

Master 2 Econométrie et Statistique, parcours Econométrie
Appliquée

Modélisation et analyse de la croissance du PIB du Canada entre 1981 et 2024

Onno Lilou
Dahmani Amel

April 20, 2025

Régressions pénalisées et sélection de variables

Résumé

Ce dossier analyse l'évolution du Produit Intérieur Brut canadien de janvier 1981 à juillet 2024, à partir de données mensuelles comprenant 523 observations et 410 variables explicatives. Il met en lumière les moteurs de croissance ainsi que l'impact des événements économiques majeurs, tels que la crise financière de 2008 et la pandémie de la Covid-19. En utilisant des techniques de modélisation avancées, telles que les régressions pénalisées Ridge, Lasso, Elastic Net, ainsi que les forêts aléatoires avec ou sans sélection de variables par SIS, cette étude explore les interactions entre divers facteurs influençant la croissance du PIB. Chacune de ces méthodes a présenté des spécificités, en se basant uniquement sur le nombre de variables sélectionnées et les paramètres de pénalisation. Les résultats montrent que les méthodes de Machine Learning, en particulier Lasso et Adaptive Lasso, sont les plus sélectives, même après l'application de SIS, mettant en évidence l'importance des variables liées à la production industrielle dans l'analyse du PIB et offrant ainsi une compréhension approfondie des dynamiques économiques.

Mots clés : Produit Intérieur Brut, Sélection de variables, Réduction de dimension, Régressions pénalisées, Canada, R.

Abstract

This study analyzes the evolution of Canada's Gross Domestic Product from January 1981 to July 2024, based on monthly data comprising 523 observations and 410 explanatory variables. It highlights the drivers of growth as well as the impact of major economic events, such as the 2008 financial crisis and the COVID-19 pandemic. Using advanced modeling techniques, including penalized regressions such as Ridge, Lasso, Elastic Net, and random forests with or without variable selection through SIS, this study explores the interactions between various factors influencing GDP growth. Each of these methods exhibits specific characteristics, making it complex to compare them based solely on the number of selected variables and penalization parameters. The results show that Machine Learning methods, particularly Lasso and Adaptive Lasso, are the most selective, even after applying SIS, underscoring the importance of variables related to industrial production in GDP analysis and thus providing a deeper understanding of Canada's economic dynamics.

Keywords: Gross Domestic Product, Variable Selection, Dimension Reduction, Penalized Regressions, Canada, R.

Sommaire

Contents

1 Analyse exploratoire et descriptive	8
1.1 Analyse des valeurs manquantes	8
1.2 Analyse des outliers	10
1.3 Analyse de la stationnarité	12
1.4 Graphique de la variable à expliquer	13
1.5 Statistiques descriptives	14
1.6 Classification	16
1.7 Corrélation	21
1.7.1 Corrélations avec la variable d'intérêt	21
1.7.2 Corrélation entre les variables explicatives	22
2 Sélection des variables	23
2.1 Approche économétrique : GETS	23
2.2 Régressions pénalisées	24
2.2.1 Ridge	24
2.2.2 Lasso	25
2.2.3 Elastic-net	26
2.2.4 SCAD	27
2.2.5 Adaptive Lasso	28
2.2.6 Récapitulatif des résultats des différentes méthodes de régularisation	29
2.3 Approche de réduction de dimension	30
2.3.1 GETS	30
2.3.2 Ridge	31
2.3.3 Lasso	31
2.3.4 Elastic-net	32
2.3.5 SCAD	32

2.3.6	Adaptive Lasso	33
2.3.7	Récapitulatif des résultats des différentes méthodes de régularisation avec réduction de dimension	33
2.4	Approche complémentaire : Random forest	34
2.4.1	Random forest sans le filtrage SIS	34
2.4.2	Random forest avec le filtrage SIS	35
2.5	Comparaison des méthodes	37
2.5.1	Avant filtrage SIS	37
2.5.2	Après filtrage SIS	39
3	Annexe	44

Liste d'abréviations et d'acronymes

PIB	Produit Intérieur Brut
ACP	Analyse en Composantes Principales
ADF	Dickey-Fuller augmenté
KPSS	Kwiatkowski-Phillips-Schmidt-Shin
PP	Phillips-Perron
GETS	General-to-Specific
SIS	Sure Independence Screening
LASSO	Least Absolute Shrinkage and Selection Operator
SCAD	Smoothly Clipped Absolute Deviation
EMP CAN	Emploi total au Canada
EMP SERV CAN	Emploi dans le secteur des services au Canada
BSI new	Indice de l'activité des petites entreprises
EMP ONT	Emploi en Ontario
TOT HRS CAN	Heures totales travaillées au Canada
SPI new	Indice des prix de détail
EMP QC	Emploi au Québec
GOOD HRS CAN	Heures de travail dans le secteur manufacturier au Canada
DM new	Mesure de la demande
UNEMP CAN	Taux de chômage au Canada

Introduction

Le Produit Intérieur Brut est l'un des indicateurs économiques les plus fondamentaux pour évaluer la santé et la dynamique économique d'un pays. En mesurant la valeur totale des biens et services produits sur un territoire donné au cours d'une période définie, il permet de suivre la richesse générée et d'observer la trajectoire économique d'un pays à travers le temps.

Depuis 1870, l'économie canadienne a connu une expansion remarquable, avec un PIB qui est passé d'environ 383 millions de dollars courants à 2,2 trillions de dollars en 2022. Cette trajectoire de croissance reflète non seulement la capacité du Canada à se diversifier et à s'adapter aux conjonctures mondiales, mais aussi l'impact de cycles d'expansion et de récession influencés par des événements économiques majeurs, tels que la crise financière mondiale de 2008, la pandémie de Covid-19, et plus récemment les tensions géopolitiques dues à la guerre en Ukraine, qui ont exacerbé les pressions inflationnistes. Classé parmi les dix plus grandes économies mondiales, le Canada s'impose grâce à ses vastes ressources naturelles, sa compétitivité en matière d'exportations et sa résilience aux crises. En tant que cinquième exportateur mondial de ressources naturelles, il occupe une position stratégique sur la scène économique internationale non négligeable.

Pour analyser en profondeur les déterminants de cette dynamique complexe, nous concentrerons notre étude sur un vaste ensemble de données mensuelles quantitatives de type Big Data, comprenant 523 observations et 410 variables, qui présentent les facteurs de croissance du PIB canadien entre janvier 1981 et juillet 2024. L'objectif est de saisir les interactions entre ces facteurs à l'aide de techniques de régression pénalisée, afin d'identifier les leviers économiques essentiels ayant un impact direct sur la croissance. Cette approche prend également en compte les spécificités politiques, économiques et conjoncturelles propres au Canada.

Notre démarche débutera par une analyse exploratoire des données, incluant la vérification et le traitement des valeurs manquantes, l'identification des valeurs aberrantes, ainsi qu'une évaluation de la stationnarité des séries temporelles et des statistiques descriptives. Cette étape permettra d'établir un diagnostic premier des données et d'orienter les étapes suivantes de la modélisation. Ensuite, nous mettrons en œuvre le modèle GETS pour raffiner notre modèle économétrique en appliquant des techniques de sélection de variables, telles que la régression pénalisée par SIS, qui permettra de réduire la complexité des modèles tout en renforçant leur robustesse. Pour compléter cette approche, un modèle de forêt aléatoire sera appliqué afin d'explorer les interactions entre nos variables de manière non paramétrique, avec et sans sélection de variables, offrant ainsi une perspective alternative et enrichie sur la croissance du PIB.

En combinant des méthodes statistiques classiques et des approches de Machine Learning, cette étude vise à fournir une analyse exhaustive de la croissance économique canadienne, à mettre en lumière les fondements structurels du PIB, et à dégager des perspectives pour l'économie future du pays.

1 Analyse exploratoire et descriptive

Nous débuterons notre analyse exploratoire et descriptive sur un jeu de données dont la variable dépendante est le PIB mensuel du Canada, couvrant la période de janvier 1981 à juillet 2024. Cette série temporelle de fréquence mensuelle comprend un ensemble de 523 observations et 410 variables explicatives quantitatives, représentant divers déterminants économiques. Les données, stationnaires équilibrées, ont été recueillies via le *site Statistique Canada*¹, qui les a obtenues de l’École des sciences de la gestion de l’Université du Québec à Montréal. Ce large éventail de variables couvre de nombreux aspects de l’économie canadienne, offrant ainsi un socle solide pour notre analyse.

Dans le cadre de notre étude, nous commencerons l’exploration de la base de données par la détection des valeurs manquantes au sein de notre variable d’intérêt, suivie de l’identification des points atypiques. Une fois cette étape accomplie, nous examinerons la stationnarité de la série, puis procéderons à des analyses statistiques descriptives pour observer les principales caractéristiques de nos données. Cette partie se conclura par une analyse en composantes principales, permettant de réduire la dimensionnalité et d’identifier les principales sources de variation, ainsi que par l’étude des relations entre les différentes variables de notre étude.

1.1 Analyse des valeurs manquantes

Dans cette première phase d’analyse de la base de données, il est primordial de détecter les valeurs manquantes, dont la présence est probable étant donné l’ampleur du jeu de données. La gestion de ces valeurs est essentielle pour garantir la fiabilité et la qualité des résultats, puisque des valeurs manquantes non traitées peuvent introduire des biais dans les analyses, tout particulièrement lorsqu’il est question de séries temporelles. Pour repérer ces valeurs, nous avons donc utilisé la fonction `colSums()` de R, qui permet de calculer le nombre de valeurs manquantes par variable.

¹Données de l’Université du Québec à Montréal, Canada.

Cette méthode nous a permis d'identifier des valeurs manquantes pour cinq variables explicatives spécifiques, détaillées dans le tableau ci-dessous. En agissant de la sorte, nous garantissons un contrôle rigoureux des données avant de poursuivre l'analyse.

Variable	Description	Valeurs manquantes
CRED_T_discontinued	Crédits totaux	46
CRED_HOUS_discontinued	Crédits aux ménages	46
CRED_MORT_discontinued	Crédits hypothécaire	46
CRED_CONS_discontinued	Crédits à la consommation	46
CRE_BUS_discontinued	Crédits aux entreprises	46

Table 1: Récapitulatif des valeurs manquantes par variables

La *Table n°1* présente cinq variables de crédit qui contiennent des valeurs manquantes dans notre jeu de données. Il s'agit de cinq indicateurs économiques : les crédits totaux, les crédits aux ménages, les crédits hypothécaires, les crédits à la consommation ainsi que les crédits aux entreprises. Chacune de ces variables présente exactement 46 valeurs manquantes, ce qui représente un total de 230 valeurs manquantes pour l'ensemble de la base. Cette absence de données, homogène entre les variables, pourrait indiquer des lacunes pour des périodes spécifiques dans les archives de données de crédit. Afin de garantir la qualité de nos analyses et la cohérence des résultats, nous avons appliqué un lissage de Kalman spécifiquement pour combler ces lacunes. Ce dernier est particulièrement adapté aux données temporelles puisqu'il permet d'estimer les valeurs manquantes en se basant sur les observations disponibles, tout en tenant compte des dynamiques sous-jacentes de la série.

Cette méthode a permis de reconstituer les valeurs manquantes de manière précise et cohérente, assurant ainsi une continuité temporelle indispensable. Après vérification avec la fonction *colSums()*, les données lissées respectent bien les tendances initiales des séries, ce qui nous fournit une base solide pour la suite de l'analyse.

1.2 Analyse des outliers

L'identification et la gestion des outliers sont des étapes cruciales pour garantir la robustesse et la fiabilité de nos résultats. Pour ce faire, nous avons entrepris une analyse approfondie visant à détecter la présence de points atypiques pour notre variable d'intérêt, le PIB canadien. En utilisant la fonction `tso()` du package `tsoutliers` dans R, nous avons pu non seulement identifier, mais également ajuster un total de 6 points atypiques significatifs. Ces points, qui peuvent influencer de manière disproportionnée les résultats de l'analyse, sont présentés dans le tableau suivant.

