Modèles statistiques pour les systèmes d'aide à la décision basés sur la réutilisation des données massives en santé : application à la surveillance syndromique en santé publique

Canelle Poirier

UNIVERSITE BIOLOGIE BRETAGNE SANTE LOIRE

UNIVERSITÉ DE RENNES 1

THESE DE DOCTORAT DE

L'UNIVERSITE DE RENNES 1
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 605
Biologie Santé
Spécialité : Génétique, Génomique, Bioinformatique

Par
Canelle Poirier

Modèles statistiques pour les systèmes d'aide à la décision basés sur la réutilisation des données massives en santé

Application à la surveillance syndromique en santé publique

Thèse présentée et soutenue à Rennes, le 13 juin 2019
Unité de recherche : Laboratoire Traitement du Signal et de l'Image
Equipe Données Massives en Santé

Rapporteurs avant soutenance :

Eric Matzner-Lober	Professeur des Universités – ENSAE Laboratoire CREST
Catherine Quantin	PU-PH – CHU de Dijon

Composition du Jury :

Président :	Mauricio Santillana	PhD – Boston Children's Hospital / Harvard Medical School
Examinateur :	Yann Le Strat	Directeur DATA – Santé publique France
Dir. de thèse :	Valérie Bertaud	Fonction Pu-PH – Université de Rennes 1
Co-dir. de thèse :	Audrey Lavenu	MCF – Université de Rennes 1

Remerciements

Les travaux présentés dans ce mémoire ont été réalisés au sein de l'équipe Données Massives en Santé du Laboratoire Traitement du Signal et de l'Image (INSERM UMR 1099 - Université Rennes 1). Je tiens à remercier Monsieur Marc Cuggia, directeur de cette équipe, de m'avoir accueilli et de m'avoir accordé de son temps depuis mon stage de Master 1. Je remercie également Madame Valérie Bertaud qui a accepté de diriger ma thèse ainsi que Madame Audrey Lavenu et Monsieur Guillaume Bouzillé pour leur encadrement, leurs conseils et leur disponibilité depuis le début. Un grand merci à tous mes collègues, notamment Pierre, Françoise, Julia, Véronique, Christian pour leur bonne humeur et leur soutien au quotidien. Merci à Audrey, Véronique, Pierre, Marie-Lisenn et Camille pour la relecture de ce mémoire.

J'exprime ma profonde gratitude à la commission franco-américaine Fulbright m'ayant permis de financer ma mobilité de 6 mois dans le laboratoire Computational Epidemiology de Boston Children's Hospital et Harvard Medical School. Merci à Monsieur John Brownstein et toute son équipe pour m'avoir si bien accueilli. Je remercie tout particulièrement Monsieur Mauricio Santillana pour son suivi irréprochable, sa grande gentillesse et sa prévenance.

Je remercie également mes professeurs de l'Université de Rennes 2 pour m'avoir transmis l'expertise statistique nécessaire à la réalisation de cette thèse. En particulier, Madame Magalie Fromont, qui a toujours été de bons conseils et m'a donné l'envie de réaliser ce doctorat.

Aux membres du jury, merci d'avoir accepté de juger ce travail.

Enfin, un immense merci à ma famille, mes parents, mes amis et mon compagnon. Merci de me soutenir dans les bons et mauvais moments et de toujours croire en moi.

Résumé

Titre Modèles statistiques pour les systèmes d'aide à la décision basés sur la réutilisation des données massives en santé : Application à la surveillance syndromique en santé publique

Résumé Depuis plusieurs années, la notion de Big Data s'est largement développée. Afin d'analyser et explorer toutes ces données, il a été nécessaire de concevoir de nouvelles méthodes et de nouvelles technologies. Aujourd'hui, le Big Data existe également dans le domaine de la santé. Les hôpitaux en particulier, participent à la production de données grâce à l'adoption du dossier patient électronique. L'objectif de cette thèse a été de développer des méthodes statistiques réutilisant ces données afin de participer à la surveillance syndromique et d'apporter une aide à la décision.

Cette étude comporte 4 axes majeurs. Tout d'abord, nous avons montré que les données massives hospitalières étaient très corrélées aux signaux des réseaux de surveillance traditionnels.
Dans un second temps, nous avons établi que les données hospitalières permettaient d'obtenir des estimations en temps réel plus précises que les données du web, et que les modèles SVM et Elastic Net avaient des performances comparables.
Puis, nous avons appliqué des méthodes développées aux Etats-Unis réutilisant les données hospitalières, les données du web (Google et Twitter) et les données climatiques afin de prévoir à 2 semaines les taux d'incidence grippaux de toutes les régions françaises. Enfin, les méthodes développées ont été appliquées à la prévision à 3 semaines des cas de gastro-entérite au niveau national, régional, et hospitalier.

Mots clés Données massives, Machine learning, Modélisation statistique, Surveillance syndromique, Aide à la décision, Santé publique

Abstract

Title Statistical models for decision support systems based on the reuse of Health Big Data : Application to syndromic surveillance in public health

Abstract Over the past few years, the Big Data concept has been widely developed. In order to analyse and explore all this data, it was necessary to develop new methods and technologies. Today, Big Data also exists in the health sector. Hospitals in particular are involved in data production through the adoption of electronic health records. The objective of this thesis was to develop statistical methods reusing these data in order to participate in syndromic surveillance and to provide decision-making support.

This study has 4 major axes. First, we showed that hospital Big Data were highly correlated with signals from traditional surveillance networks.
Secondly, we showed that hospital data allowed to obtain more accurate estimates in real time than web data, and SVM and Elastic Net models had similar performances.
Then, we applied methods developed in United States reusing hospital data, web data (Google and Twitter) and climatic data to predict influenza incidence rates for all French regions up to 2 weeks.
Finally, methods developed were applied to the 3-week forecast for cases of gastroenteritis at the national, regional and hospital levels.

Keywords Big data, Machine learning, Statistical modelling, Syndromic surveillance, Decision support, Public Health

Table des matières

Abréviations

— OMS : Organisation Mondiale de la Santé

— GFT : Google Flu Trends

— CHU : Centre Hospitalier Universitaire

— CIM-10 : Classification Internationale des Maladies

— OMS : Organisation Mondiale de la Santé

— PCC : Coefficient de corrélation de Pearson

— EQM : Erreur Quadratique Moyenne

— SVM : Support Vector Machine (Machine à Vecteurs de Support)

— RF : Random Forests (Forêts aléatoires)

— LASSO : Least Absolute Selection and Shrinkage Operator

— ARIMA : Autoregressive Integrated Moving Average

— CDC : Center for Disease Control

Table des figures

1 Introduction

1.1 Le contexte de Big Data

Depuis plusieurs années maintenant, la notion de Big Data, désignant l'ensemble des données numériques produites par l'utilisation des nouvelles technologies, d'internet et des réseaux sociaux, s'est largement développée. L'expression Big Data a été créé en 1997 selon l'Association for Computing Machinery. En 2001, le Gartner définissait le Big Data comme un regroupement de données présentant une grande variété, arrivant en volume croissant et à grande vitesse [1]. C'est ce que l'on appelle les trois "V". Aujourd'hui, cette définition reste la définition de référence. Deux autres "V" se sont ensuite ajoutés à cette définition, la valeur et la véracité des données. Les principaux enjeux du Big Data ont été de trouver des technologies capables de stocker et de traiter ces gros volumes de données afin de les analyser et en tirer de l'information. L'exploitation des données massives a ouvert de nouvelles perspectives dans de nombreux domaines comme celui de la recherche scientifique, de la finance, du commerce ou encore de la médecine. Il est possible de réaliser de l'analyse tendancielle ou prédictive, de construire des profils, d'anticiper des risques ou encore de suivre des phénomènes en temps réel.

1.2 Les modèles statistiques

Afin d'explorer et analyser ces immenses bases de données, il a été nécessaire de développer et de mettre au point des méthodes scientifiques et des outils de calcul. C'est ainsi que le "data mining" est apparu dans les années 1990 [2], également appelé "fouille de données". Il se situe entre la Statistique, permettant de lui fournir méthodes et concepts théoriques, et l'Informatique, permettant d'obtenir les données et les moyens de calcul. Le data mining peut être descriptif : mise en évidence des informations présentes dans le grand volume de données, prédictif : extrapoler de

nouvelles informations à partir de celles présentes. À la différence de la Statistique, le nombre d'individus et de variables analysés est beaucoup plus important, les données sont recueillies avant l'étude et à d'autres fins, elles sont de tous types et très souvent imparfaites (erreurs de saisies, données manquantes, individus aberrants..). Le but principal du data mining est l'aide à la décision. Il vise à l'efficacité opérationnelle en comparant performances et précisions des algorithmes en concurrence. En ce sens, il se différencie de la Statistique, qui, attachée à la branche des Mathématiques, est basée sur la notion de preuves pour valider un modèle [3]. Les méthodes utilisées en fouille de données peuvent être : des méthodes traditionnelles, comme l'analyse factorielle, la méthode des K-means, la régression linéaire, la régression logistique ; des méthodes de deep learning ou des méthodes plus récentes comme les machines à vecteurs de support (SVM), les forêts aléatoires (RF) ou les régressions régularisées de type Lasso, Ridge et Elastic Net.

1.3 La surveillance syndromique en santé publique

Aujourd'hui, à travers le monde, la surveillance sanitaire est au coeur des enjeux de santé publique. En effet, de par la mondialisation croissante et la multiplication des flux humains, le risque de propagation des maladies augmente notablement [4]. L'OMS mène une politique de prévention et mobilise une coordination et collaboration entre les pays afin d'assurer une veille sanitaire efficace. L'agence mondiale requiert auprès des états une organisation nationale de surveillance par la mise en place de réseaux d'information.

En France, c'est l'agence nationale de santé publique, Santé publique France, qui a pour mission de traiter de l'état sanitaire des collectivités et de la santé globale des populations sous tous ses aspects : curatif, préventif, éducatif et social. L'agence se définit ainsi : " Par la veille et la surveillance épidémiologiques, l'agence anticipe et alerte. Par sa maîtrise des dispositifs de prévention et de préparation à l'urgence sanitaire, elle accompagne les acteurs engagés de la santé publique. Ancrée dans les

territoires, elle mesure l'état de santé et déploie ses dispositifs au plus près des publics, dans un souci constant de fonder une connaissance juste et de proposer des réponses adaptées. " [5]
Sa mission s'articule autour de trois axes majeurs : anticiper, comprendre, agir. Ses rôles comprennent :

- L'observation épidémiologique et la surveillance de l'état de santé des populations
- La veille sur les risques sanitaires menaçant les populations
- La promotion de la santé et la réduction des risques pour la santé
- Le développement de la prévention et de l'éducation pour la santé
- La préparation et la réponse aux menaces, alertes et crises sanitaires
- Le lancement de l'alerte sanitaire

Pour cela, elle s'appuie sur différents partenaires regroupés dans un réseau, le réseau national de santé publique. Celui-ci regroupe d'un côté, les réseaux de veille et de surveillance, et de l'autre, les réseaux de prévention et de promotion de la santé. Les Agences Régionales de la Santé font partie de ce réseau. Elles ont en charge le pilotage régional du système de santé. Elles définissent et mettent en oeuvre la politique de santé en région, au plus près des besoins de la population.

Le réseau Sentinelles est également un des acteurs du réseau national de santé publique et participe activement à la veille [6]. Créé en 1984, il est composé de 1314 médecins généralistes et 116 pédiatres libéraux volontaires répartis sur le territoire métropolitain français. Il a pour mission de :

- Construire de grandes bases de données en médecine générale et en pédiatrie, à des fins de veille sanitaire et de recherche
- Développer des outils de détection et de prévision épidémique
- Mettre en place des études cliniques et épidémiologiques

Chaque semaine, les médecins Sentinelles vont transmettre les cas vus en consultations afin d'obtenir une surveillance continue sur 10 indicateurs de santé. Parmi ces 10 indicateurs de santé, 9 maladies infectieuses : la coqueluche, la diarrhée aiguë, la

maladie de Lyme, les oreillons, les syndromes grippaux, l'urétrite masculine, la varicelle et le zona et 1 indicateur non-infectieux : les actes suicidaires. À partir des données transmises, une estimation du taux d'incidence hebdomadaire est calculé pour chaque indicateur. Il est ensuite possible d'interroger la base de données du réseau et de télécharger des cartes, des séries chronologiques ou des tableaux sur l'indicateur sélectionné. Les rapports de prévision de grippe et de gastro-entérite y sont publiés chaque semaine. Cependant, ces rapports ont un délai de 1 à 3 semaines en raison du temps de traitement et d'agrégation des données. Ce décalage est problématique pour des prises de décision optimales au niveau de l'agence nationale de santé publique [7,8]. Afin d'apporter une aide à la décision, il est donc nécessaire de développer des méthodes permettant d'obtenir des estimations en temps réel et des estimations à plus long terme des taux d'incidence.

1.4 Le Big Data pour la surveillance syndromique

La surveillance syndromique a été définie par le Center for Disease Control (CDC) and Prevention d'Atlanta (homologue de Santé Publique France), comme une surveillance fondée sur une automatisation de l'enregistrement des données, permettant la mise à disposition pour le suivi et l'analyse épidémiologique en temps réel ou presque réel [9]. Afin d'apporter une solution au délai causé par les méthodes de surveillance traditionnelles, des études ont été faites réutilisant les données massives. C'est le cas de la grippe. Depuis 2014, le réseau Sentinelles, en collaboration avec l'agence nationale de Santé Publique et le Centre national de référence (CNR) des virus des infections respiratoires, a en charge la coordination nationale de la surveillance virologique des cas de syndromes grippaux vus en consultation. En effet, la grippe est un cas de santé publique majeur. Chaque année, dans le monde, jusqu'à 5 millions de cas graves et 500 000 décès peuvent être observés [7]. En France, d'après le bulletin de Santé publique France, la grippe a été responsable de 9500 décès pour l'épidémie de 2018-2019 [10]. Lors du pic épidémique, l'augmentation du nombre de visites chez

les médecins généralistes et dans le service des urgences perturbe l'organisation du système de santé. Pour réduire son impact et pour décider de mesures sanitaires adaptées, il est donc nécessaire de la surveiller afin d'anticiper, comprendre et agir. Les modèles de prévisions tels que les modèles de séries temporelles ou les modèles compartimentaux [11], sont difficilement adaptables pour la prévision des épidémies de grippe. La gravité et le début de l'épidémie peuvent évoluer d'une façon très différente d'un hiver à l'autre, en fonction du virus en circulation. C'est pour cette raison que des méthodes alternatives ont été développées, intégrant notamment des sources de données externes dans les modèles.

Un des premiers à utiliser les données massives pour la prévision des épidémies est le géant Google. En 2008, le service Google Flu Trends est apparu [12], s'appuyant sur les requêtes effectuées par les internautes sur le moteur de recherche, afin d'estimer le taux d'incidence des épidémies de grippe aux États-Unis et anticiper les indicateurs produits par le CDC. Sa méthodologie consistait en la création d'un signal au niveau national et régional à partir de 50 millions de signaux de requêtes jouées par les internautes. Google a ainsi pu montrer que les requêtes jouées par les internautes étaient fortement corrélées aux épidémies grippales. L'algorithme a très bien fonctionné pour estimer les premières épidémies, et a été étendu à d'autres pays dont la France. Malgré une modification de l'algorithme, l'épidémie de 2012-2013 a elle largement été surestimée, en raison de l'annonce d'une pandémie qui en réalité n'est pas apparue. Le manque de robustesse, dû à la sensibilité des changements de comportement des internautes et des modifications des performances du moteur de recherche, a conduit Google à arrêter son algorithme [13]. Cependant, avec plus de 3,2 milliards d'internautes, les flux de données provenant d'Internet sont nombreux et de tous types. Ils peuvent provenir de réseaux sociaux (Facebook, Twitter), de sites de consultation (YouTube, Netflix), de sites d'achat (Amazon, Cdiscount), mais aussi de sites de vente ou de location entre particuliers (Craigslist, Airbnb). Dans le cas de la grippe, certaines études ont choisi de reprendre les données de Google en les associant aux taux d'incidence historiques des syndromes grippaux, puis d'autres se

sont focalisées sur les données de Twitter ou encore Wikipedia [7, 14–19].

1.5 Les données massives hospitalières

Cependant, Internet n'est pas la seule source capable de produire des données en temps réel. Avec l'adoption du dossier patient électronique, les hôpitaux produisent également un grand nombre de données, collectées au cours de l'hospitalisation du patient. Ces données produites peuvent être non structurées, comme les comptes rendus patients ou les données d'imagerie, et structurées comme les diagnostics ou les données de laboratoire. Différents projets ont développé des technologies comme les entrepôts de données biomédicales pour pouvoir intégrer ces données et ainsi les réutiliser. Parmi eux, le plus connu, est le projet Integrating Biology and the Bedside (i2b2), développé par la faculté de médecine Harvard et maintenant utilisé à travers le monde pour la recherche clinique [20, 21]. L'entrepôt i2b2 a été créé pour accueillir des données structurées, accessibles grâce à des terminologies médicales, qui sont des ensembles de termes, rigoureusement définis, et spécifiques au domaine de la santé. Au CHU de Rennes, l'équipe Données Massives en Santé a développé un autre type d'entrepôt de données biomédicales nommé eHOP [22]. La structure de l'entrepôt a été schématisée dans la figure 1.1. L'entrepôt eHOP a été créé pour intégrer des données structurées mais aussi des données non structurées comme les données textuelles. Pour pouvoir extraire l'information pertinente contenue dans ce type de données, l'entrepôt comporte des méthodes de recherche d'information et traitement automatique du langage. Ainsi, eHOP possède un puissant moteur de recherche, capable d'identifier des patients associés à des critères spécifiques via des mots clés pour les données non structurées, ou grâce à des terminologies pour les données structurées.

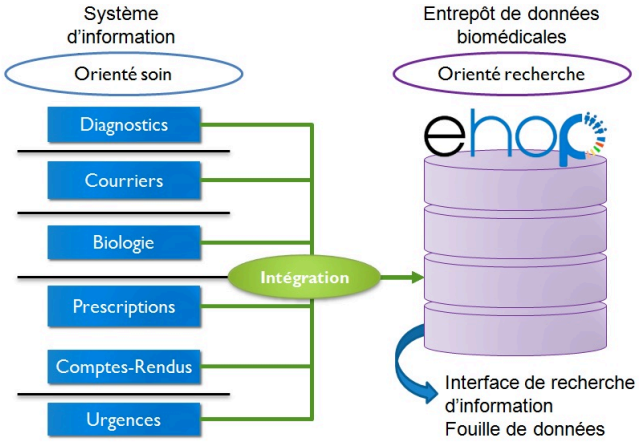


Figure 1.1: Schéma de l'entrepôt de données biomédicales eHOP

1.6 Contribution

Cette thèse s'inscrit donc dans une optique d'exploitation et de valorisation des données massives hospitalières à l'aide de modèles statistiques adaptés au Big Data afin de prévoir à plus ou moins long terme les taux d'incidence de certaines épidémies. Ceci est réalisé dans le but d'apporter une aide à la surveillance syndromique en santé publique.

Les contributions de ce travail sont multiples. Tout d'abord, cela porte sur l'évaluation des données de santé et la possibilité à en extraire des signaux pouvant permettre de prévoir les épidémies en temps réel. Cet objectif est présenté dans le Chapitre 1.

Dans un second temps, Chapitre 2, après avoir étudié les données massives hospitalières et leur intérêt potentiel pour la surveillance syndromique en temps réel, il est nécessaire de savoir si ces données sont tout aussi, voire plus efficaces, que les données du web mais également d'évaluer différents modèles statistiques permettant de réutiliser au mieux ces données.

L'objectif du Chapitre 3 est de prédire à plus long terme et également à une échelle plus fine comme l'échelle régionale.

Enfin, dans le Chapitre 4, ces méthodes développées dans le cas des syndromes grippaux sont appliquées à un autre cas d'usage, la gastro-entérite.

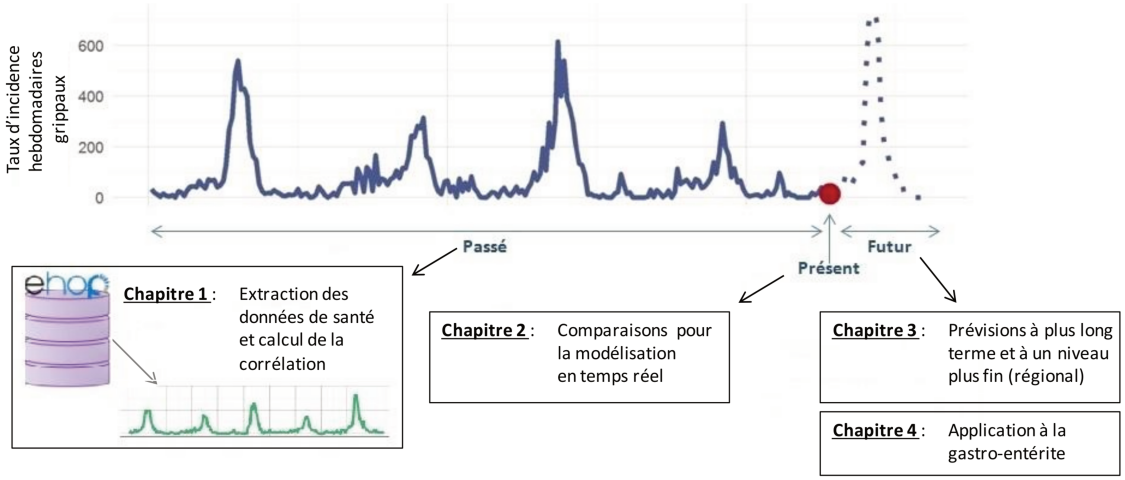


Figure 1.2: Représentation schématique du déroulement de la thèse

2 Évaluation de l'information présente dans les données massives hospitalières

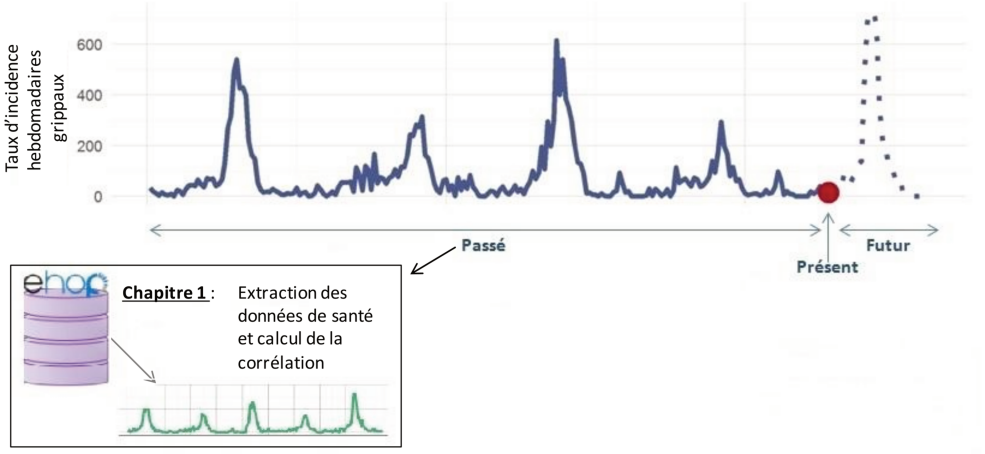


Figure 2.1: Chapitre 1 - Étude rétrospective

2.1 Problématique

Comme cela a pu être défini en introduction, il est nécessaire aujourd'hui de trouver des solutions afin d'éviter le délai de une à trois semaines nécessaire aux réseaux de surveillance traditionnels. Cette problématique est d'autant plus vraie dans le cas de la grippe, qui est un enjeu de santé publique majeur, étant responsable de 5 millions de cas graves et 500 000 décès chaque année dans le monde. Une des alternatives possible à ce délai est l'utilisation de sources de données externes. Jusqu'à présent, les études et méthodes développées se sont majoritairement concentrées sur l'exploitation des données du web mais peu d'études ont portées sur les données massives hospitalières. Au CHU de Rennes, l'équipe Données Massives en Santé a développé un entrepôt de données biomédicales nommé eHOP. L'entrepôt eHOP intègre des données structurées, comme des résultats de laboratoire ou des codes de diagnostics, et des données non structurées comme les données textuelles. Il contient 1.2 millions de patients, 45

millions de documents et 510 millions d'éléments structurés. eHOP possède un puissant moteur de recherche, capable d'identifier des patients associés à des critères spécifiques via des mots-clés pour les données non structurées ou grâce à des terminologies pour les données structurées. A l'hôpital de Rennes, l'entrepôt est utilisé quotidiennement à des fins de recherche clinique comme des études de faisabilité ou des détections de cohorte. Cet outil est également déployé dans 5 autres CHU de la région Ouest : Nantes, Brest, Angers, Poitiers et Tours.

2.2 Objectif

Ce premier axe de travail a pour but d'évaluer le potentiel des données massives hospitalières pour la surveillance des épidémies de grippe en temps réel.

