Projet de Machine Learning

Modéliser et prévoir une épidémie de grippe

Vincent Perreault, Suzanne Sigalla

Sous la direction de Yannig Goude

Université Paris Sud – EDF – le 20 février 2019

Sommaire

I Collecte et description des données **2**
I.1 Données sur la grippe . . . 2
I.1.1 Collecte . . . 2
I.1.2 Visualisation . . . 2
I.2 Données météo . . . 4
I.2.1 Collecte . . . 4
I.2.2 Visualisation . . . 4
I.3 Données de mobilité . . . 5
I.4 Données de population . . . 6
I.5 Concaténation des bases de données . . . 6

II Prévoir une épidémie de grippe **6**
II.1 Séries temporelles . . . 6
II.1.1 Modèle SARIMA par région . . . 6
II.1.2 Un autre modèle : lissage exponentiel double de Holt-Winters . . . 7
II.1.3 Courbes . . . 7
II.2 Régression linéaire, une région . . . 8
II.3 Arbre de décision par région, forêt aléatoire, une région . . . 9
II.3.1 Arbre de décision, une région . . . 9
II.3.2 Forêt aléatoire, une région . . . 10
II.4 Méthodes de régression, une région et ses voisins . . . 11
II.5 Agrégation . . . 12
II.6 Un dernier modèle : GAM . . . 13

Introduction

Dans le cadre d'un projet de machine learning avec objectif de prévision, nous avons choisi de travailler sur la prévision des épidémies de grippe. Nous avons utilisé pour cela les données du réseau Sentinelle. Créé en 1984, le réseau Sentinelle est un réseau de recherche et de veille en France, qui a été développé par l'INSERM et l'université de la Sorbonne. Ce réseau a pour objectif la constitution de larges bases de données de médecine générale et pédiatrique sur 10 maladies, notamment infectieuses, telles que la grippe, la coqueluche... ainsi que le développement d'outils de détection et de prévision d'épidémie. Ce réseau regroupe 1314 médecins libéraux et 116 pédiatres libéraux volontaires, dits "médecins Sentinelles" ; ces médecins transmettent chaque semaine les données des patients touchés par les maladies surveillées ; ces données permettent ensuite d'estimer le taux d'incidence hebdomadaire de chaque maladie surveillée. L'ensemble du réseau est coordonné par l'équipe de recherche de l'institut d'Épidémiologie et de Santé publique. Nous avons choisi d'étudier la grippe car c'est une maladie dont les enjeux en terme de santé publique sont importants, puisqu'elle est la cause de 4000 à 6000 décès chaque année selon l'INSERM.

Nous avons également travaillé à récupérer des données météorologiques dans chaque région française, car les épidémies de grippe présentent une saisonnalité annuelle et sont positivement corrélées aux températures plus froides : en effet, l'air froid et sec favorise la pénétration du virus ; celui-ci résiste mieux à la chaleur qu'au froid ; et enfin, les populations se regroupent davantage en hiver qu'en été et cette promiscuité favorise la propagation.

La grippe étant une maladie très contagieuse, les flux de populations constituent une autre variable explicative naturelle de celle-ci : nous avons donc travaillé à exploiter les données de mobilité de l'INSEE, disponibles de 2006 à 2015, afin d'améliorer nos prédictions.

I Collecte et description des données

I.1 Données sur la grippe

I.1.1 Collecte

Sur le site du réseau Sentinelle, les données hebdomadaires par région et les données hebdomadaires agrégées pour toute la France depuis 1984 sont en accès libre : on dispose ainsi, de 1984 à 2019, par semaine, par région (respectivement pour toute la France), du taux d'incidence brut, i.e. du nombre de nouveaux cas par semaine et par région (respectivement dans toute la France), et du taux d'incidence pour 100 000 habitants ; on dispose également d'un intervalle de confiance pour ces deux variables. Nous ne détaillons pas ici le calcul du taux d'incidence, mais les méthodes utilisées sont précisées à cette page.

I.1.2 Visualisation

Nous souhaitons d'abord visualiser ces données afin de mieux pouvoir les analyser par la suite. Nous travaillons d'abord sur le taux d'incidence total (i.e. sur toute la France). On trace l'évolution du taux d'incidence pour 100 000 habitants au cours du temps :

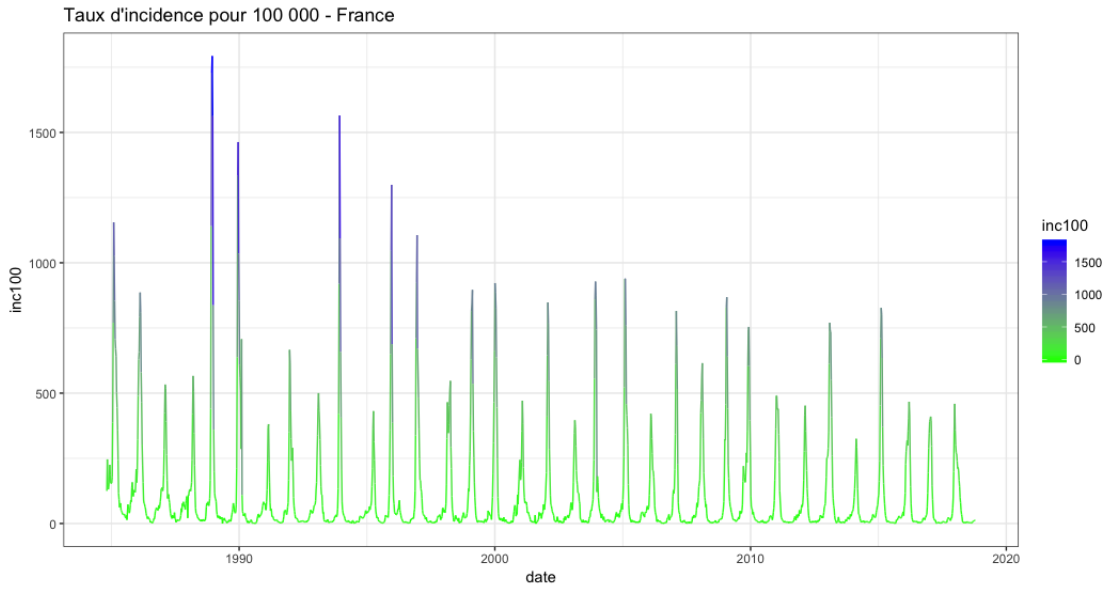


FIGURE 1 – Taux d'incidence de la grippe en France pour 100 000 habitants

Afin de lisser cette courbe, nous appliquons une transformation logarithmique des données. Nous obtenons ainsi :

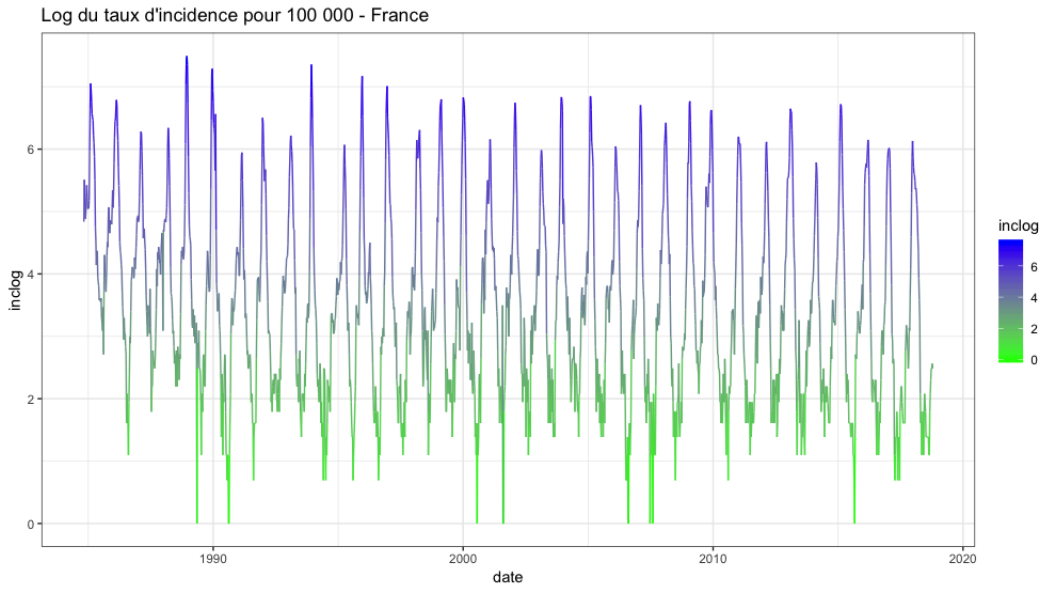


FIGURE 2 – Logarithme du taux d'incidence de la grippe en France pour 100 000 habitants

On observe de façon très claire une saisonnalité. Par la suite, nous travaillerons davantage sur le logarithme du taux d'incidence, que nous désignerons abusivement par taux d'incidence. Afin de mieux expliquer celui-ci, nous avons entrepris de collecter des données météorologiques par région, en particulier la température.