Type	Période	Coefhat	t-stat
AO	1982:10	-0.01422	-4.062
TC	2008:11	-0.01443	-5.759
TC	2020:03	-0.07064	-20.591
AO	2020:04	-0.06448	-14.809
TC	2020:05	0.07415	22.485
AO	2020:06	0.03180	7.792

Table 2: Points atypiques de la série mensuelle PIB canadien

La *Table n°2* présente l'ensemble des outliers détectés au sein de la série du Produit Intérieur Brut canadien. Nous identifions 6 points atypiques, dont 3 de type Additive Outliers (AO), qui impactent une seule observation à un moment donné, et 3 autres de type Transitory Change (TC), qui affectent temporairement la série. Ces derniers peuvent être visualisés graphiquement dans l'*Annexe n°1*.

La majorité de ces points atypiques ont été observés en 2020, ainsi qu'en 2008 et 1982. L'année 2020 est particulièrement notable en raison des répercussions économiques considérables engendrées par la pandémie de la *COVID-19*², qui a bouleversé les activités économiques à l'échelle mondiale. La pandémie a provoqué des confinements stricts, des fermetures d'entreprises, et une interruption des chaînes d'approvisionnement, entraînant une chute dramatique de la consommation et une

²"COVID-19 : impacts et réponses au Canada", site de l'Organisation internationale de la Francophonie, 2020.

augmentation du chômage. Les mesures de distanciation sociale ont également eu un impact profond sur les secteurs du tourisme, de l'hôtellerie et des loisirs, exacerbant ainsi les perturbations économiques, touchant notamment sur le PIB.

En ce qui concerne l'année 2008, l'un des outliers pourrait être attribué à la *crise des subprimes*³, qui a déclenché une crise financière majeure. Cette crise a été provoquée par l'effondrement du marché immobilier américain, caractérisé par des prêts hypothécaires à risque qui ont abouti à des défauts massifs. Les institutions financières, exposées à des actifs toxiques, ont subi d'énormes pertes, ce qui a conduit à une contraction du crédit et à une baisse de la confiance des investisseurs. En conséquence, les marchés boursiers ont chuté, entraînant une récession mondiale qui a touché de nombreux pays, y compris le Canada, avec une augmentation du chômage et une baisse de la consommation qui en partie cause une baisse du PIB.

L'outlier identifié en octobre 1982 peut quant à lui être expliqué par la *situation économique instable*⁴ résultant d'une récession qui a sévi de juillet 1981 à novembre 1982. Cette période a été marquée par une forte inflation, souvent désignée sous le terme de "stagflation", où l'inflation élevée s'est accompagnée d'une stagnation économique et d'un chômage croissant. Les hausses des taux d'intérêt mises en place par la Banque du Canada pour lutter contre l'inflation ont aggravé la situation, entraînant un ralentissement de l'investissement et de la consommation. Ce contexte a conduit à des fluctuations économiques significatives, générant des anomalies dans les données du PIB.⁵

Cette analyse nous a permis d'ajuster les points atypiques dans notre variable, assurant ainsi l'exactitude et la fiabilité de nos résultats. Cette approche rigoureuse renforce notre compréhension des fluctuations du PIB canadien.

³"Le Canada et la crise financière : un retour critique sur la thèse de l'exception canadienne" dans le volume 39, numéro 2 de la revue, publié en 2020 sur Erudit.

⁴"Pourquoi le Canada est-il si touché par la récession ?" par Pierre Paquette sur le site des Classiques des sciences sociales, 2005.

⁵"La récession de 1981-1982" sur le site de l'Histoire de la Réserve fédérale, 2020.

1.3 Analyse de la stationnarité

Avant d'analyser la stationnarité, il est essentiel de noter que toutes les variables de notre jeu de données ont été différenciées pour assurer leur stationnarité. Notre variable dépendante, le PIB canadien, a également été soumise à une transformation logarithmique avant cette différenciation, permettant ainsi une meilleure capture des variations relatives. Nous avons évalué la stationnarité à l'aide de trois tests statistiques : le test de Dickey-Fuller augmenté via la fonction `adf.test()` du package `tseries` ; le test de Kwiatkowski-Phillips-Schmidt-Shin, réalisé avec `kpss.test()` ; et le test de Phillips-Perron avec `pp.test()`.

Test	p-value
Augmented Dickey-Fuller	0.01
Kwiatkowski-Phillips-Schmidt-Shin	0.1
Phillips-Perron	0.01

Table 3: Résultats des tests de stationnarité

Le *tableau n°3* présente les résultats des tests de stationnarité, illustrés par leurs p-values respectives. Pour le test ADF, l'hypothèse nulle stipule que la série n'est pas stationnaire. Une p-value inférieure à 0,05 nous permet de rejeter H0. Dans notre analyse, la p-value observée est de 0,01, ce qui nous permet de conclure avec confiance que la variable dépendante est stationnaire. En ce qui concerne le test de KPSS, l'hypothèse nulle affirme que la série est stationnaire autour d'une tendance déterministe. Dans ce cas, une p-value supérieure à 0,05 soutient cette hypothèse. La p-value est de 0,1, ce qui nous conduit à accepter l'hypothèse nulle et à conclure que la série est stationnaire autour d'une tendance déterministe. Pour le test PP, l'hypothèse nulle stipule que la série n'est pas stationnaire, et une p-value supérieure à 0,05 confirme cette hypothèse. Dans notre cas, la p-value est également de 0,01, permettant ainsi de rejeter H0 et de dire que la série est stationnaire.

Les résultats convergents des trois tests, corroborés par l'analyse du corrélogramme présenté en *annexe n°3*, renforcent notre conclusion sur la stationnarité de la série temporelle étudiée. Ces analyses approfondies constituent une base solide.

1.4 Graphique de la variable à expliquer

Il est désormais essentiel de visualiser graphiquement l'évolution de notre variable d'intérêt, le taux de croissance du PIB canadien, après la transformation des données, afin de guider nos analyses ultérieures.

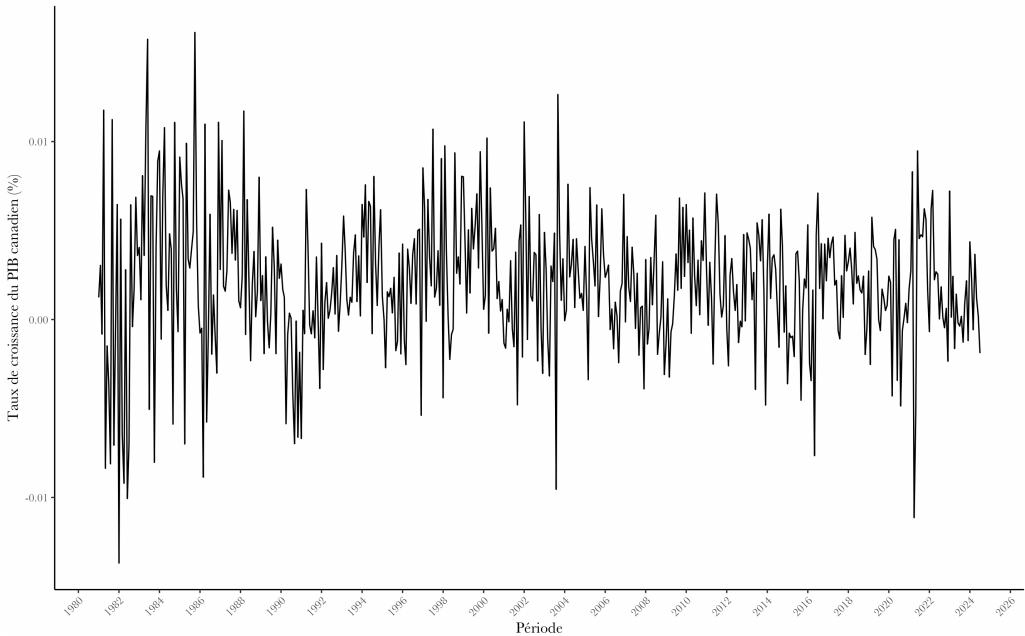


Figure 1: Taux de croissance du PIB Canadien

La *figure n°1* illustre la stationnarité du produit intérieur brut du Canada, en mettant en lumière le taux de croissance de cette variable après l'application d'une différence logarithmique. Cette approche est essentielle pour interpréter les fluctuations économiques et approfondir notre compréhension des dynamiques sous-jacentes. Sur la période de 1981 à 2024, nous observons que le taux de croissance du PIB montre une tendance à revenir vers la moyenne, ce qui indique une stabilité à long terme. De plus, le lissage appliqué atténue les points atypiques précédemment identifiés, renforçant ainsi la clarté des tendances générales. Cette stationnarité suggère que les chocs économiques ont des effets temporaires, permettant une résilience du PIB face aux diverses perturbations.

1.5 Statistiques descriptives

Dans cette section, nous présenterons les statistiques descriptives de notre variable d'intérêt, visant à analyser en profondeur sa distribution. Cette analyse nous fournira un aperçu des caractéristiques essentielles des données, notamment les mesures de tendance centrale et de dispersion, permettant une compréhension plus nuancée.

Mesure	Valeur
Médiane	0.0019373
Moyenne	0.0020078
Variance	1.531663e-05
Écart-type	0.003913647
Skewness	-0.1389993
Kurtosis	1.33407
Minimum	-0.0136872
1er quartile	-0.00011572
3ème quartile	0.0042489
Maximum	0.0161308

Table 4: Statistiques descriptives du PIB canadien

La *Table n°4* indique une moyenne des données, à 0,002, très proche de la médiane, qui s'établit à 0,002, ce qui suggère une distribution relativement symétrique. La faible variance de 1,54e-05 indique que les valeurs sont étroitement regroupées autour de cette moyenne, tandis qu'un écart-type d'environ 0,0039 renforce cette idée en confirmant une faible variabilité. Par ailleurs, notre kurtosis, mesurée à 1,33, est inférieure à 3, ce qui suggère que la distribution est plus plate que celle d'une distribution normale, avec moins de valeurs extrêmes. Ces indicateurs montrent que les données sont généralement symétriques et peu dispersées, bien qu'une légère asymétrie puisse être perceptible, possiblement en raison des anomalies observées en 2020 et 1981, comme le souligne la *figure n°1*. La concentration des valeurs autour de la médiane, également proche de la moyenne, indique une distribution relativement normale et uniforme. Le fait que la moyenne soit proche de zéro est attribué à la transformation logarithmique appliquée pour rendre les données stationnaires.

L'analyse de la distribution met en lumière une relative symétrie, confirmée par les valeurs de la kurtosis et de la skewness. La skewness, évaluée à -0,1389, révèle une légère asymétrie négative, signalant que la queue de la distribution s'étend légèrement vers la gauche de la moyenne. Cela suggère la présence de quelques valeurs inférieures à la moyenne, potentiellement influencées par des événements exceptionnels. En outre, l'écart-type, qui indique une dispersion faible autour de la moyenne, renforce l'idée que les valeurs sont étroitement regroupées, ce qui est favorable pour l'interprétation des résultats. La kurtosis, mesurée à 1,33407, révèle une distribution légèrement leptokurtique, caractérisée par une forme plutôt pointue et des queues plus épaisses que celles de la distribution normale. Cette observation suggère une prévalence de valeurs extrêmes, comme le confirment les graphiques de l'histogramme et de la boîte à moustaches présentés au sein des annexes n°4 et n°5.

Tests de Normalité

Des tests de normalité ont été réalisés pour déterminer si nos données suivent une distribution normale, condition essentielle à la validité de nos analyses antérieures.

Test	P-value
Jarque-Bera	1.018e-09
Shapiro-Wilk	9.182e-06

Table 5: Résultats des tests de normalité

Le test de Jarque-Bera évalue la normalité en analysant la skewness et la kurtosis, fournissant ainsi une mesure globale de l'asymétrie et de l'aplatissement des données. En revanche, le test de Shapiro-Wilk, bien qu'efficace pour tout type d'échantillon, est particulièrement sensible aux petites tailles d'échantillon.

Les résultats des tests de normalité montrent des p-values extrêmement faibles : 1,018e-09 pour le test de Jarque-Bera et 9,182e-06 pour le test de Shapiro-Wilk. Ces valeurs, bien inférieures au seuil critique de 0,05, entraînent le rejet de l'hypothèse nulle selon laquelle les données suivent une distribution normale. Les deux tests concordent, confirmant que les données ne suivent pas une distribution normale.

1.6 Classification

Étant donné les 410 variables explicatives présentes dans notre jeu de données, nous avons opté pour une Analyse en Composantes Principales. Afin d'obtenir un aperçu des variables les plus contributives. La *Figure n°2* ci-dessous présente les résultats de cette ACP, offrant une représentation complète de l'ensemble des variables et mettant en lumière celles qui influencent le plus la variance des données.