2.3 Considérations méthodologiques

2.3.1 Extraction des données

Les données massives hospitalières Pour évaluer le potentiel des données massives hospitalières, nous nous sommes appuyés sur les données présentes dans l'entrepôt de données cliniques eHOP du CHU de Rennes. Pour cela, il a d'abord été nécessaire d'extraire l'information présente dans l'entrepôt. Deux approches ont été testées afin d'identifier les patients souffrant de syndromes grippaux. Dans un premier temps, nous avons effectué 3 requêtes sur les données textuelles :

- Une requête en lien avec la grippe : Tous les documents contenant le mot-clé "grippe" avec absence de "vaccination" et de "grippe aviaire"
- Une requête en lien avec les symptômes : Tous les documents contenant les mots-clés "fièvre" ou "pyrexie" et "courbatures" ou "douleurs musculaires"
- Une requête en lien avec les urgences : Tous les comptes rendus des urgences ayant pour diagnostic final la grippe.

Dans un second temps, nous avons interrogé les données structurées grâce à des terminologies. Une terminologie est l'ensemble des désignations et des notions appartenant à un domaine bien défini :

— La terminologie CIM-10, qui est la classification internationale des maladies proposée par l'OMS. Elle permet de coder toutes les maladies et beaucoup de signes, symptômes, lésions ou encore traumatismes. Dans le cas de la grippe et du syndrome grippal, les codes sont : J09, J10 ou J11.

— Une terminologie locale utilisée par les laboratoires et permettant de retourner tous les résultats des tests PCR détectant la grippe.

Les données ont été extraites pour la période allant du 1er septembre 2010 au 31 août 2015. Pour chaque requête, le moteur de recherche eHOP retourne tous les documents contenant les mots clés choisis (souvent, plusieurs documents pour un même patient et un même séjour). Afin d'obtenir une incidence, nous avons conservé le document le plus ancien pour un patient et un séjour, puis calculé, pour chaque semaine, le nombre de séjours ayant au moins un document mentionnant le mot clé contenu dans la requête. Ainsi, nous avons obtenu 5 signaux extraits de l'entrepôt de données cliniques.

Les données du réseau Sentinelles : Sur le site du réseau Sentinelles, nous avons recueilli les taux d'incidence (pour 100000 habitants) des syndromes grippaux pour la région de Bretagne. Ces données ont été récupérées pour la période allant du 1er septembre 2010 au 31 août 2015. Elles sont utilisées comme signal de référence auquel on va comparer les signaux de l'entrepôt de données eHOP.

Les données du web : Afin de comparer l'apport des données hospitalières par rapport aux données du web, nous avons également récupéré les estimations faites par l'algorithme Google Flu Trends (GFT) pour la région Bretagne allant du 1er septembre 2010 au 31 août 2015.

2.3.2 Évaluation

Afin d'évaluer le pouvoir de détection des syndromes grippaux à l'aide des données hospitalières, nous avons calculé le coefficient de corrélation de Pearson (PCC) entre chacun des signaux extraits de l'entrepôt eHOP et le signal du réseau Sentinelles pour la région Bretagne.

$$PCC = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

avec x_i le taux d'incidence breton des syndromes grippaux estimé par le réseau Sentinelles pour la semaine i ; y_i l'incidence d'un des signaux extrait de l'entrepôt eHOP pour la semaine i.
Le coefficient de corrélation a également été calculé entre le signal estimé par GFT et le signal du réseau Sentinelles.

Pour détecter une épidémie, le réseau Sentinelles utilise la régression périodique de Serfling [23]. Cette méthode permet de modéliser un « niveau de base » de l'incidence, correspondant au nombre de cas de syndromes grippaux attendus une semaine donnée en l'absence d'épidémie. L'estimation de ce niveau de base prend en compte la saisonnalité. En effet, tout au long de l'année, des cas de syndromes grippaux peuvent être déclarés, pouvant être attribués aux virus grippaux ou à d'autres virus respiratoires. La circulation de tous ces virus est plus importante en hiver, augmentant l'incidence attendue pendant cette période.
L'estimation du niveau de base repose sur les incidences historiques en dehors des périodes épidémiques. Afin d'éliminer les périodes épidémiques, les taux d'incidence passés supérieurs à 279 cas pour 100 000 habitants ne sont pas pris en compte. L'équation de la régression périodique, modélisant le niveau de base de l'incidence y pour la semaine t est de la forme suivante :

$$y(t) = \mu + \alpha t + \beta_k cos(\frac{2k\pi}{T}t) + \gamma_k sin(\frac{2k\pi}{T}t) + \epsilon(t)$$

où μ est une constante, α un terme linéaire, k le nombre d'harmoniques (k=2 dans le cas de la grippe), β_k et γ_k des termes périodiques (k=1 correspond aux termes annuels, k=2 aux termes bi-annuels). La période T définie est de 52 semaines. Le terme $\epsilon(t)$ correspond à l'erreur résiduelle.
Le taux d'incidence attendu est ainsi estimé chaque semaine par la régression de Serfling. Un intervalle de confiance à 90% est calculé autour de cette prévision. Une épidémie est déclarée au niveau national si la borne supérieure de l'intervalle de confiance est dépassée pendant 2 semaines consécutives.

Nous avons donc appliqué la régression de Serfling aux signaux extraits de l'entrepôt eHOP ainsi qu'au signal de GFT. Cela nous a permis de comparer les dates de début et de fin d'épidémie estimées par les données hospitalières ou les données du web, aux dates estimées par le réseau Sentinelles.

2.4 Article

2.4.1 Résumé de l'article

Objectif : La grippe est un enjeu de santé publique majeur. Elle nécessite des moyens de surveillance à différentes échelles géographiques, coûteux et consommateurs en terme de temps. L'objectif principal est d'être capable de prédire les épidémies. Outre les systèmes de surveillance traditionnels comme le réseau Sentinelles, plusieurs études ont proposé des modèles de prévision basés sur la réutilisation des données du web. Dans cet article, nous avons évalué le potentiel des données massives hospitalières pour la surveillance des épidémies de grippe.

Méthodes : Nous avons utilisé l'entrepôt de données cliniques du CHU de Rennes, où nous avons effectué différentes requêtes, afin de récupérer l'information

pertinente présente dans les dossiers patients électroniques. Ces données nous ont permis de calculer une incidence hebdomadaire des syndromes grippaux présents à l'hôpital.

Résultats : Nous avons trouvé que la requête la plus corrélée aux estimations effectuées par le réseau Sentinelles, est la requête basée sur les comptes-rendus des urgences ayant pour diagnostic la grippe (PCC = 0.931). Les requêtes basées sur les données structurées (codes CIM-10 ou résultats de tests PCR) ont obtenu le meilleur coefficient de corrélation pour l'épidémie de 2014-2015 avec respectivement un PCC égal à 0.981 et 0.953. Cela suggère que les codes CIM-10 et les résultats des tests sont associés à des épidémies sévères (l'épidémie de 2014-2015 étant plus importante que les précédentes épidémies). Enfin, notre approche nous a permis d'obtenir des caractéristiques supplémentaires pour les patients, telles que le sexe ou les groupes d'âge, comparables à celles du réseau Sentinelles.

Conclusion : Les données massives hospitalières semblent avoir un apport pour la surveillance des épidémies de grippe en temps réel. Une telle méthode pourrait constituer un outil de surveillance complémentaire aux moyens de surveillance traditionnels. Cela pourrait permettre d'obtenir de l'information plus rapidement ou d'obtenir des caractéristiques supplémentaires sur les individus. Ce système pourrait facilement être appliqué à d'autres maladies. Cependant, une nouvelle étude est nécessaire, afin d'évaluer la réelle efficacité d'un modèle statistique réutilisant les données massives hospitalières pour prédire les épidémies de grippe.

Computer Methods and Programs in Biomedicine 154 (2018) 153–160

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Leveraging hospital big data to monitor flu epidemics

Guillaume Bouzillé[a,b,c,d,*], Canelle Poirier[a,b,f], Boris Campillo-Gimenez[a,b], Marie-Laure Aubert[f], Mélanie Chabot[f], Emmanuel Chazard[g], Audrey Lavenu[c,e], Marc Cuggia[a,b,c,d]

[a] INSERM, U1099, Rennes, F-35000, France
[b] Université de Rennes 1, LTSI, Rennes, F-35000, France
[c] CHU Rennes, CIC Inserm 1414, Rennes, F-35000, France
[d] CHU Rennes, Centre de Données Cliniques, Rennes, F-35000, France
[e] Université Rennes 1, Rennes, F-35000, France
[f] Université de Rennes 2, IRMAR, Rennes, F-35000, France
[g] Département de Santé Publique, Université de Lille EA 2694, CHU Lille, F-59000 Lille, France

ARTICLE INFO

Article history:
Received 10 February 2017
Revised 4 October 2017
Accepted 14 November 2017

Keywords:
Health big data
Clinical data warehouse
Information retrieval system
Health Information Systems
Influenza
Sentinel surveillance

ABSTRACT

Background and Objective: Influenza epidemics are a major public health concern and require a costly and time-consuming surveillance system at different geographical scales. The main challenge is being able to predict epidemics. Besides traditional surveillance systems, such as the French Sentinel network, several studies proposed prediction models based on internet-user activity. Here, we assessed the potential of hospital big data to monitor influenza epidemics.

Methods: We used the clinical data warehouse of the Academic Hospital of Rennes (France) and then built different queries to retrieve relevant information from electronic health records to gather weekly influenza-like illness activity.

Results: We found that the query most highly correlated with Sentinel network estimates was based on emergency reports concerning discharged patients with a final diagnosis of influenza (Pearson's correlation coefficient (PCC) of 0.931). The other tested queries were based on structured data (ICD-10 codes of influenza in Diagnosis-related Groups, and influenza PCR tests) and performed best (PCC of 0.981 and 0.953, respectively) during the flu season 2014–15. This suggests that both ICD-10 codes and PCR results are associated with severe epidemics. Finally, our approach allowed us to obtain additional patients' characteristics, such as the sex ratio or age groups, comparable with those from the Sentinel network.

Conclusions: Conclusions: Hospital big data seem to have a great potential for monitoring influenza epidemics in near real-time. Such a method could constitute a complementary tool to standard surveillance systems by providing additional characteristics on the concerned population or by providing information earlier. This system could also be easily extended to other diseases with possible activity changes. Additional work is needed to assess the real efficacy of predictive models based on hospital big data to predict flu epidemics.

1. Introduction

Currently, flu activity monitoring remains challenging and is a costly and time-consuming task [1]. Flu epidemics are a major public health issue because each year, they cause 250,000 to 500,000 deaths worldwide and they destabilize health care systems, resulting in overcrowding of primary care centers and emergency departments [2–4]. Many actors are involved in influenza monitoring, at the local, regional, national and international level. National surveillance systems are the cornerstone of this system. For instance, the US influenza Sentinel Provider Surveillance Network, belonging to the Center for Disease Control and Prevention (CDC), in the United States of America, and the Sentinel network in France, both provide weekly flu activity reports based on data collected from general practitioners [5,6].

Such national flu surveillance systems provide a fine-grained description of what happens at the regional or national level and allow researchers to observe inter-annual epidemic variations. However, these reports are usually available with a delay of one

* Corresponding author. LTSI - UMR Inserm - Université de Rennes 1, Equipe-Projet Données massives en santé (DMS), Campus de Villejean - Bât. 6, 35043 Rennes Cedex, France.
E-mail address: guillaume.bouzille@chu-rennes.fr (G. Bouzillé).

https://doi.org/10.1016/j.cmpb.2017.11.012
0169-2607/

to two weeks and need to be refreshed until all data from a given week have been reported. This delay in data availability limits their use for real-time monitoring purposes. Moreover, data reported by the Sentinel network provide very few details about patients, beside age or sex. Yet, it would be of great interest to better describe, for instance, the comorbidities (e.g., International Classification of Diseases, 10th revision, ICD-10, codes), or to identify subgroups of patients who are more likely to catch influenza or to develop influenza-related complications.

For these reasons, influenza surveillance now relies also on other data sources that gather additional information, such as self-reporting from patients, viral surveillance or data from emergency departments (ED) [2,7,8]. In France, the French Public Health Agency launched an additional monitoring system based on data collected from 86% of all French EDs, thus covering most of the French territory [9]. This project provides a better understanding of flu epidemic severity, especially in relation to cases that require hospitalization.

There is also a growing interest in finding other ways that rely on alternative data sources to achieve near real-time monitoring. Many studies have assessed the use of internet-user activity data because they can produce real-time indicators [10–18]. Several data sources have been explored, including Wikipedia, Twitter or Google search-engine data. For instance, Google created a project dedicated to influenza monitoring: Google Flu Trends (GFT). This project uses search queries connected with influenza-like illnesses (ILI) from Google.com to produce influenza activity estimates [2]. Since its launch in the United States in 2008, GFT predictions have proven to be very accurate when compared to CDC reports. Moreover, GFT data are available 7–10 days before those of the CDC [12]. GFT was extended to other countries and its estimates confirmed to be accurate. However, GFT yielded inaccurate data during several periods [19,20]. In 2009, it produced lower estimates at the start of the H1N1 pandemic; in 2013 its estimates were almost twice those from the CDC. As a result, GFT is currently closed to the public. GFT appeared to be sensitive to uncommon flu epidemics, to media coverage, to changes in the internet users' habits and to modifications of the algorithm in the Google search engine [11,20]. Consequently, other studies proposed to combine traditional surveillance systems and web data, to benefit from the advantages of both systems. One example is the recently published work on the ARGO model that could be considered to be a GFT update. It combine Google and CDC ILI activity data with a dynamic statistical model (least absolute shrinkage and selection operator, LASSO) to weekly redefine the best predictors for the current week and readjust their coefficients [11]. This model seems very promising because it can produce near real-time flu activity indexes that are very accurate compared with those produced by the CDC, with a correlation coefficient of predicted values for the flu seasons of the 2010–2014 period ranging from 0.928 to 0.993.

However, neither standard systems nor the current web-based models are designed to monitor flu activity at a smaller scale, such as that of a hospital. Yet, flu epidemics strongly contribute to the overcrowding of adult and pediatric EDs. A study by Dugas et al, showed a high correlation between city-level GFT data (Baltimore) and the number of patients visiting adult ($r = 0.885$) and pediatric EDs ($r = 0.652$). Specifically, GFT data correlation with standard overcrowding measures was high for pediatric EDs ($r = 0.641$ to 0.649) and moderate for adult EDs ($r = 0.421$ to 0.548) [21].

With the widespread adoption of Electronic Health Records (EHRs), hospitals also are producing a huge amount of data - collected during the course of clinical care - that offer a window into the medical care, status and outcomes of a varied population who is representative of the actual patients [22,23]. This huge amount of data holds the promise of supporting a wide range of medical and health care functions, including, among others, clinical decision-making support, disease surveillance or population health management [24].

Hospitals are currently deploying information technologies and tools intended to facilitate access to clinical data for secondary-use purposes. Among these technologies, clinical data warehouses (CDWs) come forth as one of the solutions to address Hospital Big Data (HBD) exploitation [25]. Different projects have developed CDWs with different architectures, tools and services dedicated to the reuse of patient data coming from EHRs [26–31]. Depending on their Extract-Transform and Load process, CDWs can collect data in real-time, such as the STRIDE CDW of Stanford University [30]. The most famous CDW technology is the Informatics for Integrating Biology & the Bedside project (i2b2), developed by Harvard Medical School, that is now used worldwide in clinical research and can be updated in real-time [32,33]. At our academic hospital in Rennes (France), we developed our own CDW technology, called eHOP (formerly named Roogle [31]). Structured (laboratory, prescriptions, ICD-10 diagnoses) and unstructured (discharge summaries, histopathology, operative reports) data can be integrated in eHOP in real time. Unlike i2b2 data models, eHOP integrates the chain of clinical events into its design and allows the direct access to EHRs. eHOP consists of a powerful search engine system that can identify patients who match specific criteria retrieved either from unstructured data, via keywords, or from structured data, by querying terminology-based codes. The eHOP CDW is used routinely for clinical research purposes, such as feasibility studies, cohort detection and pre-screening, at Rennes academic hospital. The eHOP technology is currently implemented in the other five academic hospitals of the Western region of France (Angers, Brest, Nantes, Poitiers and Tours). Its use will constitute a great source of health data that cover a large part of the population of the West of France who has access to health care facilities linked to eHOP (about 11 million inhabitants; 800,000 visits per year) [34].

We believe that CDWs can help to monitor influenza-like illness (ILI) thanks to their ability to provide data in near real-time and at a local scale. Moreover, the richness of the data produced during patient management will allow a better patient characterization.

In this paper, we present a feasibility study on the production of accurate near-real-time estimates of ILI activity based on the CDW eHOP.

2. Methods

We extracted data from the eHOP CDW of the academic hospital of Rennes, from September 1, 2010 to August 31, 2015. This corresponds to the last five winter seasons defined by the Sentinel network (beginning on the first day of September of every year and ending on 31 August of the following year). The data integration and storage method was the same during the entire study period. As a reference, we used French Sentinel network data on Brittany for the same period (https://websenti.u707.jussieu.fr/sentiweb/?page=table). Brittany is the French region from where most patients at Rennes academic hospital come. We also considered internet-based ILI estimates from GFT for Brittany, from September 1, 2010 to August 10, 2015 (date of GFT closure) as an additional source for comparison (https://www.google.org/flutrends/about/data/flu/fr/data.txt).

We tested two main approaches with the purpose of identifying patients who might have ILI, from data stored in eHOP (see S1 Table for a complete query description). The first approach was based on three different full-text queries to retrieve documents that match the following keywords and constraints:

- Flu query: documents matching the keywords "flu", in the absence of "flu vaccination," and "avian flu."

- Symptoms query: documents matching the keywords “fever” or “pyrexia” and “ache” or “muscle pain.”
- Emergency query: ED discharge summaries where “flu” was the final diagnosis. Only applicable to discharged patients (i.e., documents belonging to patients who were further hospitalized were not considered).

The first two queries could retrieve any kind of document, including discharge summaries of inpatients or outpatients, emergency discharge summaries, operative reports, laboratory results, Diagnosis-Related Groups (DRGs), X-ray reports or histopathology reports. The third query was focused on retrieving documents from the ED.

The second approach involved querying CDW structured data for the following appropriate terminologies:

- ICD-10 query: DRGs having at least one code belonging to the influenza-related ICD-10 chapters: J09.x, J10.x or J11.x.
- Biology query: We relied on the local terminology used by the laboratory information system to retrieve all flu PCR test results (negative and positive). The aim was to have a signal connected with ILI symptoms and not only with flu.

Given that the study purpose was not to assess query accuracy or recall, we made the assumption that potential noise was constant over time. Hence, we did not manually validate the relevance of patients retrieved by the query and we retained the entire list of patients. We then processed the weekly incidences for each query. Our definition of ILI case covered any patient visit for which a document that matched a given query was generated. The date of the case was thus the patient’s admission date. A null incidence estimate was inputted for all weeks without cases. The entire process was performed using anonymous data from the eHOP CDW.

As additional variables, we retrieved the patients’ birthdate to perform analyses based on patients’ age groups at the time of the visit: 0 to 4 years, 5 to 14 years, 15 to 64 years and 65 years and more. The aim was to assess whether the epidemic severity could be extrapolated from such data. We considered that severe epidemics might affect especially younger and/or older people among all hospitalized patients compared with the population covered by the Sentinel network. We computed the distribution of age groups on a calendar year basis, following a process similar to that of the Sentinel network, with the aim of comparing both distributions.

To evaluate ILI detection by our system, we compared our weekly ILI incidence results with the weekly incidences rates from the reference Sentinel network by calculating the Pearson’s correlation coefficient (PCC) for the entire study period and for each winter season. For comparison purposes, we did the same comparison between weekly GFT estimates and weekly incidence rates from the Sentinel network.

As an illustration of eHOP’s ability to monitor flu epidemic data, we also replicated the Serfling periodic regression analysis that is currently used by the Sentinel network to identify influenza epidemic periods [35]. We used the Sentinel’s R script, available at http://marne.u707.jussieu.fr/periodic, and the parameters currently employed in routine practice by the Sentinel network [36]: a pruning threshold corresponding to the 85th quantile, a 95th unilateral confidence interval to detect the start (when the observed data exceed this threshold for two consecutive weeks) and the end (when the observed data are below the threshold for two consecutive weeks) of ILI epidemics. We fitted the following linear regression model for the whole study period:

$$Y(t)=\mu+ \alpha \cdot t+ \beta_k \cdot \cos\left(\frac{2k\pi}{T} \cdot t\right)+\gamma_k \cdot \sin\left(\frac{2k\pi}{T} \cdot t\right)+\varepsilon(t),$$

where μ is a constant, α a linear term, k the harmonic number, β_k and γ_k are period terms. The period T is equal to 52.18 weeks and k is equal to 2. The residual error corresponds to the $\varepsilon(t)$ term.

We assessed the periodic regression performance by calculating the shift between the dates (start and end of epidemics) identified with eHOP estimates and the dates identified from Sentinel network estimates.

All analyses were performed using the R software, version 3.2.3 [37].

This study was approved by the local Ethics Committee of Rennes Academic Hospital.

3. Results

3.1. Information retrieval results

The study period included lists of patients retrieved from eHOP queries between September 1, 2010 and August 31, 2015. For this period, 14,873,482 documents were available in the eHOP CDW, as well as 2220,741 patient visits. Performing the five eHOP queries and then processing the data to produce weekly ILI estimates took approximately 7 minutes (6 m 30 s for queries on unstructured data and 30 s for queries on structured data) on a standard desktop computer. The “flu query” (the keyword “flu”, in the absence of “flu vaccination” and “avian flu”) retrieved 19,522 documents, among which there were 4604 emergency reports (24%), 3773 laboratory results (19.3%), 3344 outpatient discharge summaries (17.1%), 2882 inpatient discharge summaries (14.8%) and 798 DRGs (4%). The “symptoms query” (association of fever or pyrexia and ache or muscular pain) retrieved 2916 documents, among which there were 1436 emergency room reports (49.2%), 524 outpatient discharge summaries (18%) and 482 inpatient discharge summaries (16.5%). The remaining documents were connected with unclear or missing document types. The last three queries were connected with specific types of documents, particularly with emergency reports, laboratory results or DRGs. The patients’ distribution according to the different settings (outpatients, inpatients and ED) is illustrated in Fig. 1.

Results from queries to retrieve patients with at least one document matching the following conditions: flu query = keyword “flu” in the absence of flu vaccination and avian flu; symptoms query = keywords “fever” or “pyrexia” and “ache” or “muscle pain”; emergency query = discharge summaries from the emergency department with “flu” as final diagnosis; ICD-10 query = DRGs with at least one code belonging to the ICD-10 chapters on influenza (i.e., J09.x, J10.x or J11.x.); biology query = PCR-based flu tests (negative or positive results).

Emergency defined a stay in the emergency department without further hospitalization.

3.2. Overall estimates

Weekly ILI estimates computed from the eHOP query results are displayed in Fig. 2. During the entire study period, the ILI estimates retrieved from the query focused on ED data were the most highly correlated with the Sentinel Network’s (PCC of 0.931 compared with PCCs between 0.869 and 0.679 for other queries) (Table 1). As a comparison, the PCC for GFT with the Sentinel network was 0.925.

GFT was the data source that correlated most with the Sentinel network for the seasons 2010–11 and 2012–13 (PCC = 0.967 and 0.947, respectively). For the seasons 2011–12 and 2013–14, the eHOP query focused on EDs showed the highest correlation with the Sentinel network, but with a PCC below 0.9. For the season 2014–15, the eHOP ICD-10 query performed best, with a PCC of 0.981. The query based on symptoms was the only one with a PCC below 0.9 for this last season. For the 2013–14 flu season, both eHOP queries and GFT had PCC values below 0.9. The last complete season (2014–15) yielded the best correlations because all queries

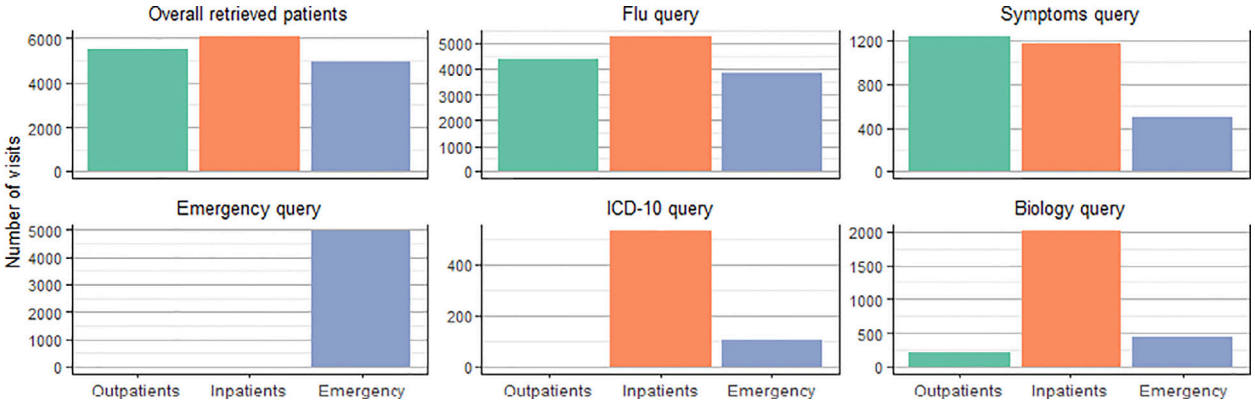


Fig. 1. Patients' settings.

