# Projet Data Mining
## Prévision de la variation du chômage aux États-Unis

Ludovic Stephan et Ludovic Schwartz

9 mars 2018

# Introduction

Dans le cadre du cours *Projet Machine Learning pour la prévision*, nous nous sommes intéressés à l'application des méthodes du Machine Learning à la prévision d'un jeu de données économique : le taux de chômage aux États-Unis.

Nous avons à cet effet récupéré dans la base de données de l'OCDE (Organisation pour la Coopération et le Développement Économiques), le taux de chômage standardisé aux Etats-Unis (Harmonised Unemployment Rate ou HUR). Ce jeu de données est mensuel et couvre la période de janvier 1955 à nos jours.

Nous avons aussi récupéré dans la base de données de l'OCDE d'autres jeux de données économiques mensuels sur la même période pour former nos covariables qui seront utilisées pour la prévision.

D'emblée, on peut faire quelques remarques sur notre jeu de données :

— L'objet que l'on cherche à prédire est une variable économique de dimension 1. C'est une variable qui dépend de façon très complexe de la situation économique du pays, et notre précision de prévision sera donc limitée.

— Le taux de chômage dépend d'un grand nombre de variables. La sélection des covariables qui seront utilisées dans nos modèles est donc un enjeu majeur du problème.

— On a un peu plus de 600 mensualités où toutes nos variables sont renseignées (période 1965-2017). Ce faible nombre d'observations permet d'utiliser des algorithmes complexes sans traitement préalable des données, mais peut être limitant pour certains modèles.

— Les valeurs du taux de chômage sont fournies avec une précision à 1 chiffre après la virgule, ce qui apporte une imprécision supplémentaire.

# 1 Récupération et exploration des données

## 1.1 Constitution du jeu de données

Toutes nos données proviennent de la base de données de l'OCDE (`https://data.oecd.org/`) ; on trouve dans cette base de données de nombreuses variables économiques sur ses pays membres. Nous avons donc effectué un premier processus de sélection pour choisir lesquelles allaient potentiellement intégrer notre modèle.

Notre variable à prédire est le taux de chômage harmonisé (Harmonised Unemployment Rate, ou HUR), qui est une mesure du chômage indépendante de la méthode de mesure utilisée dans chaque pays.


Figure 1 – Courbe du taux de chômage harmonisé aux États-Unis – © OCDE

Nous avons remarqué que les différentes crises économiques sont particulièrement visibles sur ce graphique, notamment la récession de 1981-1982 et la crise de 2008. En conséquence, nos considérations heuristiques sur le choix de covariables se sont fondées sur :

- Une relation qualitative directe avec le taux de chômage ou avec la santé économique américaine et mondiale ;
- L'existence d'observations mensuelles, remontant à suffisamment longtemps pour avoir un jeu de données d'une taille suffisante ;
- Plus heuristiquement, la possibilité d'identifier les différentes crises mentionnées sur les courbes.

Nous avons ainsi fini par sélectionner sept covariables économiques, en plus du taux de chômage :

— les taux d'intérêt à court terme des États-Unis
— l'indice de confiance des entreprises (Business Confidence Index, ou BCI)
— l'indice de confiance des consommateurs (Consumer Confidence Index, ou CCI)
— le taux de chômage harmonisé au Canada
— le prix de diverses actions sur le marché financier américain
— la quantité d'immatriculations de voitures neuves aux États-Unis
— l'indicateur composite avancé produit par l'OCDE (Composite Leading Indicator, ou CLI)

Le CLI, qui est la seule quantité "opaque" du jeu de données, est en fait un agrégat de séries temporelles sélectionnées par l'OCDE pour leur capacité à prévoir les fluctuations du PIB – plus précisément, les points de rebroussement de ces séries doivent précéder de peu ceux du PIB. Cette sélection est mise à jour régulièrement, et cet indicateur semble donc être heuristiquement notre covariable la plus utile pour le problème considéré.