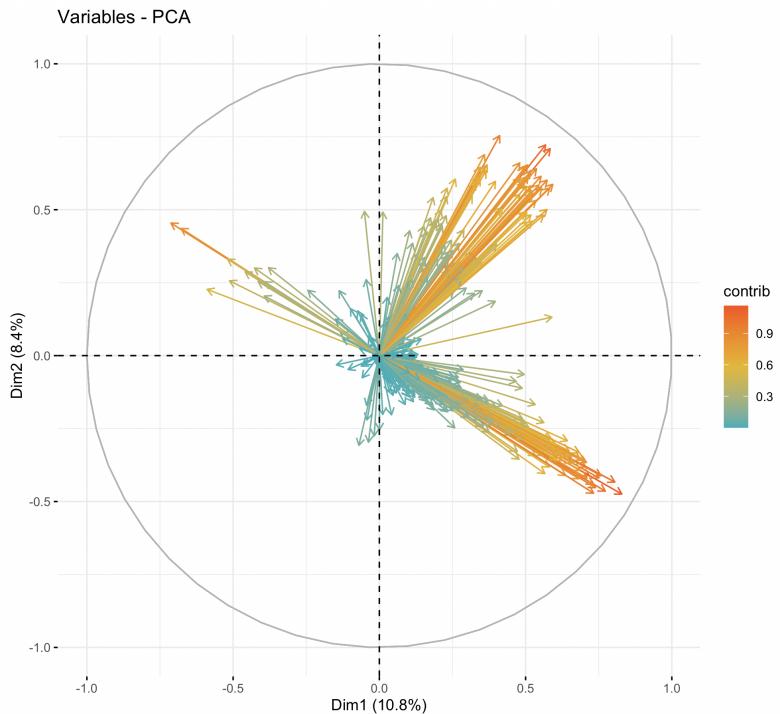


Figure 2: ACP avec l'ensemble des variables explicatives en dimension 1-2

Cette figure illustre que nos variables se regroupent principalement en deux à trois clusters distincts le long des axes 1, 2 et 4. L'inertie totale des deux premières composantes, qui atteint seulement 19,2, indique une proportion relativement faible de la variance expliquée, ce qui peut être attribué à la grande quantité de variables présentes et à la forte corrélation potentielle entre certaines d'entre elles. Par ailleurs, comme l'indique l'*annexe n°7*, nous observons une diminution continue de

l'inertie à travers les axes, suggérant que les dimensions supplémentaires apportent peu d'informations nouvelles. Cela implique que la majorité de la variance des données est capturée par les premières composantes, ce qui rend l'analyse à la fois plus simple et plus efficace.

Nous poursuivons notre étude en nous focalisant sur les 10 variables les plus contributives aux axes 1 et 2. Les deux tableaux présentés ci-dessous mettent en évidence ces variables clés afin d'identifier celles qui influencent le plus notre modèle.

Variable	Contribution
EMP_CAN	1.559701
EMP_SERV_CAN	1.471337
BSI_new	1.354451
EMP_ONT	1.304070
TOT_HRS_CAN	1.279624
SPI_new	1.239133
EMP_QC	1.217533
GOOD_HRS_CAN	1.213964
DM_new	1.169025
UNEMP_CAN	1.152471

Table 6: Dix variables les plus contributives pour l'Axe 1

L'analyse des dix variables les plus contributives à l'axe 1 met en lumière leur forte influence sur la croissance du PIB canadien. Cet axe est principalement déterminé par des variables liées aux taux d'intérêt, qui sont cruciales pour la politique monétaire : des taux bas favorisent l'emprunt et l'investissement, stimulant ainsi la croissance économique. Par ailleurs, la présence de variables relatives aux heures supplémentaires travaillées dans le secteur de la production de biens indique une demande accrue et la nécessité d'augmenter la production. De plus, les taux de prêt hypothécaire à différentes échéances suggèrent que des conditions favorables peuvent dynamiser le marché immobilier, incitant à l'achat de logements et contribuant positivement à l'économie. Ainsi, l'axe 1 illustre comment ces facteurs combinés influencent la performance économique du Canada, où les conditions du marché du travail et le

bien-être des ménages se révèlent comme des déterminants essentiels du PIB.

Variable	Contribution
CPI_MINUS_FEN_CAN	1.658253
CPI_ALL_CAN	1.521078
CPI_MINUS_FOO_CAN	1.468090
CPI_MINUS_FEN_ONT	1.384623
CPI_ALL_QC	1.272056
CPI_MINUS_FEN_MAN	1.246289
CPI_MINUS_FOO_QC	1.245151
CPI_MINUS_FEN_ALB	1.220778
CPI_MINUS_FEN_QC	1.215811
CPI_ALL_ONT	1.212247

Table 7: Dix variables les plus contributives pour l’Axe 2

L’examen des dix variables les plus contributives à l’axe 2 met en évidence les indices des prix à la consommation (IPC) dans différentes régions du Canada, soulignant l’importance de l’inflation dans l’analyse économique. Par exemple, les variables telles que l’IPC national hors énergie et l’IPC global pour le Canada montrent que les fluctuations des prix, même en excluant certains secteurs, constituent des facteurs déterminants pour l’économie. La contribution élevée d’indices régionaux, comme CPI_ONTARIO et CPI_QUEBEC, indique que les variations de prix dans des provinces majeures telles que l’Ontario et le Québec peuvent avoir un impact significatif sur les comportements de consommation et les décisions économiques. Par exemple, une inflation plus élevée peut réduire le pouvoir d’achat et peser sur la demande globale. L’axe 2 semble ainsi représenter comment l’inflation, mesurée à travers divers IPC, influence la dynamique économique et la croissance du PIB.

Il est désormais pertinent de visualiser la projection des dix variables les plus contributives de l’ACP sur le plan formé par les dimensions 1 et 2.

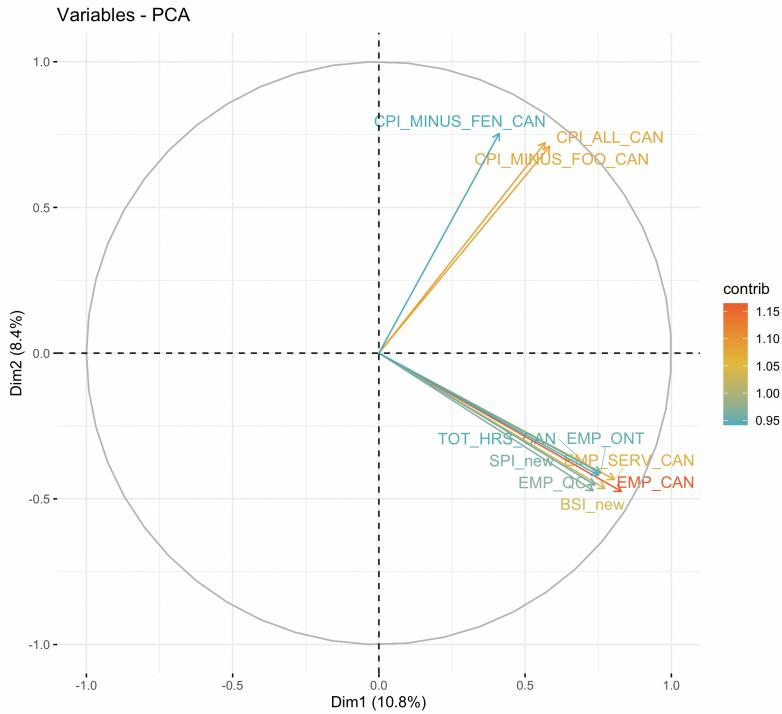


Figure 3: ACP avec focus sur les 10 variables les plus contributives

L’interprétation de l’ACP à partir des 10 variables les plus contributives aux dimensions 1 et 2 met en lumière des relations intéressantes entre l’inflation et l’activité économique au Canada. La corrélation observée entre ces indices de prix et les mesures d’activité économique suggère que des augmentations de l’emploi et des heures travaillées sont souvent accompagnées d’une pression inflationniste, témoignant d’une demande soutenue sur le marché.

Nous avons effectué une classification des variables explicatives du jeu de données, qui a révélé une répartition inégale des variables en trois clusters distincts. Cette approche nous permet d’identifier des caractéristiques spécifiques au sein de chaque groupe et d’explorer les relations potentielles entre les variables.

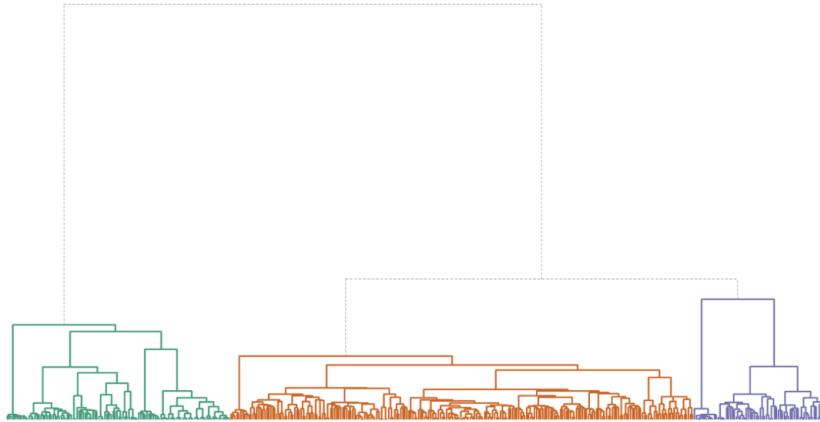


Figure 4: Classification en 3 classes de la série

Nous avons effectué une classification des variables explicatives de notre jeu de données, identifiant ainsi trois groupes distincts, reflétant les structures observées dans l'ACP intégrant toutes les variables. Le tableau ci-dessous présente le nombre de variables attribuées à chaque cluster par le modèle, offrant une vue d'ensemble de leur répartition au sein de ces catégories.

Cluster	Nombre	%	val%
Cluster 1	232	56.6	56.6
Cluster 2	66	16.1	16.1
Cluster 3	112	27.3	27.3

Table 8: Répartition des variables par cluster

Les données montrent une répartition non homogène en trois clusters : Cluster 1, avec 56.6% des variables, est dominant et pourrait indiquer une structure principale dans les données. Cluster 2 est plus restreint, ne représentant que 16.1%, ce qui suggère des caractéristiques plus spécifiques ou moins fréquentes. Enfin, Cluster 3 occupe une position intermédiaire avec 27.3% des variables, faisant potentiellement le lien entre les deux autres clusters. Cette segmentation met en lumière des sous-groupes aux caractéristiques distinctes, pouvant orienter des analyses plus ciblées.

1.7 Corrélation

Pour enrichir notre analyse exploratoire, nous allons approfondir l'étude des relations entre la variable cible et les variables explicatives, en nous concentrant sur les corrélations les plus significatives. Cette étape finale permettra de mieux comprendre les dynamiques sous-jacentes, même si toutes les variables seront incluses dans les analyses ultérieures avec les méthodes de régression pénalisée.

1.7.1 Corrélations avec la variable d'intérêt

Nous analysons les corrélations entre les variables explicatives et la variable dépendante afin d'identifier les relations significatives. Grâce au coefficient de Spearman, nous repérons les variables exerçant une influence notable sur la variable cible, pour mieux cerner les facteurs clés dans les analyses ultérieures.

