects de chaque variable.

### 2.6.2. Entre les variables explicatives

Nous terminons notre analyse descriptive par l'analyse de la corrélation entre les variables explicatives. En raison du grand nombre de variables, nous avons choisi de les analyser en les regroupant selon les clusters identifiés dans la section 2.5.2.

La figure 9 met en évidence le lien positif entre les variables du cluster 1, qui présentent des coefficients relativement élevés, supérieur à 0,5. Ce résultat était anticipé, car ce cluster regroupe des variables qui exercent un impact similaire sur le taux d'emploi. Ces variables, étant intrinsèquement liées par leur nature ou leur contexte, renforcent ainsi les conclusions que nous pouvons tirer de leur analyse conjointe.

Figure 9- Corrélations entre les variables explicatives du cluster 1




En revanche, les clusters 2 (annexe n°4) et 3 (annexe n°5) comprenant plus de variables se caractérisent par une plus grande hétérogénéité des coefficients de corrélation. En effet, dans le cluster 3, une proportion plus importante de coefficients de corrélations sont négatifs.

# 3. Sélection des variables

Dans cette section, nous appliquons plusieurs méthodes de sélection de variables pour optimiser notre modèle et gérer la haute dimensionnalité de notre base de données. La sélection de variables est cruciale pour améliorer la précision du modèle et limiter les effets de surajustement.

Pour cela, nous explorons diverses approches de sélection, en partant de la méthode économétrique GETS, puis en passant par les régularisations pénalisées comme Lasso et Elastic Net, pour finir par la réduction de dimension avec la méthode SIS à laquelle on appliquera les méthodes précédentes.

Afin d'enrichir notre base de données et de prendre en compte les effets différés, nous ajoutons des variables retardées. Chaque variable explicative est ajoutée avec un retard de 1 période. De plus, nous ajoutons la variable à expliquer avec 1, 2, 3 et 4

retards à notre base de données. Après l'ajout de ces retards, nous avons supprimé les quatre premières lignes contenant des valeurs manquantes résultant des retards. La base finale utilisée pour cette analyse comprend donc 231 variables et 519 observations.

Nous résumerons les résultats des différentes méthodes dans la quatrième partie, en présentant les valeurs d'alpha, de lambda et des coefficients bêta associés.

## 3.1. Approche économétrique : GETS

Nous démarrons la sélection de variables avec une méthode économétrique, GETS. Cette approche repose sur une méthode de sélection descendante, elle permet à chaque itération de retirer du modèle les variables qui ne contribuent pas de manière significative. Bien que notre base comporte un nombre important de variables nous avons pu réaliser le GETS sur notre base de données. Cette méthode nous permet de sélectionner 100 variables sur les 230 variables explicatives.

## 3.2. Régressions pénalisées

### 3.2.1. Ridge

Nous commençons l'analyse par régression pénalisée avec la méthode ridge. Contrairement aux autres méthodes de régressions pénalisées, cette méthode n'est pas une méthode de sélection de variable. En effet, elle réduit les coefficients des variables les moins pertinentes sans les supprimer du modèle.

Tableau 4- Résultats de la méthode Ridge

| Lambda | Nombre de variables sélectionnées |
|---|---|
| 4,328 | 230 |

Nous obtenons donc 230 variables explicatives pour un lambda optimal de 4,328 (Annexe n°6) cela signifie une forte régularisation réduisant la valeur des coefficients. Ce lambda élevé peut être expliqué par les nombreuses variables.

### 3.2.2. Lasso

La méthode Lasso se distingue par sa capacité à éliminer les variables inutiles et seulement celle-ci en rendant nul les coefficients. Cette approche favorise une sélection efficace des variables, aboutissant ainsi à un modèle plus parcimonieux et interprétable.

Tableau 5- Résultats de la méthode Lasso

| Lambda | Nombre de variables sélectionnées |
|---|---|
| 0,266 | 3 |

Dans notre analyse, la valeur lambda est de 0,266 (annexe n°7), ce qui indique un niveau modéré de régularisation. Cela permet à Lasso de conserver certaines variables tout en éliminant les autres. Cependant, dans notre cas, Lasso procède à une sélection drastique en ne retenant que trois variables. Cela démontre la puissance de Lasso pour sélectionner les variables les plus pertinentes et simplifier le modèle.

### 3.2.3. Elastic-Net (fixer à priori $\alpha$ = 0.5)

Nous poursuivons notre analyse en appliquant la méthode Elastic Net, qui ajoute une pénalisation Ridge en plus de celle de Lasso. L'un des principaux avantages de cette méthode est qu'elle permet de supprimer les variables non pertinentes en réduisant leurs coefficients à zéro, tout en étant moins sévère dans la discrimination des variables corrélées entre elles, contrairement à Lasso. En conservant certaines de ces variables, Elastic Net parvient à mieux capturer les relations entre elles, tout en réduisant le risque de surajustement.

Tableau 6- Résultats de la méthode Elastic Net (fixed)

| Alpha | Lambda | Nombre de variables sélectionnées |
|---|---|---|
| 0,5 | 0,464 | 6 |

Dans ce premier modèle Elastic Net, nous fixons alpha à 0,5, ce qui indique que nous donnons un poids égal au pénalisation de Ridge et de Lasso. Avec cette configuration, nous obtenons un lambda optimal de 0,464 (annexe n°8), et le modèle

sélectionne 6 variables explicatives. Ce lambda indique un niveau modéré de régularisation.

### 3.2.4. Elastic-Net grid search

Nous procédons une nouvelle fois à la méthode Elastic-Net mais cette fois avec une recherche de paramètres par grid search. Cette approche consiste à générer un ensemble de valeurs de λ, espacées uniformément dans une plage prédéfinie, et à tester chaque combinaison via une validation croisée. Cette méthode a pour objectif d'augmenter la robustesse du modèle, ainsi que de réduire le risque de surajustement.

Tableau 7- Résultats de la méthode Elastic Nat (grid search)

| Alpha | Lambda | Nombre de variables sélectionnées |
|---|---|---|
| 0,832 | 0,320 | 3 |

A l'issue de la grid search, nous obtenons un λ optimal de 0,320 (annexe n°9). Ce nouveau λ optimal a permis de sélectionner 3 variables explicatives.

### 3.2.5. Elastic-Net random search

Nous allons cette fois estimer les paramètres du modèle Elastic Net en appliquant cette fois une recherche de paramètre par random search. Contrairement à la grid search, qui explore systématiquement chaque combinaison de valeurs de λ sur une grille définie, la random search sélectionne aléatoirement un nombre limité de combinaisons dans l'espace de recherche des hyperparamètres.

Tableau 8- Résultats de la méthode Elastic Net (carret)

| Alpha | Lambda | Nombre de variables sélectionnées |
|---|---|---|
| 0,2 | 0,328 | 42 |

À l'issue de cette random search, nous obtenons un λ optimal de 0,328 (voir annexe n°10) et un $\alpha$ optimal de 0,2. Ce nouveau modèle a sélectionné 42 variables explicatives.

### 3.2.6. SCAD

La régression SCAD (Smoothly Clipped Absolute Deviation) est une méthode de sélection de variables dans les modèles de régression, visant à répondre à certaines limites de Lasso. Contrairement à ces dernières, SCAD permet de conserver les variables pertinentes. Elle va moins pénaliser les grands coefficients que Lasso permettant de réduire le biais mais va plus pénaliser les petits coefficients augmentant la parcimonie du modèle. Cette caractéristique permet à SCAD de réaliser une sélection plus précise des variables, offrant ainsi un équilibre optimal entre la parcimonie du modèle et sa capacité d'explication. Nous obtenons alors un lambda optimal de 0,153 (annexe n°10), avec 7 variables sélectionnées.

Tableau 9- Résultats de la méthode SCAD

| Lambda | Nombre de variables sélectionnées |
|---|---|
| 0,153 | 7 |

### 3.2.7. Adaptive Lasso

Nous terminons cette première partie sur la sélection de variables à partir des données complètes en appliquant la méthode de régression Adaptive Lasso. Cette approche vise à améliorer le modèle Lasso en introduisant un paramètre de pondération qui ajuste la pénalisation appliquée aux coefficients des variables. Dans notre analyse, nous avons obtenu un lambda optimal de 0,654 (voir annexe n°11). Cependant, il est à noter que ce modèle n'a sélectionné aucune variable explicative.

Tableau 10- Résultats de la méthode aLasso

| Lambda | Nombre de variables sélectionnées |
|---|---|
| 0,654 | 0 |

## 3.3. Régressions pénalisées avec réduction de dimension

### 3.3.1. Réduction de dimension SIS

Jusqu'à présent, nous avons appliqué diverses techniques de sélection de variables directement sur nos 230 variables explicatives. Cependant, pour affiner davantage nos résultats et obtenir un modèle plus robuste et moins complexe, nous allons effectuer une réduction de dimension en utilisant la méthode SIS (Sure Independence Screening).

Cette méthode permet de réduire efficacement le nombre de variables dans notre base de données, en conservant celles qui ont la plus forte relation avec notre cible. Ce filtrage permet de diminuer le risque de sur-apprentissage en éliminant les variables peu pertinentes, tout en améliorant l'efficacité du modèle grâce à un nombre réduit de paramètres à estimer.

Dans notre cas, l'application de la méthode SIS a permis de passer de 230 à 65 variables explicatives. Nous allons maintenant réappliquer les techniques de sélection de variables sur cette nouvelle base de données réduite, optimisant ainsi le processus pour des résultats encore plus précis et pertinents.

### 3.3.2. GETS

Après avoir appliqué la méthode SIS pour réduire le nombre de variables explicatives de 230 à 65, nous utilisons la méthode GETS pour affiner davantage notre sélection. Comme indiqué plus haut, la méthode GETS est une approche itérative qui vise à simplifier le modèle en éliminant progressivement les variables les moins significatives. L'objectif est d'aboutir à un modèle parcimonieux, ne retenant que les variables ayant un impact statistiquement significatif sur la variable cible.

Dans notre cas, la méthode GETS a permis de réduire les 65 variables obtenues avec SIS à 19 variables explicatives (voir annexe n°12), créant ainsi un modèle à la fois précis et efficient. Ce processus de simplification permet de minimiser le risque de sur-ajustement, tout en maintenant un haut niveau de pertinence pour la modélisation du taux d'emploi canadien. Les variables sélectionnées par GETS sont donc celles qui

montrent la relation la plus forte et la plus directe avec notre cible, ce qui renforce la robustesse et l'interprétabilité de notre modèle.

### 3.3.3. Ridge

À l'instar de la régression effectuée sur la base de données initiale, la méthode Ridge a retenu l'ensemble des variables restantes issues de la réduction par SIS, soit 65 variables explicatives(voir annexe n°17 ). La régression Ridge est une technique de régularisation linéaire qui limite la taille des coefficients pour réduire le risque de sur-ajustement, mais elle n'effectue pas de sélection de variables au sens strict. Plutôt que de supprimer certaines variables, la méthode Ridge minimise leur impact en contraignant leurs coefficients à se rapprocher de zéro.

Tableau 11- Résultats de la méthode SIS+Ridge

| Lambda | Nombre de variables sélectionnées |
|---|---|
| 3.594 | 65 |

Dans notre analyse, un paramètre de pénalisation λ de 3,594 a été appliqué, optimisant ainsi la régularisation. Cette approche permet de conserver toutes les variables tout en réduisant la variance du modèle, garantissant une meilleure stabilité des prédictions et une performance accrue sur les nouvelles données.

### 3.3.4. Lasso

L'approche SIS combinée à la méthode Lasso, donne un paramètre de régularisation λ de 0,3120. Le λ de cette méthode est plus élevé que dans le modèle Lasso initial, cela est logique, car la présélection des variables effectuée par SIS permet de conserver uniquement les variables les plus influentes. Le paramètre λ doit donc être plus important pour exclure des variables supplémentaires. Le nombre de variables sélectionnées par ce modèle est de 2, alors que le modèle Lasso initial retenait 3 variables, assurant une sélection encore plus ciblée et optimisée.