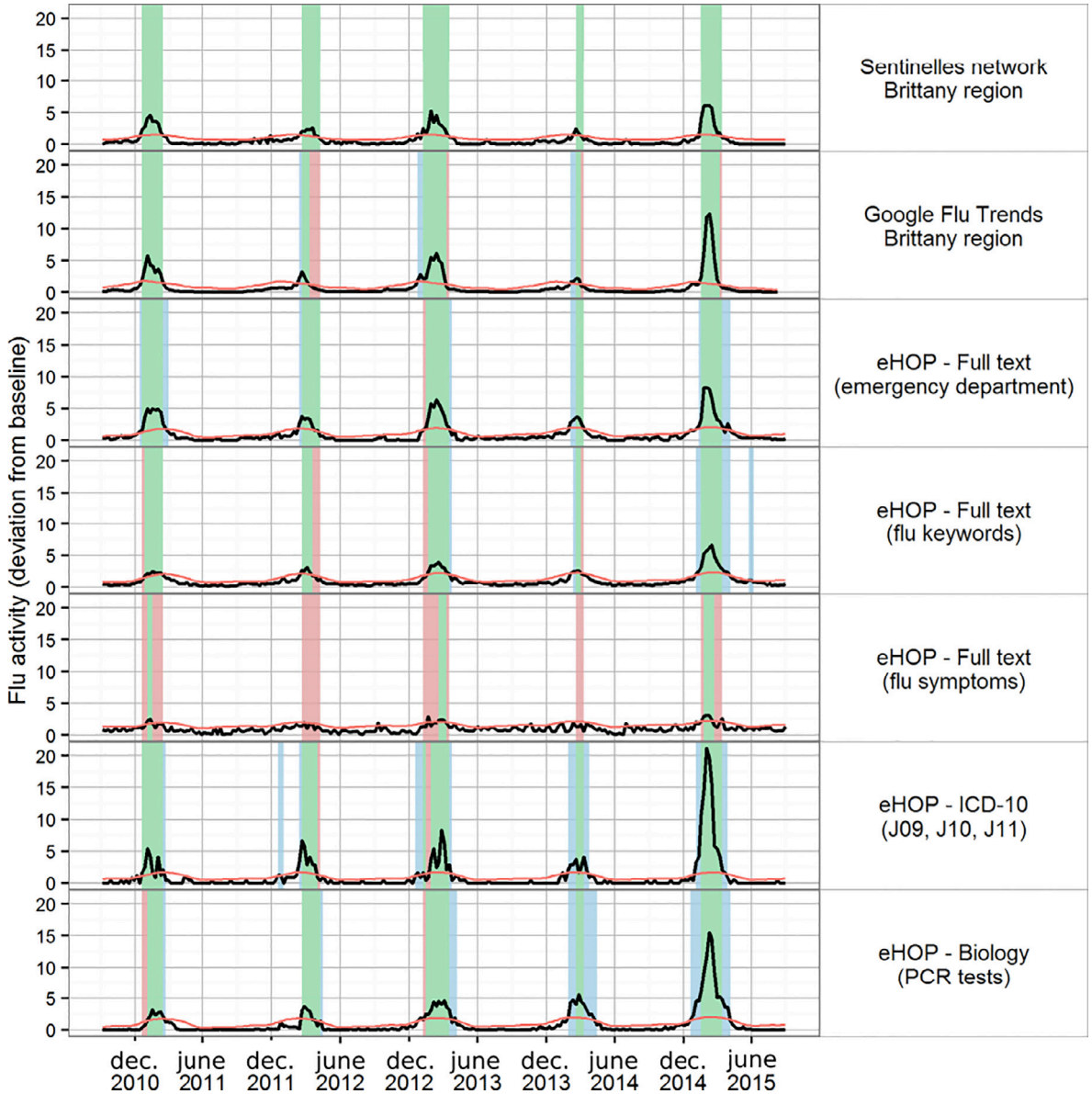


Fig. 2. Weekly influenza-like illness estimates from the different data sources and periods of detected epidemics.

Table 1
Pearson correlation coefficients between ILI activity estimates from eHOP queries or Google Flu Trends and ILI incidence rates from the Sentinel network.

Data source /query	Entire period	Winter flu seasons (from September 1 to August 31 of the following week)				
		2010–11	2011–12	2012–13	2013–14	2014–15
GFT (up to 2015–08–10)	0.925	**0.967**	0.735	**0.947**	0.850	0.940
eHOP flu	0.869	0.871	0.862	0.911	0.818	0.939
eHOP symptoms	0.679	0.784	0.664	0.652	0.298	0.837
eHOP emergency	**0.931**	0.941	**0.864**	0.933	**0.853**	0.972
eHOP ICD-10	0.829	0.854	0.789	0.758	0.732	**0.981**
eHOP biology	0.801	0.813	0.796	0.863	0.777	0.953

matched the Sentinel network data with PCC values up to 0.9, except for the symptoms query.

The reference is data from the Sentinel network for the Brittany region. Estimates from Google Flu Trends are for comparison purposes. Black curves correspond to the estimates computed from the different data sources or queries. Red curves are the upper bound of the 95% prediction interval of the periodic regression models, computed using the Serfling method to determine epidemic periods. Green areas are periods that match the Sentinel network epidemic periods. Red areas are epidemic periods not detected from data sources or queries. Blue areas are detected periods that do not match true epidemics.

In Figure months should have a capital letter at the beginning (ex., june into June).

3.3. Sex and age group estimates

In the Sentinel network data, the male to female ratio was 1, 0.96, 0.97, 0.93 and 1.01, respectively, for epidemics from 2010 to 2014. In comparison, the sex ratio observed in eHOP queries ranged from 0.94 to 3.2 in 2010, from 1.07 to 1.90 in 2011, from 0.94 to 1.36 in 2012, from 1.02 to 1.78 in 2013 and from 0.92 to 1.16 in 2014. The highest sex ratio values were found in the results obtained with the biology query, indicating that PCR tests were more often performed for male patients. There was no significant difference in the age group distribution between male and female patients for the patients retrieved with this query ($p = .41$ using the Chi-square test).

Regarding the age group distribution, eHOP queries yielded more pediatric cases (0 to 4 years), compared with the Sentinel network data (Fig. 3). The biology query retrieved more pediatric and elderly patients than the other queries.

Red bars show the age group distribution from the different eHOP queries. Green bars show the age group distribution from the Sentinel network. *p*-Values were calculated with the use of the Chi-square test or the Fisher exact test (indicated with an asterisk) See legend to Fig. 1 for a description of the eHOP queries.

3.4. Epidemic periods

For each GFT and eHOP query, we computed a periodic regression model (i.e., Serfling regression model) to detect epidemic periods, as done by the Sentinel network's current surveillance system (red line in Fig. 2). We compared epidemic periods from GFT and eHOP with reference data from the Sentinel network for the region of Brittany (Table 2).

GFT detected the beginning and the end of epidemics from 0 to 2 weeks before the Sentinel network. Among the different eHOP queries, the flu symptoms query yielded the worst results, particularly because it could not detect all epidemics. Laboratory and ICD-10 queries resulted in longer epidemics, particularly for the last two seasons: they anticipated the start of the two epidemics by 2 to 4 weeks and delayed the end by 2–5 weeks (Fig. 2). The eHOP query on flu keywords and the emergency query gave the best results. Particularly, the emergency query detected the start of epidemics from 1 to 2 weeks before the Sentinel network, except in 2013, when there was a delay of one week. For the epidemic end, the emergency query tended to produce longer epidemics, ending 0 to 3 weeks after the Sentinel network's estimates (Fig. 2).

4. Discussion

This study demonstrates the great potential of HBD for monitoring flu epidemics. CDWs, such as eHOP, allow researchers to leverage the richness of heterogeneous clinical data from EHRs. eHOP added value is that it provides the possibility of querying both structured and unstructured data that appear to be great candidate data sources for efficient monitoring of diseases activity. However, as it is the case with every information retrieval system, part of the results yielded by our system corresponds to noise, that is, patients who do not have ILI. The result precision depends partly on the query used. For instance, the "symptoms" query is particularly subject to noise and thus, does not seem to be specific enough for ILI monitoring. It also depends on the type of queried data. Unstructured data are, of course, more prone to produce noisy results. The main reasons are the mentioning of a personal or family history of influenza and the exclusion of influenza diagnoses in discharge summaries, although our system has several natural language processing capabilities, such as detection of negative sentences. Structured data are less susceptible to noise: laboratory results or ICD-10 codes ascertain the fact that the patient has ILI. The drawback is the lack of recall for such data sources, for instance, during epidemics the severity of which does not lead to hospitalization (i.e., without diagnosis-related groups), or with diagnoses that do not require any laboratory test. Thus, we cannot control the performance of our information retrieval system. This can be seen as a limitation of our approach: we cannot validate every potential case retrieved by the system, and we cannot ensure the retrieval of all patients with ILI. We could have investigated the system precision because eHOP provides the possibility to access the original documents to check whether the retrieved patients truly had ILI. However, the purpose of this study was not to assess the performance of our information retrieval system, but to show that it can produce ILI activity indexes in the same way as internet-based monitoring. Hence, our system is not intended to be as reliable as a traditional monitoring system, such as the Sentinel network, for producing weekly incidence rates. Nevertheless, it provides a good picture of weekly ILI activity in primary care through the ED data and in hospitalized patients.

We believe that the strength of our system is its capability to generate near real-time estimates from hospital big data. Our estimates are generated using health care activity suspected of being connected with ILI and, due to the proximity to actual ILI cases, they could be more reliable than internet-based indexes. Indeed, we can produce estimates based on data connected with patients who presented symptoms severe enough to require visiting the ED

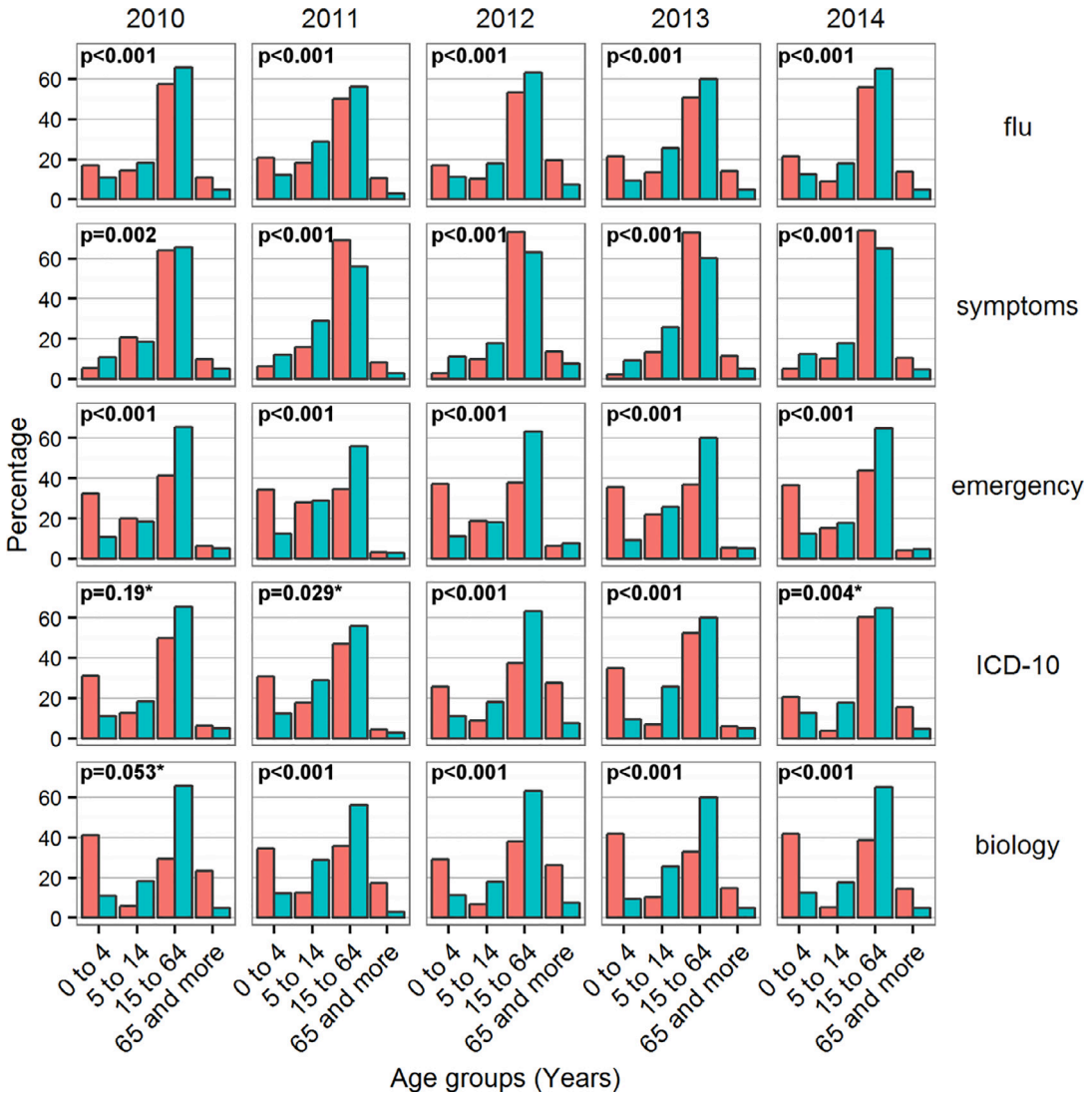


Fig. 3. Age group distributions retrieved from the different eHOP queries.

Table 2
Summary of epidemic detection delays using the different data sources or queries.

Data source/query	No. of detected epidemics	Average delay to detect the epidemic start* (week)	Average delay to detect the epidemic end* (week)
Sentinel network	5	0	0
GFT	5	-1 ± 1	-1.4 ± 1.51
eHOP emergency	5	-0.8 ± 1.09	1 ± 1.22
eHOP flu	6	0 ± 1.58	0 ± 2.24
eHOP symptoms	3	3 ± 2.64	-2.67 ± 1.53
eHOP ICD-10	7	-0.6 ± 2.30	1 ± 1.22
eHOP biology	5	-0.8 ± 2.59	2.6 ± 1.67

*Delays are related to epidemics overlapping with the true epidemic periods from the Sentinel network.

or to be transferred to hospital. On the contrary, internet-based estimates may also incorporate data from healthy internet users who can potentially be influenced by the media or are simply searching information about influenza.

The possibility to produce a fine-grained description of the diseased population is an additional strength of our system. We demonstrated this potential for simple attributes (age groups and sex ratio) that were also available in the Sentinel network annual reports, for comparison purposes. This allowed showing some differences between the population coming to hospital and the population captured by the reference system. Our system found more pediatric and geriatric cases than the Sentinel network. Particularly, the younger cases may explain the predominance of male patients found with the PCR query because it seems than male patients are more prone to respiratory infectious diseases than female patients [38].

In addition, eHOP allows a better characterization of ILI patients by using the data available in the CDW, such as comorbidities or episode severity (e.g., requiring hospitalization or intensive care), all in near-real time.

However, one must be aware that the eHOP data loading process has various delays, depending on the data source. As a result, this process involves a high degree of heterogeneity in the availability of the data used to produce ILI estimates. For instance, dis-

charge summaries are often generated several days after the patient's stay, which is not compatible with real-time monitoring. Conversely, ED discharge summaries are produced during the patient's visit and are made available as soon as the patient leaves the hospital or is transferred to a conventional unit. Similarly, laboratory results are produced during the patients' stay. Therefore, these two data sources are available in the CDW with a lag of one day, because they are uploaded in eHOP every night.

Another of the system's limitations is that we currently only have access to hospital data. This is the main cause of the differences in ILI activity compared with the Sentinel network. From the perspective of our hospital physicians working on infectious diseases this is not really a drawback, because the differences in duration and magnitude may reflect the severity of epidemics that cause more hospitalizations during a longer period. The higher estimates resulting from ICD-10 and laboratory queries also seem to be connected with more severe epidemics, as was the case in 2014–15. Moreover, local ILI activity estimates could be compared with other local indexes, such as the global hospital activity, bed occupation rates or average hospitalization length, to produce more appropriate estimates of the overcrowding risk. This is a key point for hospitals, as estimates from traditional surveillance systems do not allow them to anticipate overcrowding during severe epidemics, resulting in higher rates of hospitalization. However, we also produced estimates comparable to those of the Sentinel network, when using appropriate queries from the ED (PCC of 0.931) that correlated more closely with the Sentinel network estimates than any of the Google Correlate internet-based queries (the Google query most correlated with ILI activity from the Sentinel network for the region of Brittany and for our study period was "Tamiflu", with a PCC of 0.9265).

In our study, we were limited to the population of Rennes academic hospital that, in addition, does not entirely cover the geographical territory of Brittany. As mentioned in the Introduction, the eHOP technology is going to be deployed in all academic hospitals of the West of France. By extending the study reach, we could obtain a complete view of influenza dynamics and activity at a larger scale. We also believe that our approach is transposable to other CDW technologies, such as the i2b2 standard [32], with appropriate real-time data integration. This could allow aggregating estimates from different institutions, using a SHRINE data sharing network at different scales [39]. Indeed, the SHRINE technology allows building a multi-node, peer-to-peer infrastructure for connecting i2b2 CDWs to research networks. We are also exploring this approach by feeding an i2b2 instance with limited sets (i.e., only patients retrieved through our queries) of structured data from eHOP. Another approach could be based on the OHDSI initiative that proposes a common data model for observational studies employing other standardization procedures [40]. However, we have not yet investigated this approach.

Finally, this study only gives the proof of concept concerning the HBD potential for ILI monitoring. The next step will be to assess eHOP prediction capabilities with appropriate statistical models, using such data to predict the data generated by the Sentinel network. Several models have been explored in previous studies with promising results. Recently, Harvard University proposed an alternative model to GFT also based on Google users' activity [11]. Briefly, for each weekly ILI activity to be predicted, a model is built using predictors consisting of the 2-year history of the CDC ILI activity, submitted to an autoregressive process of order 52, and the 100 Google queries most highly correlated with the CDC ILI activity for the same period. The model uses a LASSO method to perform variable selections to only keep the most informative predictors. This kind of model could easily use our eHOP query results as covariates instead of internet-based data. Another interesting approach could be to build models that combine internet-based data and hospital data. Besides predicting ILI activity at a population level, we also want to assess whether our data can be used for predicting ED activity that might help to better manage issues connected with overcrowding. Our results also suggests that this approach could be used for monitoring the activity of other diseases that are emerging or that require precise follow-up, especially when the population is not yet worried about them.

5. Conclusions

Our study shows that HBD are a valuable data source for ILI activity monitoring. Specific data sources, such as laboratory results or DRGs, and the patient characteristics that are available in CDWs allow a fine description of epidemics. However, further investigation is necessary to assess the near real-time prediction capabilities of models that use such data sources, and to demonstrate its extensibility to other diseases.

Acknowledgments

We would like to thank the French National Research Agency (ANR), for funding this work inside the INSHARE (INtegrating and Sharing Health dAta for Research) project (grant no. ANR-15-CE19-0024).

We thank our colleagues Eric Matzner-Lober from the University of Rennes 2, Jean-Marc Chapplain from the CHU of Rennes and the COREB from the French Infectious Diseases Society who provided insight and expertise that greatly assisted the research.

We also thank the French Sentinel network for making their data publicly available.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2017.11.012.

References

[1] L. Brammer, A. Budd, N. Cox, Seasonal and pandemic influenza surveillance considerations for constructing multicomponent systems, Influenza Other Respir. Viruses 3 (2009) 51–58, doi:10.1111/j.1750-2659.2009.00077.x.

[2] J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, L. Brilliant, Detecting influenza epidemics using search engine query data, Nature 457 (2009) 1012–1014, doi:10.1038/nature07634.

[3] O.M. Araz, D. Bentley, R.L. Muelleman, Using Google flu trends data in forecasting influenza-like–illness related ED visits in Omaha, Nebraska, Am. J. Emerg. Med. 32 (2014) 1016–1023, doi:10.1016/j.ajem.2014.05.052.

[4] J.-P. Chretien, D. George, J. Shaman, R.A. Chitale, F.E. McKenzie, Influenza forecasting in human populations: a scoping review, PLOS ONE 9 (2014) e94130, doi:10.1371/journal.pone.0094130.

[5] W.W. Thompson, L. Comanor, D.K. Shay, Epidemiology of seasonal influenza: use of surveillance data and statistical models to estimate the burden of disease, J. Infect. Dis. 194 (2006) S82–S91, doi:10.1086/507558.

[6] A.J. Valleron, E. Bouvet, P. Garnerin, J. Ménarès, I. Heard, S. Letrait, J. Lefaucheux, A computer network for the surveillance of communicable diseases: the French experiment, Am. J. Public Health. 76 (1986) 1289–1292.

[7] P.M. Polgreen, Y. Chen, D.M. Pennock, F.D. Nelson, R.A. Weinstein, Using internet searches for influenza surveillance, Clin. Infect. Dis. 47 (2008) 1443–1448, doi:10.1086/593098.

[8] R. Chunara, S. Aman, M. Smolinski, J.S. Brownstein, Flu near you: an online self-reported influenza surveillance system in the USA, Online J. Public Health Inform. 5 (2013) http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3692780/ (accessed August 1, 2016).

[9] L. Josseran, J. Nicolau, N. Caillère, P. Astagneau, G. Brücker, Syndromic surveillance based on emergency department activity and crude mortality: two examples, Euro Surveill. Bull. Eur. Sur Mal. Transm. Eur. Commun. Dis. Bull. 11 (2006) 225–229.

[10] D.A. Broniatowski, M.J. Paul, M. Dredze, National and local influenza surveillance through Twitter: an analysis of the 2012–2013 influenza epidemic, PLOS ONE. 8 (2013) e83672, doi:10.1371/journal.pone.0083672.

[11] S. Yang, M. Santillana, S.C. Kou, Accurate estimation of influenza epidemics using Google search data via ARGO, Proc. Natl. Acad. Sci. U.S.A. 112 (2015) 14473–14478, doi:10.1073/pnas.1515373112.

[12] A.F. Dugas, M. Jalalpour, Y. Gel, S. Levin, F. Torcaso, T. Igusa, R.E. Rothman, Influenza forecasting with Google flu trends, PLOS ONE 8 (2013) e56176, doi:10.1371/journal.pone.0056176.