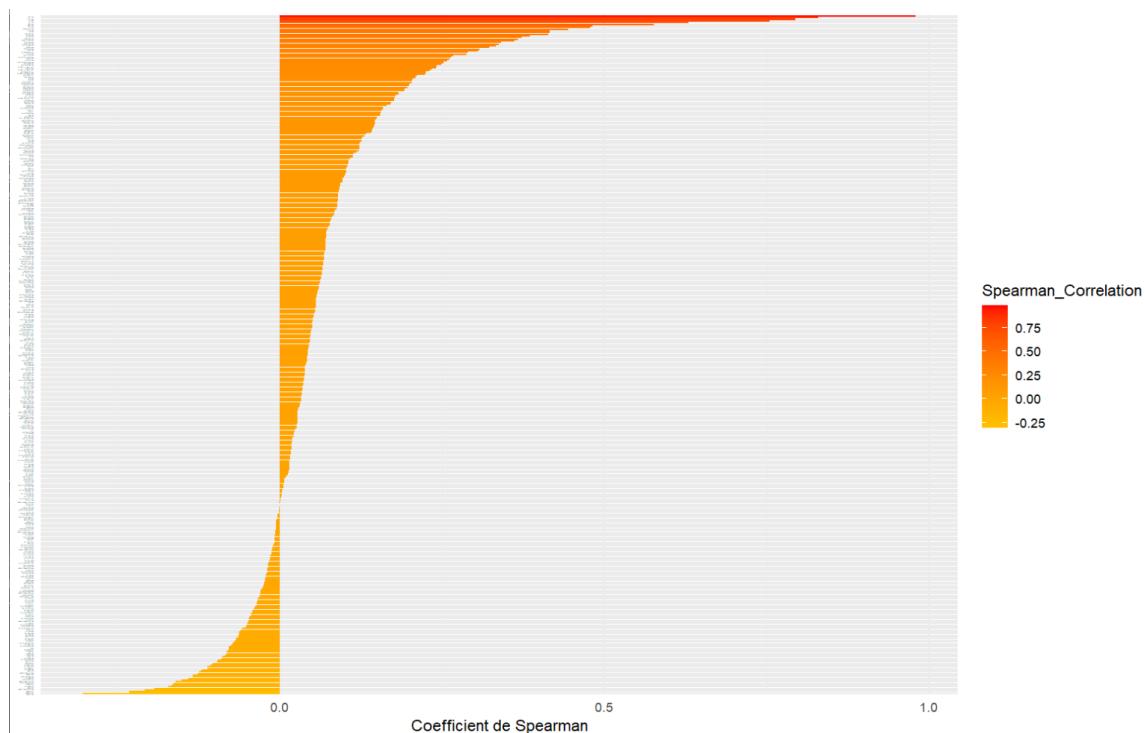


Figure 5: Corrélations avec le coefficient de Spearman

La *figure n°5* met en évidence des corrélations supérieures à 0,5 pour six de nos variables : le PIB des affaires BSI new à 0,93, le PIB des biens GPI new à 0,79, le PIB de la production industrielle IP new à 0,75, le PIB des services SPI new à 0,70, le PIB des biens durables DM new à 0,59 ainsi que le PIB des biens non durables NDM new à 0,53. Ces corrélations élevées soulèvent des préoccupations de multicolinéarité, susceptibles de compliquer les analyses. Néanmoins, les méthodes statistiques choisies sont adaptées pour atténuer l'impact de ce phénomène sur les résultats. Par ailleurs, la majorité des autres corrélations sont faibles, probablement en raison du nombre élevé de variables. Cela ne diminue cependant en rien leur potentiel impact sur la variable dépendante, le PIB canadien.

1.7.2 Corrélation entre les variables explicatives

Étant donné le nombre élevé de variables dans notre jeu de données, l'analyse des corrélations entre les variables explicatives est complexe en raison de la richesse des informations qu'elles contiennent. Néanmoins, cette étape reste essentielle pour repérer les relations significatives et évaluer la présence de multicolinéarité, offrant ainsi une première vue d'ensemble des interactions entre variables.

Notre analyse, réalisée à partir d'une matrice de corrélation de Spearman, a mis en évidence que 182 variables affichent des corrélations supérieures à 0,8. Cependant, il convient d'interpréter ce résultat avec prudence, car il reflète davantage des dynamiques temporelles que des corrélations causales significatives.

Pour clore cette section, les régressions pénalisées que nous appliquerons par la suite sont conçues pour gérer ces corrélations en éliminant les variables redondantes, permettant une analyse plus robuste des relations complexes. Cependant, une limite de la régression Lasso est son impact potentiel sur l'exclusion de certaines variables pertinentes, ce qui nécessitera une attention particulière dans l'interprétation des résultats.

2 Sélection des variables

Dans la seconde section de notre étude, nous modéliserons le taux de croissance du produit intérieur brut canadien. Étant donné l'ampleur de notre jeu de données, qui comprend 818 variables explicatives, dont 409 sont retardées et 409 non retardées, il est crucial de sélectionner les variables avec soin. Nous avons opté pour l'inclusion de quatre retards pour la variable dépendante et un retard pour les variables explicatives, afin de mieux capturer l'impact de l'historique sur la dynamique actuelle. Cette approche lisse les tendances temporelles tout en tenant compte des dépendances potentielles entre les variables. Ainsi, dans cette section, nous appliquerons plusieurs méthodes d'estimation, avec et sans filtrage des variables, et comparerons les résultats obtenus à l'aide de régressions pénalisées. Ces techniques permettront de gérer la multicolinéarité et d'identifier les variables les plus significatives.

2.1 Approche économétrique : GETS

La méthode GETS a été utilisée pour sélectionner les variables les plus pertinentes pour notre modèle. Cette technique de modélisation économétrique commence par une spécification générale intégrant toutes les variables disponibles, puis se simplifie progressivement en éliminant celles jugées peu significatives. Elle repose sur des tests statistiques et des critères de sélection visant à obtenir une structure plus simple et précise, sans variables superflues. Cependant, étant donné que notre jeu de données comporte plus de 800 variables explicatives, l'application de la méthode GETS s'est révélée inadaptée et complexe dans ce contexte. La gestion d'un nombre aussi élevé de variables nécessitait des approches plus robustes pour traiter la multicolinéarité et sélectionner efficacement les variables pertinentes.

Pour ce faire, nous appliquerons dans un premier temps les régressions pénalisées sans selection de variable, dans un second temps nous appliquerons l'approche SIS sur cette même methode afin de réduire la dimensionnalité du jeu de données. Cette demarche permettra de filtrer les variables peu pertinentes.

2.2 Régressions pénalisées

Dans cette section, nous appliquerons diverses techniques de régression pénalisée en utilisant l'ensemble de l'échantillon de notre jeu de données, sans effectuer de filtrage préalable des variables.

2.2.1 Ridge

Ridge est une méthode de régularisation, qui aide à gérer la multicolinéarité en introduisant une pénalité sur la taille des coefficients. Cette approche permet de réduire la variance des estimations et d'améliorer la précision des prévisions. En appliquant cette technique nous espérons obtenir des résultats plus robustes et stables, plus particulièrement lorsque les variables explicatives présentent des corrélations élevées. La régression Ridge a été ajusté en utilisant la fonction *glmnet()* du package *glmnet*.

	λ	Sélection de variables
Ridge	0.9770	818

Table 9: Résultats de la méthode Ridge

Pour sélectionner le meilleur paramètre de régularisation lambda une validation croisée à 10 plis a été effectuée avec une série de valeurs possibles pour lambda. Le meilleur lambda de 0.9770 a été déterminé et utilisé pour ajuster le modèle final. Cela signifie que parmi toutes les variables lambda testées, 0.9770 est celle qui minimise l'erreur quadratique moyenne. Notre paramètre lambda estimé est assez élevée, ce qui impose une forte pénalisation des coefficients, ce qui peut par conséquent diminuer leurs amplitudes. Par ailleurs il est essentiel de souligner que la méthode Ridge ne réalise pas une sélection explicite des variables. Par conséquent, l'ensemble des variables explicatives reste inclus dans le modèle.

2.2.2 Lasso

La méthode Lasso se caractérise par son approche visant à atteindre un modèle parcimonieux. Cela signifie qu'elle cherche à conserver uniquement les variables essentielles tout en éliminant celles qui sont superflues. Le Lasso applique une régularisation qui permet de réduire certains coefficients d'estimation à 0, facilitant ainsi la sélection des variables explicatives. Un des principaux avantages de cette méthode réside dans sa capacité à annuler les effets des variables explicatives peu pertinentes, contribuant ainsi à la simplification du modèle final.

Toutefois, cette méthode présente des inconvénients notables, notamment lorsque les variables explicatives sont fortement corrélées. Dans de tels cas, le Lasso a tendance à sélectionner arbitrairement une variable parmi celles corrélées, tout en annulant les coefficients des autres, ce qui peut engendrer un biais de sélection.

	λ	Sélection de variables
Lasso	0.3199	1

Table 10: Résultats de la méthode Lasso

La méthode Lasso a estimé un lambda de 0.3199, sélectionnant une unique variable explicative parmi un total de 818 candidates : IP new. Cette sélection illustre l'efficacité du Lasso dans l'identification des variables pertinentes tout en réduisant le surajustement. Il convient toutefois de s'attarder sur ce résultat ; cela peut indiquer que la variable sélectionnée a une forte relation et une significativité avec la variable cible, ou soulever des questions sur la spécificité du modèle.

Le fait de ne retenir qu'une seule variable peut être considéré comme très restrictif. Cela peut indiquer que cette variable possède une forte relation et une significativité avec la variable cible, mais cela soulève également des questions sur la spécificité du modèle et son incapacité à capturer d'autres influences potentielles.

2.2.3 Elastic-net

La régression Elastic-Net combine les avantages des régressions Ridge et Lasso, offrant une méthode robuste pour analyser des ensembles de données comportant de nombreuses variables, dont certaines peuvent être fortement corrélées. Elle est particulièrement efficace pour traiter les problèmes de multicolinéarité tout en permettant la sélection d'un sous-ensemble significatif de variables. Cette flexibilité rend Elastic-Net adaptée aux données riches en variables explicatives.

Dans notre étude, nous avons utilisé l'approche qui fixe à priori la valeur de alpha. Nous avons fixé le paramètre à 0.5, un choix qui permet d'équilibrer les bénéfices des régularisations Ridge et Lasso. Cela permet à notre modèle d'harmoniser la sélection de variables tout en atténuant la multicolinéarité, ce qui favorise une exploitation efficace des données en conservant les variables les plus pertinentes et en réduisant les effets des corrélations indésirables.

	λ	Sélection de variables
Elastic-net	0,8111308	3

Table 11: Résultats de la méthode Elastic-net

Cette valeur de alpha est particulièrement bien adaptée à la structure complexe de notre jeu de données, permettant d'atteindre un équilibre optimal entre la sélection des variables et la gestion de leurs interactions. Par ailleurs, la régression Elastic-Net a estimé un paramètre de régularisation lambda élevé de 0,8111308, un élément crucial pour déterminer la force de la pénalisation appliquée aux coefficients des variables explicatives. Ce paramètre influence la capacité du modèle à éviter le surapprentissage tout en garantissant une performance prédictive robuste. À l'issue de cette analyse, les variables retenues dans notre modèle sont BSI new, GPI new, et IP new, qui se sont avérées significatives dans l'explication de la variable cible.

2.2.4 SCAD

La régression SCAD est une méthode avancée de régression pénalisée qui se concentre sur l'estimation des coefficients de régression tout en appliquant des pénalités aux coefficients considérés comme non significatifs, les réduisant ainsi à zéro. Ce qui distingue la régression SCAD des autres techniques de régularisation, comme Lasso et Ridge, c'est sa fonction de pénalisation, qui est spécifiquement conçue pour être à la fois moins rigide que celle du Lasso et plus sévère que celle du Ridge. La fonction de pénalisation SCAD permet de maintenir un ensemble de variables explicatives pertinentes tout en éliminant efficacement celles qui ont un impact marginal sur la variable dépendante.

Cette approche non seulement favorise la sélection de variables, mais elle s'attaque également aux problèmes de multicolinéarité en réduisant les coefficients des variables corrélées de manière optimale. Par conséquent, la régression SCAD se révèle particulièrement utile dans des contextes où la précision du modèle est essentielle, tout en garantissant une interprétabilité des résultats.

	λ	Sélection de variables
SCAD	0,0510	19

Table 12: Résultats de la méthode SCAD

Avec une valeur optimale de lambda particulièrement faible, s'élevant à 0,0510, la méthode SCAD se distingue par sa capacité à sélectionner un nombre considérable de variables, surpassant ainsi d'autres techniques de sélection. En effet, 19 variables ont été identifiées et retenues grâce à cette procédure de sélection, témoignant de la richesse d'information que SCAD peut extraire des données.

2.2.5 Adaptive Lasso

La méthode de régression Adaptive Lasso représente une variante de la régression LASSO, conçue pour améliorer son efficacité en adaptant les coefficients des variables explicatives. Contrairement à la régression LASSO classique, où chaque coefficient est soumis à une pénalisation uniforme, l'Adaptive Lasso ajuste ces pénalités en attribuant des poids plus élevés à certaines variables en fonction de leur pertinence dans le modèle. Cette approche permet une meilleure prise en compte de l'importance relative des variables.