Tableau 12- Résultats de la méthode SIS+Lasso

| Lambda | Nombre de variables sélectionnées |
|---|---|
| 0.320 | 2 |

### 3.3.5. Elastic-Net (fixer à priori $\alpha$ = 0.5)

Dans l'application de la méthode Elastic Net avec une valeur de α fixée à 0,5 et précédée de la réduction dimensionnelle par SIS, nous obtenons un paramètre de régularisation λ de 0,673. Ce modèle ne garde qu'une seule variable explicative alors que le modèle Elastic Net initial sans avait retenu 6 variables.

Cette différence s'explique par le rôle de SIS, qui réduit considérablement le nombre de variables en amont, ne laissant que les plus influentes dans la base de données pour l'application de la régularisation. Cette configuration aboutit donc à un modèle encore plus simplifié, focalisé sur une seule variable clé qui explique le mieux la cible après réduction.

Tableau 13- Résultats de la méthode SIS+Elastic-Net (fixed)

| Alpha | Lambda | Nombre de variables sélectionnées |
|---|---|---|
| 0.5 | 0.673 | 1 |

### 3.3.6. Elastic-Net grid search

Nous appliquons à nouveau la méthode Elastic Net grid search, cette fois en intégrant une validation croisée sur les valeurs de λ, espacées uniformément. Avec cette configuration, combinée à la réduction de dimension par SIS, le modèle identifie un paramètre optimal de λ à 0,385, conduisant à la sélection de 2 variables explicatives (Tableau 14).

En comparaison, lorsque la méthode Elastic Net grid search avait été appliquée

directement sur la base sans SIS (Section 3.2.4), le modèle retenait un paramètre λ de 0,320, sélectionnant alors 3 variables (Tableau 7). La légère réduction du nombre de variables, de 3 à 2, montre l'effet combiné de SIS et de l'optimisation par validation croisée, qui affine encore plus le modèle se concentrant sur les variables les plus influentes avec un λ légèrement plus élevé.

Tableau 14- Résultats de la méthode SIS+Elastic-Net (grid search)

| Alpha | Lambda | Nombre de variables sélectionnées |
|---|---|---|
| 0.838 | 0.385 | 2 |

### 3.3.7. Elastic-Net random search

Nous appliquons ensuite la méthode Elastic Net avec une recherche aléatoire (random search) pour affiner les paramètres de régularisation. En utilisant cette approche sur les variables présélectionnées par SIS, nous obtenons un α de 0,2 et un λ optimal de 0,329, avec une sélection de 22 variables explicatives (Tableau 15).

Comparativement, lorsque la recherche aléatoire est appliquée sans réduction par SIS (Section 3.2.5), la méthode Elastic Net sélectionne un paramètre identique de α=0,2 et un λ=0,328, mais retient un plus grand nombre de variables, soit 42 (Tableau 8).Cette réduction de 42 à 22 variables, rendue possible par l'application conjointe de SIS et d'Elastic Net, démontre une sélection de variables plus concentrée sur les variables les plus influentes, tout en maintenant des paramètres de régularisation similaires.

Tableau 15- Résultats de la méthode SIS+Elastic-Net (random search)

| Alpha | Lambda | Nombre de variables sélectionnées |
|---|---|---|
| 0.2 | 0.329 | 22 |

### 3.3.8. SCAD

Nous poursuivons avec la méthode SCAD pour une sélection plus fine et précise. En combinant la pré-sélection de SIS avec l'approche SCAD, nous renforçons la parcimonie du modèle en privilégiant les variables les plus significatives, réduisant ainsi davantage la complexité tout en préservant sa capacité explicative.Avec un lambda optimal de 0,217, la méthode SCAD, appliquée sur cette base déjà affinée, retient uniquement 2 variables explicatives. Cette combinaison SIS+SCAD démontre une efficacité accrue dans l'identification des attributs les plus influents, assurant un équilibre optimal entre robustesse du modèle et précision de la sélection, en comparaison avec une application de SCAD sans réduction préalable (Section 3.2.6).

Tableau 16- Résultats de la méthode SIS+SCAD

| Lambda | Nombre de variables sélectionnées |
|---|---|
| 0.217 | 2 |

### 3.3.9. Adaptive Lasso

Nous terminons notre analyse avec le filtrage SIS en appliquant la méthode Adaptive Lasso. Cette approche, qui combine la sélection de variables avec une pondération adaptative, vise à affiner davantage notre modèle en conservant uniquement les variables les plus pertinentes.Dans cette configuration, l'Adaptive Lasso produit un lambda optimal de 0,588, mais, de manière similaire à la précédente analyse aLasso, aucune variable explicative n'a été sélectionnée (Tableau 17). Cette absence de sélection souligne les défis persistants pour identifier des variables explicatives significatives, même après avoir réduit notre ensemble de données par le biais de la méthode SIS.

Tableau 17- Résultats de la méthode SIS+aLasso

| Lambda | Nombre de variables sélectionnées |
|---|---|
| 0.588 | 0 |

## 3.4. Random Forest

Pour compléter notre analyse, nous avons exploré l'utilisation d'un modèle Random Forest sur notre jeu de données afin de comparer ses performances avec celles obtenues par les méthodes de régression pénalisée. Le Random Forestest un algorithme d'apprentissage automatique qui crée un ensemble d'arbres de décision indépendants et combine leurs résultats pour obtenir des prédictions plus robustes. Appliquée ici, cette méthode nous permettra d'identifier les variables les plus influentes en lien avec notre variable cible, offrant ainsi un complément d'information pour la sélection des variables pertinentes.

### 3.4.1. Random forest sans réduction SIS

Dans un premier temps, nous avons appliqué le modèle Random Forest sans recourir à une réduction de variables via l'approche SIS. Ce modèle a été paramétré pour générer 1000 arbres, avec une choix de 3 variables aléatoires à chaque division de nœud. L'objectif ici est de capturer les variables les plus déterminantes pour le taux d'emploi, et de fournir une alternative aux méthodes de régression pénalisée en évaluant la capacité de Random Forest à détecter les facteurs explicatifs clés.

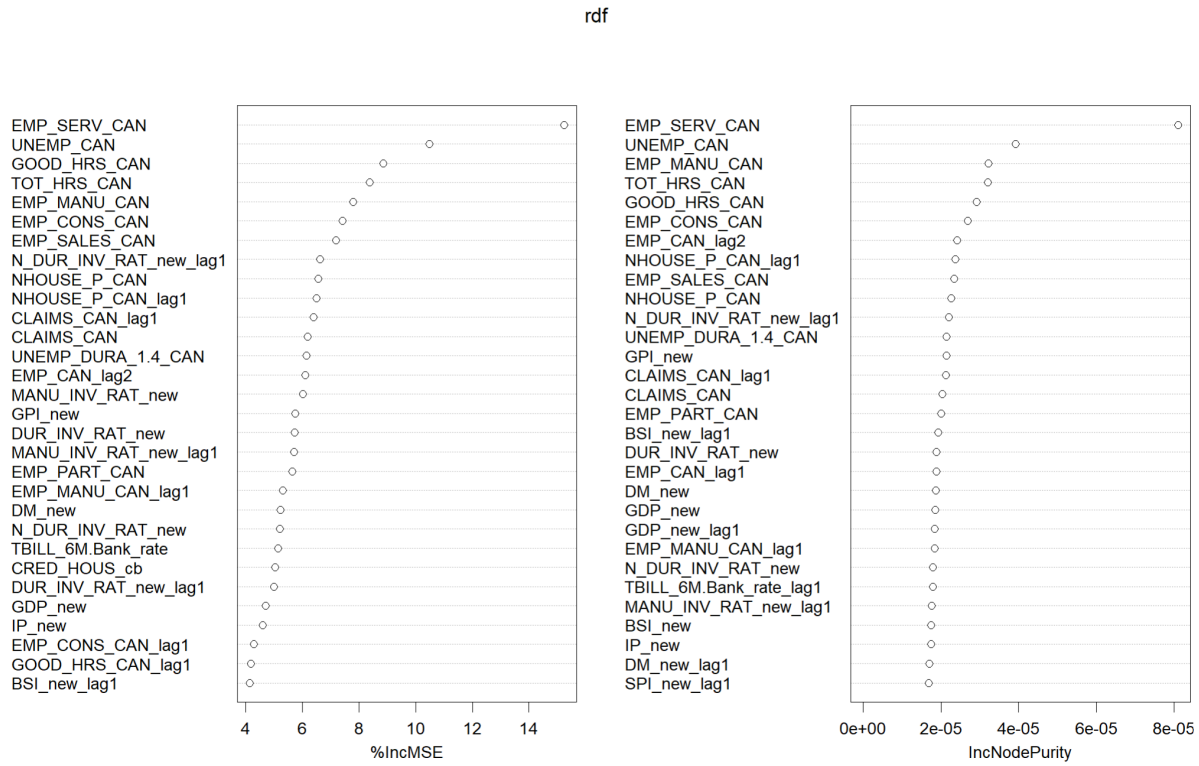
Le graphique ci-dessous présente les résultats obtenus pour les deux mesures d'importance des variables : %IncMSE et IncNodePurity. %IncMSE évalue l'impact de chaque variable sur l'erreur de prédiction, tandis que IncNodePurity indique l'impact de chaque variable sur la pureté des nœuds au sein des arbres de décision. Une valeur élevée dans l'une ou l'autre de ces mesures signale une variable ayant une influence notable sur le modèle.

L'analyse montre que les variables telles que EMP_SERV_CAN, UNEMP_CAN , et GOOD_HRS_CAN jouent un rôle prépondérant dans la prédiction du taux d'emploi canadien. Suivi par les variables TOT_HRS_CAN et EMP_MANU_CAN qui ont également une importance notable selon ces mesures. Ces résultats indiquent que les sept premières variables se distinguent comme les plus influentes, avec un léger décrochage visible après cette limite.

Ainsi, cette étape de modélisation sans réduction préalable des variables permet de mieux cerner les facteurs influents pour la prédiction du taux d'emploi, et offre un

aperçu complémentaire aux approches de régression pénalisée dans notre analyse.

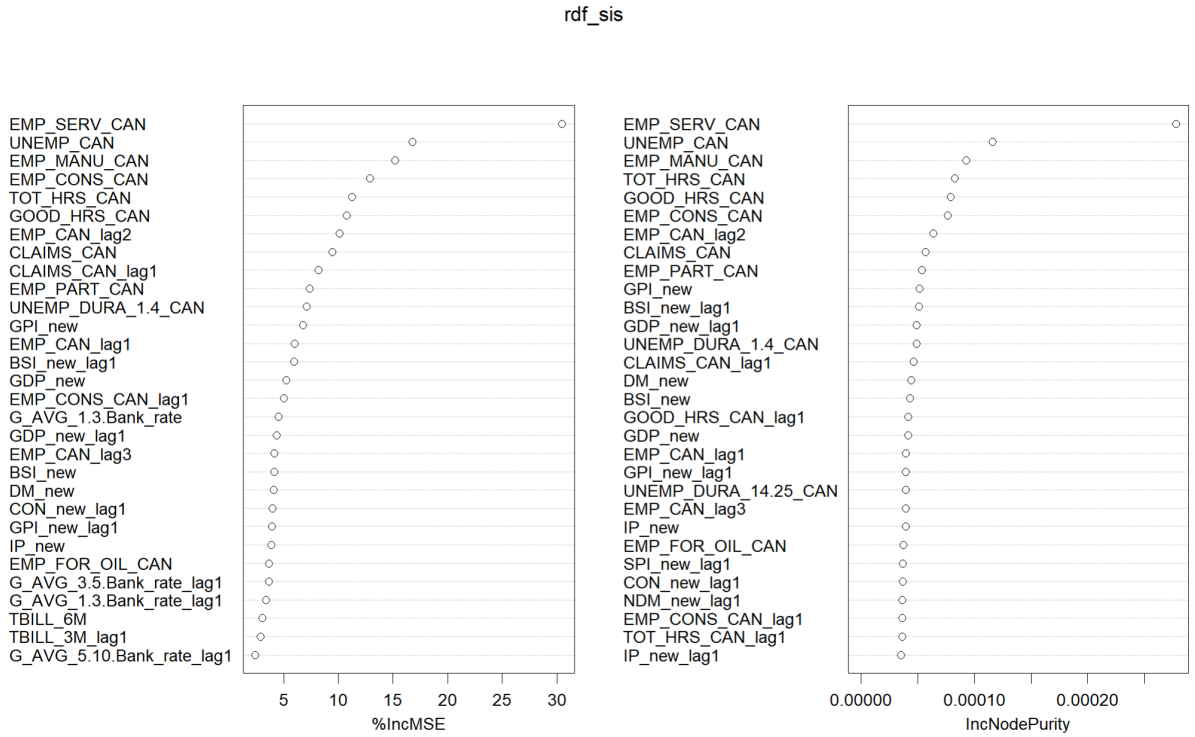
Figure 10-Random Forest sans Réduction SIS




### 3.4.2. Random forest avec réduction SIS

Dans cette seconde étape, nous avons appliqué le modèle Random Forest après une réduction préalable des variables par la méthode SIS. L'objectif de cette approche est de limiter le nombre de variables en amont, pour améliorer la précision et la simplicité de l'interprétation du modèle. Le modèle Random Forest est ici paramétré de la même manière que précédemment, avec 1000 arbres et une sélection aléatoire de 3 variables à chaque division.