I.2 Données météo

I.2.1 Collecte

Les données sur la grippe dont nous disposons sont régionales. Nous avons constaté plus haut que les épidémies de grippe constituent un phénomène saisonnier, observé généralement entre novembre et mars ; en particulier, la température semble être une variable explicative pertinente du phénomène. Nous avons donc collecté des données de température de 1984 à 2018 sur toute la France. Nous avons utilisé `Python` afin de scrapper les données du site Info Climat. Nous avons ainsi récupéré une base de données par région, comportant les minimum, maximum et moyenne par semaine des température minimales et maximales d'une grande ville dans chaque région, sur toute la période considérée (sauf valeurs manquantes pour certaines villes à certaines dates).

I.2.2 Visualisation

Nous travaillons tout d'abord région par région pour plus de précision : pour une région, nous traçons l'évolution conjointe de la température moyenne et du logarithme du taux d'incidence, chacun normalisé. On obtient, pour l'Ile-de-France (l'allure des courbes est similaire pour toutes les régions, excepté la Champagne-Ardenne, pour laquelle trop de données de température sont manquantes) :

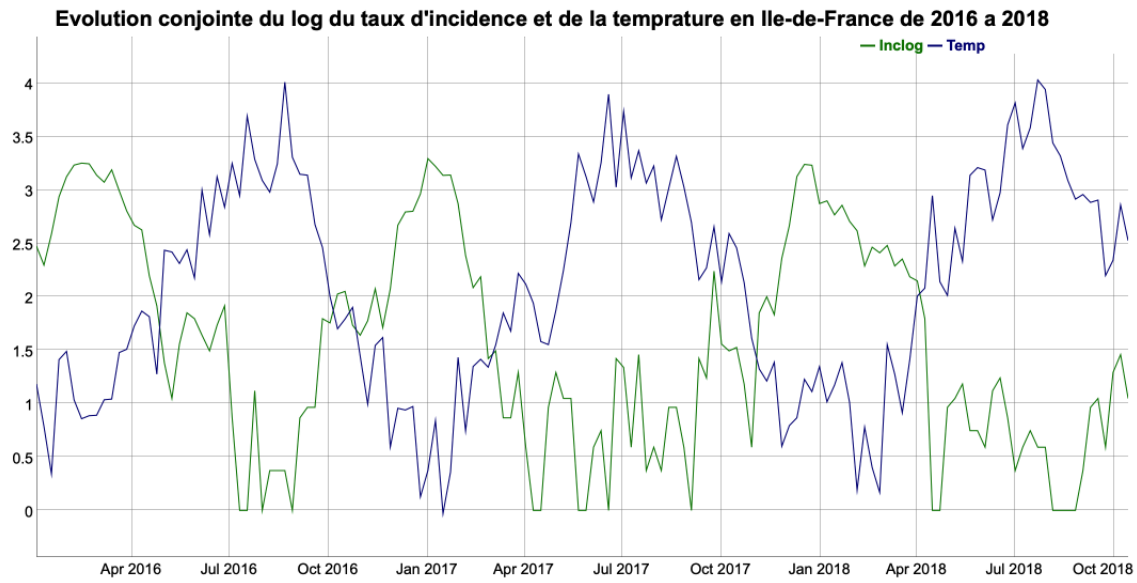


FIGURE 3 – Évolution conjointe de la température moyenne et du logarithme du taux d'incidence normalisés, en Ile-de-France, de 2016 à 2018

On note que le logarithme du taux d'incidence est compris entre 0 et 10 environ ; sa moyenne sur l'ensemble des données est de 2.5 et sa médiane de 2.6.

Comme l'on pouvait s'y attendre, la température moyenne et le taux d'incidence sont en opposition de phase ; les deux séries semblent bien corrélées. Nous allons à présent chercher à exploiter cette corrélation dans un objectif de prévision.

I.3 Données de mobilité

Il est naturel de penser que les déplacements des gens influent sur la propagation du virus sur le territoire. Les données sur les déplacements disponibles dans les bases de données nationales pourraient nous donner une information supplémentaire. Les deux types de mobilité recensées par l'État via l'INSEE sont les mobilités professionnelles, c'est-à-dire les données de déplacements des individus salariés de leur commune de domicile jusqu'à la commune de leur lieu de travail, et les mobilités scolaires, à savoir les mêmes données pour les élèves d'école primaire ou secondaire vis-à-vis de la commune de leur école.

Ces deux catégories de données ne sont disponibles que de 2006 à 2015 sur le site internet l'INSEE. Ces bases de données ont comme échelle géographique les communes. Seulement, nos données cibles, i.e. les données d'épidémies de grippes, sont elles à l'échelle régionale. Il y avait donc un besoin de convertir ces données de mobilité de l'échelle communale à l'échelle régionale.

Initialement, chaque base contenait les nombres d'individus allant d'une certaine commune à une autre, par an et par commune, pour le travail dans une base et pour l'école dans l'autre. Ces chiffres en soi n'étant pas très intéressants, nous les avons convertis en proportions en les normalisant par le nombre total de personnes allant d'une certaine commune à toutes les autres (séparément pour les deux bases, i.e. en prenant soin de séparer flux professionnels et flux scolaires). Cette transformation nous a permis d'obtenir la répartition des flux humains (pour le travail et pour l'école) pour chaque commune vers les communes d'intérêt. Il était ensuite nécessaire d'agréger ces taux au niveau régional, en transformant les flux de communes à communes en flux de région à région. Toutefois, il y avait deux grands problèmes à cette agrégation : d'abord, la grande majorité des flux provenant d'une commune se déversaient dans des communes de la même région et ainsi, l'information recherchée s'en trouve diluée, chaque flot inter-régional étant marginal ; ensuite, la manière d'agréger ces flux pour toutes les communes d'une même région, sans perdre trop d'information, est un problème délicat.

Pour le premier problème, bien que notre objectif initial était d'utiliser l'aspect réseau de flux pour modéliser la propagation du virus par les déplacements pour l'école ou pour le travail, il est devenu rapidement clair que cette propagation à l'échelle régionale ne nous donnait simplement pas une information supplémentaire pertinente. En effet, l'échelle beaucoup plus grande lui donne un aspect très tangentiel et d'effet très limité par rapport à ce qui est en jeu a priori. Toutefois, par la façon d'agréger, on peut tout de même récupérer beaucoup d'informations a priori importantes sur les déplacements intra-régionaux et comment ils se différencient parmi les communes. Cependant, les données inter-régionales que nous avons calculées par la suite n'ont finalement pas été utilisées.

Nous avons donc agrégé nos données communales en données régionales : ce processus nous a fourni 44 statistiques. Nous avons cherché à réduire la dimension de notre problème : pour cela, nous avons effectué une décomposition en composantes principales (abrégé par PCA). Ainsi, pour chaque type de mobilités, pour chaque région de départ, pour chaque région d'arrivée (même-commune, même-région et les autres régions), nous avions 11 statistiques décrivant différemment les tels taux de flux parmi les communes comprises dans la région de départ. Par le PCA, de ces 11 statistiques, nous n'en avons gardé que 3, qui expliquent 95 % de la variance. Finalement, nous obtenons 12 variables explicatives par région pour chaque année de 2006 à 2015.

I.4 Données de population

Les données de population fournies avec les données de mobilité ne sont disponibles que de 2006 à 2015, ce qui ne nous a pas permis de couvrir l'ensemble de nos données cibles de 1984 à 2018. Nous avons utilisé une autre base de l'INSEE surs les populations par tranche d'âge de 1975 à 2018.