	λ	Sélection de variables
Adaptive Lasso	0,184	2

Table 13: Résultats de la méthode Adaptive Lasso

Dans notre analyse, un paramètre de régularisation lambda optimal de 0,1844 a été déterminé. Les variables sélectionnées sont IP new, avec un coefficient de 0,371520963, et GAVG 3.5.Bank rate, dont le coefficient est de 0,009651772. Parmi ces deux variables, IPnew se distingue par son influence significative sur l'évolution du PIB canadien, indiquant ainsi son rôle crucial dans la modélisation économique. Cette capacité de l'Adaptive Lasso à sélectionner des variables pertinentes tout en tenant compte de leur importance relative en fait un outil puissant pour l'analyse des données.

2.2.6 Récapitulatif des résultats des différentes méthodes de régularisation

	λ	Sélection de variables
Ridge	0.9770	818
Elastic-net	0.8111	3
Lasso	0.3199	1
SCAD	0.0510	19
Adaptive Lasso	0.1840	2

Table 14: Résultats des différentes méthodes de régularisation

La table n°14 présente un récapitulatif des résultats des diverses méthodes estimées précédemment, mettant en évidence les valeurs de λ et le nombre de variables sélectionnées. Nous constatons que des méthodes comme le Lasso et l'Elastic Net permettent une sélection de variables plus restrictive, en sélectionnant respectivement 1 et 3 variables. Ces méthodes utilisent des pénalités qui conduisent à une réduction importante de la complexité du modèle, en éliminant les variables jugées non essentielles. En particulier, le Lasso applique une pénalité L1 qui tend à annuler complètement certains coefficients, ce qui simplifie fortement le modèle en ne conservant que les variables les plus pertinentes. L'Elastic Net, quant à lui, combine les pénalités L1 et L2, offrant un compromis en situations de forte colinéarité entre les variables, ce qui est la principale limite de la méthode Lasso et sélectionne donc un nombre limité de variables tout en améliorant la stabilité de la sélection. Cette réduction de complexité permet non seulement de limiter le sur-apprentissage, mais aussi de maintenir une performance prédictive acceptable, voire optimale, sur de nouvelles données.

2.3 Approche de réduction de dimension

Dans la section précédente, nous avons exploré diverses approches de sélection de variables pour identifier celles qui contribuent le plus à notre modèle. Pour optimiser nos résultats, nous avons choisi d'adopter la méthode SIS (Sure Independence Screening), qui sélectionne efficacement les variables pertinentes dans un vaste ensemble de données. Ce processus commence avec un ensemble vide, chaque variable étant évaluée individuellement pour déterminer son utilité. Les variables jugées significatives sont ensuite ajoutées progressivement et validées par une technique de validation croisée. Grâce à cette approche rigoureuse, nous avons réduit notre base de données de plus de 800 à 84 variables significatives, simplifiant ainsi nos modèles tout en améliorant la précision des résultats.

Nous allons maintenant appliquer les mêmes méthodes de régression pénalisée que précédemment, mais cette fois avec un ensemble de données considérablement réduit. Cette stratégie nous permettra de mieux cerner les relations entre les variables tout en optimisant la performance prédictive de nos modèles.

2.3.1 GETS

À la suite de l'application de la méthode GETS, le retour du code a généré un message d'erreur indiquant que le modèle estimé ne respecte pas les critères de sélection de la méthode. Cela peut signifier que certaines variables explicatives n'ont pas satisfait aux tests requis pour rester dans le modèle. Cette situation peut être attribuée à la colinéarité entre les variables, à un nombre insuffisant d'observations, ou à des problèmes avec les résidus du modèle, tels que l'hétéroscédasticité ou l'autocorrélation.

2.3.2 Ridge

La méthode Ridge a abouti à un lambda de 0,0284 et a retenu 84 variables, représentant ainsi la totalité de notre jeu de données après réduction de dimension, y compris les retards appliqués aux variables explicatives. Ce résultat suggère que presque toutes les variables sont jugées pertinentes pour expliquer la variable dépendante. Toutefois, cela peut également indiquer une certaine redondance ou multicolinéarité parmi les variables, ce qui nécessite une attention particulière lors de l'interprétation des résultats. Dans l'ensemble, nous observons des coefficients plus élevés, ce qui témoigne de l'influence significative de méthode SIS.

	λ	Sélection de variables
Ridge	0.0284	84

Table 15: Résultats de la méthode Ridge après filtrage

2.3.3 Lasso

La méthode Lasso a généré un lambda de 0,0722 et a retenu 14 variables. Ce lambda réduit indique une pénalité moins importante, permettant au modèle d'inclure un plus grand nombre de variables significatives. Grâce à sa capacité à identifier les variables pertinentes tout en éliminant celles jugées superflues, Lasso a permis de réduire le nombre de variables de 84 avec la méthode Ridge à seulement 14. Cette sélection contribue à une meilleure interprétabilité du modèle, facilitant ainsi l'analyse des résultats par rapport aux précédents modèles.

	λ	Sélection de variables
Lasso	0.0722	14

Table 16: Résultats de la méthode Lasso après filtrage

2.3.4 Elastic-net

Pour la régression Elastic-Net, le paramètre lambda a été fixé à 0,1048, ce qui a conduit à la sélection de 28 variables. Ce résultat reflète un compromis judicieux entre les approches Ridge et Lasso, permettant d'atténuer les effets de la multicolinéarité tout en maintenant une sélection pertinente des variables. Cette méthode offre ainsi une approche équilibrée, tirant parti des avantages des deux techniques tout en préservant une diversité de variables explicatives.

	λ	Sélection de variables
Elastic-net	0,1048	28

Table 17: Résultats de la méthode Elastic net après filtrage

2.3.5 SCAD

La méthode SCAD a conduit à un lambda de 0,0510 et a retenu 12 variables. Ce lambda, comparable à celui de la méthode précédente, est relativement faible, permettant une sélection précise des variables pertinentes tout en écartant celles dont l'impact est marginal. L'application de SCAD dans ce contexte démontre son efficacité à allier sélection de variables et gestion de la multicolinéarité, renforçant ainsi la robustesse du modèle.

	λ	Sélection de variables
SCAD	0,0510	12

Table 18: Résultats de la méthode SCAD après filtrage

2.3.6 Adaptive Lasso

Lors de cette seconde estimation de l'Adaptive Lasso, nous avons obtenu un lambda de 0,1627, ce qui a permis la sélection de trois variables clés. Cette approche se distingue par sa flexibilité et son adaptabilité, car elle ajuste les coefficients des variables en fonction de leur pertinence dans le modèle, permettant ainsi une identification précise des facteurs influents. L'Adaptive Lasso non seulement favorise une sélection plus fine, mais renforce également la robustesse et la validité des modèles.

	λ	Sélection de variables
Adaptive Lasso	0.1627	3

Table 19: Résultats de la méthode Adaptive Lasso après filtrage

2.3.7 Récapitulatif des résultats des différentes méthodes de régularisation avec réduction de dimension

Méthode	Λ	Sélection de variables
Ridge	0.0284	84
Lasso	0.0722	14
Elastic-net	0.1048	28
SCAD	0.0510	12
Adaptive Lasso	0.1627	3

Table 20: Tableau récapitulatif des résultats des méthodes de régularisation

Les résultats des modèles estiment que la méthode Adaptive Lasso effectue une sélection très stricte, n'incluant que trois variables explicatives. À l'inverse, la méthode Elastic-Net adopte une approche plus équilibrée avec 28 variables, offrant un compromis intéressant entre SCAD et Lasso. La méthode Ridge, quant à elle, conserve un total conséquent de 84 variables, ce qui est normal dans le cadre de cette technique, qui n'élimine pas les variables mais leur attribue des pondérations proches de zéro si elles sont moins significatives. Dans l'ensemble, les valeurs de lambda restent faibles, reflétant une pénalité modérée.

2.4 Approche complémentaire : Random forest

Nous avons exploré l'utilisation d'un modèle Random Forest afin de comparer ses performances avec celles des régressions pénalisées estimées. Cette technique d'apprentissage automatique construit un ensemble d'arbres de décision. Son algorithme combine ensuite les prédictions de ces arbres pour en augmenter la précision globale, ce qui est utile pour identifier les variables clés qui influencent le plus le PIB.

2.4.1 Random forest sans le filtre SIS

Dans un premier temps, nous avons construit un modèle Random Forest en utilisant la base de données complète, sans appliquer la réduction de dimension via la méthode SIS. Ce modèle est constitué de 1000 arbres de décision, où chaque arbre sélectionne aléatoirement 2 variables à chaque division. Le graphique ci-dessous présente les résultats de cette modélisation, mettant en avant les plus influentes.

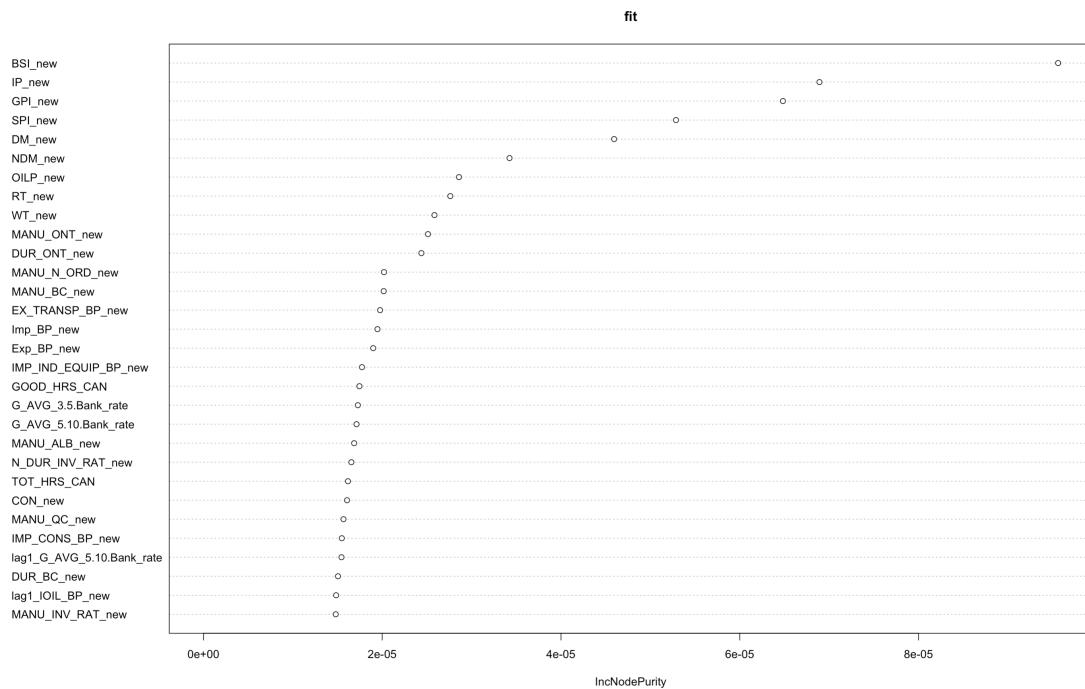


Figure 6: Random forest sans le filtre SIS

Le modèle Random Forest sans filtrage SIS montre que le PIB des affaires (BSInew) est la variable la plus influente, contribuant fortement à la réduction de l'impureté des nœuds, suivi du PIB de la production industrielle (IPnew), du PIB des biens (GPInew) et de SPLnew, également importants. Bien que le prix du pétrole (OILPnew) et RTnew aient un impact moindre, ils restent influents. Ainsi, le PIB des affaires et de la production industrielle jouent un rôle clé, suggérant leur forte corrélation avec la variable dépendante. En somme, une pureté accrue souligne leur importance dans la distinction des catégories des données.

2.4.2 Random forest avec le filtrage SIS

En appliquant les mêmes paramètres, la méthode Random Forest a été utilisée sur la base de données après le filtrage SIS. La figure ci-dessous présente les résultats de cette modélisation.