Figure 11-Random Forest avec Réduction SIS




Les graphiques ci-dessus présentent les résultats obtenus pour les deux mesures d'importance des variables, %IncMSE et IncNodePurity, dans ce contexte réduit. L'analyse montre que certaines variables, telles que EMP_SERV_CAN et UNEMP_CAN (taux de chômage), restent dominantes et importantes, même après réduction SIS. Cela confirme leur rôle crucial dans la modélisation du taux d'emploi canadien. D'autres variables, comme EMP_MANU_CAN et TOT_HRS_CAN , apparaissent également significatives, soutenant leur pertinence dans le modèle. Avec ce modèle, nous pouvons estimer qu'entre quatre variables selon %IncMSE et six variables selon IncNodePurity jouent un rôle important dans l'explication du taux d'emploi.

Cette réduction par SIS a permis de concentrer le modèle sur un ensemble plus ciblé de variables, en éliminant celles ayant moins d'influence tout en conservant les plus importantes. Cela permet une modélisation plus focalisée et potentiellement plus efficace, en offrant une alternative aux approches sans réduction de variables. Cette approche combinée nous donne ainsi une perspective complémentaire sur les déterminants clés du taux d'emploi canadien.

## 3.5. Comparaison des résultats

Nous allons maintenant résumer et comparer les résultats obtenus. Dans un premier temps, les paramètres des modèles sans la réduction par SIS montrent une large variation des valeurs de régularisation λ (Tableau 18). Par exemple, pour le modèle Ridge, le λ élevé de 4,329 indique une forte régularisation, tandis que des modèles tels que Lasso et SCAD, présentent des valeurs de λ plus faibles, avec respectivement un λ de 0,226 et 0,153.

Tableau 18 - Récapitulatif des paramètres des modèles sans SIS

| | Ridge | Lasso | Elastic Net (fixed 0.5) | Elastic Net (grid search) | Elastic Net (caret) | SCAD | Adaptive Lasso |
|---|---|---|---|---|---|---|---|
| Lambda | 4,329 | 0,266 | 0,464 | 0,320 | 0,328 | 0,153 | 0,654 |
| Alpha | | | 0,5 | 0,832 | 0,2 | | |

Après l'application de SIS, la plupart des λ augmentent (Tableau 19), à l'exception de celles de Ridge et d'aLasso. Cette hausse des λ suggère qu'en dépit d'un nombre réduit de variables, les méthodes de régression renforcent leur régularisation afin d'optimiser leurs performances. Il est également notable que les paramètres de régularisation varient là aussi considérablement, allant de 3,594 pour Ridge à 0,218 pour SCAD.

Tableau 19 - Récapitulatif des paramètres des modèles avec SIS

| | Ridge | Lasso | Elastic Net (fixed 0.5) | Elastic Net (grid search) | Elastic Net (caret) | SCAD | Adaptive Lasso |
|---|---|---|---|---|---|---|---|
| Lambda | 3.594 | 0.320 | 0.673 | 0,385 | 0.329 | 0.218 | 0,588 |
| Alpha | | | 0.5 | 0,838 | 0,2 | | |

Nous allons maintenant analyser les résultats de la sélection de variables obtenus avec les différentes méthodes, sans et avec la réduction de dimension SIS. Les variables les plus souvent sélectionnées, accompagnées de leurs valeurs de coefficient β

correspondantes, sont présentées dans le tableau 20 pour la configuration sans SIS et dans le tableau 21 pour celle avec SIS. Les résultats détaillés se trouvent en annexe 15 pour les analyses sans SIS et en annexe 16 pour celles avec SIS.

Tout d'abord, les méthodes de régularisation sans SIS conservent entre 100 variables pour GETS et aucune pour aLasso. Les modèles Lasso, SCAD, Elastic-Net avec alpha fixé et grid search retiennent un nombre limité de variables, avec respectivement 3, 7, 6 et 3 variables explicatives. En revanche, la méthode Elastic-Net random search sélectionnent un plus grand nombre de variables, avec 42 variables explicatives gardées.

La variable NHOUSE_P_CAN_lag1 (indice des prix des logements neux retardé de 1 mois) est conservée par toutes les méthodes sauf aLasso, avec un bêta positif globalement élevé en comparaison avec les autres coefficient également retenues par toutes les méthodes sauf GETS et aLasso. Les variables, EMP_CAN_lag2 (taux d'emploi retardé de 2 mois) et UNEMP_CAN (taux de chômage) quant à eu retenu par toutes les méthodes sauf GETS et aLasso. Le coefficient de EMP_CAN_lag2 est relativement élevé, tandis que celui de UNEMP_CAN se distingue comme le seul coefficient négatif parmi les résultats de ces méthodes. De plus, la méthode GETS attribue un coefficient négatif à la variable DM_new_lag1 (PIB biens durables retardé de 1 mois) alors qu'il est positif pour les autres méthodes conservant cette variable.

Tableau 21- Récapitulatif des résultats des méthodes de régularisations sans SIS

| | GETS | Ridge | Lasso | Elastic Net (fixed 0.5) | Elastic Net (grid search) | Elastic Net (caret) | SCAD | Adaptive Lasso |
|---|---|---|---|---|---|---|---|---|
| Nombre de variables sélectionnées | 100 | 230 | 3 | 6 | 3 | 42 | 7 | 0 |
| EMP_CAN_lag2 | | 0.035 | 0.047 | 0.059 | 0.045 | 9.67e-02 | 1.69e-01 | |
| G_AVG_1.3.Bank_rate_lag1 | | 0.016 | | | | 2.79e-02 | 1.77e-02 | |
| TBILL_6M.Bank_rate_lag1 | | 0.021 | | | | 3.79e-02 | 5.96e-03 | |
| NHOUSE_P_CAN_lag1 | 0.185 | 0.033 | 0.016 | 0.036 | 0.016 | 9.99e-02 | 1.06e-01 | |
| EMP_FOR_OIL_CAN | 0.199 | 0.029 | | | | 7.89e-02 | 1.07e-02 | |
| UNEMP_CAN | | -0.028 | -0.106 | -0.102 | -0.100 | -6.07e-02 | -2.37e-01 | |
| WT_new_lag1 | 0.184 | 0.026 | | 0.011 | | 8.76e-02 | 8.81e-02 | |
| DM_new_lag1 | -0.133 | 0.016 | | 0.014 | | 2.20e-02 | | |
| EMP_CAN_lag1 | | 0.022 | | 0.003 | | 2.91e-02 | | |

Avec la réduction de variable SIS, le nombre de variables sélectionnées à fortement diminuer. Nous ne retrouvons plus que 22 variables pour Elastic-Net random search, 19 variables pour GETS, 2 pour Lasso, SCAD et Elastic Net grid search. Nous n'avons plus qu'une seule variable pour Elastic Net fixed et toujours aucune pour aLasso. Les variables les plus sélectionnées sont, UNEMP_CAN qui est retenue par toutes les méthodes sauf GETS et aLasso, et EMP_CAN_lag2, qui est conservée par toutes les méthodes sauf Elastic-Net avec alpha fixé et aLasso. UNEMP_CAN affiche un coefficient négatif, tandis que EMP_CAN_lag2 présente un coefficient positif.

Tableau 22- Récapitulatif des résultats des méthodes de régularisations avec SIS

| | GETS | Ridge | Lasso | Elastic Net (fixed 0.5) | Elastic Net (grid search) | Elastic Net (caret) | SCAD | Adaptive Lasso |
|---|---|---|---|---|---|---|---|---|
| Nombre de variables sélectionnées | 19 | 65 | 2 | 1 | 2 | 22 | 2 | 0 |
| UNEMP_CAN | | -3.62e-02 | -0.059 | -0.032 | -0.054 | -9.31e-02 | -2.19e-01 | |
| EMP_CAN_lag2 | 0.104 | 5.04e-02 | 0.003 | | 0.001 | 1.31e-01 | 1.28e-01 | |

En résumé, NHOUSE_P_CAN_lag1 était la variable la plus sélectionnée par les méthodes de régularisation avant la réduction de dimension. Cependant, cette variable n'a pas été retenue dans la base lors de la réduction de dimension SIS. Par conséquent, ces variables ne sont plus présentes dans les résultats des méthodes avec la base réduite avec SIS.

En revanche, UNEMP_CAN et EMP_CAN_lag2 sont quant à eux conservés par 6 méthodes sur 8, tant avant qu'après la réduction par SIS. De plus, UNEMP_CAN se classe comme la deuxième variable la plus discriminante dans les analyses de Random Forest, tant avec que sans SIS. Tandis qu'EMP_CAN_lag2 est la septième variable la plus discriminantes pour random forest avec SIS. Cependant, la variable la plus influente selon Random Forest est EMP_SERV_CAN mais elle n'est conservée que par 4 des 8 méthodes que ce soit avec ou sans SIS. Le taux de chômage et le taux d'emploi retardé de 2 mois semblent donc jouer un rôle significatif dans l'explication du taux d'emploi.

# 4. Conclusion

À travers diverses méthodes de sélection de variables, nous avons pu mieux comprendre les déterminants du taux d'emploi et optimiser les modèles. En partant d'une base de données initiale comprenant 113 variables explicatives, nous avons intégré des variables retardées pour un total de 230 variables.

Pour sélectionner les variables les plus pertinentes, nous avons appliqué plusieurs techniques de sélection et de réduction de dimension, notamment les approches économétriques comme GETS, les régressions pénalisées (Ridge, Lasso, Elastic Net, …) et une méthode d'apprentissage supervisé, Random Forest. De plus, pour évaluer la robustesse de nos résultats, nous avons appliqué ces techniques sur la base initiale (230 variables) ainsi que sur une version réduite à 65 variables grâce à l'approche SIS.

Les résultats révèlent une variation notable dans le nombre de variables sélectionnées en fonction des méthodes employées. Dans le modèle initial à 230 variables, la méthode GETS a retenu un nombre conséquent de variables (100 sur 230), tandis que les autres méthodes en ont sélectionné nettement moins. En particulier, SCAD, Lasso, Elastic Net fixed et Elastic grid search ont fortement réduit le nombre de variables, en conservant respectivement seulement 7, 3, 6 et 3 variables. Elastic random search, a sélectionné un nombre intermédiaire de variables, avec 42 variables explicatives. De son côté, la Random Forest a permis d'identifier les variables les plus influentes pour la construction de ses nœuds de décision.

Sur la base de données réduite, les résultats montrent quelques différences notables. GETS a sélectionné un nombre plus limité de variables (19), alors qu'Elastic Net random search a conservé légèrement plus de variables (22). Lasso, Elastic Net fixed SCAD et Elastic grid search ont maintenu une sélection très restreinte, avec 2, 1, 2 et 2 variables retenues respectivement.

De l'ensemble des analyses, il ressort que deux variables semblent particulièrement déterminantes pour expliquer le taux d'emploi : le taux de chômage et

le taux d'emploi retardé de deux mois. Ces résultats indiquent que le taux de chômage joue un rôle central en tant qu'indicateur inverse du taux d'emploi, tandis que le taux d'emploi retardé traduit une dynamique d'inertie ou de persistance sur le marché de l'emploi.

En conclusion, bien que les méthodes diffèrent dans leur approche, elles réussissent à capturer les déterminants clés du taux d'emploi. Cependant, elles se distinguent par la sélection de variables, variant en nombre et en choix selon la méthode. Les méthodes GETS et random search fournissent une vision exhaustive en sélectionnant un large éventail de variables. Quant à Lasso, SCAD, Elastic Net fixed et Elastic Net grid search se démarquent par leur parcimonie et leurs sélections strictes, faisant d'eux des choix idéaux pour des modèles simplifiés, concentrés sur les variables ayant le plus fort impact.