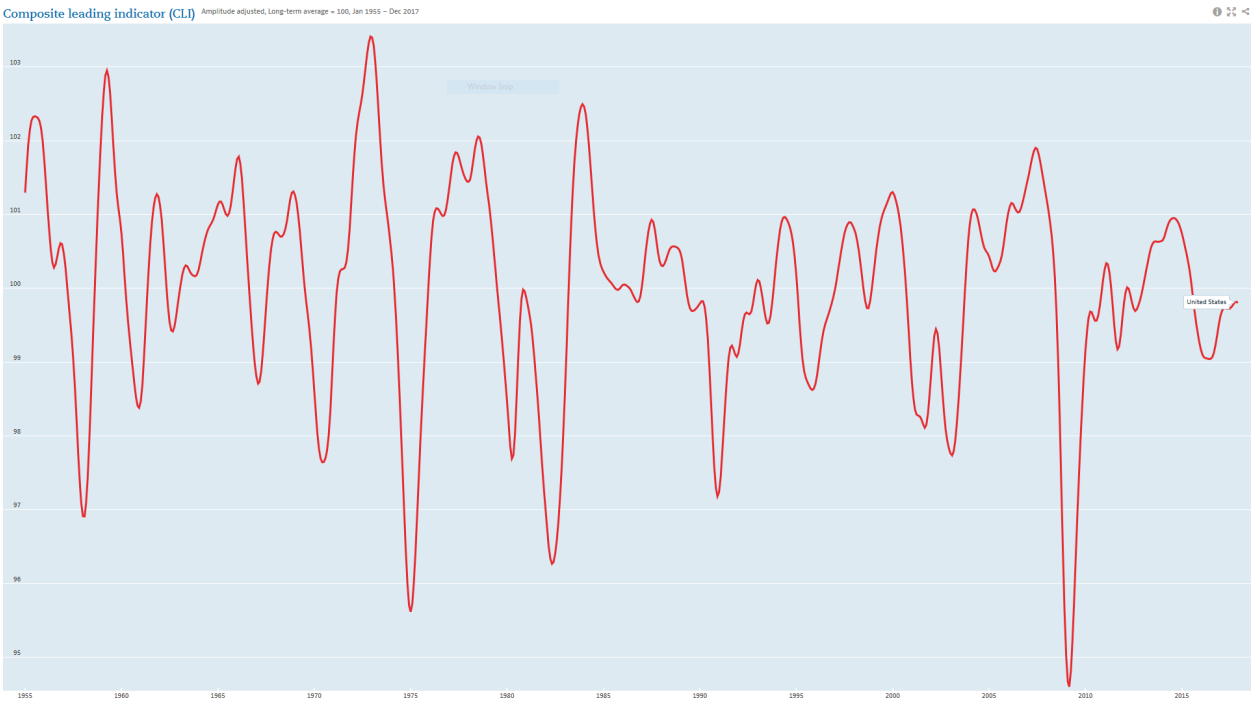



Figure 2 – Courbe du CLI aux États-Unis, normalisé tel que 2010 = 100 – © OCDE

On remarque ici encore les creux correspondant aux différentes crises et récessions économiques.

Pour voir si une covariable est utile dans la prévision, il est intéressant de voir si elle est affectée par la situation économique plus vite que le chômage. C'est à dire, qu'une modification de la situation économique va affecter notre covariable avant d'affecter le chômage, et donc que notre covariable va « détecter » le changement économique qui va influer sur le chômage futur.

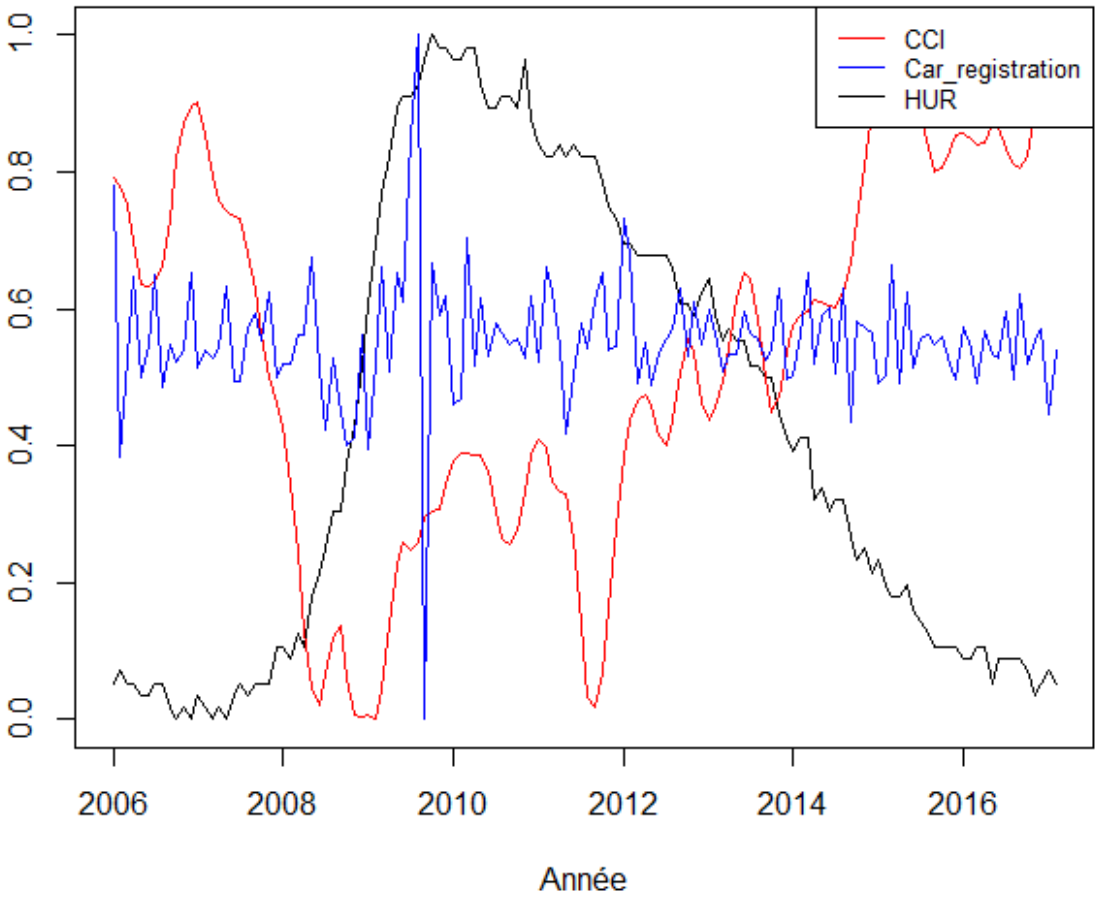



**Figure 3** – Graphes de diverses variables économiques en fonction du temps

Sur le graphique précédent, on peut voir 3 courbes : le taux de chômage en noir, l'indice de confiance des consommateurs (Consumer Confidence Index, ou CCI) en rouge et la quantité d'immatriculations de voitures neuves aux États-Unis (Car Registration ou CR) en bleu. On remarque nettement l'impact de la crise de 2008 sur ces 3 variables, cependant on remarque qu'elles ne la subissent pas en même temps, le CCI est la première, puis le taux de chômage et enfin la CR. La CR subit les effets économiques en retard par rapport au chômage et

on peut donc s'attendre à un faible pouvoir de prédiction de cette variable. Le CCI est par contre en avance, et s'annonce donc un indicateur prometteur.

## 1.2 Prise en compte des e□ets temporels

Notre objectif initial était de prévoir le taux de chômage du mois suivant. Cependant, les variations mensuelles du taux de chômage étant assez faibles, un très bon prédicteur du chômage du mois prochain se trouve être le chômage du mois actuel.

Pour contrebalancer ce phénomène, nous avons donc décidé de prévoir plutôt la variation du taux de chômage, moins stationnaire, et donc impliquant plus certaines des covariables sélectionnées précédemment.