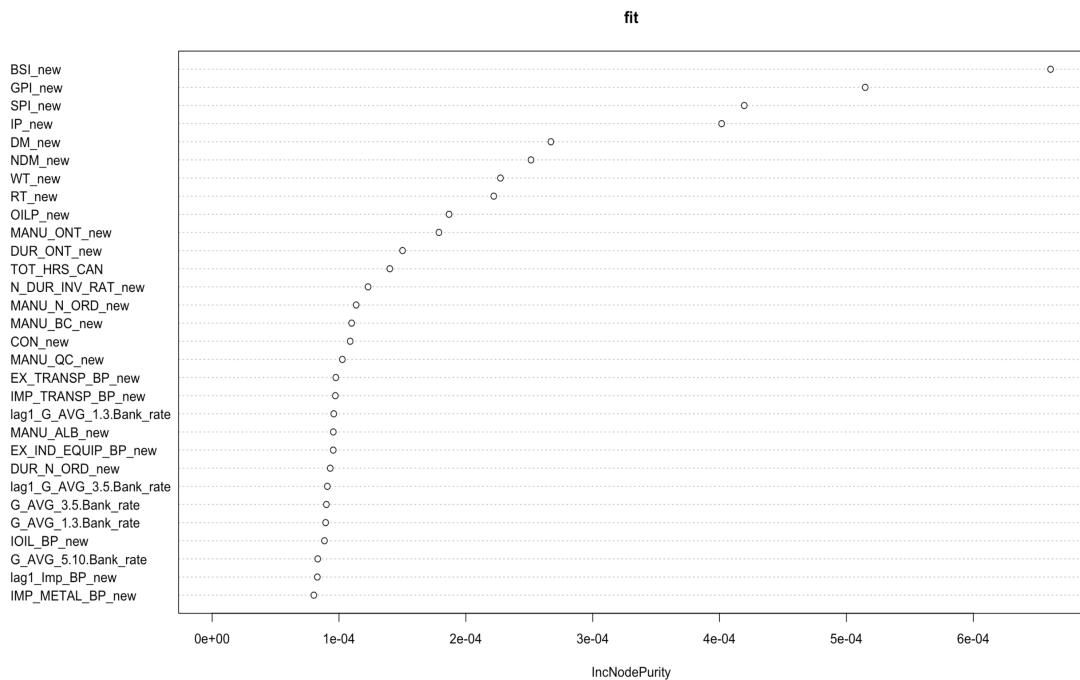


Figure 7: Random forest avec le filtrage SIS

Dans cette version du modèle Random Forest après l’application du filtrage SIS, nous constatons que la variable BSInew reste la plus influente, suivie de GPInew et SPInew, qui montrent également une importance notable. IPnew et DMnew viennent compléter les variables les plus significatives. Le filtrage SIS a permis de réduire l’ensemble des variables tout en conservant celles ayant un impact plus important sur l’impureté des noeuds. Ce processus renforce l’efficacité du modèle en ne conservant que les variables les plus pertinentes, réduisant ainsi le risque de surajustement tout en améliorant la performance prédictive.

2.5 Comparaison des méthodes

Nous allons dans cette partie introduire la comparaison de nos modèles avant et après filtrage SIS pour avoir une visualisation plus globale de nos coefficients associés aux méthodes. Nous effecturons une analyse de ces résultats pour en ressortir les plus intéressants.

2.5.1 Avant filtrage SIS

Ligne	Ridge	Lasso	Elastic-net	SCAD	aLasso
BSI new	0.099		0.002		
GPI new	0.094		0.014		
IPnew	0.100	0.238	0.097	0.688	0.372
SPInew	0.082			0.641	
EMPQC	-0.025			-0.584	
M BASE 1	0.045			0.229	
UNEMP DURA 14,25 CAN	-0.045			-0.064	
UNEMP DURA 27. CAN	-0.027			-0.062	
EMP SERV ALB	-0.029			-0.053	
IMP TRANSP BP new	-0.020			-0.041	
EMP SERV NB	-0.015			-0.038	
EMP SERV BC	-0.011			-0.029	
lag1 UNEMP DURA 27. CAN	-0.010			-0.014	
EMP PART NB	-0.019			-0.014	
lag1 IOIL BP new	-0.052			-0.009	
lag1 CRED CONS discontinued	0.015			0.018	
EMP SERV SAS	-0.028			-0.008	
lag1 SPI new	-0.017			-0.003	
CPI SERV NS	-0.021			-0.001	
G AVG 3.5 Bank rate	0.018			0.010	0.010
G AVF 1.3 Bank rate	0.022			0.006	

Table 21: Tableau comparatif des différentes méthodes sans sélection

Le *tableau n°21* présente une comparaison des méthodes de sélection de variables avant l'application du filtrage SIS, mettant en lumière les variables qui se démarquent

par leur pertinence. Certaines d'entre elles sont systématiquement retenues par la majorité des techniques, ce qui témoigne de leur importance dans le modèle. Par exemple, le PIB de la production industrielle (IPnew) se distingue particulièrement, étant sélectionné par toutes les méthodes, ce qui souligne son rôle crucial dans l'explication des variations du modèle.

Les variables PIB des affaires (BSInew) et PIB des biens (GPInew) semblent être pertinentes pour expliquer l'évolution du PIB même si elles affichent des coefficients modérés, surtout lorsqu'elles sont évaluées par les méthodes Ridge et Elastic-Net, indiquant qu'elles contribuent de manière significative mais pas prépondérante à l'analyse. La méthode Ridge, en conservant l'ensemble des variables, reflète sa nature de régularisation, mais cela se traduit par des coefficients relativement faibles des variables, ce qui pourrait masquer l'impact réel de certaines variables.

En revanche, les méthodes Lasso et aLasso se montrent plus sélectives, identifiant un nombre restreint de variables pertinentes tout en écartant celles jugées superflues. Elastic-Net et SCAD corroborent également l'importance des variables majeures, bien que des différences subtiles émergent en fonction des paramètres appliqués. Ce tableau illustre donc non seulement les dynamiques de sélection des différentes méthodes, mais également l'importance d'une approche nuancée pour identifier les variables clés dans l'analyse du PIB canadien.

Il est désormais essentiel de comparer les résultats obtenus avec l'ensemble des méthodes après l'application du filtrage SIS.

2.5.2 Après filtrage SIS

Ligne	Ridge	Lasso	Elastic-net	SCAD	aLasso
BSI new	0.459	0.425	0.322		0.003
GPI new	0.302	0.367	0.03		
SPI new	0.293		0.095	0.365	
IP new	0.342		0.357	0.702	0.386
OILP new	-0.016		0.035		
CON new	-0.144		-0.012		
RT new	0.081		0.024		
WT new	0.060	0.027	0.057	0.016	
PA new	0.121		0.026		
EMP PART CAN	-0.142	-0.12	-0.126	-0.148	
CLAIMS CAN	-0.093		-0.019		
TOT HRS CAN	-0.367	-0.014	-0.083	-0.016	
G AVG 1.3 Bank rate	0.049		0.023		
G AVG 3.5 Bank rate	-0.021	0.067	0.027	0.029	0.028
G AVG 5.10 Bank rate	0.037		0.024	0.008	
RES IMF	0.021		0.004		
EX TRANSP BP new	-0.219	-0.051	-0.069	-0.016	
IMP METAL BP new	-0.219	0.021	0.037		
IMP TRANSP BP new	-0.172	-0.095	-0.096	-0.082	
UNEMP QC	0.327	0.235	0.234	0.456	
MANU BC new	-0.012	0.008	0.029		
CPI MINUS FOO ALB	-0.07	-0.011	-0.023	-0.001	
CRED T discontinued	0.066		0.006		
CRED MORT discontinued	0.029		0.011		
lag1 G AVG 1.3 Bank rate	0.053	0.0013	0.015		
lag1 G AVG 3.5 Bank rate	-0.006		0.011		
lag1 Imp BP new	-0.080	-0.035	-0.054	-0.005	

Table 22: Tableau comparatif des différentes méthodes avec sélection SIS

En examinant la *table n°22*, qui présente une comparaison des méthodes après le filtrage SIS, nous constatons une augmentation significative du nombre de variables sélectionnées par la majorité des techniques. Par exemple, la variable "G AVG 3.5

Bank rate” est retenue par toutes les méthodes, soulignant son importance cruciale et sa pertinence dans le modèle. En revanche, la variable ”PI new”, qui avait été sélectionnée par la méthode Lasso, ne figure plus parmi celles retenues après le filtre SIS. Bien que le PIB de la production industrielle (PI new) ne soit plus inclus dans le modèle par Lasso, il se distingue néanmoins par des coefficients parmi les plus élevés, attestant de sa force explicative même en l’absence de sélection.

Cette situation peut s’expliquer par la réduction de la complexité du modèle, qui permet de se concentrer sur les variables les plus pertinentes et significatives. En effet, le filtre agit comme un prétraitement stratégique, éliminant les variables jugées moins significatives et conduisant à un modèle plus parcimonieux, axé sur l’essentiel. De surcroît, les coefficients observés après ce filtre sont considérablement plus élevés que ceux obtenus sans ce prétraitement, illustrant ainsi une amélioration notable de la robustesse et de la pertinence des variables retenues. Ce constat indique que le filtre SIS non seulement affine le modèle en mettant en lumière les variables clés, mais renforce également leur pouvoir explicatif, offrant ainsi une meilleure interprétation des relations sous-jacentes dans les données.

Conclusion

À travers ce projet, nous avons mis en œuvre une variété de techniques pour traiter notre vaste base de données en big data, qui compte 410 variables ayant un lien direct avec le PIB du Canada. Afin d'analyser plus précisément les tendances des corrélations et les décalages dans les réponses au fil du temps, nous avons appliqué un retard à nos variables indépendante ainsi que quatre retards à notre variable d'intérêt. Cette opération a considérablement augmenté la taille de notre base de données, la portant à plus de 800 variables.

Dans le cadre de notre étude, nous avons initialement tenté d'appliquer une méthode économétrique fondée sur la sélection de variables, connue sous le nom de GETS. Cependant, en raison de la complexité et du nombre important de variables dans notre base de données, cette approche n'a pas donné de résultats concluants. Nous avons donc exploré plusieurs méthodes de régressions pénalisées, notamment Ridge, Lasso, Elastic Net, SCAD et Adaptive Lasso, tout en conservant les retards appliqués. Ces modèles se sont révélés efficaces, nous permettant de sélectionner les variables les plus pertinentes tout en fournissant des coefficients spécifiques à chaque modèle, en fonction de leur contribution à l'explication de la variable dépendante.

Dans une deuxième phase, nous avons appliqué la méthode de réduction de dimension SIS pour présélectionner les variables avant d'appliquer les modèles de régression pénalisée. À partir de cette base de données à plus petite dimension, nous avons répété les étapes précédentes, notamment l'application des mêmes modèles de régression pénalisée. Cette approche a conduit à une sélection de variables significativement plus importante, illustrée par le fait que pour le modèle Elastic Net, nous sommes passés de 3 à 27 variables retenues, ce qui représente un gain substantiel en termes d'informations. Ce filtrage a également permis d'augmenter les coefficients observés, renforçant ainsi la robustesse des résultats.

Enfin, nous avons élaboré une approche complémentaire en utilisant un modèle de forêt aléatoire pour effectuer une sélection de variables de manière non linéaire. Les résultats montrent que, dans la plupart des cas, les variables sélectionnées par ce modèle coïncident avec celles choisies par les régressions pénalisées, tout en mettant en évidence les variables les plus contributives au PIB, telles que le PIB de la production industriel (IPnew), le PIB des affaires (BSInew) et Taux d'intérêt bancaire (GAVG3.5Bank Rate). Cette convergence entre les méthodes souligne la solidité des résultats obtenus.

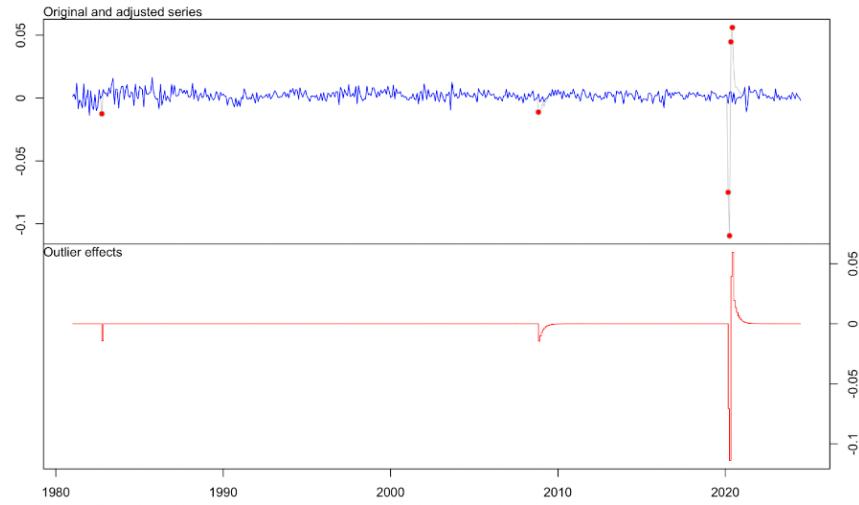
Il est également à noter que les variables qui se sont révélées comme étant les plus significatives dans nos modèles sont celles qui avaient été identifiées dans la section sur les corrélations de la partie d'analyse descriptive comme étant les plus corrélées avec notre variable dépendante. Cette concordance renforce la validité de notre analyse et souligne l'importance de ces variables dans l'explication du PIB.