Ainsi, en fonction de l'objectif de l'analyse, GETS et Elastic random search sont plus adaptés pour une exploration détaillée des déterminants du taux d'emploi. Tandis que Lasso, SCAD, Elastic Net fixed et Elastic Net grid search se révèlent plus avantageux pour des analyses nécessitant un modèle clair et concentré sur les variables essentielles.

# 5. Annexes

Annexe n°1 : Valeurs manquantes

```
> naniar::miss_var_summary(dataCAN)
# A tibble: 120 × 3
   variable                n_miss pct_miss
   <chr>                    <int>    <num>
 1 CRED_T_discontinued         46     8.80
 2 CRED_HOUS_discontinued      46     8.80
 3 CRED_MORT_discontinued      46     8.80
 4 CRED_CONS_discontinued      46     8.80
 5 CRE_BUS_discontinued        46     8.80
 6 Date                         0     0
 7 GDP_new                      0     0
 8 BSI_new                      0     0
 9 GPI_new                      0     0
10 SPI_new                      0     0
# i 110 more rows
# i Use `print(n = ...)` to see more rows
```

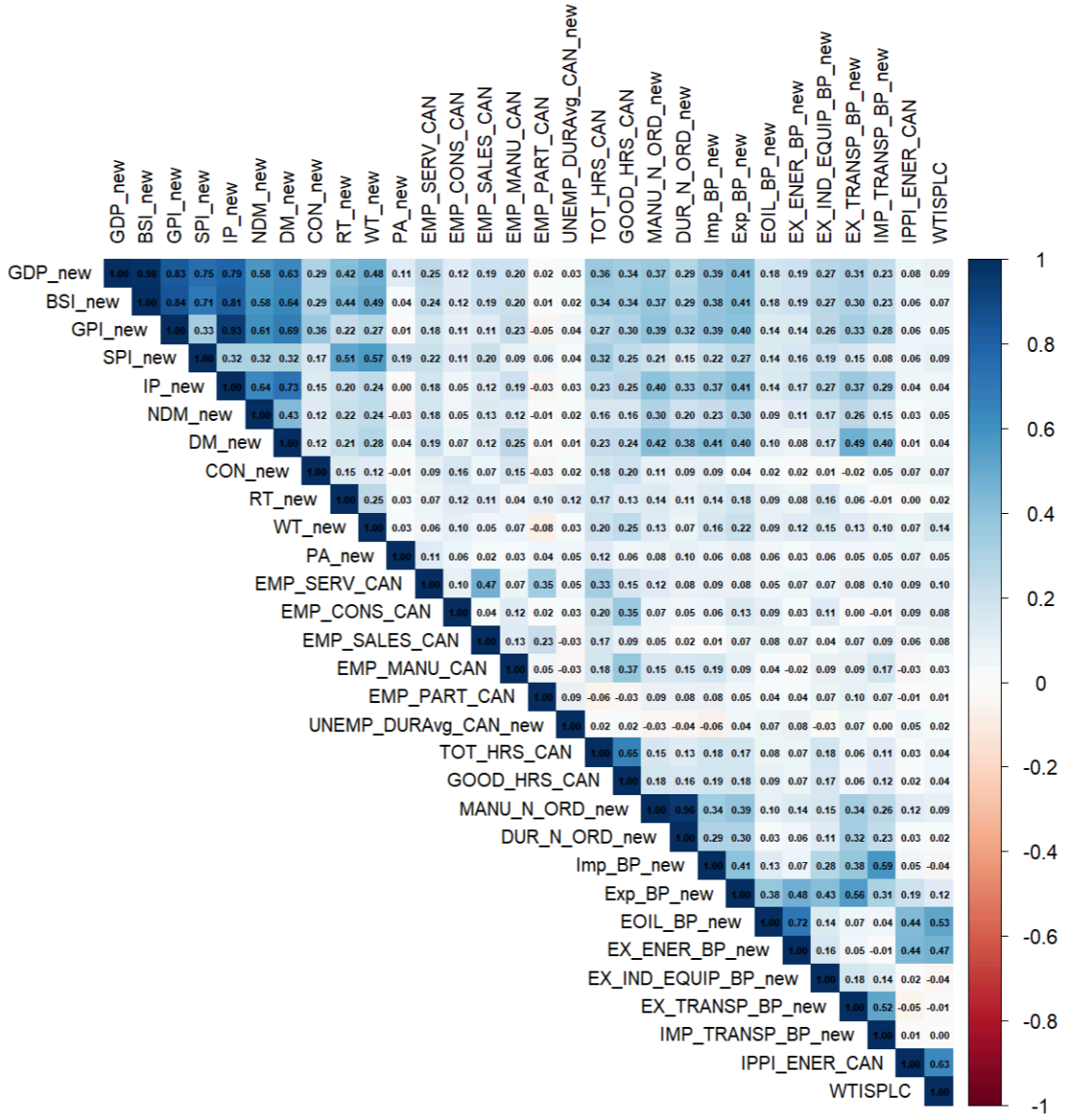
Annexe n°2 : Valeurs atypiques

| | Type | Individu | Date |
|---|---|---|---|
| 1 | TC | 471 | Mars 2020 |
| 2 | AO | 472 | Avril 2020 |
| 3 | TC | 473 | Mai 2020 |
| 4 | AO | 474 | Juin 2020 |
| 5 | AO | 477 | Septembre 2020 |
| 6 | AO | 481 | Janvier 2021 |
| 7 | TC | 482 | Février 2021 |
| 8 | TC | 484 | Avril 2021 |
| 9 | TC | 486 | Juin 2021 |
| 10 | TC | 493 | Janvier 2022 |
| 11 | AO | 494 | Février 2022 |

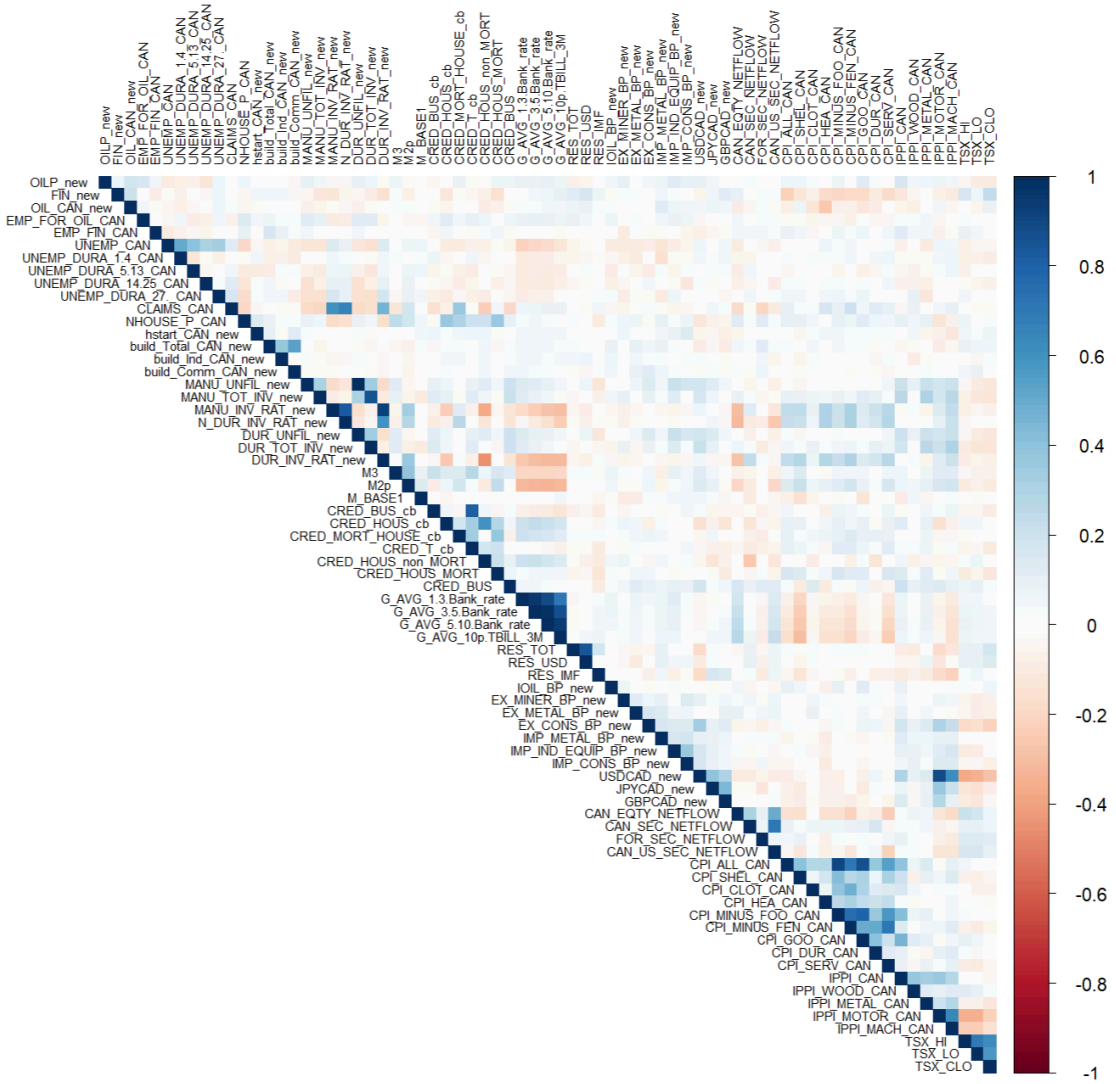
## Annexe n°3 : Statistique descriptive

```
> summary(empl2$EMP_CAN)
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-0.0082561 -0.0001179  0.0012872  0.0012107  0.0026490  0.0123738
> boxplot(empl2$EMP_CAN)
> mean(empl2$EMP_CAN)
[1] 0.001210733
> sd(empl2$EMP_CAN)
[1] 0.002322045
> var(empl2$EMP_CAN)
[1] 5.391892e-06
> library(moments)
> skewness(empl2$EMP_CAN)
[1] -0.1896643
> kurtosis(empl2$EMP_CAN)
[1] 4.597711
>
```

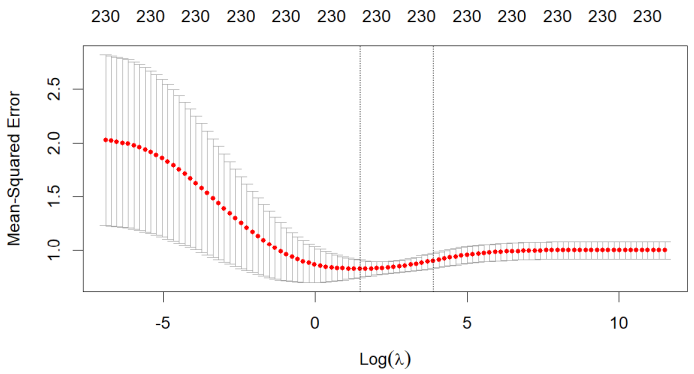
## Annexe n°4 - Corrélations entre les variables explicatives du cluster 2



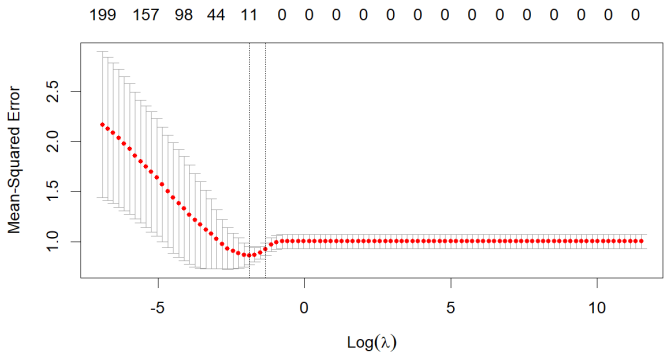
## Annexe n°5 - Corrélations entre les variables explicatives du cluster 3