Les populations sont divisées en les tranches 0-19 ans, 20-39 ans, 40-59 ans, 60-74 ans, 75 ans et plus et en total. Les populations étant toutes données en nombre d'individus et ces nombres n'ayant pas autant d'intérêt pour les tranches d'âge que leur proportion par rapport à la démographie totale, nous avons pour chaque tranche d'âge converti cette variable en taux de tranche d'âge par rapport à la population totale.

Le passage des anciennes régions aux nouvelles régions a été source d'ennuis. Nos données cibles respectent les anciennes régions et ce même jusqu'en 2018. Ainsi, nous avons dû considérer les tranches d'âges comme étant homogènes sur une nouvelle grande région pour les 3 dernières années et nous avons aussi dû inférer les populations totales des sous-régions qui ont été agrégées en ces nouvelles plus grandes régions. Pour ce faire, la proportion moyenne de population totale de ces anciennes régions sur les plus grandes ont été évaluées de 2006 à 2015 et cette proportion a été multipliée par les populations totales reportées pour les plus grandes régions pour les 3 dernières années.

I.5 Concaténation des bases de données

Une fois les données météorologiques et d'épidémie concatenées, on a pu ajouter les données de populations, c'est-à-dire les proportions de la population âgées de 0-19 ans, 20-39 ans, 40-59 ans, 60-74 ans, 75 ans et plus ainsi que les populations totales en nombre d'individus. La concaténation avec les données de météo a donné lieu à une variable `"week"` qui encode la semaine de l'année de 1 à 52 (après ajustements de la semaine qui dépasse dans l'année suivante ramenée à la 52e de l'année précédente). Cette variable `"week"` a permis une interpolation simple des données annuelles de populations au cours de l'année. En effet, les données de populations sont prises au cours de l'année et sont faites de telle sorte qu'elles approximent bien l'état de la population à la moitié de l'année (la semaine 26). Ainsi, pour chaque semaine, les données de populations par tranche et en total ont été interpolées entre les deux moitiés d'années autour de la semaine considérée. Cette interpolation plus la disponibilité des données de 1975 à 2018 a fait en sorte que ces données ne présentent pas de valeurs manquantes dans notre base de données.

Ensuite, de la même façon que pour les populations, les 12 variables de mobilités (les 3 composantes principales pour même-commune et même-région pour scolaire et professionnel) ont été interpolées entre la 26e semaine de 2006 et la 26e de 2015. La grande majorité des données considérées sont manquantes.

II Prévoir une épidémie de grippe

Nous séparons notre échantillon en deux sous-échantillons, un échantillon d'entraînement, correspondant aux années 1984 à 2015, et un échantillon test, correspondant aux années 2016 à 2018.

II.1 Séries temporelles

II.1.1 Modèle SARIMA par région

Étant donné l'allure de nos courbes, il est naturel de chercher à ajuster à nos données une série temporelle. Nous travaillons par région. Nous procédons comme suit afin d'ajuster notre modèle : nous nous restreignons à l'échantillon d'entraînement ; nous procédons d'abord à une différenciation saisonnière (au lag 52 donc, puisque nos données sont hebdomadaires), puis à une différenciation simple. Dans l'ensemble, pour

chaque région, ces opérations permettent d'obtenir une série à l'allure stationnaire. On procède à un test de Dickey-Fuller augmenté afin de rejeter l'hypothèse de non-stationnarité. On observe ensuite le graphe des autocorrélations et autocorrélations partielles afin d'ajuster les ordres maximum p, q, P, Q d'un modèle SARIMA, puis on utilise la fonction `"auto.arima"` du package `forecast` afin de sélectionner le modèle minimisant un critère AIC ou BIC. En appliquant ces étapes successives par exemple pour l'Ile-de-France, on obtient un modèle SARIMA$(p = 3, d = 1, q = 0)(P = 1, D = 1, Q = 0)$ de saisonnalité $s = 52$. Ensuite, nous utilisons ce modèle pour prédire le taux d'incidence à horizon la longueur de nos données test ; nous comparons ensuite cette prévision aux données réellement observées. Pour l'Ile-de-France, on obtient un RMSE de 2.91 et MASE de 3.71.

II.1.2 Un autre modèle : lissage exponentiel double de Holt-Winters

Nous implémentons un second modèle, le modèle double de Holt-Winters multiplicatif ; ce modèle permet de lisser les données dans le temps, tout en conservant la saisonnalité. De même que précédemment, on utilise ensuite ce modèle pour établir une prévision, que nous comparons aux données réellement observées. Nous obtenons pour l'Ile-de-France un RMSE de 2.3 et une MASE de 2.95, ce qui constitue bien une amélioration du premier modèle. Nous n'avons pas implémenté d'autres modèles de Holt-Winters car cela ne nous semblait pas pertinent, au vu de la forte saisonnalité des données.

II.1.3 Courbes

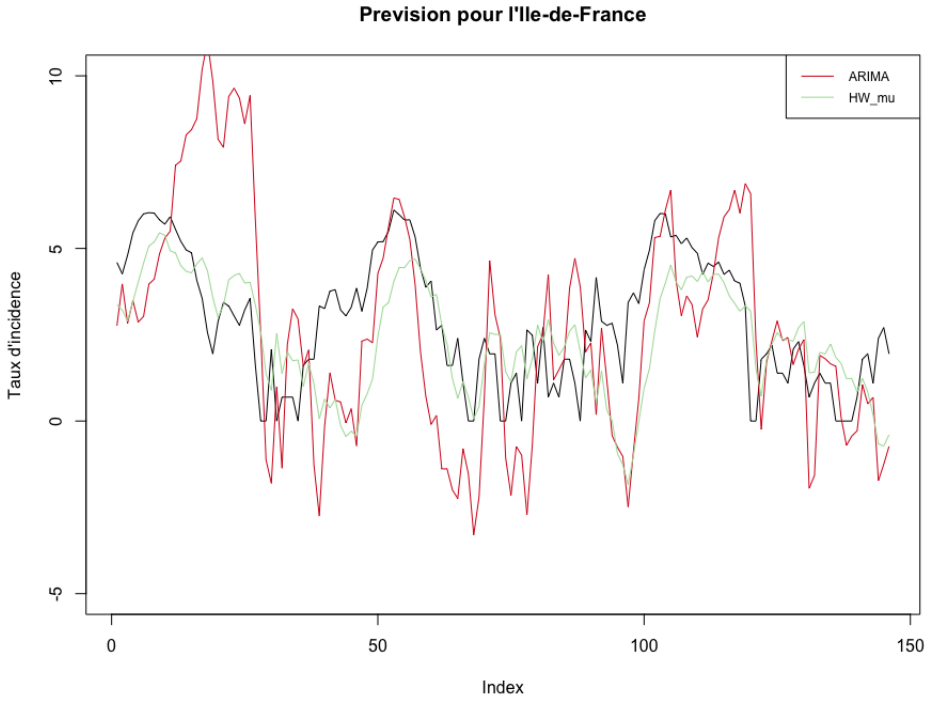


FIGURE 4 – Prévisions des modèles respectifs SARIMA$(p = 3, d = 1, q = 0)(P = 1, D = 1, Q = 0)$ et Holt-Winters double, multiplicatif, avec saisonnalité, sur les données d'Ile-de-France, sur l'échantillon test restreint aux années 2016 à 2018 – la courbe noire correspond aux données réellement observées.

On observe effectivement que le modèle à lissage exponentiel prédit les données avec davantage de précision.

II.2 Régression linéaire, une région

Il est naturel d'étudier les séries épidémiques des taux d'incidence comme une série temporelle, mais cela ne permet pas de prendre en compte des variables explicatives aussi essentielles que la température, qui peuvent apporter une grande information. Nous allons donc travailler à présent sur d'autres modèles, qui permettent la prise en compte de variables explicatives. On travaillera d'abord région par région, en se concentrant sur l'Ile-de-France en tant qu'exemple.

Nous effectuons d'abord, pour chaque région, deux régressions linéaires pénalisées, respectivement LASSO et Ridge. Pour chaque région, nous entraînons notre modèle sur notre échantillon d'entraînement, puis utilisons le modèle afin de prédire le taux d'incidence à partir de trois régresseurs : le lag du taux d'incidence, la température moyenne et la population. On calcule pour chaque région l'erreur RMSE et l'erreur MASE, plus appropriée ici que l'erreur MAPE, puisque le taux d'incidence n'est pas strictement positif. On donne à nouveau les résultats pour la région Île-de-France ; à part pour la région Champagne-Ardennes pour laquelle il manque trop de données, les allures des courbes sont les mêmes.