[13] D.R. Olson, K.J. Konty, M. Paladini, C. Viboud, L. Simonsen, Reassessing Google flu trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales, PLOS Comput. Biol. 9 (2013) e1003256, doi:10.1371/journal.pcbi.1003256.
[14] M.J. Paul, M. Dredze, D. Broniatowski, Twitter improves influenza forecasting, PLoS Curr. 6 (2014), doi:10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117.
[15] D.A. Broniatowski, M.J. Paul, M. Dredze, National and local influenza surveillance through Twitter: an analysis of the 2012–2013 influenza epidemic, PLOS ONE 8 (2013) e83672, doi:10.1371/journal.pone.0083672.
[16] K.S. Hickmann, G. Fairchild, R. Priedhorsky, N. Generous, J.M. Hyman, A. Deshpande, S.Y.D. Valle, Forecasting the 2013–2014 influenza season using Wikipedia, PLOS Comput. Biol. 11 (2015) e1004239, doi:10.1371/journal.pcbi.1004239.
[17] N. Generous, G. Fairchild, A. Deshpande, S.Y.D. Valle, R. Priedhorsky, Global disease monitoring and forecasting with Wikipedia, PLOS Comput. Biol. 10 (2014) e1003892, doi:10.1371/journal.pcbi.1003892.
[18] D.J. McIver, J.S. Brownstein, Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time, PLOS Comput. Biol. 10 (2014) e1003581, doi:10.1371/journal.pcbi.1003581.
[19] D. Lazer, R. Kennedy, G. King, A. Vespignani, The parable of Google flu: traps in big data analysis, Science 343 (2014) 1203–1205, doi:10.1126/science.1248506.
[20] D. Butler, When Google got flu wrong, Nature 494 (2013) 155–156, doi:10.1038/494155a.
[21] A.F. Dugas, Y.-H. Hsieh, S.R. Levin, J.M. Pines, D.P. Mareiniss, A. Mohareb, C.A. Gaydos, T.M. Perl, R.E. Rothman, Google flu trends: correlation with emergency department influenza rates and crowding metrics, Clin. Infect. Dis. 54 (2012) 463–469, doi:10.1093/cid/cir883.
[22] N.G. Weiskopf, C. Weng, Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research, J. Am. Med. Inform. Assoc. 20 (2013) 144–151, doi:10.1136/amiajnl-2011-000681.
[23] C. Saez, M. Robles, J.M. Garcia-Gomez, Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances, Stat. Methods Med. Res. (2014), doi:10.1177/0962280214545122.
[24] W. Raghupathi, V. Raghupathi, Big data analytics in healthcare: promise and potential, Health Inf. Sci. Syst. 2 (3) (2014).
[25] S.-Y. Shin, W.S. Kim, J.-H. Lee, Characteristics desired in clinical data warehouse for biomedical research, Healthc. Inform. Res. 20 (2014) 109–116, doi:10.4258/hir.2014.20.2.109.
[26] J.-M. Pinon, S. Calabretto, L. Poullet, Document semantic model: an experiment with patient medical records., ELPUB (1997) http://elpub.scix.net/data/works/att/97124.content.pdf (accessed April 21, 2015).
[27] D.A. Hanauer, EMERSE: The Electronic Medical Record Search Engine, AMIA. Annu. Symp. Proc. 941 (2006).
[28] S.N. Murphy, M.E. Mendis, D.A. Berkowitz, I. Kohane, H.C. Chueh, Integration of Clinical and Genetic Data in the i2b2 Architecture, AMIA. Annu. Symp. Proc. 1040 (2006).
[29] J. Rogers, C. Puleston, A. Rector, The CLEF chronicle: patient histories derived from electronic health records, in: 22nd Int. Conf. Data Eng. Workshop 2006 Proc., 2006 x109–x109, doi:10.1109/ICDEW.2006.144.
[30] H.J. Lowe, T.A. Ferris, P.M. Hernandez, S.C. Weber, STRIDE – an integrated standards-based translational research informatics platform, AMIA. Annu. Symp. Proc. (2009) 391–395.
[31] M. Cuggia, N. Garcelon, B. Campillo-Gimenez, T. Bernicot, J.-F. Laurent, E. Garin, A. Happe, R. Duvauferrier, Roogle: an information retrieval engine for clinical data warehouse, Stud. Health Technol. Inform. 169 (2011) 584–588.
[32] S.N. Murphy, G. Weber, M. Mendis, V. Gainer, H.C. Chueh, S. Churchill, I. Kohane, Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2), J. Am. Med. Inform. Assoc. 17 (2010) 124–130, doi:10.1136/jamia.2009.000893.
[33] R.W. Majeed, R. Röhrig, Automated realtime data import for the i2b2 clinical data warehouse: introducing the HL7 ETL cell, Stud. Health Technol. Inform. 180 (2012) 270–274.
[34] C. Jaglin-Grimonprez, Organiser, moderniser, innover : quelles avancées pour les patients, (2015). http://social-sante.gouv.fr/IMG/pdf/tr2_colloque-5_jaglin_20151016.pdf (accessed May 18, 2016).
[35] R.E. Serfling, Methods for current statistical analysis of excess pneumonia-influenza deaths, Public Health Rep 78 (1963) 494–506.
[36] C. Pelat, P.-Y. Boëlle, B.J. Cowling, F. Carrat, A. Flahault, S. Ansart, A.-J. Valleron, Online detection and quantification of epidemics, BMC Med. Inform. Decis. Mak. 7 (29) (2007), doi:10.1186/1472-6947-7-29.
[37] R Core Team, R: a language and environment for statistical computing, R Foundation for Statistical Computing, 2015 Vienna, Austriahttps://www.R-project.org.
[38] M. Muenchhoff, P.J.R. Goulder, Sex differences in pediatric infectious diseases, J. Infect. Dis. 209 (2014) S120–S126, doi:10.1093/infdis/jiu232.
[39] G.M. Weber, S.N. Murphy, A.J. McMurry, D. MacFadden, D.J. Nigrin, S. Churchill, I.S. Kohane, The shared health research information network (SHRINE): a prototype federated query tool for clinical data repositories, J. Am. Med. Inform. Assoc. 16 (2009) 624–630, doi:10.1197/jamia.M3191.
[40] G. Hripcsak, J.D. Duke, N.H. Shah, C.G. Reich, V. Huser, M.J. Schuemie, M.A. Suchard, R.W. Park, I.C.K. Wong, P.R. Rijnbeek, J. van der Lei, N. Pratt, G.N. Norén, Y.-C. Li, P.E. Stang, D. Madigan, P.B. Ryan, Observational health data sciences and informatics (OHDSI): opportunities for observational researchers, Stud. Health Technol. Inform. 216 (2015) 574–578.

2.5 Discussion des principaux résultats

L'étude réalisée a montré que les données massives hospitalières pouvaient apporter de l'information pertinente pour la surveillance des épidémies de grippe. En effet, la requête "grippe" mentionnée dans les comptes rendus des urgences est la requête ayant le coefficient de corrélation le plus élevé, 0.931, contre 0.925 pour le signal de GFT. Néanmoins, aujourd'hui, ce sont les données du web les plus utilisées comme alternative ou outil complémentaire aux méthodes traditionnelles.

Nous pensons que la force des données massives hospitalières, tout comme les données du web, est qu'elles peuvent être produites en temps réel ou presque réel. Elles peuvent également permettre d'avoir une description plus fine des personnes touchées, avec des caractéristiques supplémentaires comme l'âge ou le sexe. Ce découpage par classe ou par sexe est également réalisé par le réseau Sentinelles, ce qui pourrait permettre de faire des comparaisons entre les personnes se rendant à l'hôpital et celles se rendant chez les médecins généralistes. Les données hospitalières pourraient aussi permettre d'étudier la sévérité de l'épidémie avec une étude des cas nécessitant une hospitalisation ou des soins intensifs.

Une des limites de notre étude est de ne pas être en mesure de vérifier que tous les documents retournés par le moteur de recherche eHOP, correspondent bien à des cas avérés de grippe ou de syndromes grippaux. En effet, il est possible que le mot clé "grippe" soit mentionné si la personne ou un membre de sa famille, a déjà contracté le virus auparavant. Cependant, l'objectif ici n'était pas d'évaluer les performances du moteur de recherche eHOP mais de montrer que nous pouvions obtenir de l'information pertinente pour la surveillance des épidémies.
Une seconde limite de notre système est que toutes les sources de données de l'entrepôt ne sont pas chargées à la même fréquence. Par exemple, les résumés de sorties sont souvent produits et chargés dans l'entrepôt plusieurs jours après le

séjour du patient, ce qui n'est pas compatible avec une surveillance en temps réel. A l'inverse, les comptes rendus des urgences et les résultats de laboratoire sont produits immédiatement et chargés le soir même dans l'entrepôt.
De plus, les données que nous avons extraites ne proviennent que du CHU de Rennes, ce qui ne couvre pas entièrement la région Bretagne. Cependant, comme nous avons pu le mentionner dans la problématique, eHOP est en cours de développement dans les CHU de Nantes, Brest, Angers, Poitiers et Tours, ce qui pourrait nous permettre d'obtenir une vision globale de la Bretagne si nous étendions notre étude dans ces 5 CHU.

Pour finir, cette étude nous donne seulement la preuve que les données hospitalières sont pertinentes à utiliser pour la surveillance des épidémies grippales. Néanmoins, cela reste une étude descriptive, il est nécessaire de trouver un modèle approprié utilisant cette source de données afin de prédire les taux d'incidence en temps réel.

3 Comparaison des sources de données et des modèles statistiques

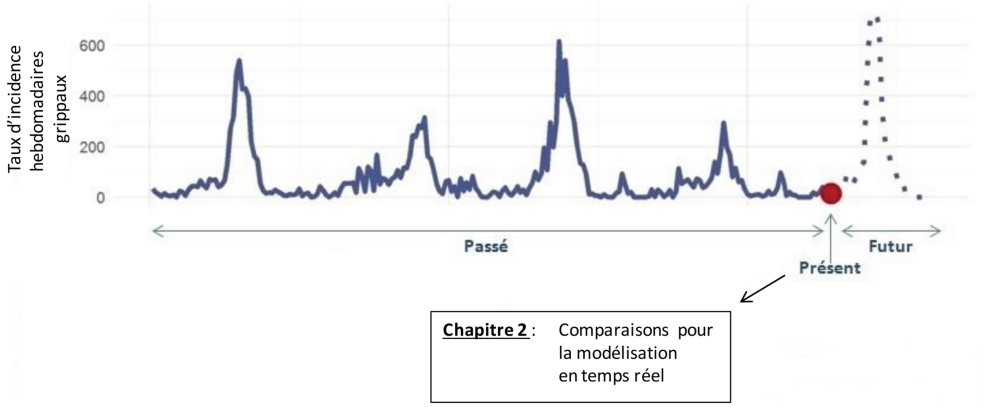


Figure 3.1: Chapitre 2 - Prévisions en temps réel

3.1 Problématique

L'étude précédente nous a permis de montrer que les données massives hospitalières étaient pertinentes et pourraient nous permettre de participer à la surveillance des épidémies grippales. Nous avons calculé la corrélation entre les signaux extraits de l'entrepôt de données cliniques et le signal estimé par le réseau Sentinelles, mais il est nécessaire de trouver un modèle statistique adapté afin de prévoir les taux d'incidence en temps réel.

Comme il a déjà été mentionné précédemment, de nombreux chercheurs ont utilisé les données du web afin de prévoir les taux d'incidence des épidémies grippales. Une étude en particulier a retenu notre attention, l'étude de Yang et al. [7], proposant une approche basée sur les données de Google, les données historiques du CDC et un modèle de régression linéaire pénalisée LASSO. A l'échelle nationale, le signal estimé est corrélé à 0.98 avec le signal réel estimé par le CDC.

Une étude plus récente, celle de Santillana et al. [8], se base sur les données massives hospitalières pour prévoir les taux d'incidence de la grippe en temps réel à l'échelle régionale. Pour cela, ces chercheurs utilisent un modèle SVM et obtiennent des coefficients de corrélation situés entre 0.90 et 0.99 en fonction de la région des États-Unis étudiée.

3.2 Objectif

L'objectif de cette étude est de déterminer si les données massives hospitalières permettent d'obtenir des résultats similaires ou de meilleurs résultats que les données du web ou si au contraire les données du web sont plus pertinentes.
Le second objectif, tout aussi important, est de trouver le meilleur modèle statistique permettant d'estimer en temps réel les taux d'incidence grippaux. Les prévisions ont été réalisées à l'échelle de la France et à l'échelle de la région Bretagne.

3.3 Considérations méthodologiques

3.3.1 Les sources de données

Les données cliniques de l'entrepôt eHOP

Tout comme pour l'étude précédente, afin d'étudier les données massives hospitalières, nous nous sommes appuyés sur les données présentes dans l'entrepôt de données cliniques eHOP du CHU de Rennes. Ici, les deux approches ont été conservées mais des requêtes supplémentaires ont été effectuées. Les requêtes sur données textuelles sont :

- Des requêtes en lien avec la grippe ou les syndromes grippaux avec les mots-clés :
  - "grippe"
  - "syndrome grippal"
  - "grippe" ou "syndrome grippal"

    - "grippe" ou "syndrome grippal" avec absence de "vaccin grippe" ou "vaccination"
    - "vaccin grippe"
    - "grippe" ou "syndrome grippal" aux urgences
- Des requêtes en lien avec les symptômes :
    - "fièvre" ou "pyrexie"
    - "courbatures" ou "douleurs musculaires"
    - fièvre" ou "pyrexie" ou "courbatures" ou "douleurs musculaires"
    - fièvre" ou "pyrexie" et "courbatures" ou "douleurs musculaires"
- Des requêtes en lien avec les médicaments :
    - "Tamiflu"

Les requêtes sur données structurées sont restées identiques mais chaque code CIM-10 et chaque test effectué en laboratoire ont fait l'objet d'un signal différent.
Au total, en combinant les requêtes précédentes avec les mots clés "OU" et "ET", nous avons extrait 34 signaux de l'entrepôt de données eHOP. Les données ont été récupérées pour la période allant du 14 décembre 2003 au 24 octobre 2016.

Les données Google

Le moteur de recherche Google met à disposition un service appelé Google Correlate [24] permettant d'obtenir les 100 requêtes jouées par les internautes les plus corrélées à un signal de référence. Ainsi, sur l'interface web, nous fournissons le signal national des taux d'incidence de grippe estimé par le réseau Sentinelles, et nous obtenons la fréquence par semaine des 100 requêtes les plus corrélées. Les 100 requêtes les plus corrélées au signal breton ont également été récupérées. La corrélation a été calculée entre le mois de Janvier 2004 et le mois d'Octobre 2016 et les fréquences hebdomadaires ont été recueillies pour la période allant du 4 janvier 2004 au 24 octobre 2016.

Les données du réseau Sentinelles

Sur le site du réseau Sentinelles, nous avons recueilli les taux d'incidence (pour 100 000 habitants) des syndromes grippaux pour la région France et la région bretonne. Ces données ont été récupérées pour la période allant du 28 décembre 2002 au 24 octobre 2016. Ces données sont utilisées comme signal de référence mais aussi comme variables historiques lors de la construction de nos modèles.

3.3.2 Préparation des jeux de données

En s'appuyant sur les articles présentés dans la problématique [7, 8] et utilisant des jeux de données avec un nombre de variables explicatives très différent, nous avons construit pour chaque source de données (Google et eHOP), deux jeux de données : un jeu de données avec un grand nombre de variables explicatives et un jeu de données avec un nombre plus restreint. Ces jeux de données ont été construits à l'échelle nationale et régionale. Chacun de ces jeux de données ont été complétés avec les données historiques du réseau Sentinelles. Nous avons donc au niveau national et au niveau régional les jeux de données suivant :

- eHOP Complet : Ce jeu de données contient les 34 variables extraites d'eHOP ainsi que 52 semaines d'historique pour les données du réseau Sentinelles.
- eHOP Restreint : Ce jeu de données contient les 3 variables extraites d'eHOP les plus corrélées au signal de référence ainsi que 2 semaines d'historique pour les taux d'incidence grippaux.
- Google Complet : Ce jeu de données contient les 100 variables extraites de Google Correlate ainsi que 52 semaines d'historique pour les données du réseau Sentinelles.
- Google Restreint : Ce jeu de données contient les 3 variables de Google les plus corrélées au signal de référence ainsi que 2 semaines d'historique pour les taux d'incidence grippaux.

3.3.3 Les modèles statistiques comparés

Notre période de test débutait le 28 décembre 2009 et se terminait le 24 octobre 2016. Les modèles ont été ajustés en utilisant un jeu d'apprentissage correspondant aux 6 années qui précèdent la semaine que l'on cherche à prédire. Tous les modèles ont été recalibrés chaque semaine afin d'incorporer les nouvelles informations à disposition. Trois modèles ont ainsi été comparés :

Elastic Net

Le modèle Elastic Net est un modèle de régression linéaire pénalisée, permettant de prendre en compte le grand nombre de variables explicatives et également la corrélation pouvant être présente entre ces variables [25]. Pour cela, elle combine les avantages de deux autres régressions linéaires pénalisées, les méthodes LASSO et Ridge [26, 27]. À l'origine, le modèle de régression linéaire multiple s'écrit sous la forme :

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$$

où $\beta_0, \beta_1, ..., \beta_p$ sont les paramètres du modèle à estimer et où les résidus ϵ_i vérifient $\mathbb{E}[\epsilon_i] = 0$, $cov(\epsilon_i, \epsilon_j) = 0 \ \forall i \neq j$ et $var(\epsilon_i) = \sigma^2$. Les résidus sont supposés gaussiens pour l'inférence statistique. Les paramètres $\beta_0, \beta_1, ..., \beta_p$ sont estimés par la méthode des moindres carrés ordinaires. Pour cela, les estimateurs $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ minimisent le critère empirique :

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ip})^2.$$

Le principe de la régression linéaire pénalisée, est d'ajouter une contrainte sur les coefficients à estimer afin de pouvoir maîtriser l'amplitude de leurs valeurs. Ainsi, dans le cas de la régression Ridge, le critère à minimiser est de la forme :

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{jp})^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

où λ est un hyper-paramètre à fixer afin de contrôler l'impact de la pénalité. Dans le cas de la régression LASSO, le critère à minimiser est de la forme :

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{jp})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

A la différence de la régression Ridge, la régression LASSO peut permettre d'effectuer une sélection de variables, en attribuant la valeur nulle à certains coefficients β_j, comme le montre la figure 3.2.

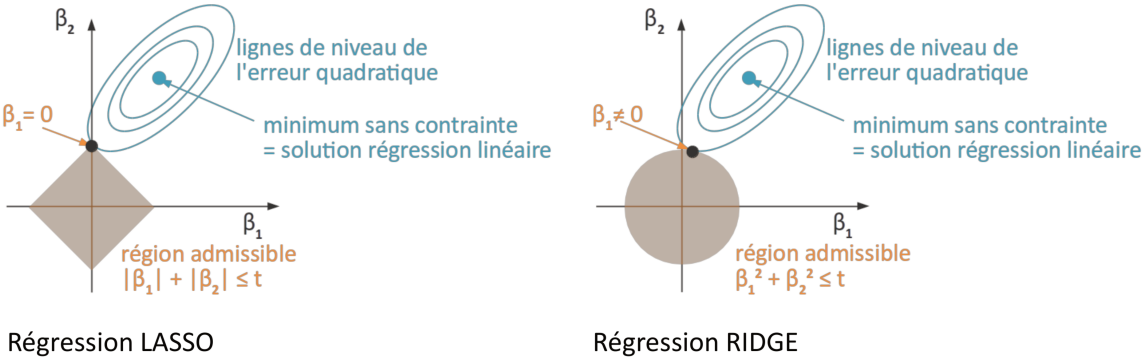


Figure 3.2: Contraintes régressions Ridge et Lasso

Afin de prévoir les épidémies de grippe à l'aide des données massives hospitalières, notre modèle Elastic Net s'écrit sous la forme :

$$y_t = \mu_y + \sum_{j=1}^{N} \eta_j y_{t-j} + \sum_{i=1}^{K} \beta_i X_{i,t} + g(x_t, \epsilon_t)$$

où y_t est le taux d'incidence du syndrome grippal pour la semaine t, $X_{i,t}$ correspond aux données du web ou données hospitalières et $g(x_t, \epsilon_t)$ correspond aux résidus, éventuellement modélisés par un modèle ARIMA(p,d,q). Afin d'estimer les paramètres $\eta_1, ..., \eta_N$, $\beta_1, ..., \beta_K$, le critère à minimiser est de la forme :

$$\sum_{i=1}^{n}(y_t - \mu_y - \sum_{j=1}^{N} \eta_j y_{t-j} - \sum_{i=1}^{K} \beta_i X_{i,t})^2 +$$
$$\lambda[\alpha(\sum_{j=1}^{N} |\eta_j| + \sum_{i=1}^{K} |\beta_i|) + (1-\alpha)/2(\sum_{j=1}^{N} \eta_j^2 + \sum_{i=1}^{K} \beta_i^2)]$$

où λ est l'hyper-paramètre à fixer et α est le paramètre de mélange Elastic Net. $\alpha = 1$, correspond au cadre d'une régression Lasso et $\alpha = 0$ correspond à une régression Ridge. Afin d'optimiser les paramètres, nous avons utlisé une validation croisée 10 blocs. Le modèle a été réalisé grâce au package *glmnet* du logiciel R [28, 29].

Les forêts aléatoires (RF)

La méthode des forêts aléatoires est une technique d'apprentissage, basée sur

l'agrégation d'arbres de régression ou de discrimination, appelés aussi arbres de décision [30]. Comme pour la régression, les arbres de décision sont utilisés pour la prédiction ou l'explication d'une variable cible Y, à partir d'un ensemble de variables explicatives X. Le principe des arbres est de diviser l'ensemble des données d'apprentissage successivement en sous-groupes. Les sous-groupes doivent être le plus homogènes possibles, les divisions se font grâce aux variables explicatives, qui, à chaque étape, discriminent au mieux la variable cible. Les sous-groupes intermédiaires de la variable Y sont appelés noeuds et les sous-groupes finaux sont appelés feuilles. Dans notre étude, nous avons utilisé des arbres de régression car nous cherchions à prévoir une variable quantitative, les taux d'incidence de grippe. Les sous-groupes ont été construits grâce aux variables issues du réseau Sentinelles et les variables issues de eHOP et Google. Afin d'obtenir des sous-groupes les plus homogènes possibles, il est nécessaire de choisir en priorité les variables qui vont diminuer la variance intra-classe. Les valeurs des noeuds et des feuilles vont correspondre à la moyenne des taux d'incidence de grippe composant les sous-groupes. La figure 3.3 montre un exemple d'arbre de régression obtenu à partir de 4 variables quantitatives : $X1$, $X2$, $X3$ et $X4$.

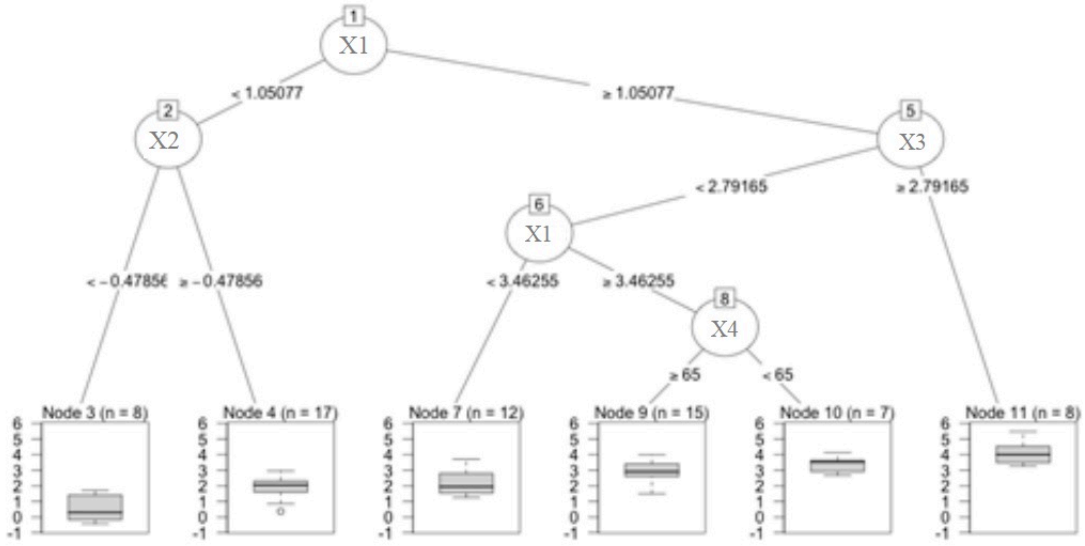


Figure 3.3: Exemple arbre de régression

Cependant, les arbres souffrent d'une grande instabilité. Lorsque le nombre de variables explicatives grandit, les données deviennent éparses et éloignées, ce qui pose problème pour obtenir des groupes d'individus homogènes (fléau de la dimension). De plus, les arbres est une méthode sensible à l'échantillonnage. c'est pour cette raison qu'il est nécessaire d'utiliser une méthode d'agrégation afin de diminuer la variance.
La forêt aléatoire (RF) est une méthode basée sur le "bagging" d'arbres. Le bagging a été introduit par Léo Breiman en 1996 et vient de la contraction de Boostrap Aggregating, qui consiste à agréger un nombre B d'estimateurs. Dans notre cas, les estimateurs correspondaient aux arbres de régression, nous en avons construit 1500. Chaque arbre de régression prédisait un taux d'incidence grippal puis l'estimation finale de notre taux d'incidence correspondait à la moyenne des 1500 prédictions :

$$\hat{y} = \frac{1}{1500} \sum_{b=1}^{1500} \hat{y}_b(X)$$

où $\hat{y}_b(X)$ est l'estimation du bième arbre de régression. Chaque arbre se distingue par le sous-échantillon de données sur lequel il est entraîné. En effet, pour chaque arbre, nous construisions un échantillon bootstrap en tirant aléatoirement avec remise, un nombre N d'observations identique à celui des données d'origine. De plus, une partie aléatoire est ajoutée, en ne choisissant que m variables explicatives parmi toutes les variables disponibles. Ce nombre m a été choisi par validation croisée 10 blocs. Le modèle a été réalisé grâce au package randomForest du logiciel R [28, 31].

Les machines à vecteurs supports (SVM)

La méthode des machines à vecteurs supports ou Support Vector Machine (SVM) fait partie de la famille des algorithmes d'apprentissage supervisé pour des problèmes de discrimination ou de régression [32]. Initialement construits pour de la classification binaire, l'objectif est de rechercher une règle de décision basée sur une séparation par hyperplan de marge optimale. En effet, si les données sont linéairement séparables, il existe une infinité d'hyperplans séparateurs et donc une infinité de règles de discrimination linéaires potentielles. Celle ayant les meilleures propriétés de

généralisation afin d'éviter un phénomène de sur-ajustement correspond à l'hyperplan séparateur de marge maximale γ.