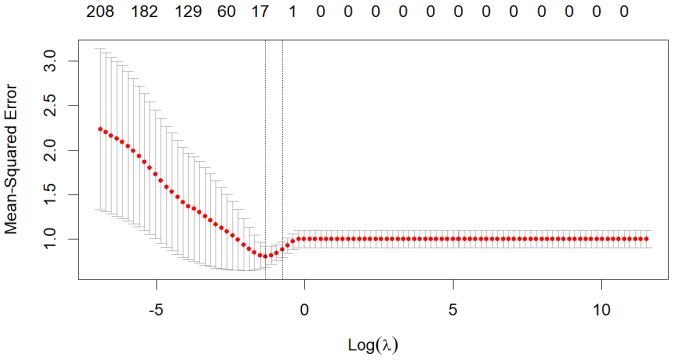

## Annexe n°6 - Valeur de lambda pour Ridge



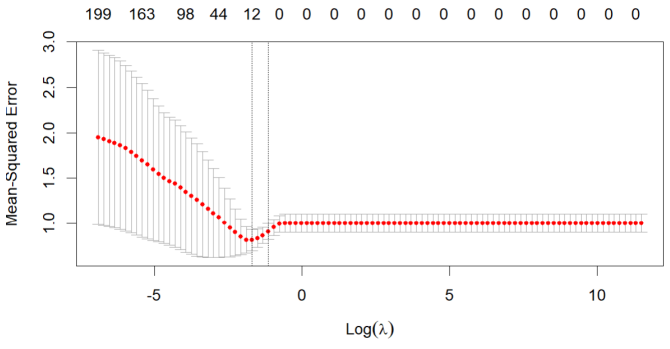
## Annexe n°7 - Valeur de lambda pour Lasso



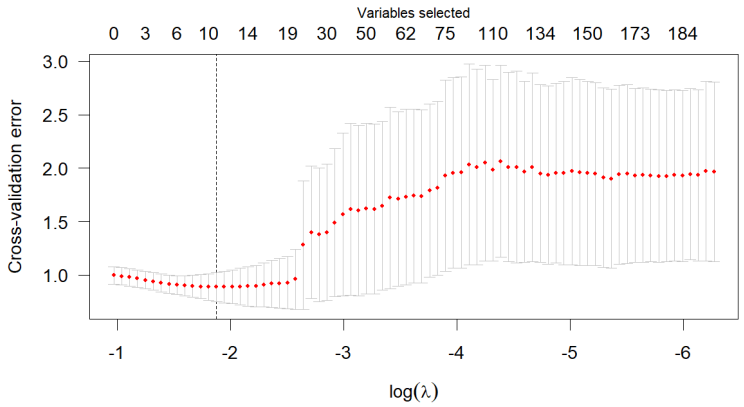

## Annexe n°8 - Valeur de lambda pour Elastic Net (fixed)



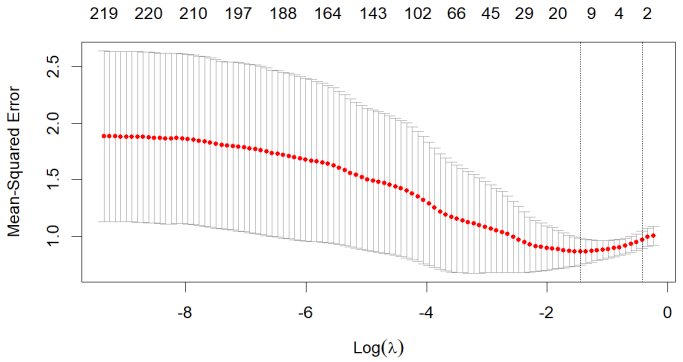

## Annexe n°9 - Valeur de lambda pour Elastic Net (grid search)




## Annexe n°10 - Valeur de lambda pour SCAD



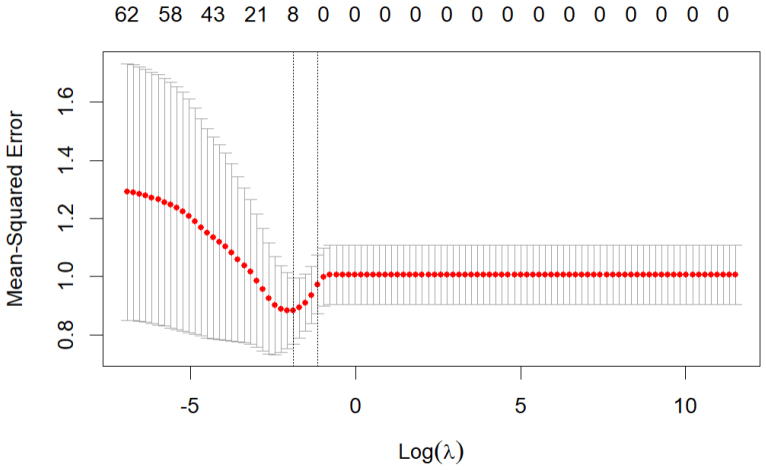

Annexe n°11 - Valeur de lambda pour Adaptive Lasso




Annexe n°12 - Méthode GETS-SIS

```
> names_mX_getsm2_sis
 [1] "ar1"                       "BSI_new"
 [3] "IP_new"                    "EMP_SERV_CAN"
 [5] "EMP_FOR_OIL_CAN"           "EMP_CONS_CAN"
 [7] "EMP_MANU_CAN"              "UNEMP_DURA_14.25_CAN"
 [9] "TSX_CLO"                   "GDP_new_lag1"
[11] "BSI_new_lag1"              "OILP_new_lag1"
[13] "EMP_CONS_CAN_lag1"         "UNEMP_DURA_1.4_CAN_lag1"
[15] "CLAIMS_CAN_lag1"           "TOT_HRS_CAN_lag1"
[17] "GOOD_HRS_CAN_lag1"         "G_AVG_1.3.Bank_rate_lag1"
[19] "G_AVG_3.5.Bank_rate_lag1" "EMP_CAN_lag2"
```

Annexe n°13 - Méthode Lasso-SIS




```
> # Get the name of relevant variables
> which(! coef(model_cv_sis_lasso) == 0, arr.ind = TRUE)
             row col
(Intercept)    1   1
UNEMP_CAN     14   1
EMP_CAN_lag2  65   1
```

Beta :

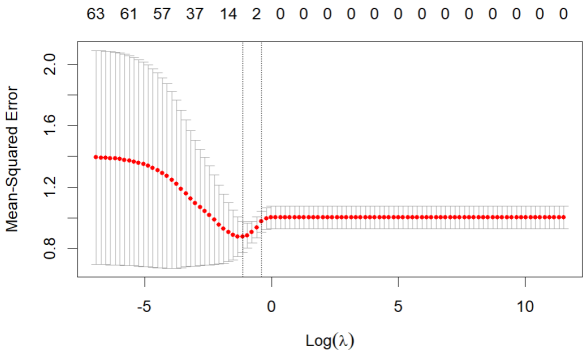
EMP_CAN_lag2 0.002992574

UNEMP_CAN -0.059456173

lambda_cv_sis_lasso

[1] 0.3199267

## Annexe n°14 - Méthode Elastic-Net (fixer à priori $\alpha$ = 0.5)-SIS




```
> which(! coef(model_cv_sis_en) == 0, arr.ind = TRUE)
            row col
(Intercept)   1   1
UNEMP_CAN    14   1
```

beta :

UNEMP_CAN -0.03219427

lambda_cv_sis_en

[1] 0.6734151

Annexe n°15 - Résultats de la sélection de variable sans la sélection SIS

| | GETS | Ridge | Lasso | Elastic Net (fixed 0.5) | Elastic Net (grid search) | Elastic Net (caret) | SCAD | Adaptive Lasso |
|---|---|---|---|---|---|---|---|---|
| Nombre de variables sélectionnées | 100 | 230 | 3 | 6 | 3 | 42 | 7 | 0 |
| EMP_CAN_lag2 | | 0.035 | 0.047 | 0.059 | 0.045 | 9.67e-02 | 1.69e-01 | |
| G_AVG_1.3.Bank_rate_lag1 | | 0.016 | | | | 2.79e-02 | 1.77e-02 | |
| TBILL_6M.Bank_rate_lag1 | | 0.021 | | | | 3.79e-02 | 5.96e-03 | |
| NHOUSE_P_CAN_lag1 | 0.185 | 0.033 | 0.016 | 0.036 | 0.016 | 9.99e-02 | 1.06e-01 | |
| EMP_FOR_OIL_CAN | 0.199 | 0.029 | | | | 7.89e-02 | 1.07e-02 | |
| UNEMP_CAN | | -0.028 | -0.106 | -0.102 | -0.100 | -6.07e-02 | -2.37e-01 | |
| WT_new_lag1 | 0.184 | 0.026 | | 0.011 | | 8.76e-02 | 8.81e-02 | |
| DM_new_lag1 | -0.133 | 0.016 | | 0.014 | | 2.20e-02 | | |
| EMP_CAN_lag1 | | 0.022 | | 0.003 | | 2.91e-02 | | |
| EMP_SERV_CAN | 1.008 | 0.025 | | | | 5.59e-02 | | |
| EMP_CONS_CAN | 0.272 | 0.022 | | | | 2.64e-02 | | |
| EMP_MANU_CAN | 0.361 | 0.026 | | | | 5.09e-02 | | |
| UNEMP_DURA_1.4_CAN | -0.118 | -0.024 | | | | -5.42e-02 | | |
| UNEMP_DURA_14.25_CAN | -0.071 | -0.016 | | | | -1.13e-02 | | |
| UNEMP_DURA_27._CAN | 0.178 | -0.030 | | | | -7.01e-02 | | |
| CLAIMS_CAN | | -0.018 | | | | -2.82e-02 | | |
| NHOUSE_P_CAN | -0.076 | 0.021 | | | | 1.76e-04 | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| hstart_CAN_new | 0.106 | 0.014 | | | | 9.04e-03 | | |
| build_Ind_CAN_new | | 0.017 | | | | 2.15e-02 | | |
| build_Comm_CAN_new | | 0.003 | | | | 1.61e-02 | | |
| M_BASE1 | 0.204 | 0.015 | | | | 3.83e-02 | | |
| G_AVG_1.3.Bank_rate | 0.135 | 0.014 | | | | 7.26e-03 | | |
| TBILL_6M.Bank_rate | 0.120 | 0.021 | | | | 2.85e-02 | | |
| RES_TOT | | 0.012 | | | | 3.03e-04 | | |
| EX_MINER_BP_new | | 0.017 | | | | 2.59e-02 | | |
| EX_CONS_BP_new | | 0.011 | | | | 1.10e-03 | | |
| IMP_METAL_BP_new | | 0.014 | | | | 5.75e-03 | | |
| CPI_SERV_CAN | | -0.009 | | | | -9.34e-03 | | |
| IPPI_WOOD_CAN | | 0.011 | | | | 3.34e-03 | | |
| OILP_new_lag1 | | -0.016 | | | | -3.50e-04 | | |
| CON_new_lag1 | | 0.015 | | | | 2.94e-03 | | |
| PA_new_lag1 | 0.102 | -0.013 | | | | -1.10e-02 | | |
| EMP_CONS_CAN_lag1 | | 0.018 | | | | 3.28e-02 | | |
| EMP_MANU_CAN_lag1 | -0.009 | 0.018 | | | | 3.66e-02 | | |
| UNEMP_DURA_5.13_CAN_lag1 | 0.120 | -0.016 | | | | -1.45e-02 | | |
| CLAIMS_CAN_lag1 | -0.126 | -0.018 | | | | -3.33e-02 | | |
| build_Comm_CAN_new_lag1 | 0.116 | 0.012 | | | | 3.82e-0 | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| DUR_INV_RAT_new_lag1 | -0.639 | -0.010 | | | | -1.87e-03 | | |
| IMP_TRANSP_BP_new_lag 1 | | 0.015 | | | | 1.86e-02 | | |
| TSX_HI_lag1 | 0.092 | 0.011 | | | | 1.94e-03 | | |
| EMP_CAN_lag3 | | 0.023 | | | | 3.24e-02 | | |
| EMP_CAN_lag4 | 0.097 | 0.023 | | | | 2.23e-02 | | |