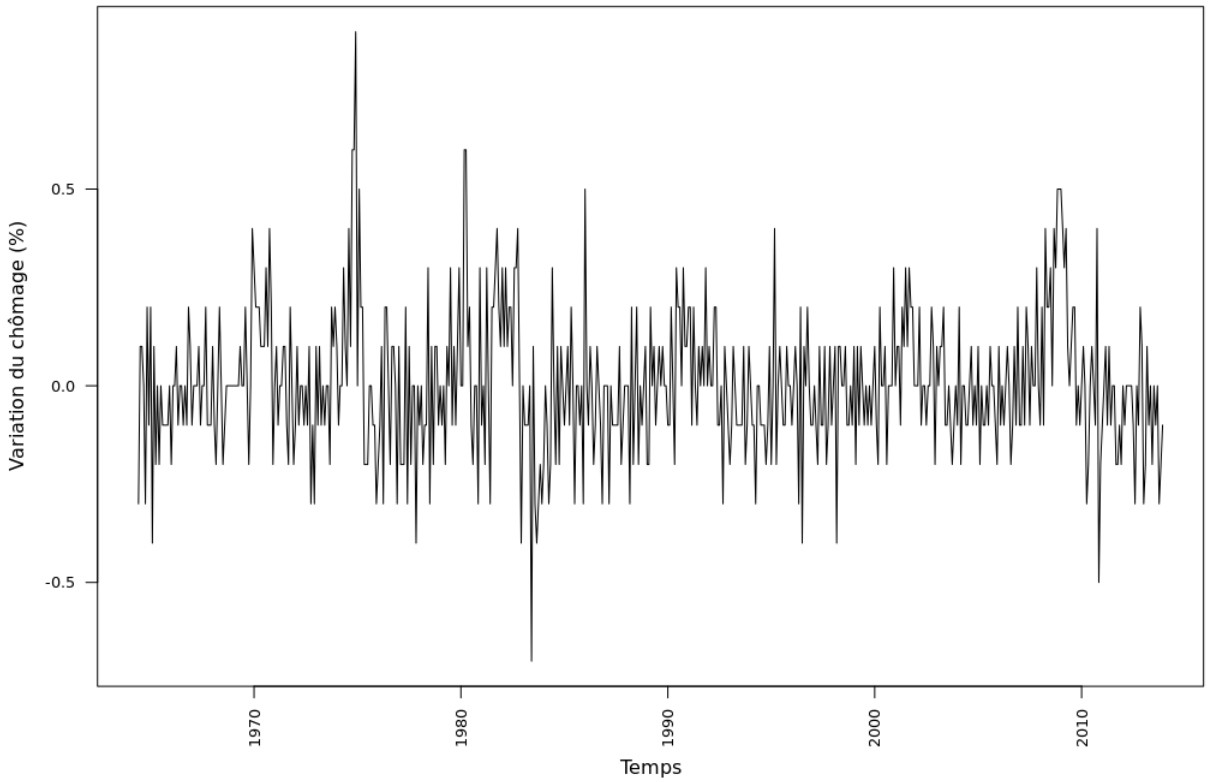
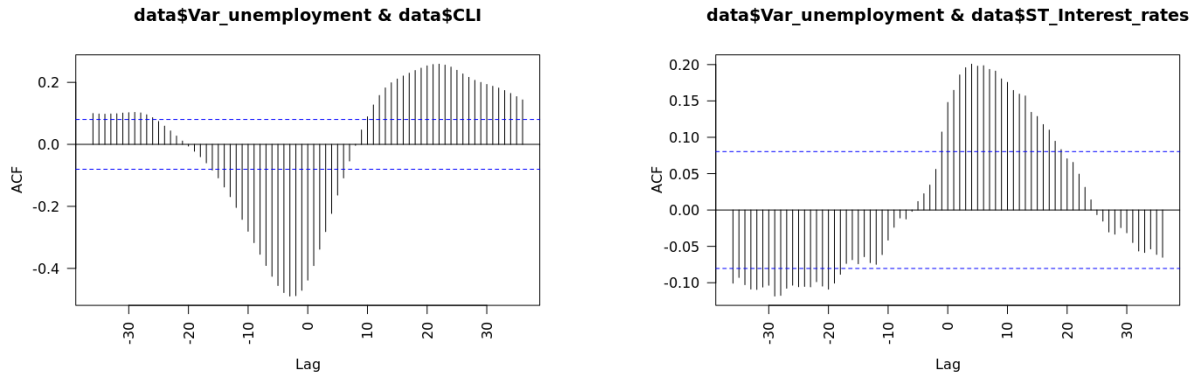



**Figure 4** – Courbe de la variation du taux de chômage aux États-Unis

Nous disposons donc maintenant de nombreuses covariables, qui sont toutes reliées à la situation économique soit des États-Unis, pour la plupart, soit du monde, comme par exemple le chômage canadien. Comme nous l'avons remarqué précédement, nous sommes uniquement intéressés pour la prévision aux variables qui sont affectées par la situation économique avant le taux de chômage, puisque tout effet postérieur n'est d'aucune utilité pour prévoir ce taux.

Nous avons donc cherché une méthode simple pour observer ce phénomène, qui nous a été apportée par la covariance croisée (`ccf` dans R).

Il s'agit en fait simplement d'un calcul de corrélation entre la variation du taux de chômage et une autre covariable, en décalant celle-ci vers le passé ou le futur à chaque fois. Intuitivement, en observant le point où la corrélation entre la covariable décalée et le taux de chômage est maximale, on peut alors déterminer si la situation économique impacte d'abord le chômage (décalage positif), auquel cas la covariable risque d'être peu utile, ou bien d'abord l'autre indicateur (décalage négatif), qui aura alors un intérêt pour la prévision.




(a) `ccf` avec le CLI

(b) `ccf` avec les taux d'intérêts

`Figure` 5 – Graphes de la fonction `ccf` entre le taux de chômage et deux covariables différentes

On remarque par exemple ici que la corrélation maximale entre le CLI et le taux de chômage est atteinte pour un décalage de 3, ce qui semble indiquer que cet indicateur nous sera utile par la suite ; cela n'est bien sûr pas une surprise, puisqu'il a été construit pour avoir une capacité de prédiction. Au contraire, les taux d'intérêts à court terme ont tendance à avoir une corrélation maximale pour un décalage positif, ce qui pourrait impliquer qu'ils sont affectés par la situation économique avec un temps de retard par rapport au chômage, et donc qu'ils auront un faible pouvoir de prévision.