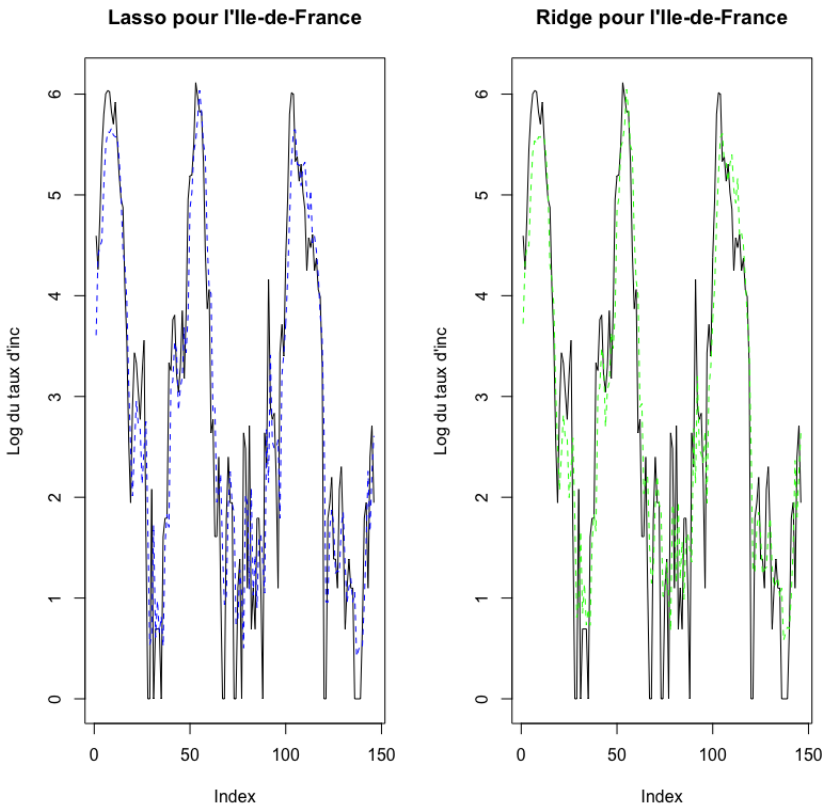


FIGURE 5 – Prévisions des modèles respectifs Ridge et LASSO, sur les données d'Ile-de-France, sur l'échantillon test restreint aux années 2016 à 2018. La courbe réelle est en noire, les courbes prédites par les modèles respectifs en pointillés et en couleur.

On observe que le modèle semble prédire correctement le taux d'incidence en hiver, mais moins bien en été. Pour l'Ile-de-France, on obtient :

— LASSO : RMSE de 0.79 et MAPE de 0.93 ;

— Ridge : RMSE de 0.79 et MAPE de 0.97.

Ces résultats sont bien meilleurs que ceux obtenus précédemment pour les séries temporelles.

II.3 Arbre de décision par région, forêt aléatoire, une région

II.3.1 Arbre de décision, une région

Suite logique d'un modèle de régression linéaire, nous avons également tenté d'implémenter des arbres de décision pour modéliser nos données. Nous travaillons toujours sur le taux d'incidence ; nous considérons comme variables explicatives plusieurs variables liées à la température par région (les minimum, maximum et moyenne par semaine des température minimales et maximales pour une grande ville par région), ainsi que le taux d'incidence retardé d'une semaine et la population par région. Cette dernière variable n'est que peu informative puisque c'est une moyenne annuelle, mais nous la conservons tout de même. Nous étudions à nouveau la région Ile-de-France. Nous ajustons un arbre sur nos données d'entraînement et opérons une prédiction à partir de l'échantillon test. Nous traçons ci-dessous les courbes respectives des taux d'incidence réel et prédit en utilisant l'arbre précédemment ajusté :

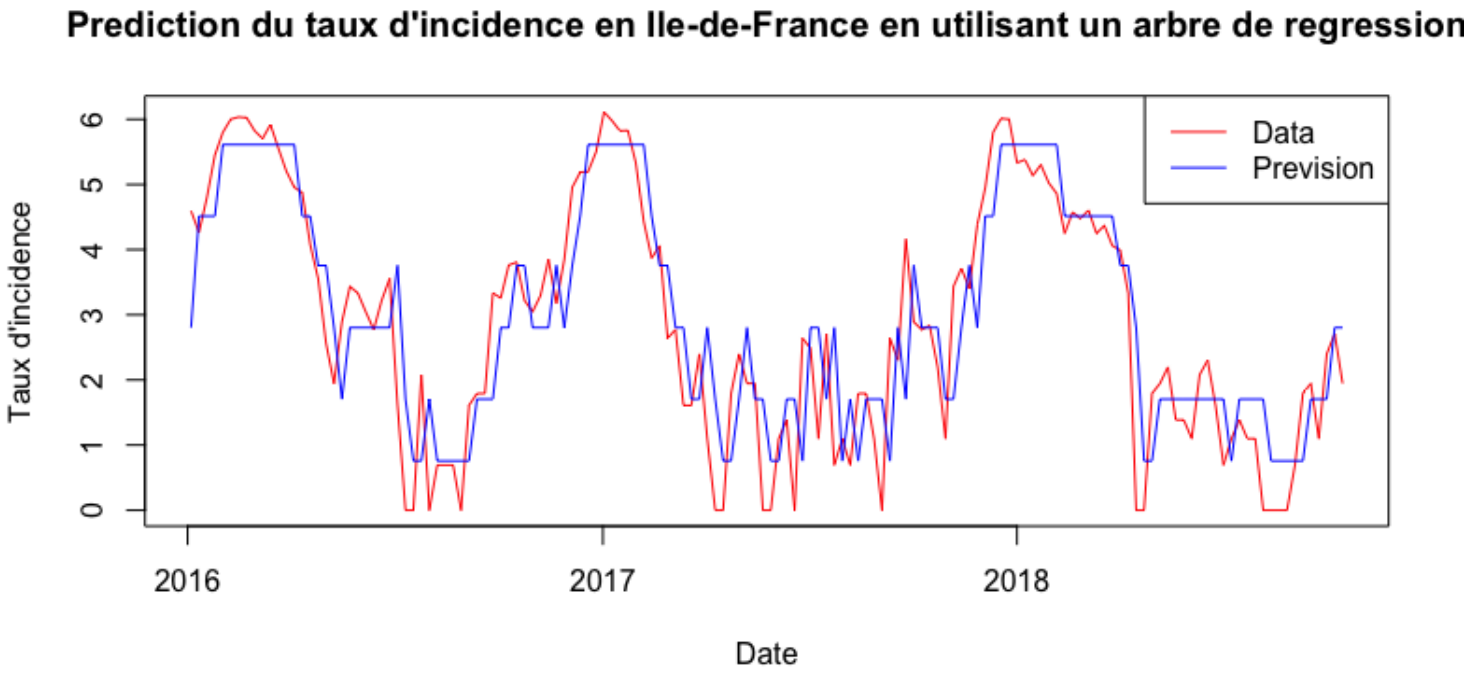


FIGURE 6 – Prévision par arbre de décision sur les données d'Ile-de-France, sur l'échantillon test restreint aux années 2016 à 2018.

On constate que ce modèle n'est pas apte à bien prédire les valeurs les plus extrêmes. On obtient pour l'Ile-de-France une erreur RMSE de 1.28 et MAPE de 1.75. Ces erreurs sont plus élevées que dans le cas de la régression LASSO ou Ridge, mais restent meilleures que celles obtenues en modélisant nos données par séries temporelles. Nous essayons donc d'ajuster un meilleur modèle en implémentant une forêt aléatoire.