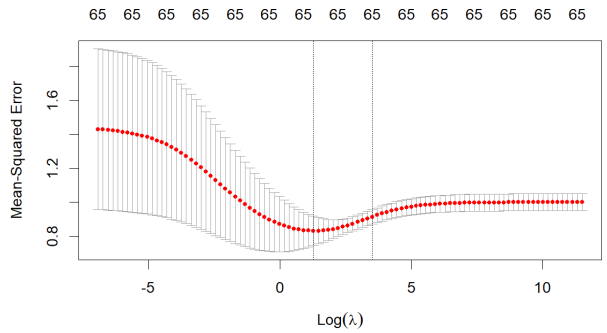
Annexe n° 16- Résultats de la sélection de variable avec la sélection SIS

| | GETS | Ridge | Lasso | Elastic Net (fixed 0.5) | Elastic Net (grid search) | Elastic Net (caret) | SCAD | Adaptive Lasso |
|---|---|---|---|---|---|---|---|---|
| Nombre de variables sélectionnées | 19 | 65 | 2 | 1 | 21 | 22 | 2 | 0 |
| UNEMP_CAN | | -3.62e-02 | -0.059 | -0.032 | -0.097 | -9.31e-02 | -2.19e-01 | |
| EMP_CAN_lag2 | 0.104 | 5.04e-02 | 0.003 | | 0.124 | 1.31e-01 | 1.28e-01 | |
| BSI_new | -0.513 | | | | | | | |
| IP_new | 0.199 | | | | | | | |
| EMP_SERV_CAN | 0.424 | 2.71e-02 | | | 0.031 | 3.34e-02 | | |
| EMP_FOR_OIL_CAN | 0.163 | 3.67e-02 | | | 0.074 | 8.25e-02 | | |
| EMP_CONS_CAN | 0.215 | 2.50e-02 | | | 0.011 | 1.51e-02 | | |
| EMP_MANU_CAN | 0.196 | 3.11e-02 | | | 0.049 | 5.10e-02 | | |
| UNEMP_DURA_1.4_CAN | | -2.65e-02 | | | -0.024 | -3.56e-02 | | |
| UNEMP_DURA_14.25_CAN | -0.10 | -2.39e-0 | | | -0.006 | -1.67e-02 | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 3 | 2 | | | | | | |
| TSX_CLO | 0.088 | | | | | | | |
| CLAIMS_CAN | | -2.68e-02 | | | -0.039 | -4.66e-02 | | |
| GOOD_OVT_HRS_CAN | | 2.03e-02 | | | 0.010 | 1.26e-02 | | |
| G_AVG_1.3.Bank_rate | | 2.21e-02 | | | 0.023 | 2.66e-02 | | |
| GDP_new_lag1 | -1.259 | | | | | | | |
| BSI_new_lag1 | 1.358 | 1.94e-02 | | | 0.013 | 1.24e-02 | | |
| GPI_new_lag1 | | 1.94e-02 | | | 0.003 | 5.36e-03 | | |
| NDM_new_lag1 | | 1.97e-02 | | | 0.023 | 3.07e-02 | | |
| OILP_new_lag1 | -0.133 | -1.97e-02 | | | | -6.17e-03 | | |
| CON_new_lag1 | | 2.38e-02 | | | 0.031 | 3.53e-02 | | |
| EMP_CAN_lag1 | | 3.57e-02 | | | 0.054 | 5.42e-02 | | |
| EMP_CONS_CAN_lag1 | 0.113 | 3.05e-02 | | | 0.061 | 6.75e-02 | | |
| UNEMP_DURA_1.4_CAN_lag1 | 0.139 | | | | | | | |
| CLAIMS_CAN_lag1 | -0.116 | -2.43e-02 | | | -0.018 | -2.53e-02 | | |
| TOT_HRS_CAN_lag1 | -0.357 | | | | | | | |
| GOOD_OVT_HRS_CAN_lag1 | | 2.08e-02 | | | 0.006 | 1.01e-02 | | |
| GOOD_HRS_CAN_lag1 | 0.309 | | | | | | | |
| G_AVG_1.3.Bank_rate_lag1 | 0.435 | 2.54e-0 | | | 0.047 | 4.90e-02 | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 2 | | | | | | |
| G_AVG_3.5.Bank_rate_lag1 | -0.331 | | | | | | | |
| EMP_CAN_lag3 | | 3.35e-02 | | | 0.047 | 5.11e-02 | | |

Annexe n° 17- Ridge-SIS




lambda_cv_sis_ridge

[1] 3.593814

model_cv_sis_ridge$beta

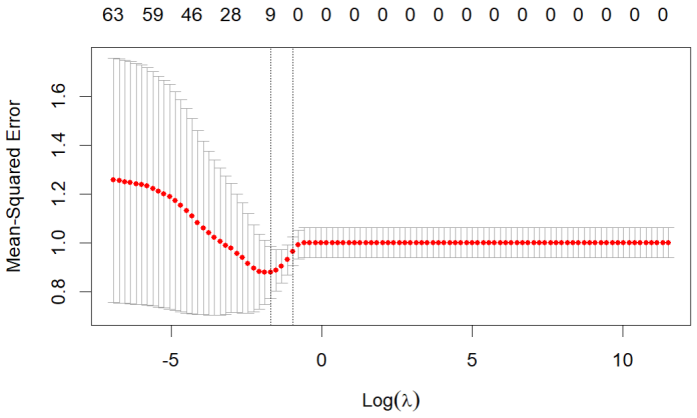
65 x 1 sparse Matrix of class "dgCMatrix"

s0

| | | | |
|---|---|---|---|
| GDP_new | 7.976444e-04 | UNEMP_DURA_14.25_CAN | -2.388486e-02 |
| BSI_new | 4.608160e-04 | CLAIMS_CAN | -2.677665e-02 |
| GPI_new | 4.711340e-03 | TOT_HRS_CAN | 1.002560e-02 |
| IP_new | 8.267118e-03 | GOOD_HRS_CAN | 6.713241e-03 |
| NDM_new | 9.134154e-03 | GOOD_OVT_HRS_CAN | 2.025515e-02 |
| DM_new | 3.265977e-03 | MANU_TOT_INV_new | 5.237478e-03 |
| EMP_SERV_CAN | 2.710254e-02 | DUR_TOT_INV_new | 1.447051e-02 |
| EMP_FOR_OIL_CAN | 3.665191e-02 | CRED_BUS_cb | 5.005135e-03 |
| EMP_CONS_CAN | 2.503809e-02 | CRED_T_cb | 2.694300e-03 |
| EMP_FIN_CAN | 8.790197e-03 | CRED_BUS | 1.003372e-02 |
| EMP_MANU_CAN | 3.111315e-02 | TBILL_6M | 1.248988e-02 |
| EMP_PART_CAN | 1.543270e-02 | G_AVG_1.3.Bank_rate | 2.203485e-02 |
| UNEMP_CAN | -3.617321e-02 | G_AVG_5.10.Bank_rate | 7.365231e-03 |
| UNEMP_DURA_1.4_CAN | -2.649956e-02 | RES_IMF | -8.932429e-03 |

| | |
|---|---|
| IOIL_BP_new | -2.583378e-03 |
| TSX_CLO | 1.110000e-02 |
| GDP_new_lag1 | 1.515076e-02 |
| BSI_new_lag1 | 1.938073e-02 |
| GPI_new_lag1 | 1.934491e-02 |
| SPI_new_lag1 | 9.478849e-03 |
| IP_new_lag1 | 1.481096e-02 |
| NDM_new_lag1 | 1.970242e-02 |
| OILP_new_lag1 | -1.936767e-02 |
| CON_new_lag1 | 2.382915e-02 |
| RT_new_lag1 | 1.521334e-02 |
| OIL_CAN_new_lag1 | -5.696317e-04 |
| EMP_CAN_lag1 | 3.574803e-02 |
| EMP_FOR_OIL_CAN_lag1 | -4.399727e-03 |
| EMP_CONS_CAN_lag1 | 3.050305e-02 |
| EMP_FIN_CAN_lag1 | -7.683244e-03 |
| EMP_PART_CAN_lag1 | -7.336164e-03 |
| UNEMP_DURA_1.4_CAN_lag1 | 1.328718e-02 |
| CLAIMS_CAN_lag1 | -2.428434e-02 |
| TOT_HRS_CAN_lag1 | 3.576079e-03 |
| GOOD_HRS_CAN_lag1 | 1.958974e-02 |
| GOOD_OVT_HRS_CAN_lag1 | 2.075198e-02 |
| MANU_TOT_INV_new_lag1 | 8.808555e-03 |
| DUR_TOT_INV_new_lag1 | 8.629302e-03 |
| TBILL_3M_lag1 | 1.131210e-02 |
| TBILL_6M_lag1 | 1.393818e-02 |
| G_AVG_1.3.Bank_rate_lag1 | 2.537088e-02 |
| G_AVG_3.5.Bank_rate_lag1 | 1.546747e-02 |
| G_AVG_5.10.Bank_rate_lag1 | 9.585710e-03 |
| RES_IMF_lag1 | -1.326480e-03 |
| IOIL_BP_new_lag1 | -8.531296e-05 |
| EX_IND_EQUIP_BP_new_lag1 | -1.214187e-02 |
| IMP_IND_EQUIP_BP_new_lag1 | 7.522136e-03 |
| IPPI_MACH_CAN_lag1 | -6.748067e-03 |
| TSX_CLO_lag1 | 6.664614e-03 |
| EMP_CAN_lag2 | 5.044204e-02 |
| EMP_CAN_lag3 | 3.352193e-0 |

## Annexe n° 18-Elastic-Net grid search-sis




```
> elasticnet_cvm_sis_gr
[1] 0.8376623
> lambda_cv_sis_gr
[1] 0.3853529
```

model_cv_sis_gr$beta
65 x 1 sparse Matrix of class "dgCMatrix"
s0

| | |
|---|---|
| GDP_new | . |
| BSI_new | . |
| GPI_new | . |

| | | | |
|---|---|---|---|
| IP_new | . | IP_new_lag1 | . |
| NDM_new | . | NDM_new_lag1 | 0.023175562 |
| DM_new | . | OILP_new_lag1 | . |
| EMP_SERV_CAN | 0.031371352 | CON_new_lag1 | 0.030843379 |
| EMP_FOR_OIL_CAN | 0.073963768 | RT_new_lag1 | . |
| EMP_CONS_CAN | 0.011259133 | OIL_CAN_new_lag1 | . |
| EMP_FIN_CAN | . | EMP_CAN_lag1 | 0.053888623 |
| EMP_MANU_CAN | 0.049331866 | EMP_FOR_OIL_CAN_lag1 | . |
| EMP_PART_CAN | . | EMP_CONS_CAN_lag1 | 0.061118316 |
| UNEMP_CAN | -0.096987263 | EMP_FIN_CAN_lag1 | . |
| UNEMP_DURA_1.4_CAN | -0.023702327 | EMP_PART_CAN_lag1 | . |
| UNEMP_DURA_14.25_CAN | -0.006446257 | UNEMP_DURA_1.4_CAN_lag1 | . |
| CLAIMS_CAN | -0.039096086 | CLAIMS_CAN_lag1 | -0.017997006 |
| TOT_HRS_CAN | . | TOT_HRS_CAN_lag1 | . |
| GOOD_HRS_CAN | . | GOOD_HRS_CAN_lag1 | . |
| GOOD_OVT_HRS_CAN | 0.009538289 | GOOD_OVT_HRS_CAN_lag1 | 0.006118947 |
| MANU_TOT_INV_new | . | MANU_TOT_INV_new_lag1 | . |
| DUR_TOT_INV_new | . | DUR_TOT_INV_new_lag1 | . |
| CRED_BUS_cb | . | TBILL_3M_lag1 | . |
| CRED_T_cb | . | TBILL_6M_lag1 | . |
| CRED_BUS | . | G_AVG_1.3.Bank_rate_lag1 | 0.046524759 |
| TBILL_6M | . | G_AVG_3.5.Bank_rate_lag1 | . |
| G_AVG_1.3.Bank_rate | 0.022970191 | G_AVG_5.10.Bank_rate_lag1 | . |
| G_AVG_5.10.Bank_rate | . | RES_IMF_lag1 | . |
| RES_IMF | . | IOIL_BP_new_lag1 | . |
| IOIL_BP_new | . | EX_IND_EQUIP_BP_new_lag1 | . |
| TSX_CLO | . | IMP_IND_EQUIP_BP_new_lag1 | . |
| GDP_new_lag1 | . | IPPI_MACH_CAN_lag1 | . |
| BSI_new_lag1 | 0.013484015 | TSX_CLO_lag1 | . |
| GPI_new_lag1 | 0.002994779 | EMP_CAN_lag2 | 0.124212735 |
| SPI_new_lag1 | . | EMP_CAN_lag3 | 0.046820922 |