Notons cependant qu'à ce moment là de notre étude, nous n'avions pas encore effectué de sélection de variables : bien que très informatifs, ces graphiques n'apportent que des intuitions sur nos données que nous devrons confirmer par le calcul.

Au vu de ces considérations, nous avons donc décidé de générer de nouvelles covariables, constituées des indicateurs précédents décalés de quelques mois (entre un et trois mois pour chaque covariable concernée). Nous avons appliqué ce traitement aux indicateurs `BCI`, `CCI`, et `CLI`, ainsi qu'à notre variable de taux de chômage et à sa variation. Nous avons aussi rajouté des variables contenant respectivement le mois et l'année, afin de chercher à exploiter par la suite le fait d'avoir une série temporelle.

Une fois tout ce traitement effectué, notre base de données était donc constituée de 27 variables, pour un total de 633 observations.

# 2 Modèles de prévision

Afin de pouvoir tester par la suite l'efficacité de nos modèles, nous avons mis à part les années 2014 à 2017 comme données de test ; cela nous laisse avec un ensemble d'apprentissage contenant 595 observations.

## 2.1 Modèles simples ; sélection de variables

Afin de confirmer par l'expérience ce que nous avons remarqué avec la fonction `ccf`, nous avons commencé par lancer des modèles de prédiction classiques pour avoir une idée plus précise de l'importance des variables.

### 2.1.1 Régression linéaire

Comme prévu pour un modèle complexe, une simple régression linéaire n'atteint pas une précision importante ($R^2 = 0.3315$), mais nous donne tout de même des informations intéressantes. On a 3 variables significatives, le BCI, et le taux de chômage du mois actuel et un peu moins important, le chômage du mois précédent. Cela ne signifie pas que les autres variables ne sont pas significatives, juste que si elles le sont, leur effet n'est pas linéaire. On remarque de plus sur le graphe de la régression un phénomène qui va rester présent sur presque tous les modèles : on récupère la tendance générale aisément, mais il reste un bruit difficile à récupérer.

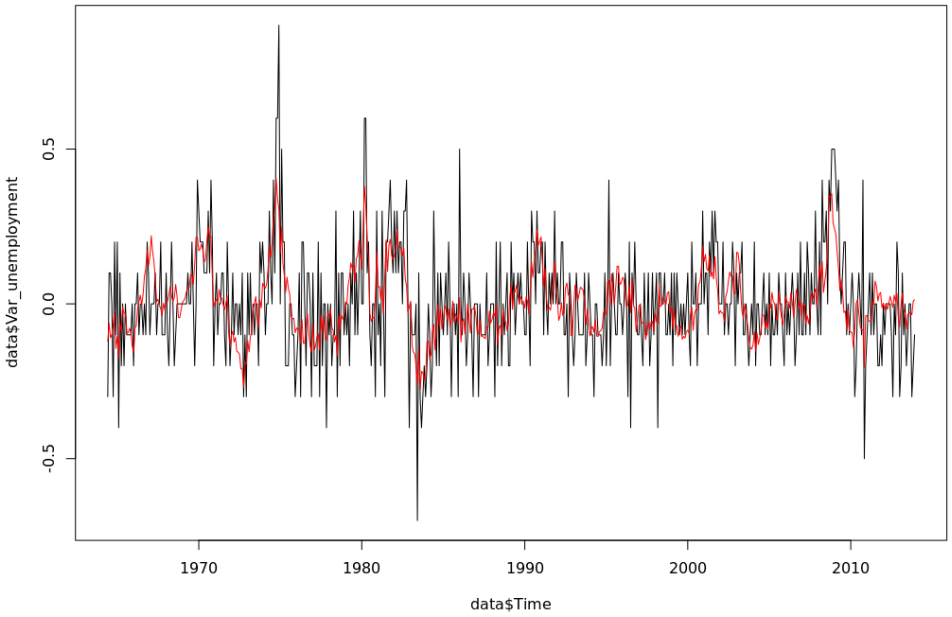



Figure 6 – Régression linéaire simple

Il est ici difficile d'estimer à quel point ce bruit restant est dû à l'arrondi des chiffres du chômage effectué par l'OCDE, qui peut amener une erreur très importante – 0.05 points d'erreur sur une amplitude maximale de 0.5 est énorme –, ou à des phénomènes économiques induisant un bruit. Il est à noter que contrairement à ce que nous pensions au premier abord, on n'observe pas de comportement cyclique : la variation de chômage ne dépend presque pas du mois de l'année, alors que cela aurait été une hypothèse logique.

### 2.1.2 Forêts aléatoires, importance des variables

Notre deuxième modèle général, et celui qui nous a donné le plus d'informations, est une forêt aléatoire simple : en effet, ce modèle, en plus d'une prédiction, fournit aussi une notion d'importance de variables qui va nous permettre de faire de la sélection.

Les résultats d'importance de ce modèle sont plus conformes à ce que nous attendions : les variables que nous avions repéré comme peu intéressantes (car évoluant plus « lentement » que le chômage), comme les taux d'intérêt à court terme ou les nombres de nouvelle imatriculation ont une importance très faible.