II.3.2 Forêt aléatoire, une région

Nous conservons les mêmes variables explicatives que précédemment et implémentons, pour chaque région, une forêt de 500 arbres aléatoires. Nous traçons de nouveau les courbes respectives des taux d'incidence réel et prédit :

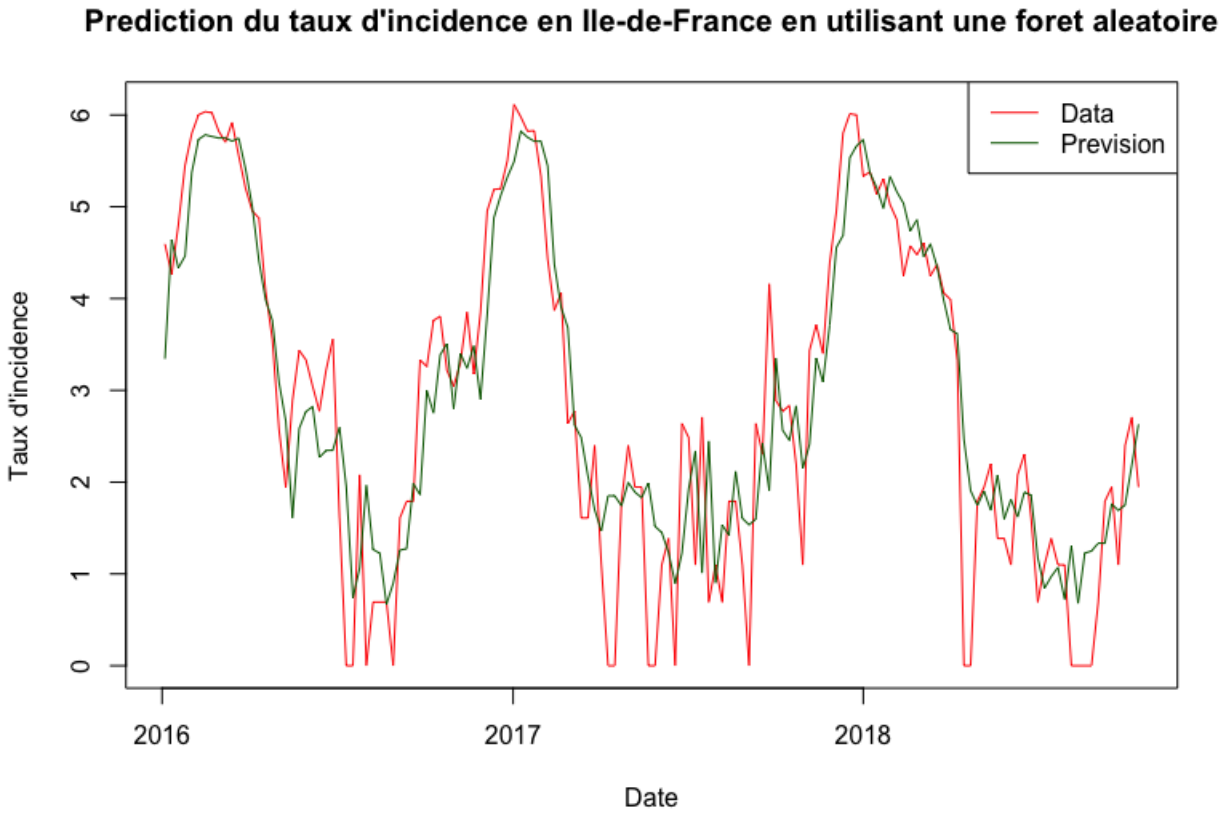


FIGURE 7 – Prévision par random forest sur les données d'Ile-de-France, sur l'échantillon test restreint aux années 2016 à 2018.

Ce modèle est plus proche des données réelles, notamment il s'ajuste mieux aux pics d'incidence élevée et moins bien aux pics d'incidence basse, ces derniers pics étant davantage bruités ; cependant, d'un point de vue pratique, l'essentiel est justement que le modèle prédise les pics d'incidence élevée puisque l'objectif sous-jacent est entre autres de mieux anticiper les périodes où l'épidémie est la plus virulente, plutôt que les périodes où la grippe ne touche personne. Pour l'Ile-de-France, on obtient une erreur RMSE de 0.79 et une erreur MASE de 0.94. Ces résultats sont du même ordre que ceux obtenus précédemment pour les régressions LASSO et Ridge.

II.4 Méthodes de régression, une région et ses voisins

Jusque là, nous n'avons considéré le problème que d'un point de vue exclusivement temporel, région par région. Cependant, la propagation d'une épidémie a bien évidemment un caractère spatial. Nous souhaitons à présent introduire cette dimension, en considérant, pour chaque région, les régions qui lui sont voisines. Par "régions voisines", on entend régions qui ont une frontière commune. Ainsi, pour les méthodes utilisant des variables explicatives considérées jusqu'à présent, on souhaite intégrer, en plus du taux d'incidence retardé d'une semaine de la région considéré, le taux d'incidence retardé d'une semaine des régions qui lui sont voisines. On travaille à nouveau sur l'Ile-de-France en guise d'exemple ; les régions voisines de l'Ile-de-France sont la Picardie, la Haute-Normandie, le Centre, la Bourgogne et la Champagne-Ardenne.

Nous effectuons à nouveau deux régressions LASSO et Ridge, sur les données de la région Ile-de-France divisé en un échantillon d'entraînement et un échantillon test, données auxquelles on ajoute le taux d'incidence retardé de ses régions voisines. On obtient alors :

— pour le LASSO, une erreur RMSE de 0.78 et une erreur MASE de 0.95 ;

— pour la régression Ridge, une erreur RMSE de 0.78 et une erreur MASE de 0.96 ;

Comparativement, la prise en comte des voisins améliore très légèrement le résultat dans le cadre de la régression Ridge, vis-à-vis des erreurs RMSE et MASE ; mais cela détériore légèrement le résultat pour le LASSO si l'on s'en tient à l'erreur MASE. C'est donc décevant : la prise en compte de la dimension spatiale ne semble pas, a priori, améliorer le modèle. Cependant, nous tentons d'ajuster également un arbre de régression en prenant cette dimension spatiale en compte. On obtient, pour un arbre de régression, une erreur RMSE de 1.28 et une erreur MASE de 1.75. Le fait d'inclure les résultats de régions voisines n'a donc pas d'influence. Enfin, en ajustant une forêt aléatoire, on obtient une erreur RMSE de 0.80 et une erreur MASE de 0.96. La prise en compte des régions voisines n'induit donc pas une amélioration du modèle. On peut le comprendre de la manière suivante : il est probable que le découpage spatial soitrev trop grossier pour pouvoir réellement considérer l'effet de contagion d'une région à une autre.

II.5 Agrégation

On décide à présent d'agréger les experts introduits jusqu'à présent. On utilise pour cela le package `Opera`. On travaille à nouveau région par région, on donne à nouveau l'exemple de l'Ile-de-France. On agrège alors les six experts que l'on a déjà considérés, à savoir la régression LASSO, la régression Ridge, l'arbre de décision, la forêt aléatoire, le modèle ARIMA et le modèle ARIMA à poids exponentiels. On obtient la figure ci-dessous.