Soit l'espace $\mathcal{X} = \mathbb{R}^d$ muni du produit scalaire usuel. Les données observées $d_1^n = (x_1, y_1), .., (x_n, y_n)$ sont dites linéairement séparables si, $\forall i = 1, .., n$, il existe (w, b) tel que :

- $y_i = 1$ si $< w, x_i > +b > 0$
- $y_i = -1$ si $< w, x_i > +b < 0$

avec $< w, x > +b = 0$ l'équation de l'hyperplan séparateur de vecteur orthogonal w.
La distance entre un point x et l'hyperplan est donnée par : $d(\mathcal{X}, x) = \frac{|<w,x>+b|}{||w||}$
Soit 2 entrées de l'ensemble d'apprentissage notées x_1 et x_{-1} de sorties respectives 1 et -1, se situant sur les frontières définissant la marge. Ces entrées sont appelées vecteurs supports. L'hyperplan séparateur correspondant se situe à mi-distance entre x_1 et x_{-1}. La marge s'exprime alors : $\gamma = \frac{1}{2}\frac{|<w,x_1-x_{-1}>|}{||w||} = \frac{1}{||w||}$
Trouver l'hyperplan séparateur de marge maximale revient donc à trouver le couple (w, b) tel que $||w||^2$ soit minimal sous la contrainte $y_i(< w, x_i > +b) \geq 1$. Ce problème d'optimisation est résolu grâce à la méthode des multiplicateurs de Lagrange. Une représentation schématique de la méthode SVM est présentée en figure 3.4. Celle-ci a été extraite d'un cours en ligne [33].

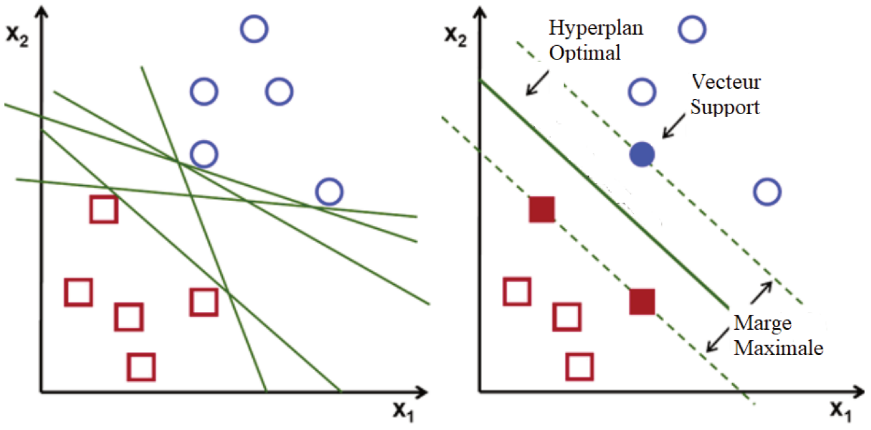


Figure 3.4: Représentation schématique SVM

Lorsque les données ne sont pas linéairement séparables, la solution est d'autoriser quelques vecteurs à être bien classés mais dans la région définie par la marge,

ou mal classés. On parle alors de marge souple ou marge relaxée. La contrainte $y_i(< w, x_i > +b) \geq 1$ devient $y_i(< w, x_i > +b) \geq 1 - \xi_i$ avec $\xi_i > 0$. Les variables ξ_i sont appelées variables ressorts. Cependant, au risque d'obtenir une marge maximale infinie, il est nécessaire de pénaliser les grandes valeurs de ξ_i. Le problème d'optimisation revient alors à minimiser $||w||^2 + C \sum_i^n \xi_i$ sous les contraintes $y_i(< w, x_i > +b) \geq 1 - \xi_i$ et $\xi_i > 0$.

Cependant, lorsque les données sont difficilement linéairement séparables, une SVM linéaire donnera une très mauvaise discrimination avec un nombre de vecteurs supports et un nombre de variables ressorts très élevé. Pour cela, il est possible de réaliser une SVM non linéaire en appliquant l'astuce du noyau ("kernel trick"). L'idée est d'envoyer les données dans un espace de plus grande dimension voire de dimension infinie, afin qu'elles deviennent linéairement séparables. L'espace de plus grande dimension est appelé espace de Hilbert $\mathcal{H}$ et est muni du produit scalaire $< ... >_{\mathcal{H}}$. Les données sont envoyées dans cet espace grâce à une fonction ϕ puis une SVM linéaire est appliquée aux nouvelles données $(\phi(x_i), y_i)$. Pour ne pas avoir à déterminer explicitement $\mathcal{H}$ et ϕ, l'astuce du noyau est employée. Un noyau est une fonction $k : \mathcal{X}$ x $\mathcal{X} \to \mathbb{R}$ telle que $k(x, x') =< \phi(x), \phi(x') >_{\mathcal{H}}$ pour une fonction $\phi : \mathcal{X} \to \mathcal{H}$. Il existe des noyaux classiques comme le noyau polynomial, le noyau gaussien ou le noyau laplacien.

Dans le cas de la régression, le principe reste le même, mais l'objectif n'est pas de trouver un hyperplan mais une fonction de régression la plus plate possible ayant au plus une déviation ϵ par rapport aux exemples d'apprentissage. Une représentation schématique de la SVR est présentée en figure 3.5. Cette figure a été extraite d'une publication de S.Lahiri et al. [34]

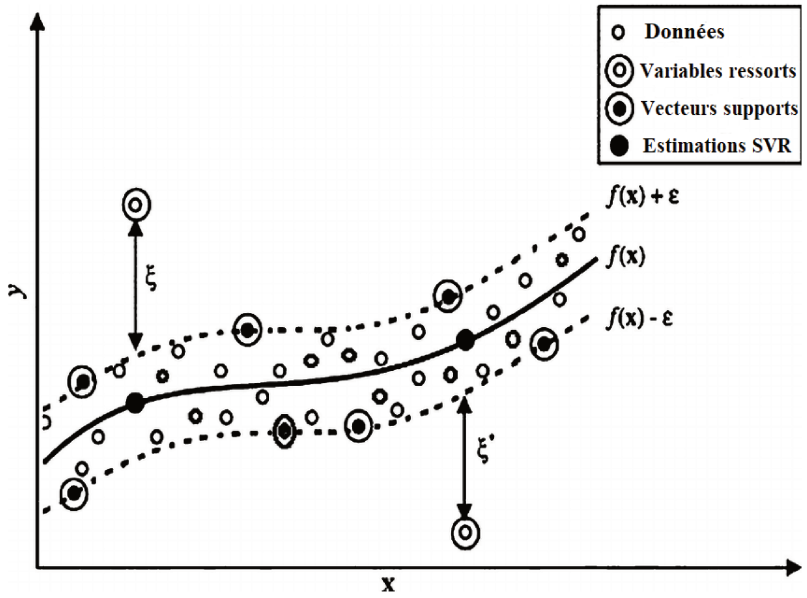


FIGURE 3.5: Représentation schématique SVR

En se basant sur l'étude de Santillana et al. [8], nous avons réalisé une SVR linéaire et optimisé la contrainte C via une validation croisée 10 blocs. Le modèle a été réalisé grâce au package R e1071 [28, 35].

3.3.4 Évaluation

Afin de comparer les estimations obtenues en fonction des modèles et des sources de données utilisés, nous avons calculé quatre indicateurs :

- La différence de hauteur entre le pic estimé et le pic du réseau Sentinelles (ΔH).
- La différence en nombre de semaines entre le pic que nous avons estimé et celui du pic du réseau Sentinelles (ΔL).
- L'erreur quadratique moyenne : $EQM = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- Le coefficient de corrélation : $PCC = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2}}$

avec y_i le taux d'incidence du réseau Sentinelles pour la semaine i ; $\hat{y}_i$ le taux d'incidence estimé pour la semaine i.

3.4 Article

3.4.1 Résumé de l'article

Objectif : Après avoir montré que les signaux extraits des entrepôts de données biomédicaux étaient corrélés aux signaux des réseaux de surveillance traditionnels, l'objectif est de montrer l'intérêt des données massives hospitalières par rapport aux données du web. Nous avons également évalué la performance de différents modèles statistiques pour la prévision des épidémies de grippe en temps réel.

Méthodes : Nous avons utilisé les données de Google afin d'évaluer les données du web et les données de l'entrepôt biomédical eHOP afin d'évaluer les données massives hospitalières. Nous avons comparé 3 modèles statistiques : Les forêts aléatoires (RF), les machines à vecteurs supports (SVM) et un modèle de régression linéaire Elastic Net.

Résultats : Les meilleures estimations au niveau national et régional ont été obtenues grâce aux données hospitalières et au modèle SVM, même si Elastic Net donne des résultats similaires. Le coefficient de corrélation obtenu était égal à 0.98 et l'EQM était égale à 866. Au niveau de la région Bretagne, le meilleur coefficient de corrélation était égal à 0.923 et l'EQM était égale à 2364.

Conclusion : Nous avons pu montrer que les données hospitalières combinées aux données historiques du réseau Sentinelles permettaient d'obtenir des estimations plus précises que les données du web au niveau national et régional, peu importe le modèle statistique utilisé. De plus, le modèle SVM et le modèle Elastic Net ont des performances comparables.

Original Paper

Real Time Influenza Monitoring Using Hospital Big Data in Combination with Machine Learning Methods: Comparison Study

Canelle Poirier[1,2], MSc; Audrey Lavenu[3], PhD; Valérie Bertaud[1,2,4], DMD, PhD; Boris Campillo-Gimenez[2,5], MD, MSc; Emmanuel Chazard[6,7], MD, PhD; Marc Cuggia[1,2,4], MD, PhD; Guillaume Bouzillé[1,2,4], MD, MSc

[1]Laboratoire Traitement du Signal et de l'Image, Université de Rennes 1, Rennes, France
[2]INSERM, U1099, Rennes, France
[3]Centre d'Investigation Clinique de Rennes, Université de Rennes 1, Rennes, France
[4]Centre Hospitalier Universitaire de Rennes, Centre de Données Cliniques, Rennes, France
[5]Comprehensive Cancer Regional Center, Eugene Marquis, Rennes, France
[6]Centre d'Etudes et de Recherche en Informatique Médicale EA2694, Université de Lille, Lille, France
[7]Public Health Department, Centre Hospitalier Régional Universitaire de Lille, Lille, France

Corresponding Author:
Canelle Poirier, MSc
Laboratoire Traitement du Signal et de l'Image
Université de Rennes 1
2 rue Henri Le Guilloux
Rennes, 35033
France
Phone: 33 667857225
Email: canelle.poirier@outlook.fr

Abstract

Background: Traditional surveillance systems produce estimates of influenza-like illness (ILI) incidence rates, but with 1- to 3-week delay. Accurate real-time monitoring systems for influenza outbreaks could be useful for making public health decisions. Several studies have investigated the possibility of using internet users' activity data and different statistical models to predict influenza epidemics in near real time. However, very few studies have investigated hospital big data.

Objective: Here, we compared internet and electronic health records (EHRs) data and different statistical models to identify the best approach (data type and statistical model) for ILI estimates in real time.

Methods: We used Google data for internet data and the clinical data warehouse eHOP, which included all EHRs from Rennes University Hospital (France), for hospital data. We compared 3 statistical models—random forest, elastic net, and support vector machine (SVM).

Results: For national ILI incidence rate, the best correlation was 0.98 and the mean squared error (MSE) was 866 obtained with hospital data and the SVM model. For the Brittany region, the best correlation was 0.923 and MSE was 2364 obtained with hospital data and the SVM model.

Conclusions: We found that EHR data together with historical epidemiological information (French Sentinelles network) allowed for accurately predicting ILI incidence rates for the entire France as well as for the Brittany region and outperformed the internet data whatever was the statistical model used. Moreover, the performance of the two statistical models, elastic net and SVM, was comparable.

(JMIR Public Health Surveill 2018;4(4):e11361) doi:10.2196/11361

KEYWORDS

electronic health records; hospital big data; internet data; influenza; machine learning; Sentinelles network

Introduction

Background

Influenza is a major public health problem. Outbreaks cause up to 5 million severe cases and 500,000 deaths per year worldwide [1-5]. During influenza peaks, large increase in visits to general practitioners and emergency departments causes health care system disruption.

To reduce its impact and help organize adapted sanitary responses, it is necessary to monitor influenza-like illness (ILI; any acute respiratory infection with fever ≥ 38°C, cough, and onset within the last 10 days) activity. Some countries rely on clinical surveillance schemes based on reports by sentinel physicians [6], where volunteer outpatient health care providers report all ILI cases seen during consultation each week. In France, ILI incidence rate is then computed at the national or regional scale by taking into account the number of sentinel physicians and medical density of the area of interest. ILI surveillance networks produce estimates of ILI incidence rates, but with a 1- to 3-week delay due to the time needed for data processing and aggregation. This time lag is an issue for public health decision making [2,7]. Therefore, there is a growing interest in finding ways to avoid this information gap. Nsoesie et al [8] reviewed methods for influenza forecasting, including temporal series and compartmental methods. The authors showed that these models have limitations. For instance, influenza activity is not consistent from season to season, which is a problem for temporal series. Alternative strategies have been proposed, including using different data sources, such as meteorological or demographic data, combined with ILI surveillance network data [9-11] or big data, particularly Web data [12]. With over 3.2 billion Web users, data flows from the internet are huge and of all types; they can be from social networks (eg, Facebook and Twitter), viewing sites, (eg, YouTube and Netflix), shopping sites, (eg, Amazon and Cdiscount), but also from sales or rentals website between particulars (eg, Craigslist and Airbnb). In the case of influenza, some studies used data from Google [2,4,9,13-16], Twitter [17,18], or Wikipedia [19-21]. The biggest advantage of Web data is that they are produced in real time. One of the first and most famous studies on the use of internet data for detecting influenza epidemics is Google Flu Trends [13,22], a Web service operated by Google. They showed that internet users' searches are strongly correlated with influenza epidemics. However, for the influenza season 2012-2013, Google Flu Trends clearly overestimated the flu epidemic due to the announcement of a pandemic that increased the internet users' search frequency, whereas the pandemic finally did not appear. The lack of robustness, due to the sensitivity to the internet users' behavioral changes and the modifications of the search engine performance led to stop the Google Flu Trends algorithm [2,23,24].

Some authors updated the Google Flu Trends algorithm by including data from other sources, such as historical flu information for instance or temperature [2,13-16]. Yang et al [2] proposed an approach that relies on Web-based data (Centers for Diseases Control ILI activity and Google data) and on a dynamic statistical model based on a least absolute shrinkage and selection operator (LASSO) regression that allows overcoming the aforementioned issues. At the national scale, the correlation between predictions and incidence rates was 0.98.

The internet is not the only data source that can be used to produce information in real time. With the widespread adoption of electronic health records (EHRs), hospitals also produce a huge amount of data that are collected during hospitalization. Moreover, many hospitals are implementing information technology tools to facilitate the access to clinical data for secondary-use purposes. Among these technologies, clinical data warehouses (CDWs) are one of the solutions for hospital big data (HBD) exploitation [25-28]. The most famous is the Informatics for Integrating Biology & the Bedside (i2b2) project, developed by the Harvard Medical School, which is now used worldwide for clinical research [29,30]. In addition, it has been shown that influenza activity changes detected retrospectively with EHR-based ILI indicators are highly correlated with the influenza surveillance data [31,32]. However, few HBD-based models have been developed to monitor influenza [7,33]. Santillana et al proposed a model using HBD and a machine learning algorithm (support vector machine [SVM]) with a good performance at the regional scale [7]. The correlation between estimates and ILI incidence rates ranges from 0.90 to 0.99, depending on the region and season.

Objectives

It would be interesting to determine whether HBD gives similar, better, or lower results than internet data with these statistical models (machine learning and regression). To this aim, we first evaluated HBD capacity to estimate influenza incidence rates compared with internet data (Google data). Then, we aim to find the best statistical model to estimate influenza incidence rates at the national and regional scales by using HBD or internet data. As these models have been described in the literature, we focused on two machine learning algorithms, random forest (RF) and SVM, and a linear regression model, elastic net.

Methods

Data Sources

Clinical Data Warehouse eHOP

At Rennes University Hospital (France), we developed our own CDW technology called eHOP. eHOP integrates structured (laboratory test results, prescriptions, and International Classification of Diseases 10th Revision, ICD-10, diagnoses) and unstructured (discharge letter, pathology reports, and operative reports) patients data. It includes data from 1.2 million in- and outpatients and 45 million documents that correspond to 510 million structured elements. eHOP consists of a powerful search engine system that can identify patients with specific criteria by querying unstructured data with keywords, or structured data with querying codes based on terminologies. eHOP is routinely used for clinical research. The first approach to obtain eHOP data connected with ILI was to perform different full-text queries to retrieve patients who had, at least, one document in their EHR that matched the following search criteria:

1. Queries directly connected with flu or ILI were as follows:
   - "flu"
   - "flu" or "ILI"
   - "flu" or "ILI", in the absence of "flu vaccination"
   - "flu vaccination"
   - "flu" or "ILI", only in emergency department reports
2. Queries connected with flu symptoms were as follows:
   - "fever" or "pyrexia"
   - "body aches" or "muscular pain"
   - "fever or pyrexia" or "body aches or muscular pain"
   - "flu vaccination"
   - "fever or pyrexia" and "body aches or muscular pain"
3. Drug query was as follows:
   - "Tamiflu"

The second approach was to leverage structured data with the support of appropriate terminologies:

1. ICD-10 queries were as follows: J09.x, J10.x, or J11.x (chapters corresponding to influenza in ICD-10). We retained all diagnosis-related groups with these codes.
2. Laboratory queries were as follows: influenza testing by reverse transcription polymerase chain reaction; we retained test reports with positive or negative results because the aim was to evaluate more generally ILI symptom fluctuations and not specifically influenza.

In total, we did 34 queries. For each query, the eHOP search engine returned all documents containing the chosen keywords (often, several documents for 1 patient and 1 stay). For query aggregation, we kept the oldest document for 1 patient and 1 stay and then calculated, for each week, the number of stays with, at least, one document mentioning the keyword contained in the query. In this way, we obtained 34 variables from the CDW eHOP. Multimedia Appendix 1 shows the queries and the number of concerned stays. We retrieved retrospective data for the period going from December 14, 2003 to October 24, 2016. This study was approved by the local Ethics Committee of Rennes Academic Hospital (approval number 16.69).

Google Data

For comparison with internet data, we obtained the frequency per week of the 100 most correlated internet queries (Multimedia Appendices 2 and 3) by French users from Google Correlate [34], and we used this information to retrieve Google Trends data. Unlike Google Correlate, Google Trends data [35] are available in real time, but we had to use Google Correlate to identify the most correlated queries to a signal. The times series passed into Google Correlate are the national flu time series and the regional flu time series (Brittany region) obtained from the French Sentinelles network (see below). The time period used to calculate the correlation is from January 2004 to October 2016. We used the R package gtrendsR to obtain automatically Google Trends data from January 4, 2004 to October 24, 2016 [36,37].

Sentinelles Network Data

We obtained the national (Metropolitan France) and regional (Brittany region, because Rennes University Hospital, from which EHR data were obtained, is situated in this region) ILI incidence rates (per 100,000 inhabitants) from the French Sentinelles network [38-40] from December 28, 2002 to October 24, 2016. We considered these data as the gold standard and used them as independent historical variables for our models.

Data Preparation

Based on previous studies that included datasets with very different numbers of explanatory variables according to the used statistical model [2,7], we built two datasets (one with a large number of variables and another with a reduced number of selected variables) from eHOP and Google data, for both the national and regional analyses (Figure 1).

Each one of these four datasets was completed with historical Sentinelles data. Therefore, for this study, we used the following:

1. eHOP Complete: this eHOP dataset included all variables from eHOP and the historical data from the Sentinelles network with the ILI estimates for the 52 weeks that preceded the week under study (thus, from t-1 to t-52).
2. eHOP Custom: this eHOP dataset included the 3 most correlated variables between January 2004 and October 2016 from eHOP for the ILI signal for week t, −1 (t-1), and −2 (t-2), and historical information from the Sentinelles network with ILI estimates for t-1 and t-2.
3. Google Complete: this Google dataset included the 100 most ILI activity-correlated queries from Google Trends and historical information from the Sentinelles network with ILI estimates for t-1 to t-52.
4. Google Custom: this Google dataset included the 3 most ILI activity-correlated queries between January 2004 and October 2016 from Google Trends for t, (t-1), and (t-2) and historical data from the Sentinelles network with ILI estimates for (t-1) and (t-2).

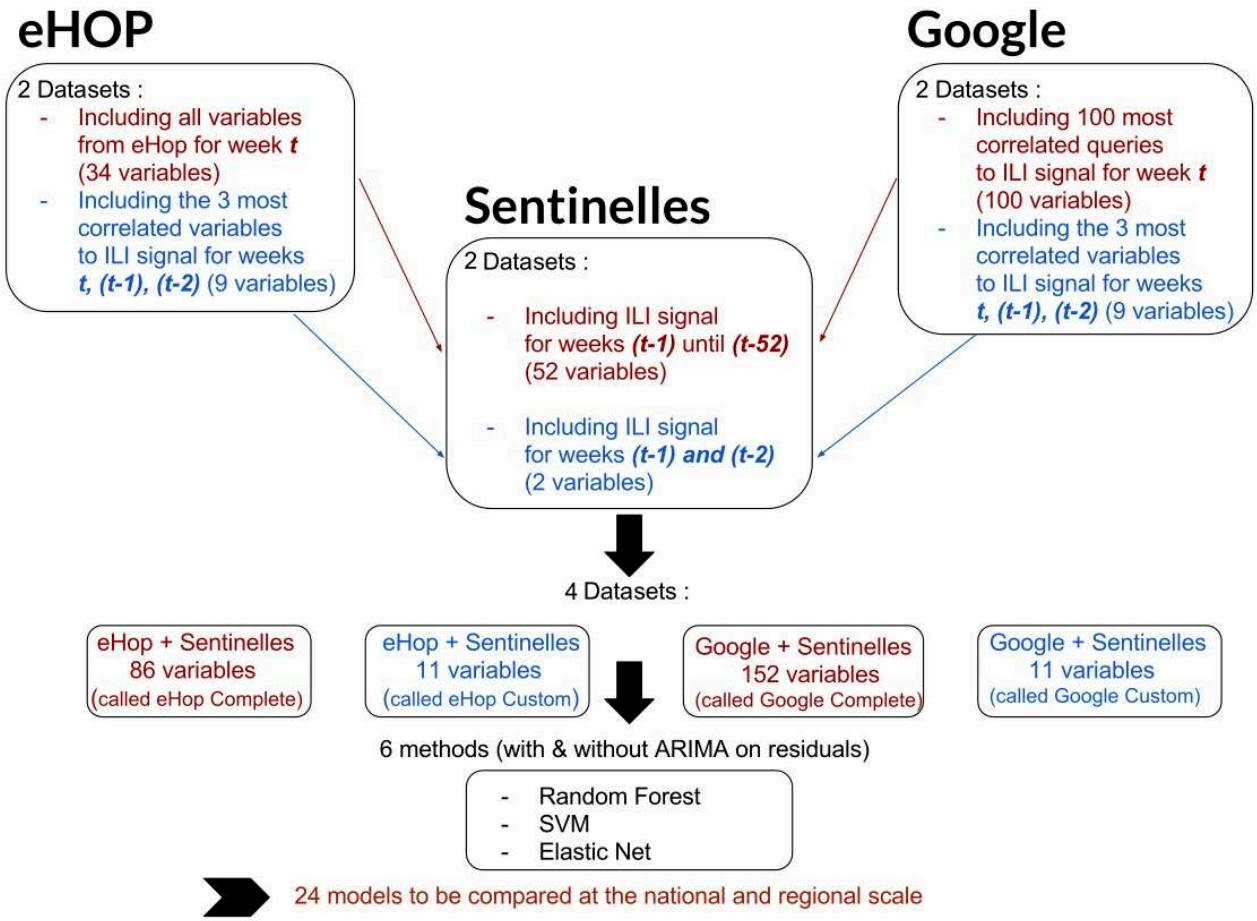
Statistical Models

Our test period started on December 28, 2009 and finished on October 24, 2016. We fitted our models using a training dataset that corresponded to the data for the previous 6 years. Each model was dynamically recalibrated every week to incorporate new information. For instance, to estimate the ILI activity fluctuations for the week starting on December 28, 2009, the training data consisted of data from December 21, 2003 to December 21, 2009.

Elastic Net

Elastic net is a regularized regression method that takes into account the correlation between explanatory variables and also a large number of predictors [41]. It combines the penalties of the LASSO and Ridge methods, thus allowing keeping the advantages of both methods and overcoming their limitations [42,43]. With datasets that may have up to 152 potentially correlated variables, we performed the elastic net regression analysis using the R package glmnet and the associated functions [36,44]. We fixed a coefficient alpha equal to.5 to give the same importance to the LASSO and Ridge constraints. We optimized the shrinkage parameter lambda via a 10-fold cross validation.