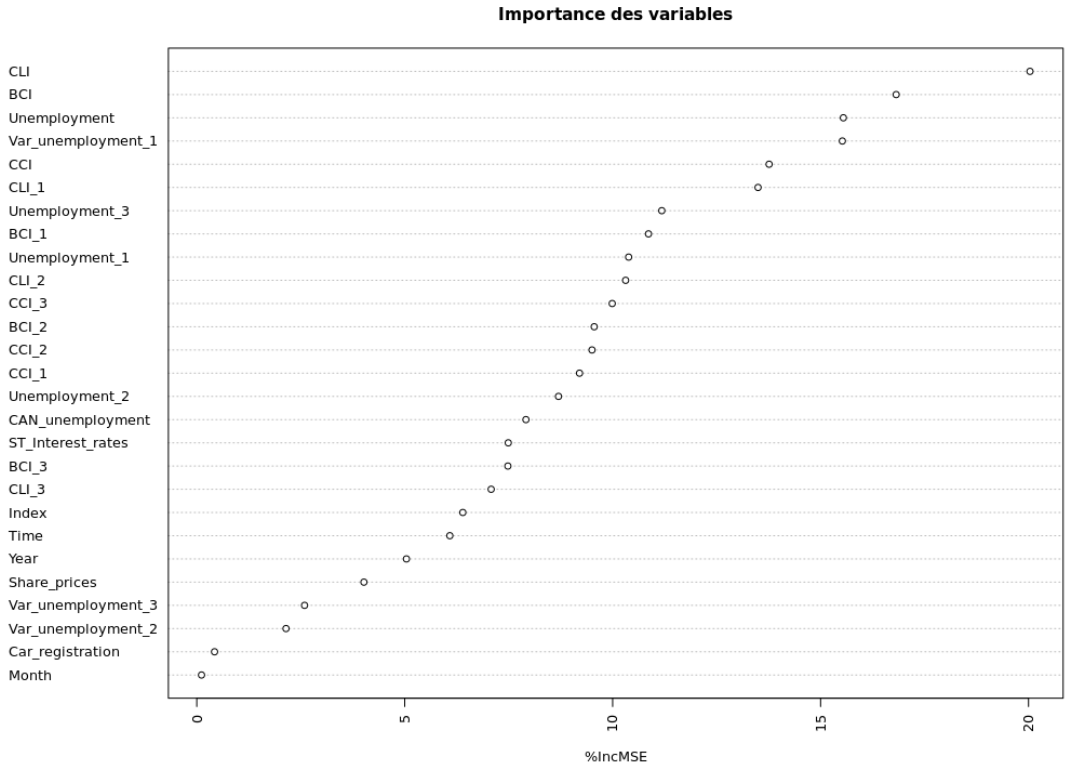



Figure 7 – Importance des variables calculée par forêt aléatoire

Les variables que nous décidons de conserver sont le CLI (et sa version laggée), le BCI, et le CCI, nos 3 indicateurs, ainsi que le taux de chômage du mois passé et sa variation précédente. Notre modèle sera donc basé sur ces 6 covariables.

Il est à noter que cette sélection est effectuée certes à partir d'informations qualitatives, mais toujours à vue d'oeil ; une méthode de sélection automatique à partir de modèles GAM emboîtés, présente dans le paquet R associé, sera présentée par la suite.

## 2.2 Modèle Random Forest

### 2.2.1 Optimisation des paramètres

Un modèle de prévision intéressant est celui des Random Forest. On entraîne notre forêt sur notre ensemble d'apprentissage, et on fait notre prédiction en passant les nouvelles données dans notre modèle.

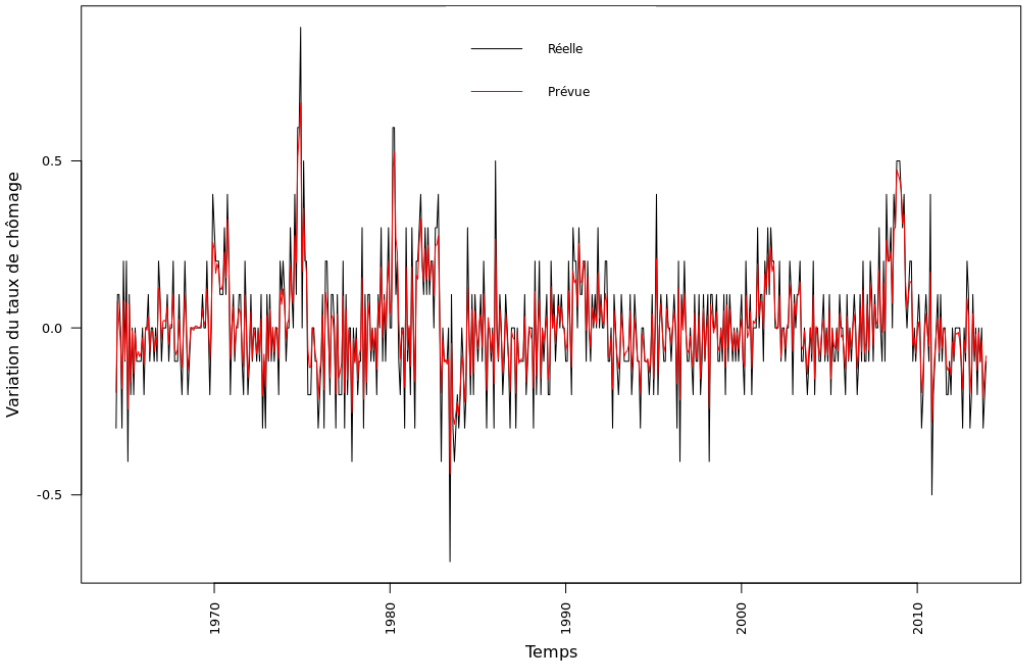



Figure 8 – Forêt aléatoire classique

Les résultats sont encourageants : même si l'erreur reste importante, on arrive déjà à capturer l'allure du graphe de notre variable, notamment les points de rebroussement. Cependant, ce résultat reste améliorable : les forêts aléatoires dépendent de plusieurs paramètres que l'on peut chercher à optimiser dans notre modèle :

— `ntree`, le nombre d'arbres que contient la forêt

— `mtry`, le nombre de variables prises en considération lors des séparations de l'arbre

On optimise ces deux critères l'un après l'autre par validation croisée pour chercher à réduire notre erreur de prédiction. Il est à noter que cette optimisation prenant un temps important (surtout pour les grandes valeurs de `ntree`), il est difficile de générer plusieurs forêts et le résultat aura donc une variance importante.