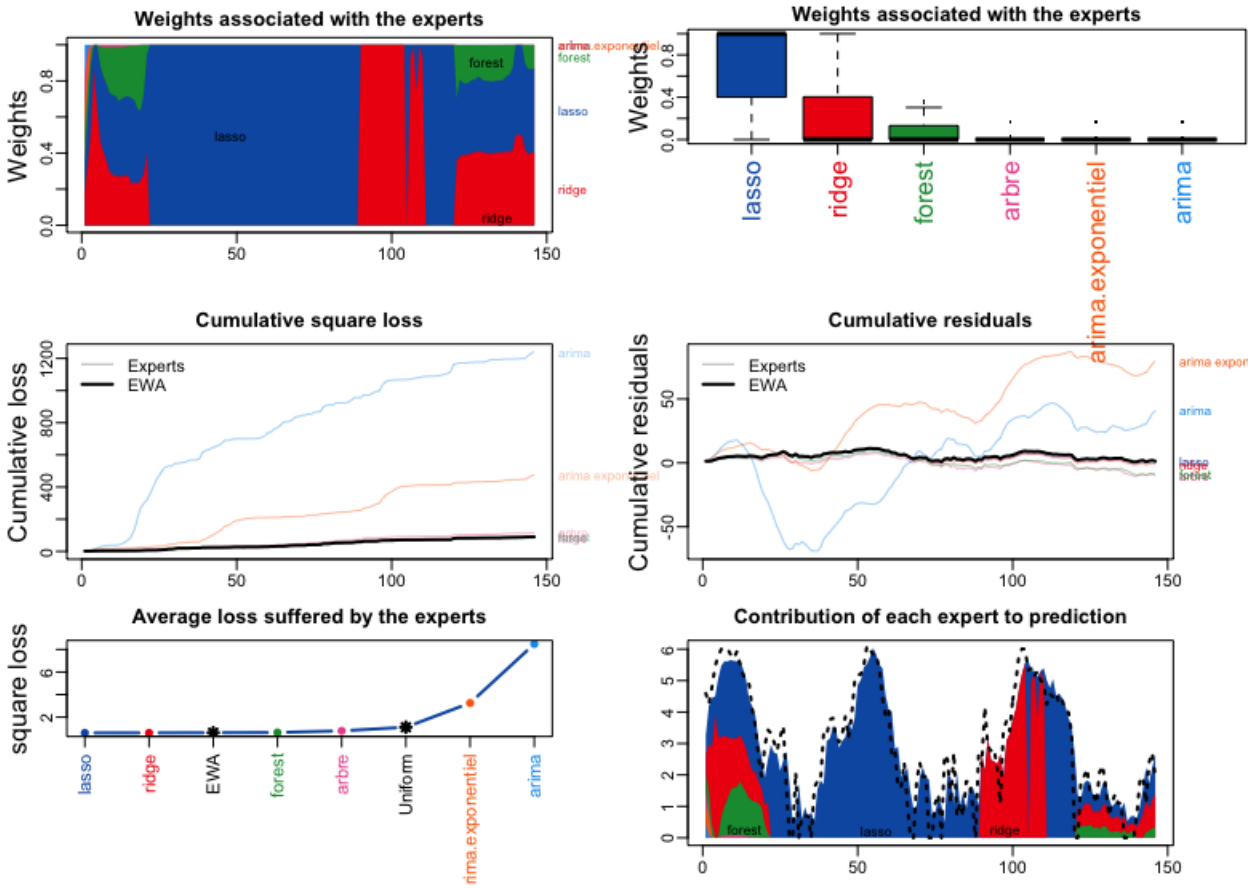


FIGURE 8 – Agrégation de six experts pour la prévision du taux d'incidence en Ile-de-France : LASSO, Ridge, arbre de décision, forêt aléatoire, ARIMA, ARIMA à poids exponentiel

Il est intéressant de noter qu'au début, le LASSO éclipse presque entièrement tous les autres experts, puis est éclipsé par le prédicteur Ridge, puis "revient" un peu plus tard – sans disparaître. Les prédicteurs les plus déterminants dans ce modèle sont les prédicteurs LASSO, Ridge et la forêt aléatoire, et les poids des modèles ARIMA sont très faibles comparativement aux autres experts : on peut interpréter cela comme l'indice d'une forte corrélation entre le taux d'incidence et les variables explicatives de nos modèles.

L'erreur RMSE du modèle d'experts agrégés à poids exponentiels est de 0.76, ce qui est meilleur que notre plus petite erreur RMSE obtenue jusqu'à présent ! (à savoir celle du LASSO). Finalement, en superposant les courbes de prédiction et en traçant celle du modèle agrégé, on obtient :

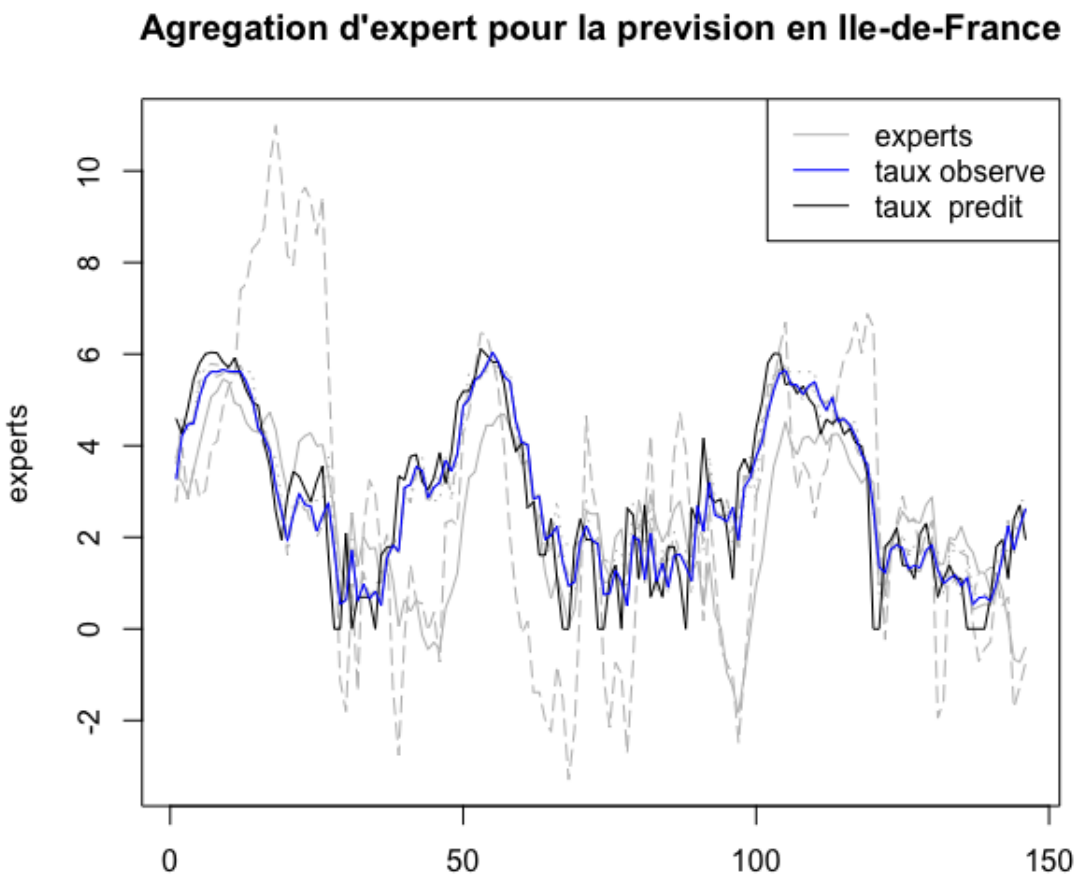


FIGURE 9 – Prédiction par agrégation de six experts pour la prévision du taux d'incidence en Ile-de-France : LASSO, Ridge, arbre de décision, forêt aléatoire, ARIMA, ARIMA à poids exponentiel

On observe que la courbe prédite est proche de la courbe observée ; on a à nouveau que les pics hauts sont mieux modélisés que les pics bas, mais de façon plus atténuée que pour l'arbre de régression.

II.6 Un dernier modèle : GAM

Vu la grande quantité et variété de co-variables dont nous disposons pour expliquer notre quantité cible, nous avons enfin choisi d'appliquer l'algorithme GAM à notre base de données. Nous travaillons ici sur le taux d'incidence non transformé (i.e. sur la variable initiale et non la variable logarithmique). Nous souhaitons expliquer le taux d'incidence par les différentes statistiques de température ; les variables obtenues dans le cadre de notre PCA sur les mobilités professionnelles et scolaires ; la semaine ; la proportion de population âgée de 0 à 19 ans et celle âgée de 60 ans et plus, car ce sont les plus touchées par la grippe ; enfin, la population totale. Plusieurs de ces variables explicatives comportent des données manquantes. Par conséquent, nous avons construit le GAM en trois étapes. Le premier GAM (auquel on réfère par la suite par "GAM1") est évaluable pour toute semaine dans la base de donnée et s'évalue à partir des variables

mentionnées ci-dessus exceptées les données de température. La deuxième couche, entraînée sur les résidus du premier GAM là où les données sont disponibles, spécifie la prédiction par rapport aux statistiques de température. La troisième et dernière couche ne peut être utilisée qu'entre 2006 et 2015. Il s'agit de la prédiction supplémentaire sur les résidus restant des deux premiers (où le résidu initial est gardé là où les données disponibles ne permettent pas d'évaluer la deuxième couche) faite sur les 12 statistiques de mobilités mentionnées précédemment. Une telle construction nous permet de toujours pouvoir utiliser au moins une couche ("GAM1") pour faire une prédiction et d'en plus pouvoir la raffiner quand la météo est disponible ("GAM2") et/ou quand la mobilité est disponible ("GAM3").