Figure 1. Schematic representation of the study design, including the data preparation and data modeling steps. ILI: influenza-like illness; SVM: support vector machine; ARIMA: autoregressive integrated moving average.



Random Forest

RF model combines decision trees constructed at training time using the general bootstrap aggregating technique (known as bagging) [45]. We used the R package randomForest to create RF models with a number of decision trees equal to 1500 [36,46].

Support Vector Machine

SVM is a supervised machine learning algorithm that can be used for classification or regression analyses [47]. Unlike multivariate regression models, SVM can learn nonlinear functions with the kernel trick that maps independent variables in a higher dimensional feature space. As Santillana et al [7], we used the linear kernel and optimized the cost parameter via a 10-fold cross validation with the R package e1071 [36,48].

Validity

Elastic net is a model that fulfills some assumptions on residuals. Means and variances must be constant, and residuals must be not correlated. Thus, residuals are called white noise. To test the stationarity and whiteness, we used Dickey Fuller's and Box-Pierce's tests available from the R packages tseries and stats [36,49]. When assumptions were not respected, we fitted residuals with a model of temporal series, called autoregressive integrated moving average (ARIMA) model. For RF and SVM, assumptions on residuals are not required. However, for comparison purpose, we tested them with the ARIMA model on residuals (Multimedia Appendices 4 and 5). We also assessed the calibration of the models by plotting the estimates against the real observations and by adding the regression line [50] (Multimedia Appendices 6 and 7).

Evaluation

We compared our ILI estimates with ILI incidence rates from the Sentinelles network by calculating different indicators. The mean squared error (MSE); Pearson correlation coefficient (PCC); variation in the height of the epidemic peak (ΔH), which corresponds to the difference between the height of the ILI incidence rate peak during the epidemic period estimated by the models and the height estimated by the Sentinelles network; and prediction lag (ΔL), which corresponds to the time difference between the ILI incidence rate peak estimated by the models and the peak estimated by the Sentinelles network, were calcuated. For the global comparison (ie, the entire study period), we calculated only the MSE and PCC. We calculated the four metrics only for the epidemic periods (plus 2 weeks before the start and after the end of the epidemic). The start and end date of epidemics were obtained from the Sentinelles network [39]. Indeed, clinicians want to know when an epidemic starts and finishes, as well as its amplitude and severity. Therefore, interepidemic periods are less important. We also calculated the mean of each indicator for each influenza season to assess

the model robustness. We also added two indicators to the mean of (ΔH) and (ΔL): the mean of |ΔH| and |ΔL|. We used the mean of (ΔH) to assess whether the models tended to underestimate or overestimate the peak calculated by the Sentinelles network, and the mean of (ΔL) to determine whether the predictions made by our models were too late or too in advance relative to the Sentinelles data. The mean of |ΔH| and |ΔL| allowed us to assess the estimate variability.

Results

Principal Results

Here, we show the results we obtained with the four datasets and three models—RF, SVM, and elastic net+residuals fitted by ARIMA (ElasticNet+ARIMA). The model on residuals was required to fulfill the assumptions for elastic net but not for the RF and SVM models. All results are presented in Multimedia Appendices 4 and 5. Moreover, we present two influenza outbreaks, including the 2010-2011 season (flu outbreak period for which the best estimates were obtained with all models) and the 2013-2014 season (flu outbreak period for which the worst estimates were obtained with all models; Multimedia Appendix 8). The calibration plots are in presented in Multimedia Appendices 7 and 9.

National Analysis

Dataset Comparison

PCC ranged from 0.947 to 0.980 when using the eHOP datasets (Multimedia Appendix 8) and from 0.937 to 0.978 with the Google datasets. MSE ranged from 2292 to 866 for the eHOP and from 2607 to 968 for the Google datasets. The mean PCC values during epidemic periods varied from 0.90 to 0.96 for the eHOP and from 0.87 to 0.96 for the Google datasets. The mean MSE values ranged from 7597 to 2664 for the eHOP and from 9139 to 2805 for the Google datasets.

Model Comparison

The eHOP Custom dataset gave the best results with the SVM model and ElasticNet+ARIMA (Multimedia Appendix 8). The SVM model and ElasticNet+ARIMA showed similar performance concerning the global activity (PCC=0.98; MSE, <900) and also during epidemic periods (mean values), although PCC decreased (0.96) and the MSE increased (> 2500). Both models tended to overestimate the height of the epidemic peaks (ΔH=6 with SVM; ΔH=26 with ElasticNet+ARIMA), but the SVM model was slightly more accurate (|ΔH|=19 for SVM; |ΔH|=30 for the ElasticNet+ARIMA model). Conversely, the SVM model showed a larger prediction lag (ΔL=+0.83). Figure 2 illustrates the estimates obtained with the best models (SVM and ElasticNet+ARIMA with the dataset eHop Custom).

The same figure with the dataset Google Custom is presented in Multimedia Appendix 10. In the same way, there is a figure with eHOP Custom and Google Custom datasets with the model ElasticNet+ARIMA presented in Multimedia Appendix 11.

For the outbreak of 2010-2011, eHOP Custom using ElasticNet+ARIMA gave the best PCC (0.98) and the best MSE (1222). With this model, there was a slight overestimation of the height of the epidemic peak (ΔH=23) and a prediction lag of 1 week. For the 2013-2014 outbreak, eHOP Custom using SVM gave the best PCC (0.95) and MSE (996), as well as the best ΔH (19) and prediction lag (1 week; Multimedia Appendix 8).

Regional Analysis

Figure 3 shows that ILI incidence rate variations were more important at the regional than the national level. For this reason, PCC decreased and MSE increased by the order of magnitude. The same figure with the dataset Google Custom is presented in Multimedia Appendix 12.

Dataset Comparison

PCC ranged from 0.911 to 0.923 (Multimedia Appendix 8) with the eHOP and from 0.890 to 0.912 with the Google datasets. MSE varied from 2906 to 2364 and from 3348 to 2736 for the eHOP and Google datasets, respectively. During epidemic periods, the mean PCC value ranged from 0.83 to 0.86 and from 0.70 to 0.83 for the eHOP and Google datasets, respectively. The mean MSE values ranged from 7423 to 5893 for the eHOP and from 9598 to 7122 for the Google datasets.

Model Comparison

Like at the national scale, eHOP Custom allowed obtaining the best PCC and MSE, and the SVM (PCC=0.923; MSE=2364) and ElasticNet+ARIMA (PCC=0.918; MSE=2451) models showed similar performances (Multimedia Appendix 8). Similar results were obtained also for the mean values during epidemic periods. Nevertheless, the PCC decreased (0.86 for SVM and 0.84 for ElasticNet+ARIMA), and the MSE increased (6050 for SVM and 5999 for ElasticNet+ARIMA). Both models tended to underestimate the height of the epidemic peaks (ΔH=−60 with SVM; ΔH=−32 with ElasticNet+ARIMA). The SVM model gave better PCC and MSE than the ElasticNet+ARIMA model, but ElasticNet+ARIMA was slightly more accurate for the epidemic peak height (|ΔH|=60 for SVM; |ΔH|=38 for the ElasticNet+ARIMA model). Although both models had a prediction lag (ΔL=+0.3), the ElasticNet+ARIMA model absolute lag value was smaller than that of SVM (|ΔL|=0.7; |ΔL|=1). For the 2010-2011 outbreak, eHOP Complete using the RF model gave the best PCC (0.92) and MSE (4263); with this model, there was a slight peak underestimation (ΔH=−40) but no prediction lag. For the 2013-2014 epidemic, the best PCC (0.78) and MSE (2113) were obtained with the Google Complete dataset and the ElasticNet+ARIMA model; there was a slight epidemic peak height underestimation (ΔH=−26) and 1 week of prediction lag.

Figure 2. National influenza-like illness (ILI) activity retrospective estimates obtained using the eHOP Custom dataset and the elastic net model with residuals fitted or the support vector machine model compared with the ILI activity levels from the French national Sentinelles networks. Global signal and 2010-2011 and 2013-2014 outbreaks are presented. SVM: support vector machine.

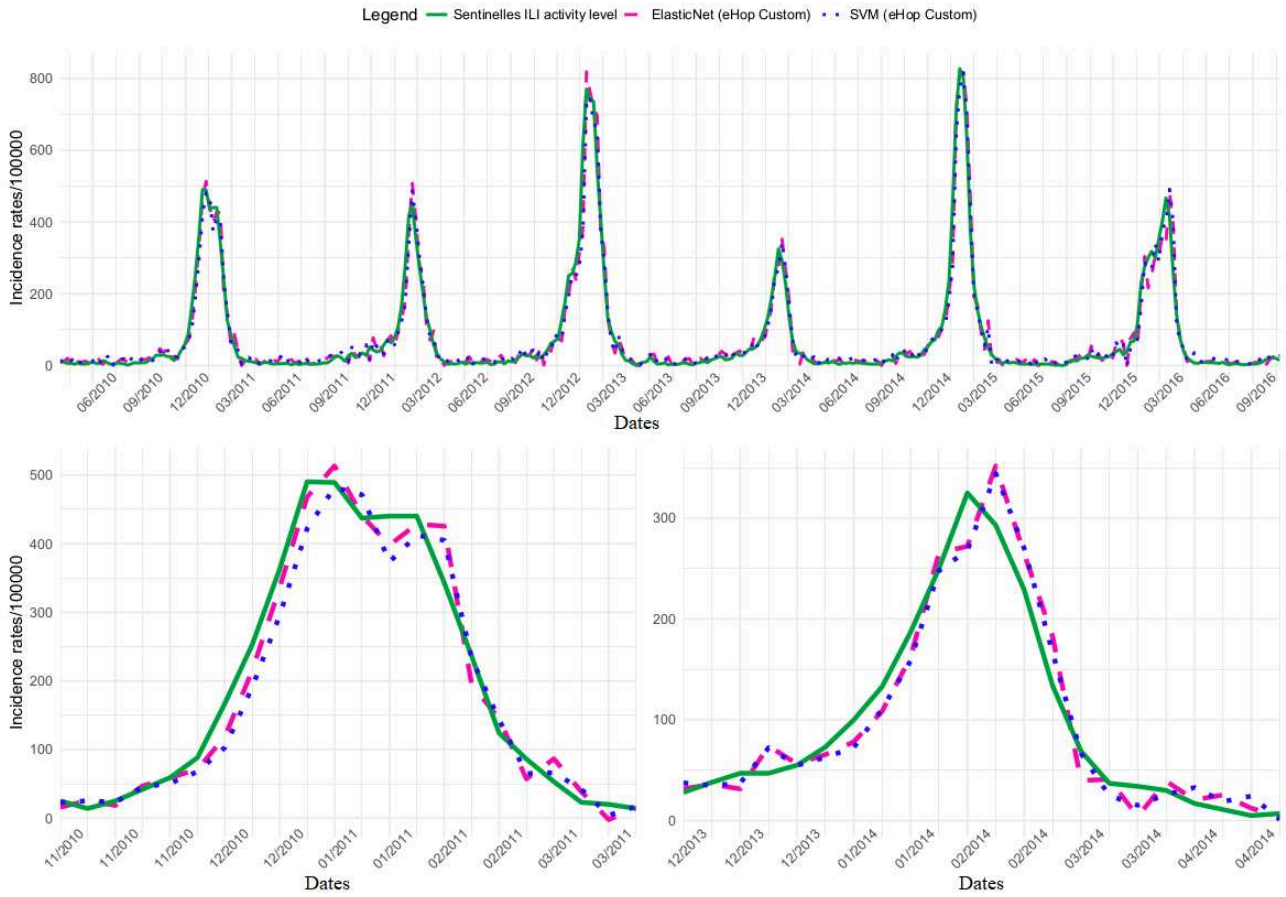
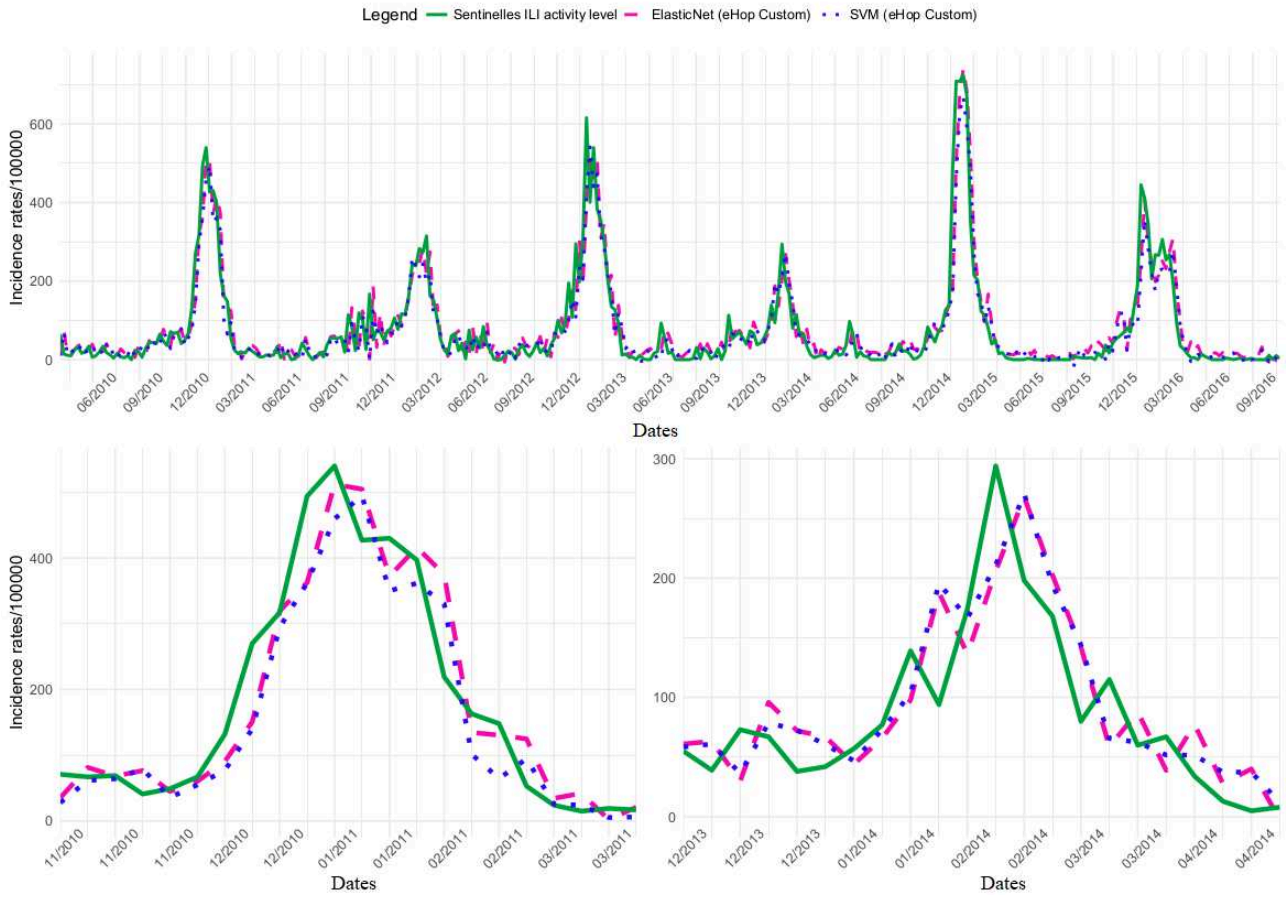


Figure 3. Regional influenza-like illness (ILI) activity retrospective estimates obtained using the eHOP Custom dataset and the elastic net model with residuals fitted or the support vector machine model compared with the ILI activity levels from the French regional Sentinelles networks. Global signal and 2010-2011 and 2013-2014 outbreaks are presented. SVM: support vector machine.



Discussion

Data

Here, we show that HBD in combination with flu activity-level data from a national surveillance network allows accurately predicting ILI incidence rate at the national and regional scale and outperform Google data in most cases. The correlation coefficients obtained for the French data are comparable to those reported by studies on US data [2,7]. At the national and regional level, the best PCC and the best MSE during the entire study period or during epidemics were obtained using the eHOP Custom dataset. Moreover, the PCC and MSE values obtained with the eHOP datasets were better than those obtained with the Google datasets, particularly at the regional level (PCC 0.911-0.923 vs 0.890-0.912; MSE 2906-2364 vs 3348-2736, respectively; Multimedia Appendix 8). However, the national signal is smoother and less noisy than the regional signal; the contribution of other data sources, such as hospital data or Web data, in addition to historical influenza data is more important at the regional level (Multimedia Appendices 4 and 5). The contribution of these external sources being less important at the national level, the differences observed between hospital data and Web data at this scale could be more significant.

Like internet data, some HBD can be obtained in near real time, especially records from emergency departments that are available on the same day or the day after. This is the most important data source for our models using eHOP datasets. Some other data, such as laboratory results, are available only on a weekly basis; however, they are not the most important data source for our models.

Moreover, in comparison to internet data, HBD have some additional advantages. First, data extracted from CDWs are real health data can give information that cannot be extracted from internet data, particularly information about patients (sex, age, and comorbidities) [51]. In addition, an important clinical aspect is to determine the epidemic severity. With HBD, it is possible to gauge this parameter by taking into account the number of patients who were admitted in intensive care or died as the result of flu. Second, some CDW data (particularly emergency department discharge summaries and laboratory test results) can confirm that people were really affected by influenza or ILI symptoms. On the other hand, people can make internet queries not because they are ill, but for other people, for prevention purposes or just because it is a topical subject. Third, HBD could also be used to estimate the incidence rates of diseases that do not generate internet activity (eg, diseases without or with little media coverage or that are not considered interesting by the general population). Fourth, there is a spatial decorrelation between internet data and the regional estimates that were not observed with the eHOP data. It is quite reasonable that hospital-based data give a better estimate of regional epidemics, although currently, we have only data from Rennes University

Hospital that might not be representative of the entire Brittany region.

A major HBD limitation is that, generally, clinical data are not publicly available. In our case, we could only access the Rennes University Hospital HBD. However, the epidemic peak in Brittany could have occurred earlier or later relative to the national peak, and this could have introduced a bias in our estimation. We can hypothesize that ILI estimates, particularly nationwide, might be improved if we could extract information from HBD in other regions. In the United States, a patient research system allows aggregating patient observations from a large number of hospitals in a uniform way [52]. In France, several initiatives have been developed to create search systems. For instance, an ongoing project (Réseau interrégional des Centres de Données Cliniques) [53] in the Northwest area of France associates six University Hospital centers (Angers, Brest, Nantes, Poitiers, and Rennes et Tours) and Orleans Regional Hospital Centre, thus collecting data on patients in the Bretagne, Centre-Val de Loire, and Pays de la Loire regions. This corresponds to 15.5% of Metropolitan France and 14.4% of the entire French population. Another way to aggregate patient data could be a cloud-based platform, and we are currently setting up this kind of architecture; this platform will integrate two University Hospital centers, Brest and Rennes, the French health reimbursement database (Système national d'information interrégimes de l'Assurance Maladie) and registries, such as the birth defect registry or cancer registry.

Statistical Models

Regarding the statistical models, we show that SVM and elastic net with ARIMA model are fairly comparable with PCC ranging from 0.970 to 0.980 at the national scale and from 0.890 to 0.923 at the regional scale. The SVM and elastic net models in combination with the eHOP custom dataset were the most robust models, although they did not always give the best results. Indeed, they showed the best performance in term of PCC and MSE for the global signal and also for the mean values. Nevertheless, these models have some limits. The main limitation of the SVM model is the very slow parameter optimization when there are many variables. With the SVM model, it can be important to preselect the important variables to reduce the dataset size to improve the optimization speed. For this, one needs a good knowledge of the available data, which may be difficult when using big data. On the other hand, elastic net shows good performance with many variables, which is an advantage when the most relevant variables to estimate ILI incidence rates are not known in advance. The elastic net model is a parametric model that fulfills certain assumptions on residuals, differently from the SVM model. With elastic net, residuals must be fitted to have a statistically valid model. Nevertheless, if we had to choose a model, we would prefer SVM with the eHOP Custom dataset because it has a better PCC than elastic net at the regional scale.

Another limitation is that indicators are better for the global period than for epidemic periods. This implies that models are less efficient during flu outbreaks, while clinical concerns are higher during epidemics when good estimates of the outbreak starting date, amplitude, and end are needed.

Finally, the results of our models with Web data may have been overestimated due to the way we obtained data from Google Correlate. Indeed, Google Correlate used information that we did not have at the beginning of our test period. The time period for our time series passed into Google Correlate is from January 2004 to October 2016. But, the beginning of our test period for our models is January 2010. To be more precise, we should recalculate the correlation coefficients for each week to predict with the data available at that time.

In the same way, to custom datasets, we calculated the 3 most correlated variables on a time period including our test period. To compare the results, we built another dataset from eHOP, including the 3 most correlated variables to ILI regional signal between December 2003 and December 2009 (before our test period), and we applied an ElasticNet+ARIMA model. In this way, we kept 2 variables on the 3 present in the eHOP custom dataset. The difference does not seem significant (Multimedia Appendix 6), but it would be interesting to test this hypothesis with all models at the national and regional scale with Google and eHOP custom datasets.

Perspectives

Future research could address clinical issues not only nationally or regionally but also at finer spatial resolutions such as a city like Lu et al did [54], a health care institution or in subpopulations. Indeed, by predicting epidemics, it will be possible to organize hospitals during epidemics (eg, bed planning and anticipating overcrowding). Moreover, in this study, we compared internet and HBD data; however, hybrid systems could be developed to take advantage of multiple sources [55,56]. For instance, internet data might avoid the limit of the local source linked to the choice or availability of HBD. Data collected by volunteers who self-report symptoms in near real time could be exploited [57]. Similarly, by combining models, we could retain the benefits of each of them and improve the estimates of ILI incidence rates. For example, we could use another algorithm, such as stacking [58], to concomitantly use the SVM and elastic net models. We could also test other kernels than the linear kernel for SVM models. Finally, we carried out a retrospective study using various models with clinical data in combination with the flu activity from the Sentinelles network to estimate ILI incidence rates in real time. Our models need now to be tested to determine whether they can anticipate and predict ILI incidence rates.

Conclusions

Here, we showed that HBD is a data source that allows predicting the ILI activity as well or even better than internet data. This can be done using two types of models with similar performance—SVM (a machine learning model) and elastic net (a model of regularized regression). This is a promising way for monitoring ILI incidence rates at the national and local levels. HBD presents several advantages compared with internet data. First, they are real health data and can give information about patients (sex, age, and comorbidities). This could allow for making predictions on ILI activity targeted to a specific group of people. Second, hospital data can be used to determine the epidemic severity by taking into account the number of patients who were admitted in intensive care or died as a result

of flu. Third, hospital data (particularly the emergency department discharge summaries and laboratory test results) can confirm that people were really affected by influenza. Finally, HBD could also be used to estimate the incidence rates of diseases that do not generate internet activity. Although massive data cannot take the place of traditional influenza surveillance methods at this time, they could be used to complete them. For instance, real-time forecasting is necessary for decision making. It can also be used to manage the patients' flow in general practitioners' offices and hospitals, particularly emergency departments.

Acknowledgments

We would like to thank the French National Research Agency for funding this work inside the Integrating and Sharing Health Data for Research Project (Grant No. ANR-15-CE19-0024). We thank Magalie Fromont Renoir and Ronan Le Gue´vel from the University of Rennes 2 who provided insight and expertise that greatly assisted the research. We also thank the French Sentinelles network for making their data publicly available.

Authors' Contributions

CP, GB, AL, and BCG conceived the experiments; CP conducted the experiments and analyzed the results.

Conflicts of Interest

None declared.

Multimedia Appendix 1

eHOP queries (with the number of concerned hospital stays from 2003 to 2016).

[PDF File (Adobe PDF File), 21KB - publichealth_v4i4e11361_app1.pdf]

Multimedia Appendix 2

The 100 most correlated Google queries at national level.

[PDF File (Adobe PDF File), 15KB - publichealth_v4i4e11361_app2.pdf]

Multimedia Appendix 3

The 100 most correlated Google queries at regional level.

[PDF File (Adobe PDF File), 14KB - publichealth_v4i4e11361_app3.pdf]

Multimedia Appendix 4

Accuracy metrics for all seasons obtained with all models for the national scale.

[PDF File (Adobe PDF File), 72KB - publichealth_v4i4e11361_app4.pdf]

Multimedia Appendix 5

Accuracy metrics for all seasons obtained with all models for the regional scale.

[PDF File (Adobe PDF File), 80KB - publichealth_v4i4e11361_app5.pdf]

Multimedia Appendix 6

Comparison between two datasets with ElasticNet + ARIMA model: Dataset 1 corresponds to the dataset called eHOP Custom used in the paper and including the 3 most correlated variables to ILI signal between December 2009 to October 2016 (our test period). Dataset 2 includes the 3 most correlated variables to ILI signal between December 2003 to December 2009 (before our test period).

[PDF File (Adobe PDF File), 25KB - publichealth_v4i4e11361_app6.pdf]

Multimedia Appendix 7

National calibration.