Pour conclure, nos résultats indiquent que les méthodes Lasso et Adaptive Lasso, tant avant qu'après l'application de la méthode SIS, se révèlent être les plus sélectives. En revanche, SCAD et Elastic Net montrent une sélectivité plus faible dans les deux configurations, ce qui pourrait limiter leur efficacité dans certains contextes. Pour approfondir cette analyse, il serait judicieux de mener des prévisions sur les modèles afin d'évaluer et de comparer les performances des différentes approches. Cette démarche nous permettrait non seulement de confirmer les résultats observés, mais aussi d'acquérir une compréhension plus fine des dynamiques sous-jacentes de notre jeu de données, ouvrant ainsi la voie à des applications futures plus ciblées et efficaces dans le domaine de l'économétrie.

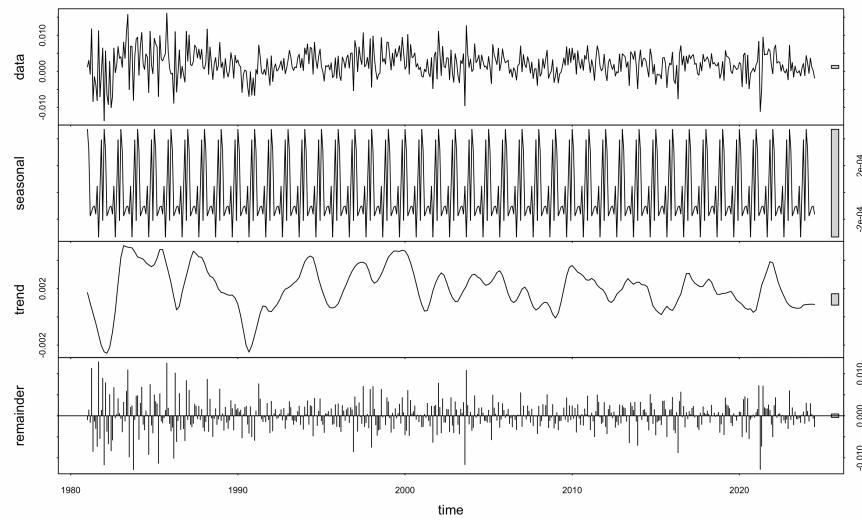
Bibliographie

- https://www.stevanovic.uqam.ca/DS_LCMD.html
- <https://www.francophonie.org/canada-covid19>
- <https://www.erudit.org/fr/revues/ps/2020-v39-n2-ps05351/1070038ar/>
- http://classiques.uqac.ca/contemporains/Paquette_Pierre/Pourquoi_Canada_si_touche_par_recession/Pourquoi_Canada_si_touche_par_recession_texte.html

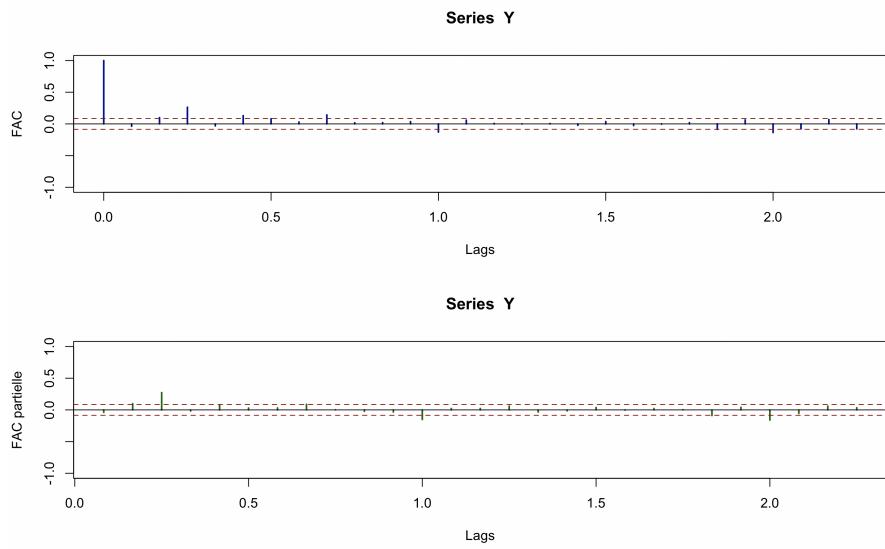
3 Annexe



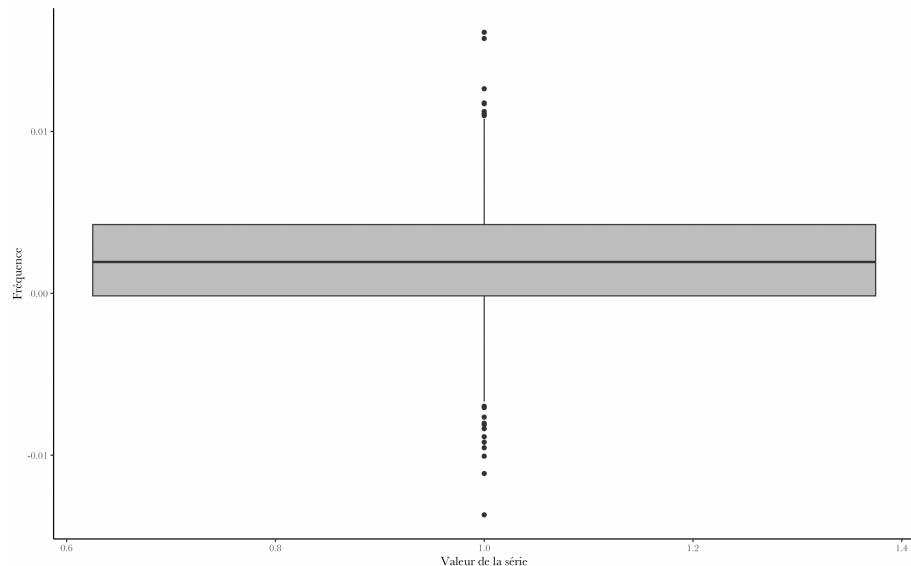
Annexe 1: Graphique de détection des outliers pour la série du PIB canadien par tsoutliers



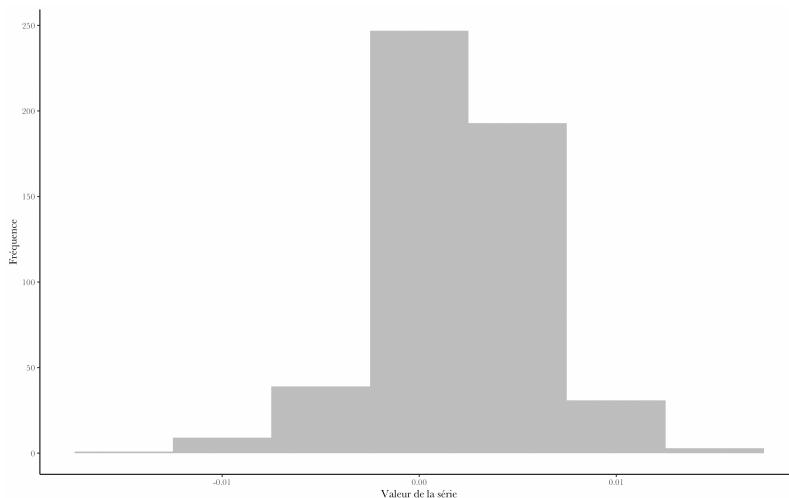
Annexe 2: Décomposition de la série PIB canadien



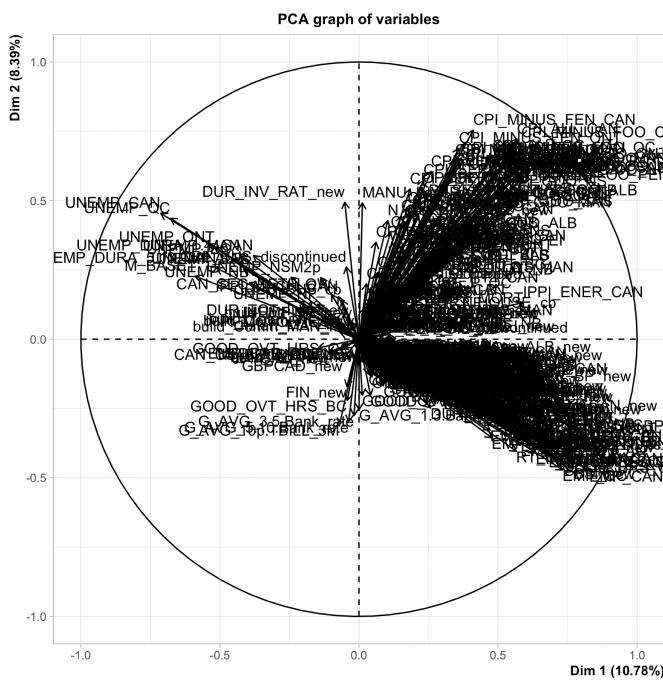
Annexe 3: Corrélogramme de la série PIB Canadien



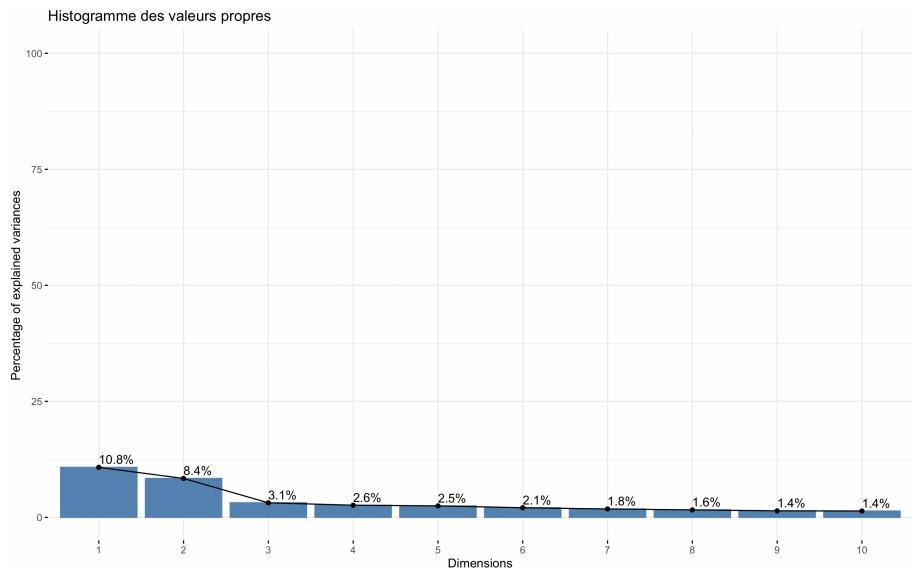
Annexe 4: Boxplot de la série PIB Canadien



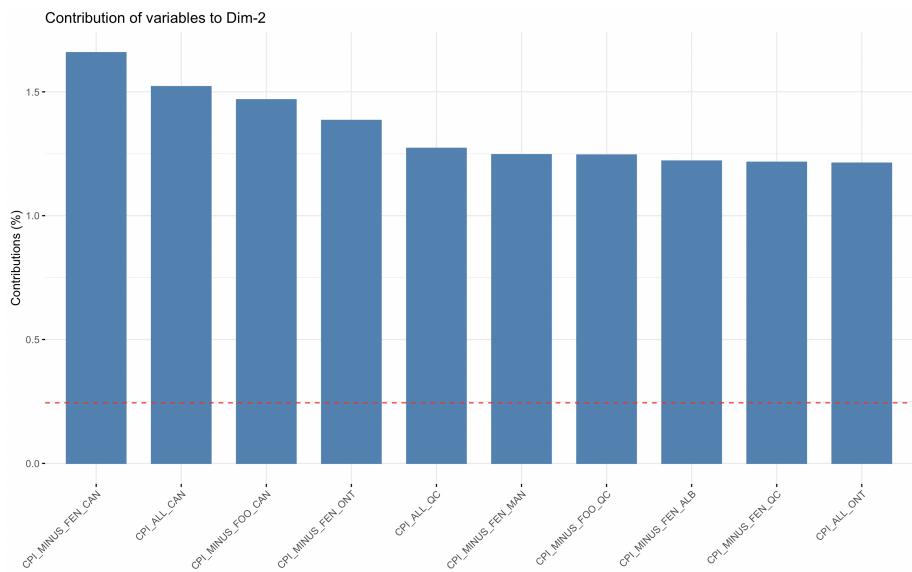
Annexe 5: Histogramme de la série PIB Canadien