On trouve alors que `ntree` = 200 est une valeur largement suffisante (d'autres valeurs plus haut conduisent parfois à des erreurs plus faibles, mais aussi à des temps de calcul trop importants), et `mtry` = 3 environ.

### 2.2.2 RF à fenêtre glissante

Une idée pour la prévision du chômage est que plus notre modèle s'entraîne sur un modèle économique récent(donc proche), plus il a de chances de bien comprendre le chômage actuel. On peut donc penser à utiliser le modèle de « prévision à fenêtre » : pour un $n$ donné, on entraîne le modèle sur les $n$ années précédentes, et on calcule son erreur de prédiction sur le mois (ou l'année) suivante. Pour des raisons de temps de calcul, nous avons préféré de décaler la fenêtre d'année en année.

Intuitivement, si on prends une fenêtre trop grande, on entraîne notre modèle sur une période ou la situation économique était différente, et où les relations entre les variables sont changées. On a un optimum à trouver entre avoir suffisament de données pour entraîner correctement notre modèle, et avoir des données récentes qui vont nous apprendre correctement les relations entre nos variables.

On peut alors, comme pour les paramètres précédents, optimiser la taille $n$ de la fenêtre d'apprentissage : on trouve alors une taille optimale d'environ 35 ans, ce qui signifie qu'au delà de cette taille, notre modèle ira cherchr des données représentatives d'une situation économique plus éloignée de la situation économique actuelle et perdra en performance. On peut voir ce phénomène sur le graphe ci-dessous qui représente l'erreur de prediction (RMSE) sur les 90 derniers mois de notre ensemble d'apprentissage selon la taille de la fenêtre utilisé pour le modèle. Il serait intéressant de poursuivre la courbe vers la droite, mais malheureusement, nous sommes à la limite des données contenues dans le dataset.

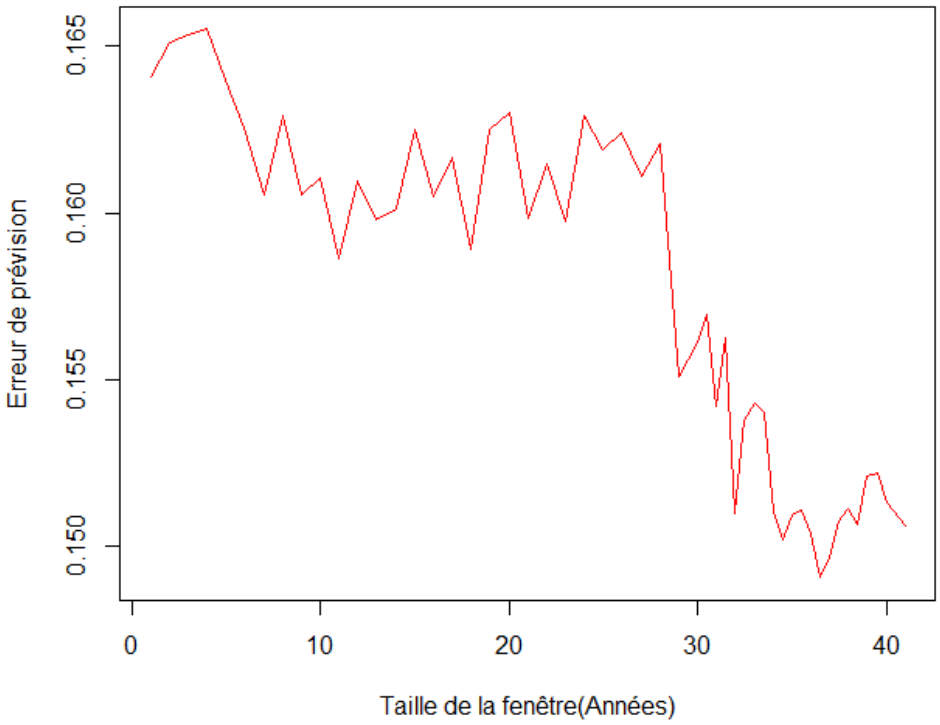



Figure 9 – Erreur en fonction de la taille de la fenêtre

### 2.2.3 Résultats

On remarque ici que malgré une optimisation plutôt longue en terme de temps de calcul, on gagne finalement peu de précision entre une forêt naïve et une forêt à fenêtre optimisée : les variations sont toujours relativement bien suivies, mais le modèle a tendance à être de trop faible amplitude par rapport aux fluctuations réelles.

## 2.3 Modèle GAM incrémental

Pour essayer de mieux capturer ces fluctuations, nous avons décidé de tenter un modèle additif GAM ; cependant, le design de l'équation correspondante et l'optimisation des paramètres est plus difficile ici que dans le cas précédent.

### 2.3.1 Équation du modèle

À l'intérieur même de l'équation d'un modèle GAM, il y a plusieurs paramètres à gérer : le type de base de splines et la dimension de chaque base. Pour éviter un temps de calcul exponentiel en le nombre de variables, nous avons décidé d'effectuer un modèle GAM