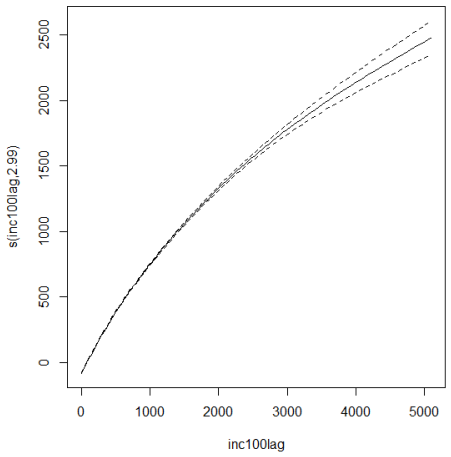
L'apprentissage a été fait dans l'ordre au sens où les paramètres du "GAM1" ont été ajustés avant les paramètres du "GAM2" sur les résidus et ainsi de suite pour le "GAM3". Les paramètres ont également été appris une famille de données à la fois, c'est-à-dire, pour le "GAM1", les paramètres (le nombre de noeud k et la famille des splines) ont été appris, dans l'ordre, pour le taux d'incidence retardé d'une semaine, puis pour la semaine, la population totale (non significative et donc enlevée du modèle par la suite), la proportion dans la tranche d'âge 0-19 ans "Pop0_19" et enfin la proportion dans les tranches d'âges 60 ans et plus. Pour le "GAM2", les paramètres ont été appris, dans l'ordre, simultanément pour les 3 statistiques pour la température minimale (`TMin.min`, `TMin.moy`, `TMin.max` ayant tous les mêmes paramètres), puis simultanément pour les 3 statistiques pour la température maximale (`TMax.min`, `TMax.moy`, `TMax.max`). Pour le "GAM3", les paramètres ont été appris, dans l'ordre, simultanément pour les 4 statistiques pour la 1ere composante principale (`Mobilite_Sc_MCm_1`, `Mobilite_Sc_MRg_1`, `Mobilite_Pr_MCm_1` et `Mobilite_Pr_MRg_1` ayant tous les mêmes paramètres que `Mobilite_Pr_MRg_1` était significatif alors c'est le seul qui a été gardé pour la suite), simultanément pour les 4 statistiques pour la 2e composante principale (`Mobilite_Sc_MCm_2`, `Mobilite_Sc_MRg_2`, `Mobilite_Pr_MCm_2` et `Mobilite_Pr_MRg_2` ayant tous les mêmes paramètres ; aucune n'était significative et donc aucune n'a été gardée pour la suite) ainsi que simultanément pour les 4 statistiques pour la 3e composante principale (`Mobilite_Sc_MCm_3`, `Mobilite_Sc_MRg_3`, `Mobilite_Pr_MCm_3` et `Mobilite_Pr_MRg_3` ayant tous les mêmes paramètres ; aucune n'était significative et donc aucune n'a été gardée pour la suite).

Les paramètres ont été appris par validation croisée à 10 blocs où chaque combinaison possible des k et des bases de splines est considérée. Le nombre de "noeuds" k considéré a toujours d'abord été éprouvé parmi la sélection $\{3, 5, 10, 15, 20\}$ puis par incrément de 1 entre des bornes obtenues après le premier tour. Les bases de splines considérées (hormis exception) ont toujours été dans un premier temps les splines de régression cubiques, les splines de régression cubiques avec rétrécissement et les splines de régression de plaque mince. Après le premier tour, une sous-sélection de ces derniers était considérée. Pour la co-variable de semaine, le caractère périodique devait être retrouvé dans la base de splines utilisée et donc les splines de régression cubiques cycliques ainsi que les P-splines cycliques ont été essayées.

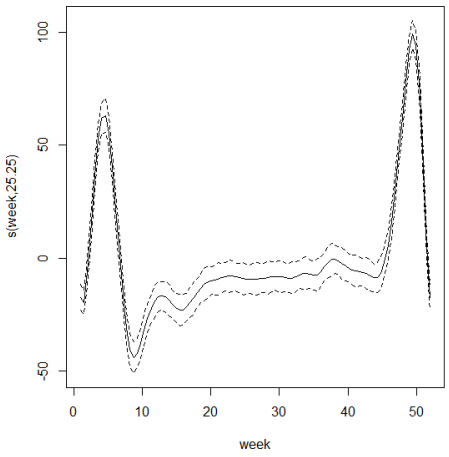
L'entraînement a été réalisé sur la base de données jusqu'à l'année 2014 incluse, ce qui correspond à à peu près 90 % des données pour toutes les couches considérées (en terme de données disponibles à l'évaluation). Le dernier bloc a été utilisé pour l'évaluation. Toutefois, pour réaliser l'évaluation finale, les données exactes pour les températures ont été utilisées (ne pouvant pas fournir des estimations de l'époque) ainsi que pour la mobilité et la population (qui seraient autrement extrapolées des données passées). Nous précisons que la variable cible (le taux d'incidence) a une moyenne d'environ 90 et un écart-type de 215 ; RMSE résultant sur les 3 années restantes est d'environ 65. À défaut du MAPE – puisque nos données valent fréquemment 0 – si on considère la moyenne des différences absolues entre la cible et l'estimation GAM renormalisée par la moyenne du taux d'incidence, nous obtenons une différence moyenne de 60 % de la moyenne de la variable à estimer.

Si l'on inspecte maintenant le modèle obtenu, nous nous attendons à reconnaître certains comportements. Dans la figure 10, on retrouve la première couche de notre GAM à travers les fonctions lisses que l'algorithme a appris pour nos différentes co-variables. De la même façon, dans les figures 12 et 11, on retrouve respectivement les fonctions lisses apprises pour la deuxième et la troisième couche de notre GAM.

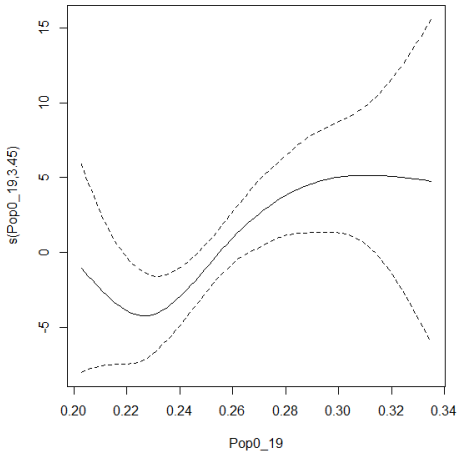
Dans la première sous-figure, on peut remarquer que les épidémies semblent dépendre proportionnellement des épidémies décalées d'une semaine, avec un adoucissement de l'effet pour de grandes valeurs. Cela n'est pas surprenant, d'autant plus que la durée pendant laquelle un malade est contagieux est de l'ordre de grandeur d'une semaine. On voit aussi dans la sous-figure à sa droite l'influence de l'hiver aux extrémités du calendrier. Dans les sous-figures juste en-dessous, on peut voir que les jeunes (via l'école a priori) et les personnes âgées (du fait de leur système immunitaire) sont ceux qui contribuent le plus à l'accroissement du nombre de malades.



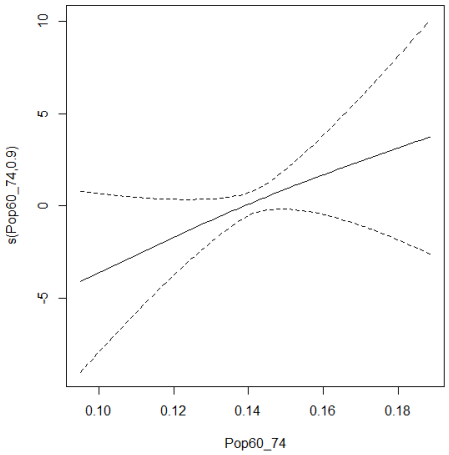
(a) Taux d'incidence retardé d'une semaine



(b) Semaine de l'année



(c) Proportion dans la tranche d'âge 0-19 ans



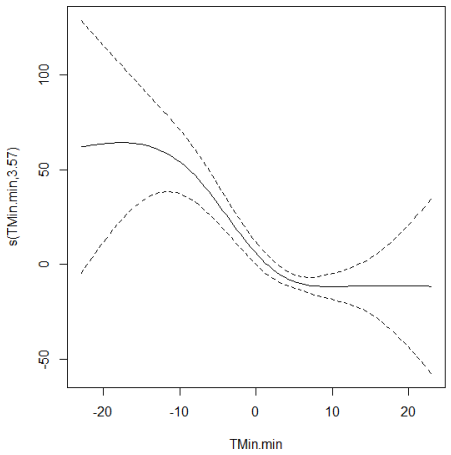
(d) Proportion dans la tranche d'âge 60 ans et plus

FIGURE 10 – Dépendance apprise entre le taux d'incidence et les co-variables.