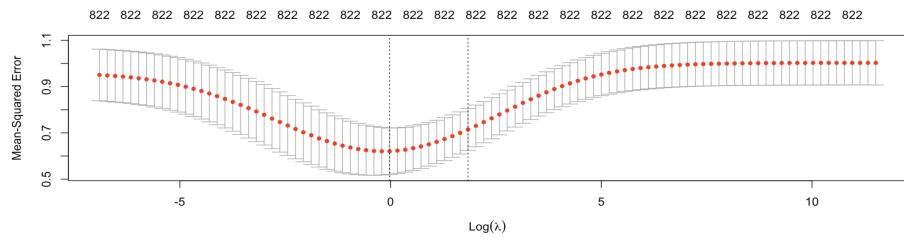
Annexe 6: ACP de la série PIB Canadien avec toutes les variables



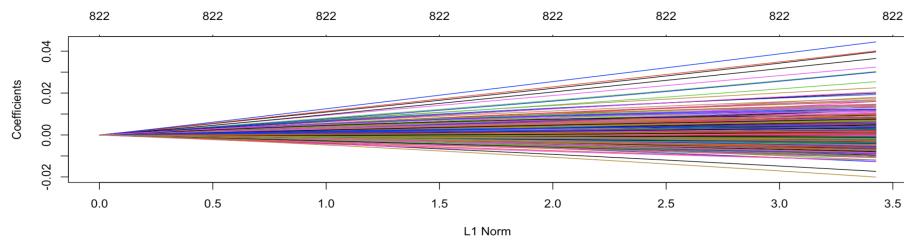
Annexe 7: Histogramme des valeurs propres



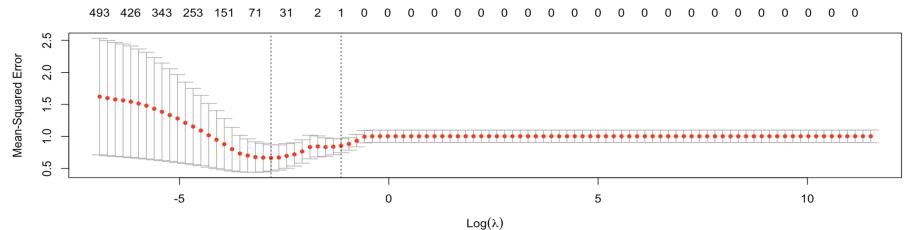
Annexe 8: ACP histogramme



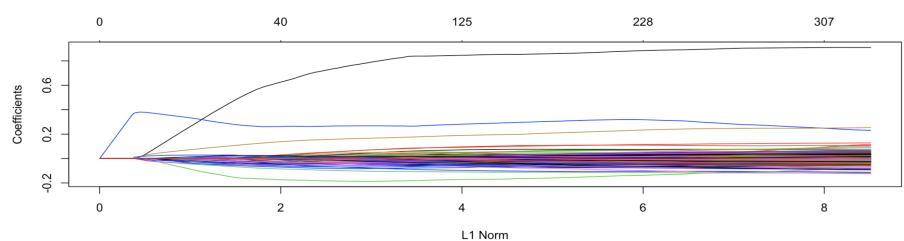
Annexe 9: Courbe Ridge



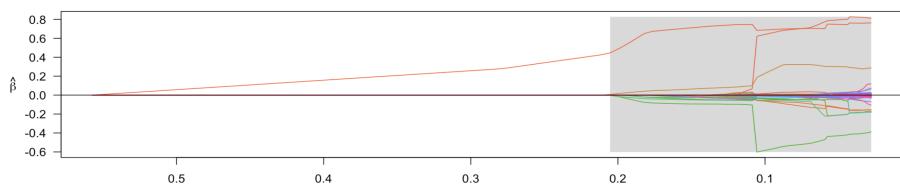
Annexe 10: Graphique Ridge



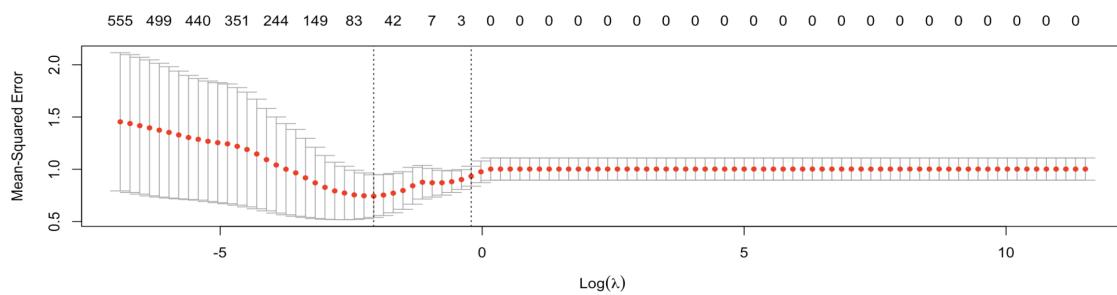
Annexe 11: Courbe Lasso



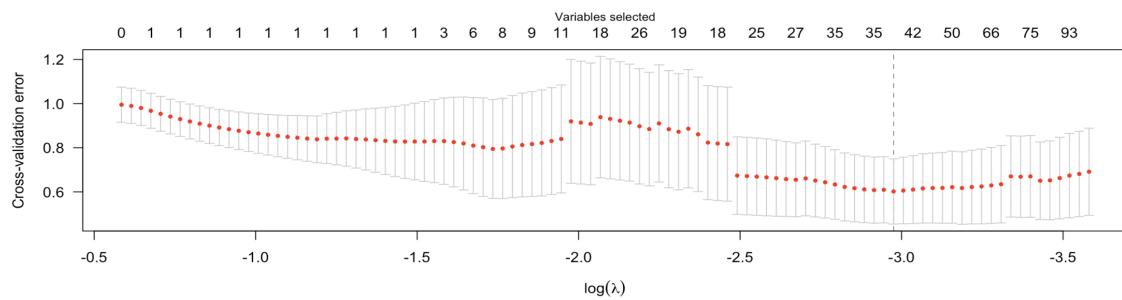
Annexe 12: Graphique Lasso



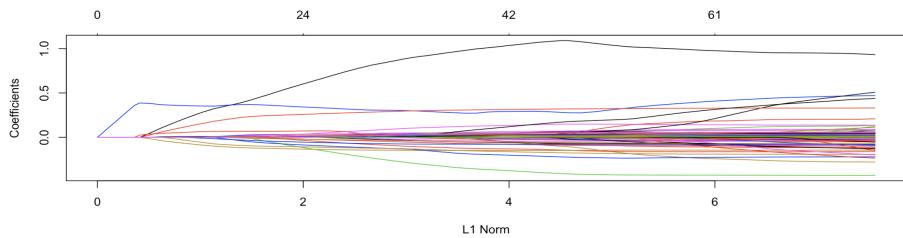
Annexe 15: Graphique SCAD



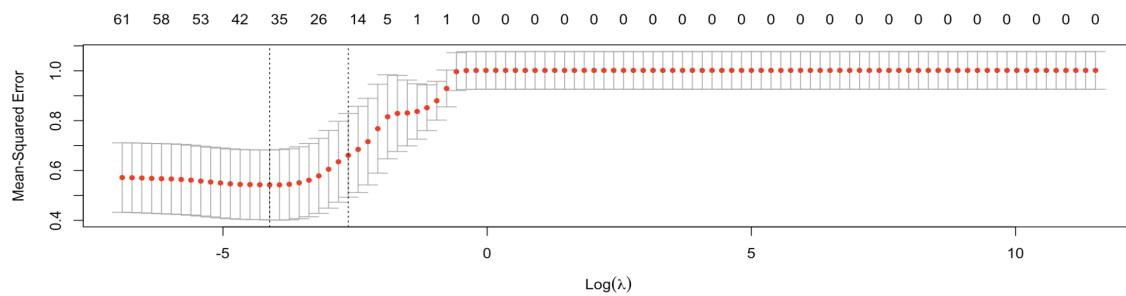
Annexe 13: Courbe Elastic-net



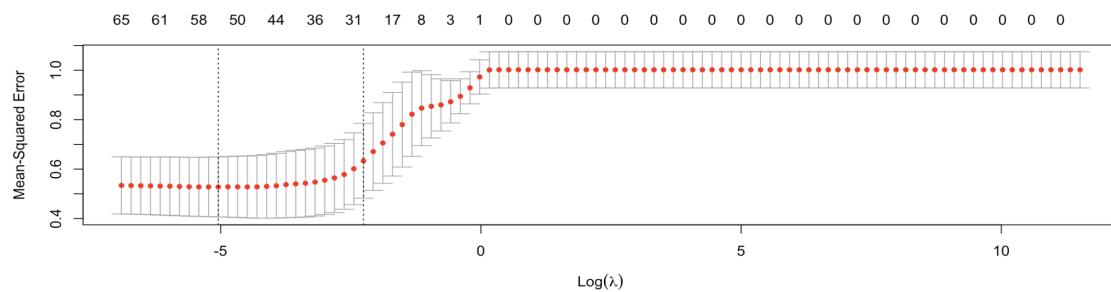
Annexe 14: Courbe SCAD



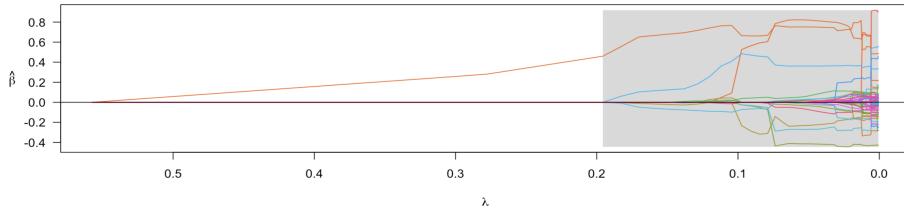
Annexe 17: Graphique Lasso après sélection de variable



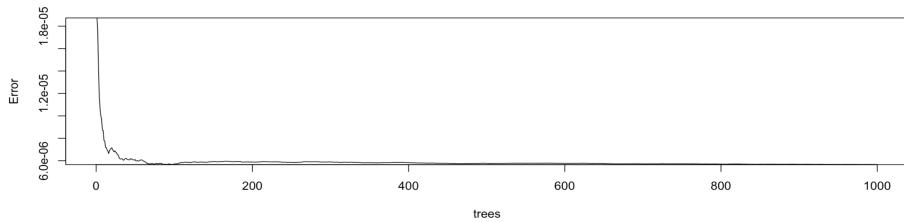
Annexe 16: Lasso après sélection de variable



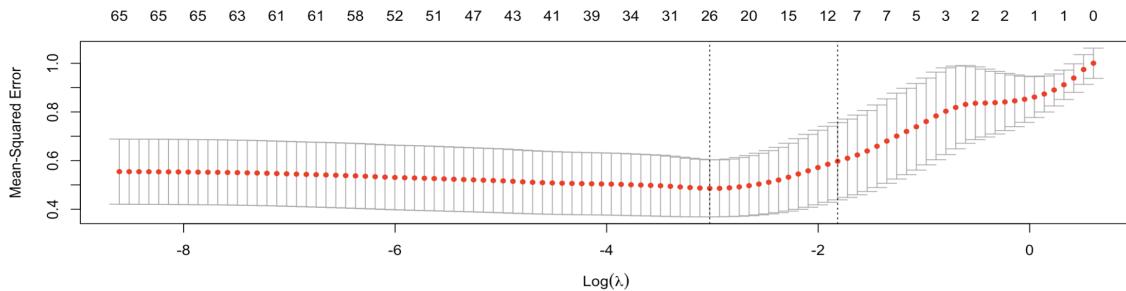
Annexe 18: Elastic-net après sélection de variable



Annexe 19: Graphique SCAD après sélection de variable



Annexe 21: Random forest



Annexe 20: Adaptive lasso après sélection de variable