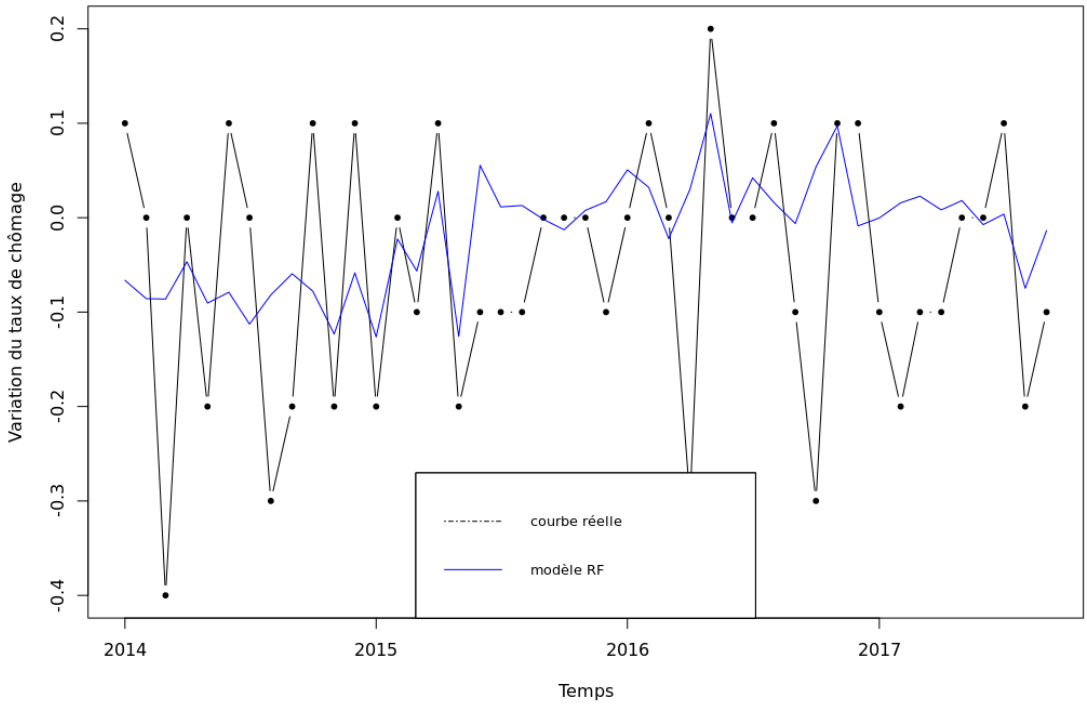



Figure 10 – Prévision du modèle RF

incrémental : une fois les variables ordonnées par ordre d'importance (par exemple via une forêt aléatoire), on les ajoute une par une au modèle, en optimisant les paramètres de la base de spline à chaque fois – ou, le cas échéant, en gardant un modèle linéaire. Cela permet d'optimiser un maximum les variables les plus importantes, tout en réduisant le temps de calcul.

On peut aussi remarquer que cela nous donne une méthode classique de sélection de variables : on obtient ainsi une suite de modèles imbriqués de dimension de plus en plus grande, qu'on peut ensuite sélectionner par de la validation croisée. C'est ce qui est effectué dans la fonction `select.variables` du paquet.

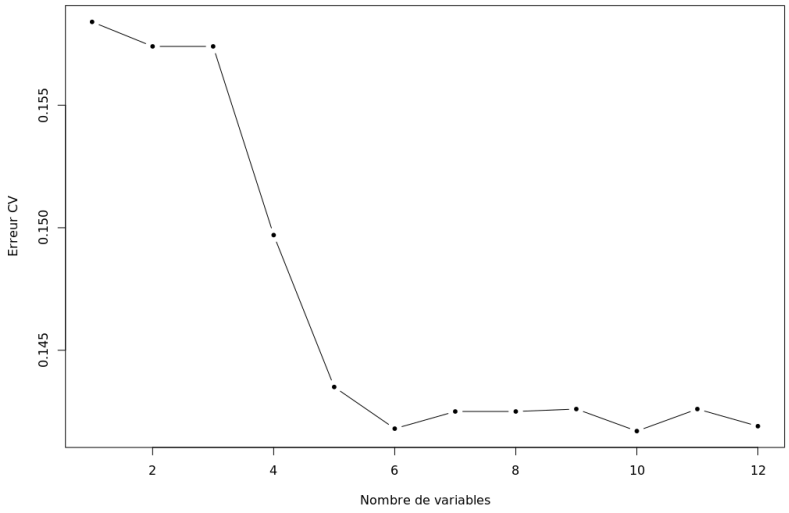



**Figure** 11 – Erreur CV en fonction de la taille du modèle

Ici, cette procédure confirme la sélection faite précédemment : les 6 variables sélectionnées correspondent exactement à celles mentionnées plus haut.

### 2.3.2 Prise en compte du caractère temporel

Tout comme pour les forêts aléatoires, nous avons fait l'hypothèse qu'un modèle économique serait plus précis s'il était entraîné sur une préiode moindre que l'intégralité des données, pour tenir compte de la différence de situation économique entre 1960 et 2014. Ainsi, nous avons ici encore optimisé un paramètre correspondant à la taille de la fenêtre d'entraînement.

De façon peut-être contre-intuitive, le paramètre optimal de fenêtre est bien différent de celui pour les forêts aléatoires : il est ici d'environ 25 ans, soit 10 ans de moins que pour les forêts. Il est relativement difficile d'expliquer ce changement, puisque les deux modèles ont un fonctionnement très différent.

En plus d'un paramètre de fenêtre, on dispose aussi d'un paramètre `weights` dans le modèle GAM, qui contrôle l'importance relative des observations et peut servir à donner plus de poids aux observations récentes. Nous avons choisi de prendre des poids exponentiels ($w(t)$ `/` $\exp(\alpha t)$), et d'optimiser le paramètre $\alpha$ via validation croisée.

On trouve un paramètre optimal $\alpha_0$ d'environ 0.05, ce qui correspond à un poids 4.5 fois plus important sur les observations récentes.

Enfin, nous avons appliqué un modèle ARIMA sur les résidus des prédictions, afin d'améliorer encore la performance du modèle ; il est intéressant de noter que l'algorithme ARIMA appliqué aux résidus des forêts aléatoires n'a eu aucun effet (modèle $(0, 0, 0)$ sélectionné), alors qu'il est relativement efficace sur le modèle GAM.