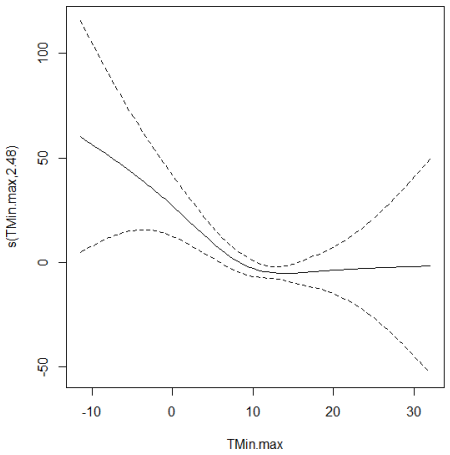
Dans la deuxième figure, on peut voir que des basses températures minimales semblent corrélées avec l'augmentation du nombre de malades. On lit l'effet inverse pour les températures maximales (toujours quelque peu mitigé par sa moyenne).

Finalement, dans la dernière figure, on peut observer l'effet complexe de la mobilité professionnelle dans la même région à laquelle on retranche les déplacements intra-commune. Toutefois, cet effet est d'un ordre de grandeur moindre comparé au taux d'incidence retardé, la semaine de l'année et les températures. On retrouve quelque chose de plus près des amplitudes pour les démographies des jeunes et des plus vieux. La première composante principale est plus ou moins alignée à l'opposée de la moyenne pondérée des taux de mobilité. Ainsi il semble y avoir un creux dans les valeurs moyennes alors qu'aux extrémités (très peu ou

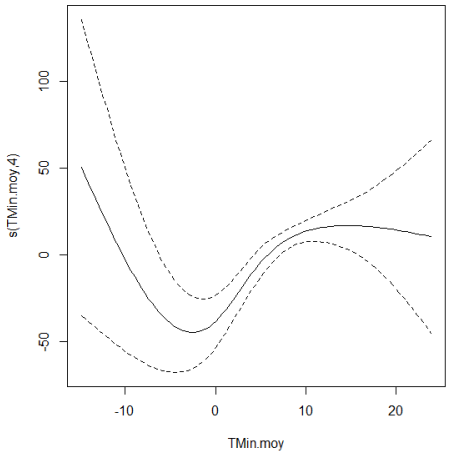
FIGURE 12 – Dépendance apprise entre la cible et les co-variables de météo.



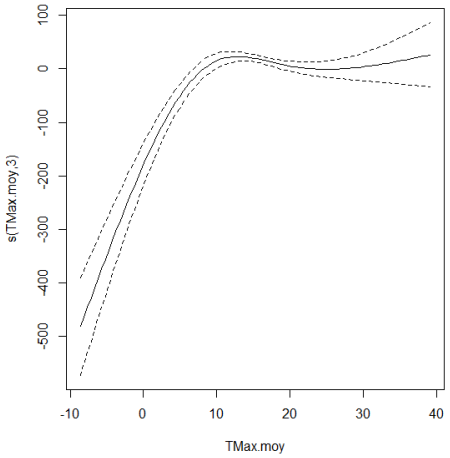
(a) Température minimale minimale



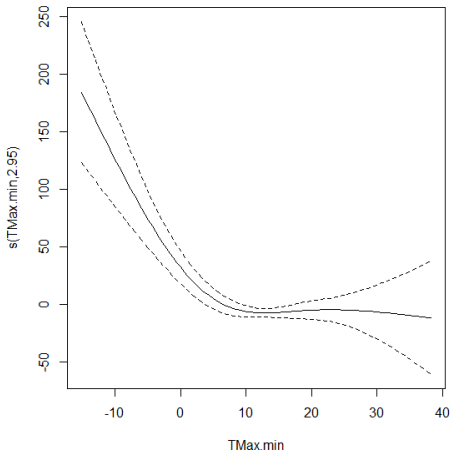
(b) Température minimale maximale



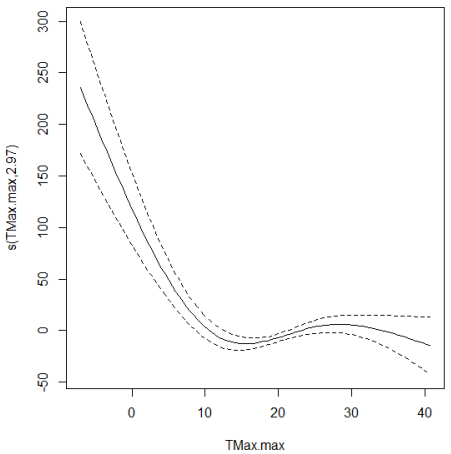
(c) Température minimale moyenne



(d) Température maximale moyenne



(e) Température maximale minimale



(f) Température maximale maximale

beaucoup de gens travaillant dans la même région mais à l'extérieur de leur commune) ; il semble y avoir un effet légèrement positif.

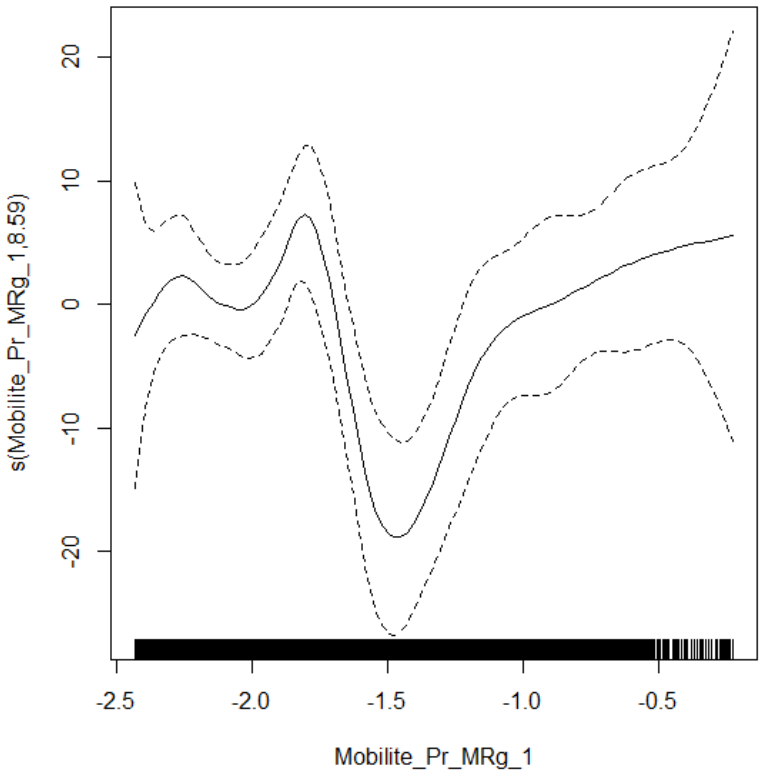


FIGURE 11 – Dépendance apprise par rapport à la première composante principale de la mobilité professionnelle vers la même région (excluant vers la même commune).

Conclusion

Les méthodes de machine learning appliquées aux séries temporelles d'épidémies de grippe ont donné des résultats satisfaisants : un premier modèle purement temporel permettait d'obtenir une modélisation convenable des données, que nous avons ensuite raffinées, d'abord par les méthodes de régression linéaire du type LASSO et Ridge, puis par les méthodes d'arbres et de forêt aléatoires. Nous avons également implémenté un modèle GAM afin de pouvoir totalement exploiter les données collectées, notamment les données mobilités. Nous avons cependant conscience que nos résultats peuvent encore être améliorés, en termes de méthodes et de précision, et que nous avons appliqué seulement une partie des méthodes pertinentes pour la modélisation et prévision. Également, le caractère spatial des données aurait pu être mieux exploité. Cependant, ce projet nous a permis de mieux comprendre et maîtriser certaines méthodes essentielles du machine learning, et également certains aspects importants de la collecte de données - étape chronophage. Il était finalement satisfaisant d'obtenir des résultats une fois ce travail-là terminé, en particulier pour les régressions Ridge et LASSO et pour l'agrégation d'experts. Le code utilisé pour ce projet est disponible à cette page.