## Annexe n° 19- Elastic-Net random search-sis

```
model$bestTune
  alpha    lambda
20   0.2 0.3287355
```

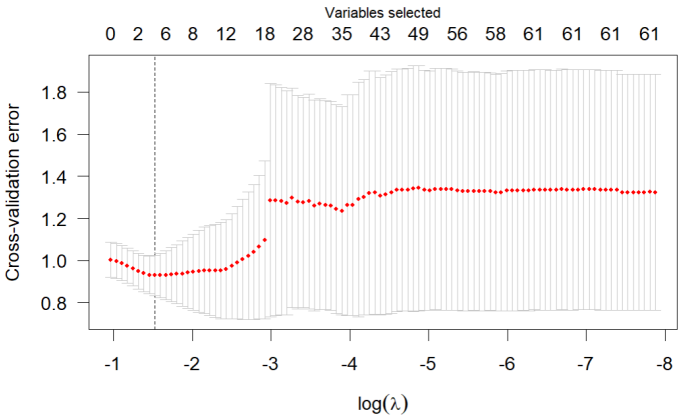
```
coef(model$finalModel, model$bestTune$lambda)
```

66 x 1 sparse Matrix of class "dgCMatrix"

| | s1 | | |
|---|---|---|---|
| (Intercept) | -1.802267e-17 | X6 | . |
| X1 | . | X7 | 3.340503e-02 |
| X2 | . | X8 | 8.253588e-02 |
| X3 | . | X9 | 1.505818e-02 |
| X4 | . | X10 | . |
| X5 | . | X11 | 5.103489e-02 |

| | | | |
|---|---|---|---|
| X12 | . | X39 | . |
| X13 | -9.313031e-02 | X40 | . |
| X14 | -3.559362e-02 | X41 | 5.417317e-02 |
| X15 | -1.671650e-02 | X42 | . |
| X16 | -4.655687e-02 | X43 | 6.747557e-02 |
| X17 | . | X44 | . |
| X18 | . | X45 | . |
| X19 | 1.258887e-02 | X46 | . |
| X20 | . | X47 | -2.530293e-02 |
| X21 | . | X48 | . |
| X22 | . | X49 | . |
| X23 | . | X50 | 1.005442e-02 |
| X24 | . | X51 | . |
| X25 | . | X52 | . |
| X26 | 2.657618e-02 | X53 | . |
| X27 | . | X54 | . |
| X28 | . | X55 | 4.899878e-02 |
| X29 | . | X56 | . |
| X30 | . | X57 | . |
| X31 | . | X58 | . |
| X32 | 1.236614e-02 | X59 | . |
| X33 | 5.362371e-03 | X60 | . |
| X34 | . | X61 | . |
| X35 | . | X62 | . |
| X36 | 3.065969e-02 | X63 | . |
| X37 | -6.165264e-03 | X64 | 1.305314e-01 |
| X38 | 3.527252e-02 | X65 | 5.113304e-02 |

Annexe n° 19- scad-sis




lambda_SCAD_sis

[1] 0.2172845

SCAD_Final_sis$beta

0.2173

(Intercept) -9.396600e-18

GDP_new 0.000000e+00

BSI_new 0.000000e+00

GPI_new 0.000000e+00

IP_new 0.000000e+00

NDM_new 0.000000e+00

DM_new 0.000000e+00

EMP_SERV_CAN 0.000000e+00

EMP_FOR_OIL_CAN 0.000000e+00

EMP_CONS_CAN 0.000000e+00

EMP_FIN_CAN 0.000000e+00

EMP_MANU_CAN 0.000000e+00

EMP_PART_CAN 0.000000e+00

UNEMP_CAN -2.191794e-01

UNEMP_DURA_1.4_CAN 0.000000e+00

UNEMP_DURA_14.25_CAN 0.000000e+00

CLAIMS_CAN 0.000000e+00

TOT_HRS_CAN 0.000000e+00

GOOD_HRS_CAN 0.000000e+00

GOOD_OVT_HRS_CAN 0.000000e+00

MANU_TOT_INV_new 0.000000e+00

DUR_TOT_INV_new 0.000000e+00

CRED_BUS_cb 0.000000e+00

CRED_T_cb 0.000000e+00

CRED_BUS 0.000000e+00

TBILL_6M 0.000000e+00

G_AVG_1.3.Bank_rate 0.000000e+00

G_AVG_5.10.Bank_rate 0.000000e+00

RES_IMF 0.000000e+00

IOIL_BP_new 0.000000e+00

TSX_CLO 0.000000e+00

GDP_new_lag1 0.000000e+00

BSI_new_lag1 0.000000e+00

GPI_new_lag1 0.000000e+00

SPI_new_lag1 0.000000e+00

IP_new_lag1 0.000000e+00

NDM_new_lag1 0.000000e+00

OILP_new_lag1 0.000000e+00

CON_new_lag1 0.000000e+00

RT_new_lag1 0.000000e+00

OIL_CAN_new_lag1 0.000000e+00

EMP_CAN_lag1 0.000000e+00

EMP_FOR_OIL_CAN_lag1 0.000000e+00

EMP_CONS_CAN_lag1 0.000000e+00

EMP_FIN_CAN_lag1 0.000000e+00

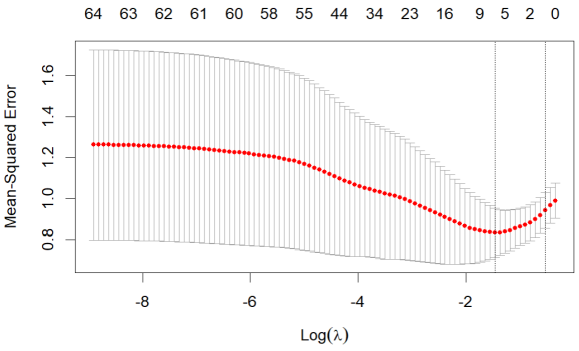
EMP_PART_CAN_lag1 0.000000e+00

UNEMP_DURA_1.4_CAN_lag1 0.000000e+00

CLAIMS_CAN_lag1 0.000000e+00
TOT_HRS_CAN_lag1 0.000000e+00
GOOD_HRS_CAN_lag1 0.000000e+00
GOOD_OVT_HRS_CAN_lag1 0.000000e+00
MANU_TOT_INV_new_lag1 0.000000e+00
DUR_TOT_INV_new_lag1 0.000000e+00
TBILL_3M_lag1 0.000000e+00
TBILL_6M_lag1 0.000000e+00
G_AVG_1.3.Bank_rate_lag1 0.000000e+00
G_AVG_3.5.Bank_rate_lag1 0.000000e+00
G_AVG_5.10.Bank_rate_lag1 0.000000e+00
RES_IMF_lag1 0.000000e+00
IOIL_BP_new_lag1 0.000000e+00
EX_IND_EQUIP_BP_new_lag1 0.000000e+00
IMP_IND_EQUIP_BP_new_lag1 0.000000e+00
IPPI_MACH_CAN_lag1 0.000000e+00
TSX_CLO_lag1 0.000000e+00
EMP_CAN_lag2 1.277170e-01
EMP_CAN_lag3 0.000000e+00

```
> which(! coef(SCAD_Final_sis) == 0, arr.ind = TRUE)
 (Intercept)   UNEMP_CAN EMP_CAN_lag2
           1          14          65
```

Annexe n° 20- alasso-sis

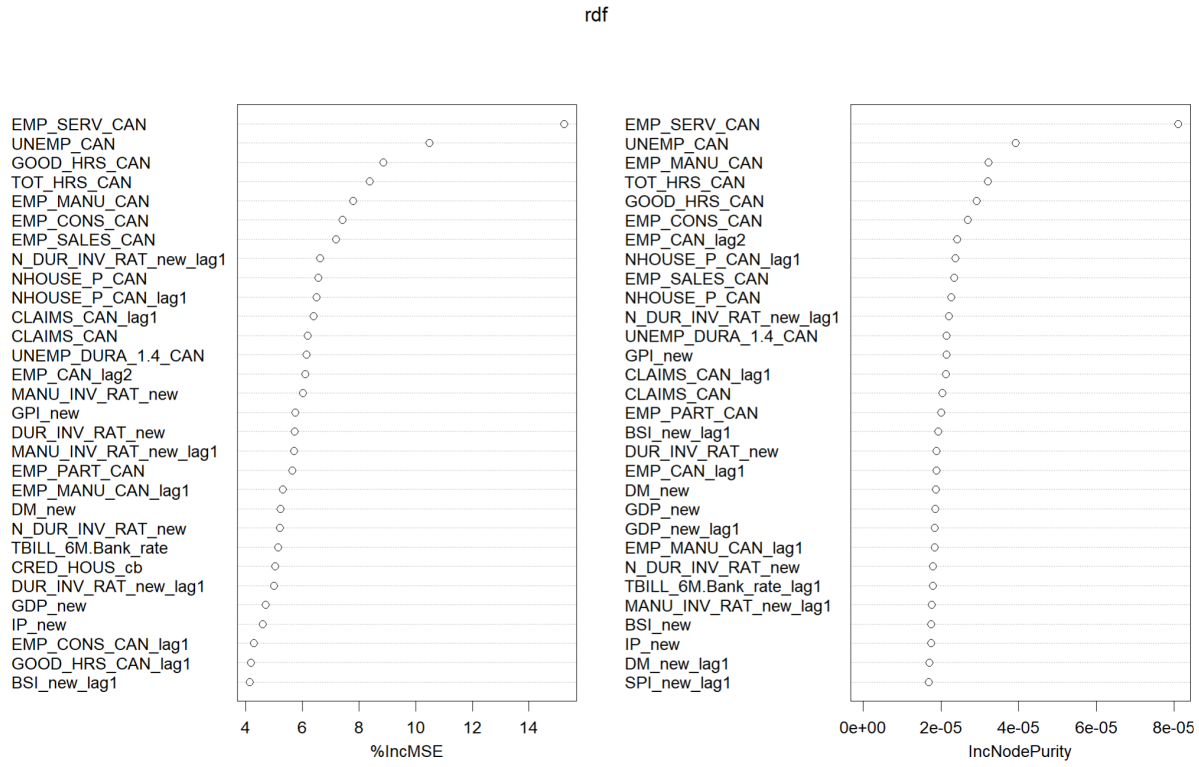



```
> which(! coef(model_cv_sis_alasso) == 0, arr.ind = TRUE)
             row col
(Intercept)   1   1

> lambda_cv_sis_alasso
[1] 0.5879425
```

Annexe n° 21- Random-forest

1. Random forest sans réduction SIS




```
rdf$importance[order(rdf$importance[, 1], decreasing = TRUE), ]
```

| | %IncMSE | IncNodePurity |
|---|---|---|
| EMP_SERV_CAN | 3.035784e-07 | 8.099995e-05 |
| UNEMP_CAN | 1.130575e-07 | 3.917623e-05 |
| GOOD_HRS_CAN | 8.101166e-08 | 2.919787e-05 |
| TOT_HRS_CAN | 7.906883e-08 | 3.210169e-05 |
| EMP_MANU_CAN | 6.930744e-08 | 3.220700e-05 |
| EMP_SALES_CAN | 5.603920e-08 | 2.331221e-05 |
| NHOUSE_P_CAN_lag1 | 5.064619e-08 | 2.363040e-05 |
| N_DUR_INV_RAT_new_lag1 | 4.977158e-08 | 2.203091e-05 |
| NHOUSE_P_CAN | 4.867879e-08 | 2.260460e-05 |
| EMP_CONS_CAN | 4.820477e-08 | 2.686181e-05 |
| MANU_INV_RAT_new_lag1 | 4.719125e-08 | 1.756391e-05 |
| MANU_INV_RAT_new | 4.496425e-08 | 1.453458e-05 |
| CLAIMS_CAN | 4.265923e-08 | 2.038640e-05 |
| EMP_CAN_lag2 | 4.104233e-08 | 2.418119e-05 |
| CLAIMS_CAN_lag1 | 4.089436e-08 | 2.128180e-05 |
| DUR_INV_RAT_new | 3.853610e-08 | 1.885238e-05 |
| GPI_new | 3.642233e-08 | 2.139188e-05 |
| UNEMP_DURA_1.4_CAN | 3.579388e-08 | 2.143674e-05 |
| EMP_PART_CAN | 3.542779e-08 | 2.008249e-05 |
| CRED_HOUS_cb | 3.313090e-08 | 1.576116e-05 |