[PNG File, 185KB - publichealth_v4i4e11361_app7.png]

Multimedia Appendix 8

Accuracy metrics for the 2010-2011 (flu outbreak period for which the best estimates were obtained with all models) and 2013-2014 (flu outbreak period for which the worst estimates were obtained with all models) seasons. PCC and MSE for the global period (Global) and mean values (Means) of all indicators for each model during the epidemic periods. In bold, the best results for each dataset. a. Data for the whole France. b. Data for the Brittany region.

[PDF File (Adobe PDF File), 52KB - publichealth_v4i4e11361_app8.pdf]

Multimedia Appendix 9

Regional calibration.

[PNG File, 202KB - publichealth_v4i4e11361_app9.png]

Multimedia Appendix 10

National ILI activity retrospective estimates obtained using the Google Custom dataset and the Elastic Net model with residuals fitted (pink dashed line) or the SVM model (blue dotted line) compared with the ILI activity levels from the French national Sentinelles networks (green solid line). a. Global signal. b. 2010-2011 and c. 2013-2014 outbreaks.

[PNG File, 142KB - publichealth_v4i4e11361_app10.png]

Multimedia Appendix 11

National ILI activity retrospective estimates obtained using the Google Custom dataset and the Elastic Net model (blue dotted line) or eHOP Custom dataset and the Elastic Net model (pink dashed line) compared with the ILI activity levels from the French national Sentinelles networks (green solid line). a. Global signal. b. 2010-2011 and c. 2013-2014 outbreaks.

[PNG File, 142KB - publichealth_v4i4e11361_app11.png]

Multimedia Appendix 12

Regional ILI activity retrospective estimates obtained using the Google Custom dataset and the Elastic Net model with residuals fitted (pink dashed line) or the SVM model (blue dotted line) compared with the ILI activity levels from the French regional Sentinelles networks (green solid line). a. Global signal. b. 2010-2011 and c. 2013-2014 outbreaks.

[PNG File, 159KB - publichealth_v4i4e11361_app12.png]

References

1. Ferguson NM, Cummings DAT, Fraser C, Cajka JC, Cooley PC, Burke DS. Strategies for mitigating an influenza pandemic. Nature 2006 Jul 27;442(7101):448-452. [doi: 10.1038/nature04795] [Medline: 16642006]
2. Yang S, Santillana M, Kou S. Accurate estimation of influenza epidemics using Google search data via ARGO. Proceedings of the National Academy of Sciences 2015 Nov 24:14473. [doi: 10.1038/srep25732]
3. Si-Tahar M, Touqui L, Chignard M. Innate immunity and inflammation--two facets of the same anti-infectious reaction. Clin Exp Immunol 2009 May;156(2):194-198 [FREE Full text] [doi: 10.1111/j.1365-2249.2009.03893.x] [Medline: 19302246]
4. Yang W, Lipsitch M, Shaman J. Inference of seasonal and pandemic influenza transmission dynamics. Proc Natl Acad Sci USA 2015 Feb 17;112(9):2723-2728. [doi: 10.1073/pnas.1415012112] [Medline: 25730851]
5. Nichol KL. Cost-benefit analysis of a strategy to vaccinate healthy working adults against influenza. Arch Intern Med 2001 Mar 12;161(5):749-759. [Medline: 11231710]
6. Fleming DM, Van Der Velden J, Paget WJ. M. Fleming WJP J van der Velden. The evolution of influenza surveillance in Europe and prospects for the next 10 years. Vaccine ? 2003;21:1753.
7. Santillana M, Nguyen AT, Louie T, Zink A, Gray J, Sung I, et al. Cloud-based Electronic Health Records for Real-time, Region-specific Influenza Surveillance. Sci Rep 2016 Dec 11;6:25732 [FREE Full text] [doi: 10.1038/srep25732] [Medline: 27165494]
8. Nsoesie E, Brownstein J, Ramakrishnan N. A systematic review of studies on forecasting the dynamics of influenza outbreaks. Influenza and Other Respiratory Viruses ? 2014;8:316.
9. Chretien J, George D, Shaman J, Chitale RA, McKenzie FE. Influenza forecasting in human populations: a scoping review. PLoS One 2014;9(4):e94130 [FREE Full text] [doi: 10.1371/journal.pone.0094130] [Medline: 24714027]
10. Soebiyanto RP, Adimi F, Kiang RK. Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters. PLoS One 2010 Mar 01;5(3):e9450 [FREE Full text] [doi: 10.1371/journal.pone.0009450] [Medline: 20209164]

11. Shaman J, Karspeck A, Yang W, Tamerius J, Lipsitch M. Real-time influenza forecasts during the 2012-2013 season. Nat Commun 2013;4:2837 [FREE Full text] [doi: 10.1038/ncomms3837] [Medline: 24302074]
12. Milinovich GJ, Williams GM, Clements ACA, Hu W. Internet-based surveillance systems for monitoring emerging infectious diseases. Lancet Infect Dis 2014 Feb;14(2):160-168. [doi: 10.1016/S1473-3099(13)70244-5] [Medline: 24290841]
13. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009 Feb 19;457(7232):1012-1014. [doi: 10.1038/nature07634] [Medline: 19020500]
14. Shaman J, Karspeck A. Forecasting seasonal outbreaks of influenza. Proceedings of the National Academy of Sciences 2012 Nov 26;109(50):20425-20430. [doi: 10.1073/pnas.1208772109] [Medline: 23184969]
15. Olson DR, Konty KJ, Paladini M, Viboud C, Simonsen L. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. PLoS Comput Biol 2013;9(10):e1003256 [FREE Full text] [doi: 10.1371/journal.pcbi.1003256] [Medline: 24146603]
16. Zhang Y, Bambrick H, Mengersen K, Tong S, Hu W. Using Google Trends and ambient temperature to predict seasonal influenza outbreaks. Environment International 2018;117:91.
17. Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. PLoS One 2013;8(12):e83672 [FREE Full text] [doi: 10.1371/journal.pone.0083672] [Medline: 24349542]
18. Paul MJ, Dredze M, Broniatowski D. Twitter improves influenza forecasting. PLoS Curr 2014 Oct 28;6 [FREE Full text] [doi: 10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117] [Medline: 25642377]
19. Hickmann KS, Fairchild G, Priedhorsky R, Generous N, Hyman JM, Deshpande A, et al. Forecasting the 2013-2014 influenza season using Wikipedia. PLoS Comput Biol 2015 May;11(5):e1004239 [FREE Full text] [doi: 10.1371/journal.pcbi.1004239] [Medline: 25974758]
20. Generous N, Fairchild G, Deshpande A, Del Valle SY, Priedhorsky R. Global disease monitoring and forecasting with Wikipedia. PLoS Comput Biol 2014 Nov;10(11):e1003892 [FREE Full text] [doi: 10.1371/journal.pcbi.1003892] [Medline: 25392913]
21. McIver DJ, Brownstein JS. Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. PLoS Comput Biol 2014 Apr;10(4):e1003581 [FREE Full text] [doi: 10.1371/journal.pcbi.1003581] [Medline: 24743682]
22. Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis 2009 Nov 15;49(10):1557-1564. [doi: 10.1086/630200] [Medline: 19845471]
23. Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. Science 2014 Mar 14;343(6176):1203-1205. [doi: 10.1126/science.1248506] [Medline: 24626916]
24. Butler D. When Google got flu wrong. Nature 2013 Feb 14;494(7436):155-156. [doi: 10.1038/494155a] [Medline: 23407515]
25. Hanauer DA. EMERSE: The Electronic Medical Record Search Engine. 2006 Presented at: AMIA Annual Symposium Proceedings; 2006/11/11; Washington p. 941.
26. Murphy SN, Mendis ME, Berkowitz DA. Integration of Clinical and Genetic Data in the i2b2 Architecture. 2006 Presented at: AMIA Annual Symposium Proceedings; 2006; Washington p. 1040.
27. Lowe HJ, Ferris TA, Hernandez PM. STRIDE ? An Integrated Standards-Based Translational Research Informatics Platform. 2009 Presented at: AMIA Annual Symposium Proceedings; 2009; San Francisco p. 391.
28. Cuggia M, Garcelon N, Campillo-Gimenez B. Roogle: an information retrieval engine for clinical data. Studies in Health Technology and Informatics 2011;169:8. [doi: 10.3233/978-1-60750-806-9-584]
29. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17(2):124-130 [FREE Full text] [doi: 10.1136/jamia.2009.000893] [Medline: 20190053]
30. Murphy S, Wilcox A. Mission and Sustainability of Informatics for Integrating Biology and the Bedside (i2b2). EGEMS (Wash DC) 2014;2(2):1074 [FREE Full text] [doi: 10.13063/2327-9214.1074] [Medline: 25848608]
31. Viboud C, Charu V, Olson D, Ballesteros S, Gog J, Khan F, et al. Demonstrating the use of high-volume electronic medical claims data to monitor local and regional influenza activity in the US. PLoS One 2014;9(7):e102429 [FREE Full text] [doi: 10.1371/journal.pone.0102429] [Medline: 25072598]
32. Bouzillé G, Poirier C, Campillo-Gimenez B, Aubert ML, Chabot M, Chazard E, et al. Leveraging hospital big data to monitor flu epidemics. Computer Methods and Programs in Biomedicine 2018:160.
33. Santillana M, Nsoesie EO, Mekaru SR, Scales D, Brownstein JS. Using clinicians' search query data to monitor influenza epidemics. Clin Infect Dis 2014 Nov 15;59(10):1446-1450 [FREE Full text] [doi: 10.1093/cid/ciu647] [Medline: 25115873]
34. Google Correlate. URL: https://www.google.com/trends/correlate [accessed 2018-06-19] [WebCite Cache ID 70IClAsSD]
35. Google Trends. URL: https://trends.google.fr/trends/?geo=FR [accessed 2018-06-20] [WebCite Cache ID 70JjgMxmh]
36. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing 2015 [FREE Full text]
37. Massicotte P, Eddelbuettel D. gtrendsR: Perform and Display Google Trends Queries. https://github.com/PMassicotte/gtrendsR 2017 [FREE Full text]

38. Valleron AJ, Bouvet E, Garnerin P. A computer network for the surveillance of communicable diseases: the French experiment. American Journal of Public Health 1986;76:92.
39. Flahault A, Blanchon T, Dorléans Y, Toubiana L, Vibert JF, Valleron AJ. Virtual surveillance of communicable diseases: a 20-year experience in France. Stat Methods Med Res 2006 Oct;15(5):413-421. [doi: 10.1177/0962280206071639] [Medline: 17089946]
40. Réseau Sentinelles. URL: https://websenti.u707.jussieu.fr/sentiweb [accessed 2018-06-19] [WebCite Cache ID 70IEHtetc]
41. Zou H, Hastie T. Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society 2005;67:320.
42. Kennard EH. Ridge regression: biaised estimation for nonorthogonal problems. Technometrics 1970;1.
43. Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society 1996;58:267-288.
44. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software 2010;33:1-22.
45. Breiman L. Random Forests. Machine Learning 2001;45:5-32. [doi: 10.1023/A:1010933404324]
46. Liaw A, Wiener M. Classification and Regression by randomForest. R News 2002;2:18-22.
47. Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995;20(3):273-297. [doi: 10.1007/BF00994018]
48. Meyer D, Dimitriadou E, Hornik K. e1071: Misc Functions of the Department of Statistics. Probability Theory Group (Formerly: E1071) https://CRAN.R-project.org/package=e1071 2015.
49. Trapletti A, Hornik K. tseries: Time Series Analysis and Computational Finance. http://CRAN.R-project.org/package=tseries 2015.
50. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology 2010:128-138.
51. Olson D, Heffernan R, Paladini M, Konty K, Weiss D, Mostashari F. Monitoring the Impact of Influenza by Agemergency Department Fever and Respiratory Complaint Surveillance in New York City. PLOS Medicine 2007;4(8).
52. McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, et al. SHRINE: enabling nationally scalable multi-site disease studies. PLoS One 2013;8(3):e55811 [FREE Full text] [doi: 10.1371/journal.pone.0055811] [Medline: 23533569]
53. Bouzillé G, Westerlynck R, Defossez G. Sharing health big data for research - A design by use cases: the INSHARE platform approach. Studies in Health Technology and Informatics 2017.
54. Lu F, Hou S, Baltrusaitis K, Shah M, Leskovec J, Sosic R. Accurate Influenza Monitoring and Forecasting Using Novel Internet Data Streams: A Case Study in the Boston Metropolis. JMIR Public Health Surveillance 2018;4(1).
55. Groupment Interrégional de Recherche Clinique et d'Innovation Grand Ouest. URL: https://www.girci-go.org/ [accessed 2018-06-20] [WebCite Cache ID 70JklABe6]
56. Simonsen L, Gog JR, Olson D, Viboud C. Infectious Disease Surveillance in the Big Data Era: Towards Faster and Locally Relevant Systems. J Infect Dis 2016 Dec 01;214:S380-S385 [FREE Full text] [doi: 10.1093/infdis/jiw376] [Medline: 28830112]
57. Bansal S, Chowell G, Simonsen L, Vespignani A, Viboud C. Big Data for Infectious Disease Surveillance and Modeling. J Infect Dis 2016 Dec 01;214:S375-S379 [FREE Full text] [doi: 10.1093/infdis/jiw400] [Medline: 28830113]
58. Wolpert DH. Stacked generalization. Neural Networks 1992.

Abbreviations

ARIMA: autoregressive integrated moving average
CDW: clinical data warehouse
EHR: electronic health record
HBD: hospital big data
ILI: influenza-like illness
LASSO: least absolute shrinkage and selection operator
MSE: mean squared error
PCC: Pearson correlation coefficient
RF: random forest
SVM: support vector machine
ΔH: epidemic peak
ΔL: prediction lag

Edited by G Eysenbach; submitted 21.06.18; peer-reviewed by B Polepalli Ramesh, F Lu; comments to author 08.08.18; revised version received 10.09.18; accepted 10.09.18; published 17.12.18

Please cite as:
Poirier C, Lavenu A, Bertaud V, Campillo-Gimenez B, Chazard E, Cuggia M, Bouzillé G
Real Time Influenza Monitoring Using Hospital Big Data in Combination with Machine Learning Methods: Comparison Study
JMIR Public Health Surveill 2018;4(4):e11361
URL: http://publichealth.jmir.org/2018/4/e11361/
doi:10.2196/11361
PMID:

3.5 Discussion des principaux résultats

Les sources de données

Au niveau des données, nous avons montré que les données hospitalières, combinées aux données historiques du réseau Sentinelles, permettaient d'obtenir dans la plupart des cas, des estimations plus précises que les données du web, au niveau national et régional. Que ce soit au niveau national ou au niveau régional, le jeu de données qui a permis d'otenir le meilleur PCC et la meilleure EQM est le jeu de données "eHOP restreint" composé des 3 signaux de eHOP les plus corrélés au signal de Sentinelles ainsi que 2 semaines d'historique pour les taux d'incidence. Nous avons également pu voir que la contribution des sources de données externes était plus importante au niveau régional et que l'apport des données hospitalières en comparaison aux données du web était plus significatif.

Comme il a déjà été mentionné précédemment, l'avantage des données hospitalières et des données du web est qu'elles peuvent être obtenues en quasi-temps réel. Cependant, contrairement aux données du web, nous pouvons avoir une certitude plus importante sur le fait que le patient soit atteint de la grippe ou d'un syndrome grippal. Il est également possible d'obtenir des caractéristiques supplémentaires sur les individus comme l'âge ou le sexe et de connaître la sévérité d'une épidémie en prenant en compte le nombre de patients admis en soins intensifs ou décédés suite à la grippe. Grâce aux données hospitalières, il pourrait aussi être possible d'effectuer des estimations pour des maladies non génératrices d'activité sur internet. Enfin, la corrélation spatiale est plus importante avec les données hospitalières qu'avec les données du web. Ce qui pourrait permettre de justifier le fait que nous obtenons de meilleures estimations au niveau régional avec les données hospitalières qu'avec les données du web. Cependant, nous utilisons seulement les données du CHU de Rennes, qui ne sont peut être pas entièrement représentatives de la région Bretagne.

Une des limites principales de notre étude est le fait que nous utilisons seulement les données hospitalières du CHU de Rennes. En effet, nous pouvons penser que les estimations au niveau national pourraient être améliorées si les données provenaient de plusieurs régions ou d'autres hôpitaux. Au CHU de Rennes, l'équipe données massives en santé développe un entrepôt de données de santé avec d'autres CHU du grand Ouest mais également avec les données du Système national d'information interrégimes de l'Assurance Maladie (SNIIRAM). Cet outil pourrait nous permettre d'obtenir plus de données de santé et couvrir un territoire plus large.

Les modèles statistiques

Au niveau des modèles statistiques, nous avons pu montrer que deux modèles avaient des performances comparables, le modèle SVM et le modèle Elastic Net. La principale limite du modèle SVM, est le temps d'optimisation des paramètres du modèle lorsque le nombre de variables explicatives est important. Pour ce modèle, il pourrait donc être judicieux de sélectionner les variables pertinentes au prélable. Pour le modèle Elastic Net, tout comme pour la régression linéaire simple, nous avons effectué des tests sur les résidus et nous les avons modélisé par un processus ARIMA(p,d,q) si nous n'obtenions pas des bruits blancs. Cette étape n'est pas nécessaire avec le modèle SVM.

Pour finir, notre étude réalise une comparaison entre les données hospitalières et les données du web. Cependant, si nous combinions ces 2 sources de données, il est possible que cela nous permettrait d'améliorer nos prévisions. De plus, nous effectuons de la prévision en temps réel mais il pourrait être intéressant de prédire à plus long terme et à une échelle plus fine. C'est pour cette raison que le travail suivant a porté sur la combinaison de sources de données externes afin de prévoir les taux d'incidence grippaux à 2 semaines de toutes les régions de France.

4 Combinaison des sources de données pour la prédiction

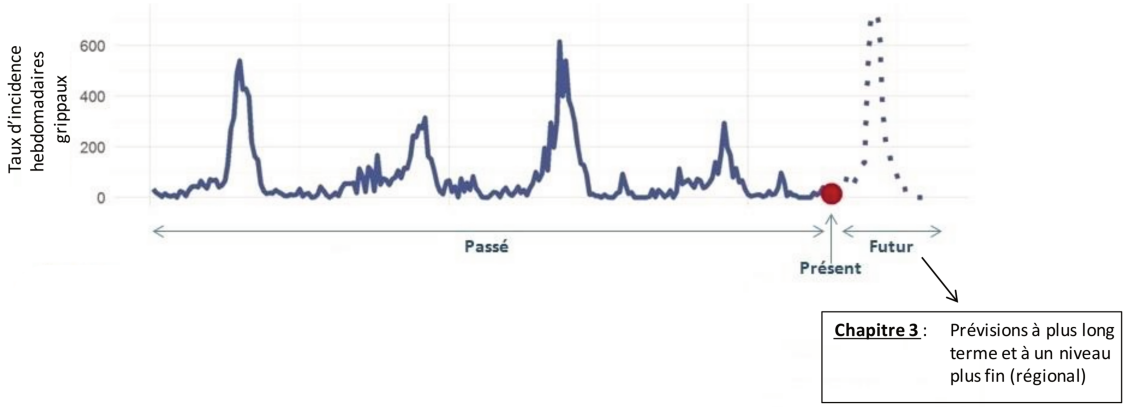


Figure 4.1: Chapitre 3 - Prévisions à plus long terme

4.1 Problématique

L'étude précédente nous a permis de montrer que les données massives hospitalières permettaient d'obtenir dans la plupart des cas, des prévisions en temps réel plus précises que les données du web. De plus, nous avons pu voir que deux modèles avaient des performances comparables, le modèle SVM et le modèle de régression linéaire pénalisé Elastic Net.

Cependant, il pourrait être important de savoir dans quelle mesure les estimations pourraient être améliorées si nous combinions ces 2 sources de données ou les résultats de différents modèles statistiques. De plus, toujours dans une optique d'aide à la décision, il serait intéressant de prédire les épidémies à plus long terme et à une échelle de résolution plus fine. L'équipe de Fred S.Lu et al [36]. a repris le modèle ARGO de Yang et al. [7] présenté dans le chapitre 2 et basé sur un modèle de régression linéaire LASSO réutilisant les données de Google et les données historiques du CDC. Ils ont adapté ce modèle afin d'estimer en temps réel les épidémies de grippe au

niveau de chaque état des États-Unis, en réutilisant les données de Google, les données historiques du CDC ainsi que les données de dossiers patients électroniques provenant d'une compagnie américaine Athenahealth. Dans cette même étude, ces chercheurs ont développé deux autres modèles, le modèle Net et un modèle d'ensemble ARGONet. Le modèle Net est un modèle se basant sur la corrélation des épidémies entre les différents états et le modèle ARGONet et un modèle utilisant les estimations des deux autres approches ARGO et Net.

4.2 Objectif

L'objectif de cette partie est d'étendre aux régions françaises, les approches ARGO, Net et ARGONet développées aux États-Unis. Pour cela, nous souhaitons utiliser les données de Google, les données massives hospitalières provenant de l'entrepôt de données biomédicales eHOP et les données historiques du réseau Sentinelles. Nous souhaitons également étudier l'ajout d'autres sources de données comme les données climatiques et les données de Twitter. Enfin, notre but est d'effectuer des prévisions jusqu'à 2 semaines.

4.3 Considérations méthodologiques

4.3.1 Les sources de données

Les données du réseau Sentinelles

Nous avons recueilli sur le site du réseau Sentinelles les taux d'incidence (pour 100 000 habitants) des syndromes grippaux pour les 12 régions de France métropolitaine (hors Corse). Ces données ont été récupérées en août 2018, pour la période allant du 05 janvier 2004 au 13 mars 2017. Comme pour notre étude précédente, ces données sont utilisées comme signal de référence mais aussi comme variables historiques lors de la construction de nos modèles.

Les données massives hospitalières

Tout comme l'étude précédente, nous avons réitéré les mêmes requêtes sur l'entrepôt de données eHOP afin de récupérer les 34 signaux pour la période allant du 05 janvier 2004 au 13 mars 2017. Cependant, pour chaque région française, nous avons également récupéré les 100 signaux les plus corrélés provenant d'une base de données contenant les séries temporelles construites à partir des données structurées de l'entrepôt eHOP. Au total, pour chaque région, nous avons donc obtenu 134 variables issues de l'entrepôt de données biomédical eHOP dont au minimum 34 variables sont communes à toutes les régions.

Les données de Google

Pour cette étude également, grâce au service Google Correlate, nous avons récupéré la fréquence par semaine des 100 requêtes les plus corrélées à chaque région de France. Ces données ont été récupérées en août 2018, pour la période allant du 05 janvier 2004 au 13 mars 2017.

Les données climatiques

Sur le site Info Climat [37], nous avons pu obtenir les températures et les précipitations pour chaque plus grande ville de chaque région. Nous avons ensuite calculé la moyenne par semaine de ces indicateurs afin d'obtenir des données hebdomadaires. Ces données ont été récupérées en août 2018, pour la période allant du 05 janvier 2004 au 13 mars 2017.

Les données Twitter

Les données Twitter ont été extraites de l'ensemble de données Geotweet de Boston Children's Hospital qui utilise l'API de diffusion de Twitter permettant de récupérer tous les tweets en fonction de coordonnées géographiques. Les tweets pertinents ont été extraits de ce jeu de données en filtrant par date (entre 2013 et 2017), par lieu (la France) et par mots clés en lien avec la grippe ("grippe", "grippé", "syndrome grippal", "fièvre", "toux", "congestion", "malade", "faiblesse", "courbatures", "tamiflu", "la crève").

Pour chaque mot clé, le nombre de tweets par semaine a été calculé afin d'obtenir des incidences hebdomadaires. Ainsi, nous avons obtenu 11 variables provenant de Twitter. Ces données ont été récupérées en décembre 2018, pour la période allant du 05 janvier 2004 au 13 mars 2017.

4.3.2 Les modèles statistiques

Le modèle ARGO

Le modèle ARGO est un modèle de régression linéaire pénalisée utilisant la méthode LASSO présentée précédemment. La formulation de notre modèle est :

— Estimation en temps réel :

$$y_{it} = \sum_{j=1}^{52} \eta_j y_{it-j} + \sum_{k=1}^{100} \alpha_k x_{kit} + \sum_{l=1}^{134} \beta_l z_{lit} + \sum_{p=1}^{11} \gamma_p v_{pit} + \sum_{m=1}^{2} \delta_m w_{mit} + \epsilon_{it}$$

— Prévision à une semaine :

$$y_{it+1} = \sum_{j=1}^{52} \eta_j y_{it-j} + \sum_{k=1}^{100} \alpha_k x_{kit} + \sum_{l=1}^{134} \beta_l z_{lit} + \sum_{p=1}^{11} \gamma_p v_{pit} + \sum_{m=1}^{2} \delta_m w_{mit} + \epsilon_{it+1}$$

— Prévision à deux semaines :

$$y_{it+2} = \sum_{j=1}^{52} \eta_j y_{it-j} + \sum_{k=1}^{100} \alpha_k x_{kit} + \sum_{l=1}^{134} \beta_l z_{lit} + \sum_{p=1}^{11} \gamma_p v_{pit} + \sum_{m=1}^{2} \delta_m w_{mit} + \epsilon_{it+2}$$

où y_{it} correspond au taux d'incidence de la grippe pour la semaine t et la région i, $\sum_{j=1}^{52} \eta_j y_{it-j}$ correspond aux taux d'incidence historiques de la grippe pour la région i, $\sum_{k=1}^{100} \alpha_k x_{kit}$ correspond aux données de Google pour la région i, $\sum_{l=1}^{134} \beta_l z_{lit}$ correspond aux données hospitalières pour la région i, $\sum_{p=1}^{11} \gamma_p v_{pit}$ correspond aux données de Twitter pour la région i, $\sum_{m=1}^{2} \delta_m w_{mit}$ correspond aux données climatiques pour la région i, ϵ_t correspond aux résidus.