## 2.4 Comparaison

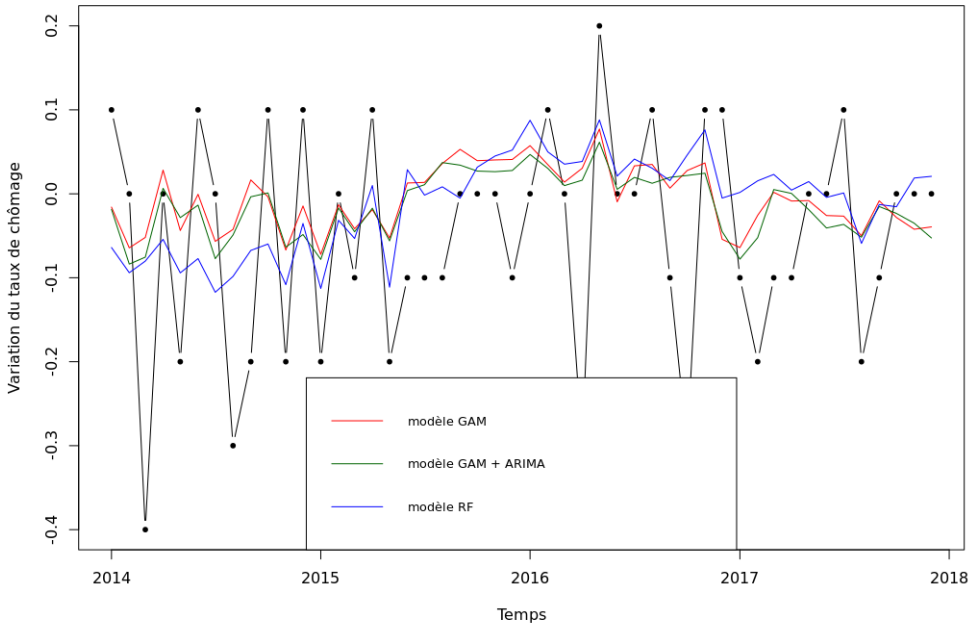



**Figure** 12 – Courbes des différentes prévisions

En comparant les différents modèles, on peut noter plusieurs remarques :

— Les différentes prévisions ont une variance très importante : les forêts ont une composante aléatoire inhérente, et puisqu'on les utilise pour déterminer les importances relatives, cela se transmet aux modèles GAM. En général, les erreurs quadratiques tournent tout de même autout de 0.131.

— Comme pour les modèles précédents, les variations de la variable sont relativement bien récupérées, mais tous les modèles sous-estiment complètement leur amplitude. Cela peut être dû aux arrondis sur les données, qui affectent uniquement l'amplitude et non le sens des variations.

| Modèle | Erreur quadratique moyenne |
|---|---|
| Régression linéaire simple | 0.1457 |
| Forêt aléatoire simple | 0.1338 |
| Forêt aléatoire optimisée | 0.1315 |
| GAM seul | 0.1301 |
| GAM + ARIMA | 0.1276 |

**Table** 1 – Erreurs pour différents modèles

### 2.4.1 Agrégation

Ayant à notre disposition une multitude d'« experts », une idée naturelle serait d'en faire une agrégation. Cependant, lorsque qu'on regarde nos différents modèles, on peut voir qu'ils sont très proches les uns des autres, et qu'une agrégation d'expert n'apporterait probablement pas grand chose, ce qui est d'ailleurs confirmé par le graphique suivant :

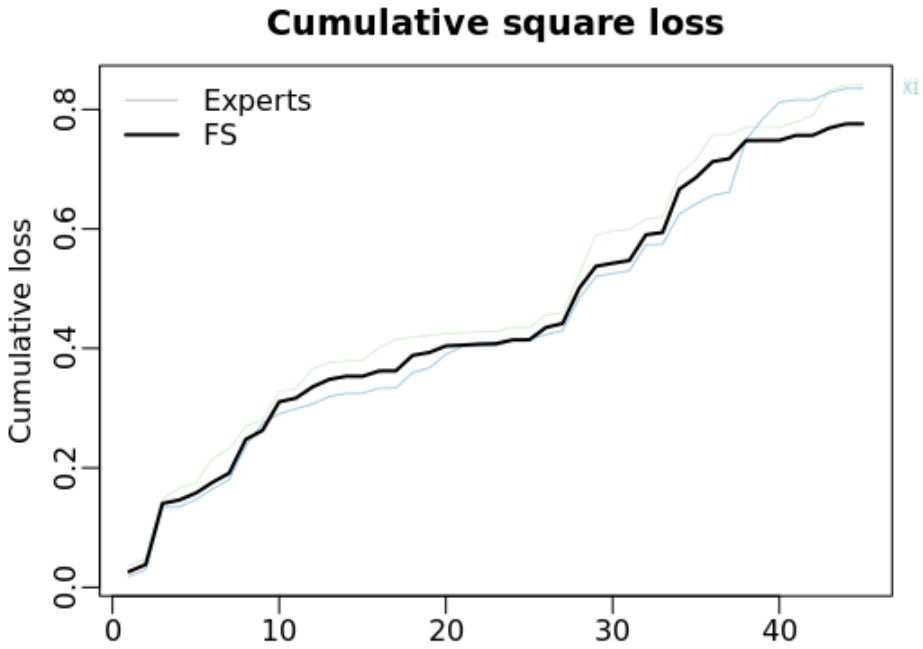



Figure 13 – Erreurs quadratiques des modèles et de l'agrégation

# Conclusion

En résumé, comme nous nous y attendions, même notre meilleur modèle n'était pas très précis. Cependant, nous avons quand même réussi à faire diminuer significativement notre erreur de prévision par rapport à des modèles basiques. De plus, nos dernier modèles ne dépendent que de 6 covariables, ce qui signifie que nous avons réussi à mettre en avant des variables avec un fort pouvoir prédicteur du chômage. C'est un succès quand on sait à quel point il s'agit d'un mécanisme économique complexe, et cela démontre également l'efficacité de nos méthodes de sélection de variable, et de nos heuristiques sur les variables.

De plus, nous avons été limité dans certaines directions de recherche par le nombre d'observations à notre disposition, et il serait donc intéressant de reprendre nos méthodes dans quelques années pour voir si on peut en tirer de nouvelles choses.

Nos différents modèles ont des difficultés à prévoir les grosses variations mensuelles de chômage, et il pourrait être intéressant de chercher de nouveaux modèles ou de nouvelles covariables qui permettraient d'améliorer notre prévision à cet égard.