DM_new 3.192399e-08 1.873230e-05
TBILL_6M.Bank_rate 2.957996e-08 1.602191e-05
N_DUR_INV_RAT_new 2.803628e-08 1.791151e-05
IP_new 2.801059e-08 1.745417e-05
DUR_INV_RAT_new_lag1 2.770593e-08 1.423598e-05
GDP_new 2.732127e-08 1.848944e-05
EMP_MANU_CAN_lag1 2.628121e-08 1.831201e-05
EMP_CAN_lag1 2.504282e-08 1.874967e-05
GOOD_HRS_CAN_lag1 2.500457e-08 1.420253e-05
EMP_CAN_lag3 2.396997e-08 1.568063e-05
BSI_new 2.320363e-08 1.749286e-05
TBILL_6M.Bank_rate_lag1 2.284626e-08 1.790315e-05
BSI_new_lag1 2.259943e-08 1.930003e-05
CRED_HOUS_non_MORT 1.983085e-08 1.464332e-05
G_AVG_3.5.Bank_rate 1.874042e-08 1.111963e-05
WT_new_lag1 1.836233e-08 1.611045e-05
EMP_CONS_CAN_lag1 1.795033e-08 1.215569e-05
UNEMP_DURA_5.13_CAN 1.675417e-08 1.322539e-05
EMP_CAN_lag4 1.646080e-08 1.463623e-05
CON_new_lag1 1.628620e-08 1.365333e-05
DM_new_lag1 1.625358e-08 1.692777e-05
NDM_new 1.515276e-08 1.051336e-05
G_AVG_3.5.Bank_rate_lag1 1.495016e-08 1.025091e-05
CPI_SERV_CAN_lag1 1.435813e-08 1.017992e-05
GPI_new_lag1 1.391073e-08 1.526285e-05
BANK_RATE_L_lag1 1.334741e-08 8.576872e-06
TBILL_6M 1.287901e-08 1.097726e-05
G_AVG_1.3.Bank_rate_lag1 1.244981e-08 1.578178e-05
CRED_HOUS_cb_lag1 1.239544e-08 1.305171e-05
GDP_new_lag1 1.229303e-08 1.842590e-05
SPI_new 1.223253e-08 1.305897e-05
DUR_TOT_INV_new_lag1 1.184142e-08 7.446757e-06
EMP_FOR_OIL_CAN 1.129161e-08 1.502239e-05
CPI_MINUS_FEN_CAN 1.094807e-08 9.252862e-06
IMP_TRANSP_BP_new 1.035101e-08 1.373568e-05
RES_TOT 1.022429e-08 9.233393e-06
GOOD_OVT_HRS_CAN 1.017566e-08 1.261455e-05
TSX_HI_lag1 9.710342e-09 9.413004e-06
WTISPLC_lag1 9.448897e-09 1.108156e-05
RT_new_lag1 9.393304e-09 1.107978e-05

MANU_UNFIL_new 9.186789e-09 9.142918e-06

G_AVG_5.10.Bank_rate_lag1 9.078427e-09 9.384055e-06

TBILL_6M_lag1 8.872106e-09 9.955313e-06

DUR_UNFIL_new 8.595704e-09 1.129645e-05

TBILL_3M_lag1 8.562940e-09 1.061906e-05

IMP_IND_EQUIP_BP_new 8.454947e-09 1.035263e-05

MANU_N_ORD_new 8.424239e-09 8.930393e-06

IOIL_BP_new_lag1 8.398961e-09 9.006873e-06

RES_USD 8.267039e-09 8.959242e-06

GOV_AVG_3_5Y_lag1 8.185493e-09 9.309173e-06

GOV_AVG_3_5Y 8.123201e-09 1.043257e-05

G_AVG_5.10.Bank_rate 8.036850e-09 1.035035e-05

DUR_TOT_INV_new 8.023884e-09 9.928857e-06

UNEMP_CAN_lag1 7.740924e-09 1.280252e-05

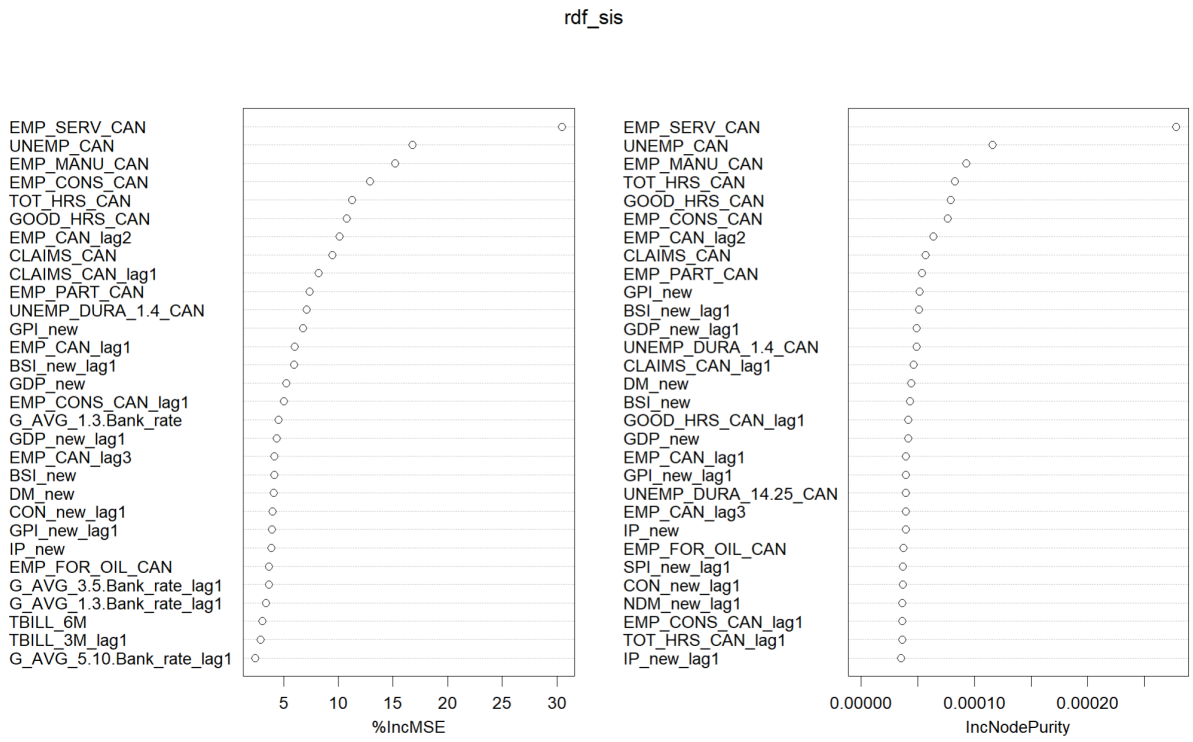
CPI_MINUS_FEN_CAN_lag1 7.656126e-09 1.449478e-05

CPI_ALL_CAN_lag1 7.607003e-09 1.212170e-05

FOR_SEC_NETFLOW 7.319203e-09 8.200658e-06

G_AVG_1.3.Bank_rate 7.243208e-09 1.383636e-05

## 2. Random forest avec réduction SIS




```
rdf_sis$importance[order(rdf_sis$importance[, 1], decreasing = TRUE), ]
```

%IncMSE IncNodePurity

EMP_SERV_CAN 9.971610e-07 2.776071e-04
UNEMP_CAN 2.753239e-07 1.157696e-04
EMP_MANU_CAN 1.750505e-07 9.284485e-05
GOOD_HRS_CAN 1.520871e-07 7.892931e-05
EMP_CONS_CAN 1.376415e-07 7.644816e-05
TOT_HRS_CAN 1.343025e-07 8.258805e-05
CLAIMS_CAN 9.774274e-08 5.695726e-05
EMP_CAN_lag2 9.380824e-08 6.374788e-05
CLAIMS_CAN_lag1 6.681367e-08 4.652865e-05
UNEMP_DURA_1.4_CAN 6.226522e-08 4.906204e-05
EMP_PART_CAN 5.830403e-08 5.393537e-05
GPI_new 5.775576e-08 5.177744e-05
BSI_new_lag1 4.849609e-08 5.131811e-05
EMP_CAN_lag1 3.978492e-08 3.989284e-05
GDP_new 3.916339e-08 4.169021e-05
GDP_new_lag1 3.826228e-08 4.939986e-05
G_AVG_1.3.Bank_rate 3.599069e-08 3.469354e-05
EMP_CONS_CAN_lag1 3.232251e-08 3.639012e-05
IP_new 3.189595e-08 3.958671e-05
BSI_new 3.118433e-08 4.317348e-05
DM_new 3.084112e-08 4.430816e-05
EMP_CAN_lag3 2.761476e-08 3.964336e-05
G_AVG_1.3.Bank_rate_lag1 2.634173e-08 3.405588e-05
GPI_new_lag1 2.619867e-08 3.971656e-05
CON_new_lag1 2.571409e-08 3.679669e-05
G_AVG_3.5.Bank_rate_lag1 2.365627e-08 2.491145e-05
EMP_FOR_OIL_CAN 2.344067e-08 3.735983e-05
TBILL_6M 1.953435e-08 3.267004e-05
GOOD_HRS_CAN_lag1 1.642163e-08 4.192009e-05
G_AVG_5.10.Bank_rate_lag1 1.613818e-08 2.412953e-05
TBILL_3M_lag1 1.521541e-08 2.820360e-05
UNEMP_DURA_14.25_CAN 1.513147e-08 3.969142e-05
NDM_new 1.502629e-08 2.621686e-05
NDM_new_lag1 1.205138e-08 3.651652e-05
TBILL_6M_lag1 1.138874e-08 2.320807e-05
SPI_new_lag1 1.128559e-08 3.699592e-05
GOOD_OVT_HRS_CAN 1.122480e-08 3.235204e-05
EMP_PART_CAN_lag1 9.783580e-09 3.127044e-05
TOT_HRS_CAN_lag1 9.730514e-09 3.630299e-05
G_AVG_5.10.Bank_rate 9.728718e-09 2.347383e-05

DUR_TOT_INV_new 9.250961e-09 2.844885e-05
IP_new_lag1 8.690763e-09 3.554675e-05
CRED_BUS 8.191027e-09 2.204107e-05
EMP_FIN_CAN_lag1 6.745009e-09 2.368510e-05
EMP_FIN_CAN 6.725834e-09 2.262251e-05
CRED_T_cb 5.023782e-09 2.539850e-05
MANU_TOT_INV_new 4.841781e-09 2.723577e-05
IPPI_MACH_CAN_lag1 2.915177e-09 2.283638e-05
DUR_TOT_INV_new_lag1 2.758742e-09 2.173823e-05
IOIL_BP_new_lag1 2.568053e-09 2.369296e-05
TSX_CLO 2.416736e-09 2.057632e-05
MANU_TOT_INV_new_lag1 1.436485e-09 2.365780e-05
OILP_new_lag1 1.076290e-09 2.341087e-05
RES_IMF_lag1 9.090321e-10 1.826164e-05
EX_IND_EQUIP_BP_new_lag1 7.553403e-10 2.452483e-05
OIL_CAN_new_lag1 2.926780e-10 1.904272e-05
RT_new_lag1 -3.082107e-10 3.047958e-05
EMP_FOR_OIL_CAN_lag1 -1.611316e-09 2.156856e-05
CRED_BUS_cb -2.055685e-09 1.876717e-05
UNEMP_DURA_1.4_CAN_lag1 -2.221976e-09 2.048161e-05
IOIL_BP_new -4.210443e-09 2.107524e-05
TSX_CLO_lag1 -4.974402e-09 2.397876e-05
GOOD_OVT_HRS_CAN_lag1 -5.041244e-09 2.524832e-05
RES_IMF -6.911487e-09 1.971712e-05
IMP_IND_EQUIP_BP_new_lag1 -7.520672e-09 2.350553e-05

# 6. Table des matières