Comme pour les résidus d'une régression linéaire, nous avons vérifié la moyenne, la variance et la corrélation pour savoir si nous obtenions des bruits blancs. Si ce n'était pas le cas, les résidus pouvaient être modélisés par un processus ARIMA. Pour tester la stationnarité et la blancheur de nos résidus, nous avons utilisé les tests de Dickey Fuller et Box Pierce disponibles dans les packages tseries et stats du logiciel R. Comme pour la régression pénalisée Elastic Net, afin d'estimer les coefficients du modèle, le critère à minimiser est :

$$\sum_{i=1}^{n}(y_{it} - \sum_{j=1}^{52}\eta_j y_{it-j} - \sum_{k=1}^{100}\alpha_k x_{kit} - \sum_{l=1}^{134}\beta_l z_{lit} - \sum_{p=1}^{11}\gamma_p v_{pit} - \sum_{m=1}^{2}\delta_m w_{mit})^2 + \lambda(\sum_{j=1}^{52}|\eta_j| + \sum_{k=1}^{100}|\alpha_k| + \sum_{l=1}^{134}|\beta_l| + \sum_{p=1}^{11}|\gamma_p| + \sum_{m=1}^{2}|\delta_m|)$$

où λ est l'hyper-paramètre à fixer et α est le paramètre de mélange Elastic Net.L'hyper-paramètre λ a été optimisé grâce à une validation croisée 10 blocs. Nous avons appliqué ces modèles pour chaque région de France grâce au package caret du logiciel R et la méthode glmnet associée [28, 38]. Les modèles ont été recalibrés chaque semaine en incorporant les nouvelles données disponibles, mais en conservant toutes les données précédentes. De cette façon, la taille de notre jeu d'apprentissage augmente chaque semaine. Nous avons effectué nos prévisions entre le mois de janvier 2011 et le mois de mars 2017.

Le modèle Net

Le modèle Net est un modèle LASSO utilisant la relation existante entre les différentes régions. En effet, le calcul du coefficient de corrélation de Pearson entre chaque signal de chaque région, montre qu'il existe une synchronicité entre les différentes zones.

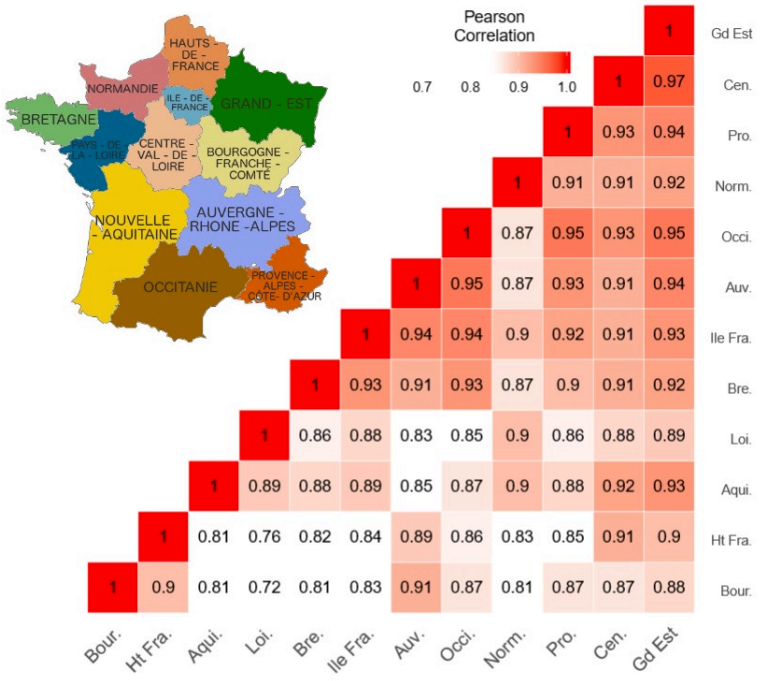


Figure 4.2: Corrélation entre les régions

Afin d'estimer le taux d'incidence d'une région donnée, nous avons utilisé les prévisions en temps réel obtenues par le modèle ARGO pour les autres régions, ainsi que deux semaines d'historique pour toutes les régions y compris celle que l'on cherche à prédire. Notre modèle s'écrit alors :

— Estimation en temps réel :
$$y_{it} = \sum_{l=1}^{2}\sum_{j=1}^{12} \alpha_j y_{jt-l} + \sum_{j=1_{j\neq i}}^{12} \beta_j \hat{y}_{jt} + \epsilon_{it}$$
— Prévision à une semaine :
$$y_{it+1} = \sum_{l=1}^{2}\sum_{j=1}^{12} \alpha_j y_{jt-l} + \sum_{j=1_{j\neq i}}^{12} \beta_j \hat{y}_{jt} + \epsilon_{it+1}$$
— Prévision à deux semaines :
$$y_{it+2} = \sum_{l=1}^{2}\sum_{j=1}^{12} \alpha_j y_{jt-l} + \sum_{j=1_{j\neq i}}^{12} \beta_j \hat{y}_{jt} + \epsilon_{it+2}$$

où y_{it} correspond au taux d'incidence de la grippe pour la semaine t pour la région i, $\sum_{l=1}^{2}\sum_{j=1}^{12} \alpha_j y_{jt-l}$ correspond aux 2 semaines d'historique pour toutes les régions de France, $\sum_{j=1_{j\neq i}}^{12} \beta_j \hat{y}_{jt}$ correspond aux prédictions obtenues par le modèle ARGO pour toutes les régions excepté la région i et enfin ϵ_t correspond aux résidus de notre

modèle.

Nous avons appliqué ces modèles pour chaque région en utilisant un jeu d'apprentissage de 2 ans. Le modèle est recalibré chaque semaine en incorporant les nouvelles informations. Ainsi, nous avons effectué des estimations entre le mois de janvier 2013 et le mois de mars 2017.

Le modèle ARGONet

Le modèle ARGONet est une méthode d'ensemble combinant les prévisions effectuées par le modèle ARGO et le modèle Net. Nous avons testé 3 méthodes :

- Pour une semaine donnée, l'estimation ARGONet est celle du modèle ARGO, si le modèle ARGO donne l'erreur moyenne la plus faible sur les K estimations qui la précèdent. Sinon, l'estimation ARGONet est celle du modèle Net. K peut prendre les valeurs 1, 2, 3 ou 4 semaines.
- Pour une semaine donnée, l'estimation ARGONet est la moyenne des estimations obtenues par le modèle ARGO et le modèle Net.
- Enfin, pour une semaine donnée, l'estimation ARGONet est le résultat d'une régression linéaire entre les estimations des modèles ARGO et Net. Le modèle de régression linéaire est entraîné sur une période de 2 ans.

Le modèle Autorégressif

Pour évaluer l'importance des sources de données externes, nous avons construit un modèle autorégressif d'ordre 52 (AR(52)). Pour cela, nous avons repris le modèle LASSO en utilisant comme seules variables explicatives 52 semaines d'historique des taux d'incidence grippaux. Nos modèles s'écrivent alors :

- Estimation en temps réel :
  $y_{it} = \sum_{j=1}^{52} \alpha_j y_{it-j} + \epsilon_{it}$
- Prévision à une semaine :
  $y_{it+1} = \sum_{j=1}^{52} \alpha_j y_{it-j} + \epsilon_{it+1}$
- Prévision à deux semaines :
  $y_{it+2} = \sum_{j=1}^{52} \alpha_j y_{it-j} + \epsilon_{it+2}$

où y_{it} correspond au taux d'incidence de la grippe pour la semaine t et pour la région i, $\sum_{j=1}^{52} \alpha_j y_{it-j}$ correspond aux 52 semaines précédant la semaine t pour la région i, ϵ_t correspond aux résidus.
Nous avons appliqué ce modèle pour chaque région et nous avons utilisé un jeu d'apprentissage de 6 ans. Ce modèle a également été recalibré chaque semaine.

4.3.3 Évaluation

Notre période de test débute au mois de janvier 2015 et se termine au mois de mars 2017.

Les comparaisons
Comme pour l'étude précédente, afin d'évaluer la performance de nos modèles, nous avons choisi de calculer 2 indicateurs : le coefficient de corrélation de Pearson (PCC) et l'erreur quadratique moyenne (MSE).

Afin d'évaluer l'ajout de sources de données externes, nous avons effectué les comparaisons suivantes :

- MSE et PCC du modèle autorégressif et du modèle ARGO incluant les données historiques plus les 10 variables les plus corrélées provenant des données hospitalières et des données de Google. La contribution individuelle ayant déjà été montrée précédemment.
- MSE et PCC du modèle autorégressif et du modèle ARGO incluant les données historiques plus les données climatiques.
- MSE et PCC du modèle autorégressif et du modèle ARGO incluant les données historiques plus les données de Twitter.

Enfin, nous avons comparé le modèle autorégressif AR(25), le modèle ARGO utilisant toutes les sources de données, le modèle Net et le modèle ARGONet.

4.4 Article

4.4.1 Résumé de l'article

Objectif : Ayant constaté que les données hospitalières permettaient d'obtenir, dans la plupart des cas, des prévisions en temps réel plus précises que les données du web, il convient désormais d'affiner un modèle statistique permettant de prédire à plus long terme les taux d'incidence grippaux et à une échelle plus fine que le niveau national.

Méthodes : Pour cela, nous avons utilisé 3 modèles statistiques, le modèle ARGO, basé sur un modèle de régression LASSO réutilisant les données hospitalières, les données du web et les données climatiques. Le modèle Net, basé également sur un modèle de régression LASSO et réutilisant la synchronicité observée entre les différentes régions. Enfin, une méthode d'ensemble, ARGONet, combinant les prévisions de ces deux modèles. Nous avons comparé ces 3 modèles grâce à deux indicateurs, le PCC et l'EQM.

Résultats : En temps réel, en fonction des régions, le meilleur PCC varie entre 0.852 et 0.971 et l'EQM varie entre 0.309 et 0.057. Pour la prévision à une semaine, le meilleur PCC varie entre 0.768 et 0.958 et l'EQM varie entre 0.460 et 0.083. Pour la prévision à 2 semaines, le meilleur PCC varie entre 0.779 et 0.957 et la meilleure EQM varie entre 0.438 et 0.086. Dans la plupart des cas, ces résultats sont obtenus avec le modèle ARGONet.

Conclusion : Nous avons montré que, pour toutes les régions de France, le modèle ARGONet était le modèle le plus robuste pour prévoir les taux d'incidence grippaux jusqu'à 2 semaines. De plus, nous avons montré que l'utilisation de toutes les sources de données externes permettait d'améliorer les prévisions et plus particulièrement les prévisions à plus long terme. Pour conclure, la méthode d'ensemble ARGONet développée aux États-Unis pourrait être utilisée afin de compléter les méthodes de surveillance traditionnelles.

Influenza forecasting for the French regions by reusing hospital, web and climatic data sources with an ensemble approach ARGONet

Canelle Poirier[1,2,*], Yulin Hswen[3,4], Guillaume Bouzillé[1,2,5], Marc Cuggia[1,2,5], Audrey Lavenu[6,7], John Brownstein[4,8], Thomas Brewer[4], and Mauricio Santillana[8,9,*]

[1]INSERM, U1099, Rennes, F-35000, France;
[2]Université de Rennes 1, LTSI, Rennes, F-35000, France;
[3]Department of Social and Behavioral Sciences, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[4]Informatics Program, Boston Children's Hospital, Boston, MA, USA
[5]CHU Rennes, Centre de Données Cliniques, Rennes, F-35000, France;
[6]Université de Rennes 1, Faculté de médecine, Rennes, F-35043, France;
[7]INSERM CIC 1414, Université de Rennes 1, Rennes, F-35043, France
[8]Department of Pediatrics, Harvard Medical School, Boston, MA, USA;
[9]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA;
[*]**Correspondence**: canelle.poirier@univ-rennes1.fr and msantill@g.harvard.edu

ABSTRACT

Background : Traditional surveillance systems produce estimates of influenza-like illness (ILI) incidence rates, but with 1- to 3-week delay. This time lag is an issue for public health decision making. Today, there is a growing interest in finding ways to avoid this information gap and to predict influenza epidemics in near real time and in longer term.
Objective : Here, we extend methods developed in US to the France. We use french regional historical data, Google data and HBD to predict ILI incidence rates in each french region. We also add climatic data and Twitter data. Estimates are not only made in near real time ("nowcasting") but up to 2 weeks.
Methods: We used 3 different statistical models, ARGO, using all the data sources with a LASSO regression. Net, using the synchronicity observed historically in flu activity between each French regions. ARGONet, an ensemble approach combining estimates from ARGO and Net. We compared these models with the mean squared error (MSE) and Pearson coefficient correlation (PCC).
Results: In real time, depending on the region, the best PCC varies from 0.852 and 0.971 and the best MSE varies from 0.309 to 0.057. For one-week forecast, the best PCC is between 0.768 and 0.958 and the best MSE is between 0.460 and 0.083. For two-week forecast, the best PCC varies from 0.779 to 0.957 and the best MSE varies from 0.438 to 0.086. In most cases, the best results are obtained with ARGONet model.
Conclusions: We found that the ARGONet model is the most robust model to estimate ILI activity up to two-weeks at the french regional level. Moreover, we show that external data sources allow to improve estimates and more particularly for longer-term forecasts. To conclude, the ensemble approach ARGONet developed in United States could be used to complete traditional influenza surveillance method in France.

Introduction

Influenza is a major public health problem with up to 5 million severe cases and 500,000 deaths per year worldwide[1–3]. In France, the epidemic of 2018-2019 causes 9500 deaths. Moreover, during epidemic peaks, there is a large increase of visits to general practitioners and to emergency departments which causes healthcare system disruption. To reduce its impact and to organize adapted sanitary response, it is essential to monitor flu outbreaks and Influenza-Like-Illness (ILI; any acute respiratory infection with fever 38 C, cough and onset within the last 10 days).

In France, the principal monitoring tool is the Sentinelles network[4,5]. Each week, 1314 general practitioners and 116 liberal paediatricians report all ILI cases seen during consultation. An estimation of the ILI incidence rate is then computed, at the national and regional scale, by taking into account the number of sentinel physicians and the medical density of the area of interest. Nevertheless, estimates of ILI incidence rates are produced with a one to three-week delay due to the time needed for data processing and aggregation. This time lag is an issue for public health decision-making.[2,6]

It is therefore necessary to find a way to avoid this delay and to predict outbreaks in order to reduce the impact of the flu. Numerous studies combining traditional statistical methods, like temporal series or compartmental methods, and data sources, like meteorological or demographic data[7]. But in recent years, in the Big Data context, other methods have been developed to try to take advantage from other data sources.

With over 3.2 billion web users, data flows from the internet are huge and of all types. Some studies used data from Google[2,3,8–10], Twitter[11–13] or Wikipedia[14–17]. One of the first and most famous studies on the use of internet data for detecting influenza epidemics is Google Flu Trends (GFT)[18]. This web service, created in 2009 and operated by Google, used the volume of selected Google search terms to estimate ILI activity in real time. But after a large overestimation in 2012-2013, the algorithm is forced to be stopped[19]. Some authors updated the Google Flu Trends algorithm. This is the case of Shihao Yang et al[2]. who proposed an approach called ARGO using ILI activity data (from Centers for Diseases Control) and Google data associated to a dynamical least absolute shrinkage and selection operator (LASSO) regression.

Another data source, less used, is hospital big data (HBD). With the adoption of electronic health records (EHRs), hospitals also produce a huge amount of data that are collected during hospitalization[20]. In United States, Mauricio Santillana et al.[6] proposed a model using HBD and a machine learning algorithm. This study showed good performances at the regional level. In France, Canelle Poirier et al.[21] proposed a study comparing HBD and web data and different statistical models to estimate ILI incidence rates in real time at the national level.

Finally, Fred S. Lu et al.[22] extended the ARGO approach developed by Shihao Yang et al. to each states of the United States. They included Google data, EHRs and historical flu trends. They developed also a spatial network approach, called Net, using the synchronicity observed historically in flu activity between each states. At last, they developed ARGONet, an ensemble approach combining estimates from ARGO and Net.

Our contribution. In this study, we extend ARGO, Net and ARGONet approaches to the France. We use french regional historical data, Google data and HBD to predict ILI incidence rates in each french region. We also add climatic data and Twitter data. Estimates are not only made in near real time ("nowcasting") but up to 2 weeks.

Methods

Data sources

Sentinelles network data

We obtained weekly ILI incidence rates (per 100000 inhabitants) for the French regions (12) from the French Sentinelles network (websenti.u707.jussieu.fr/sentiweb). We retrieved these data in August 2018 from 05 January 2004 to 13 March 2017. We considered these data as the gold standard and used them also as independent historical variables in our models.

Hospital Data

We retrieved hospital data from the clinical data warehouse (CDW) of Rennes University Hospital (France),This CDW, called eHOP, integrates structured (laboratory test results, prescriptions, ICD-10 diagnoses) and unstructured (discharge letter, pathology reports, operative reports) patients' data. It includes data from 1.2 million inpatients and outpatients and 45 million documents that correspond to 510 million structured elements. eHOP consists of a

powerful search engine system that can identify patients with specific criteria by querying unstructured data with keywords, or structured data with querying codes based on terminologies.
The first approach to obtain eHOP data connected with ILI was to perform different manual queries to retrieve patients who had at least one document in their EHR that matched the following search criteria:

- Queries directly connected with flu or ILI with the keywords "flu" or "ILI".
- Queries connected with flu symptoms with the keywords ”fever”, ”pyrexia”, ”body aches” or ”muscular pain”.
- Queries connected with flu drugs with the keyword ”Tamiflu”.
- Queries with the ICD-10 terminology.
- Queries connected with flu tests, positive or negative results.

In total, we did 34 manual queries. For each query, the eHOP search engine returned all documents containing the chosen keywords (often, several documents for one patient and one stay). For query aggregation, we kept the oldest document for one patient and one stay and then calculated, for each week, the number of stays with at least one document mentioning the keyword contained in the query.
The second approach was to retrieve the 100 most correlated signals to ILI signal for each region, according to pearson distance, from a database containing the time series constructed from the structured data. Because our test period is from 05 January 2015 to 20 February 2017, we calculated the correlation between January 2004 and December 2014.
In this way, for each region, we obtained 134 variables from the CDW eHOP. We retrieved retrospective data in August 2018 for the period going from 03 January 2005 to 13 March 2017.

Google Data

We obtained the frequency per week of the 100 most correlated internet queries by French users from Google Correlate (https://www.google.com/trends/correlate). Because our predict period is from 05 January 2015 to 20 February 2017, we inserted ILI signal for each French region, from January 2004 to December 2014 into Google Correlate. We retrieved Google Correlate data in August 2018 for the period going from 05 January 2004 to 13 March 2017.

Climatic Data

We obtained temperatures and precipitations per day for the largest city of each region from the French climatological website Info Climat (https://www.infoclimat.fr). Then, we calculated the mean per week of temperatures and precipitations to have weekly data. We retrieved climatic data in August 2018 for the period going from 07 January 2008 to 13 March 2017.

Twitter Data

Twitter data is extracted from Boston Children’s Hospital Geotweet dataset, which uses Twitter’s streaming API to gather any tweet attached to geographical coordinates. Coordinates are either provided directly by user, or by a Twitter side classification model which estimates location.
Relevant tweets are extracted from the dataset by filtering by year (2013-2017), by location (France), and by keywords pertaining to influenza (“grippe”, “grippé”, “syndrome grippal”, “fièvre”, “toux”, “congestion”, “malade”, “faiblesse”, “courbatures”, “tamiflu”, “la crève”). From there we can aggregate tweets to get weekly counts. In this way, we obtained 11 variables from Twitter. We retrieved Twitter data in December 2018 for the period going from 30 December 2013 to 13 March 2017.

Statistical models

The ARGO model

The ARGO model is a regularized regression using the LASSO method[23]. We performed the LASSO regression with the R package caret and the associated function fit with the method glmnet[24,25]. We optimized the shrinkage

parameter lambda via a 10-fold cross-validation. As for residuals of a linear regression, we checked means, variances and correlation to know if we obtained white noises. To test the stationarity and whiteness, we used Dickey Fuller's and Box-Pierce's tests available from the R packages tseries and stats[26]. The formulation of our model is :

- Real time estimates:
$$y_{it} = \sum_{j=1}^{52} \eta_j y_{it-j} + \sum_{k=1}^{100} \alpha_k x_{kit} + \sum_{l=1}^{134} \beta_l z_{lit} + \sum_{p=1}^{11} \gamma_p v_{pit} + \sum_{m=1}^{2} \delta_m w_{mit} + \varepsilon_{it}$$

- One-week ahead forecast:
$$y_{it+1} = \sum_{j=1}^{52} \eta_j y_{it-j} + \sum_{k=1}^{100} \alpha_k x_{kit} + \sum_{l=1}^{134} \beta_l z_{lit} + \sum_{p=1}^{11} \gamma_p v_{pit} + \sum_{m=1}^{2} \delta_m w_{mit} + \varepsilon_{it+1}$$

- Two-week ahead forecast:
$$y_{it+2} = \sum_{j=1}^{52} \eta_j y_{it-j} + \sum_{k=1}^{100} \alpha_k x_{kit} + \sum_{l=1}^{134} \beta_l z_{lit} + \sum_{p=1}^{11} \gamma_p v_{pit} + \sum_{m=1}^{2} \delta_m w_{mit} + \varepsilon_{it+2}$$

where y_{it} corresponding to the flu incidence rate at time t for the region i, $\sum_{j=1}^{52} \eta_j y_{it-j}$ corresponding to the historical flu incidence rates for the region i, $\sum_{k=1}^{100} \alpha_k x_{kit}$ corresponding to Google data for the region i, $\sum_{l=1}^{134} \beta_l z_{lit}$ corresponding to hospital data for the region i, $\sum_{p=1}^{11} \gamma_p v_{pit}$ corresponding to Twitter data, $\sum_{m=1}^{2} \delta_m w_{mit}$ corresponding to climatic data for the region i, ε_t corresponding to residuals.
We applied this model for each region. The model was dynamically recalibrated every week by incorporating all the previous data available, depending on the data sources. In this way, the size of our training dataset increases every week. We obtained estimates from January 2011 to March 2017.

The Net model

The Net model is a LASSO model using the relationship between the regions. Indeed, Figure S1 (Heatmap of pairwise correlations between all regions) shows synchronicity between the different areas. For each region, we used historical data of all regions and estimates obtained with ARGO model for all regions expected the region to be predicted.
The formulation of our model is :

- Real time estimates:
$$y_{it} = \sum_{l=1}^{2} \sum_{j=1}^{12} \alpha_j y_{jt-l} + \sum_{j=1_{j\neq i}}^{12} \beta_j \hat{y}_{jt} + \varepsilon_{it}$$

- One-week ahead forecast:
$$y_{it+1} = \sum_{l=1}^{2} \sum_{j=1}^{12} \alpha_j y_{jt-l} + \sum_{j=1_{j\neq i}}^{12} \beta_j \hat{y}_{jt} + \varepsilon_{it+1}$$

- Two-week ahead forecast:
$$y_{it+2} = \sum_{l=1}^{2} \sum_{j=1}^{12} \alpha_j y_{jt-l} + \sum_{j=1_{j\neq i}}^{12} \beta_j \hat{y}_{jt} + \varepsilon_{it+2}$$

where y_{it} corresponding to the flu incidence rate at time t for the region i, $\sum_{l=1}^{2} \sum_{j=1}^{12} \alpha_j y_{jt-l}$ corresponding to two weeks of historical flu incidence rates for all regions, $\sum_{j=1_{j\neq i}}^{12} \beta_j \hat{y}_{jt}$ corresponding to ARGO predictions for all regions excepted the region i to be predicted and ε_t corresponding to residuals.
We applied this model for each region. We used a two years' training dataset. The model was dynamically recalibrated every week to incorporate new information. We obtained estimates from January 2013 to March 2017.

The ARGONet model

The ARGONet model is an ensemble approach combining the predictive power of ARGO and Net models. For this model we tested three methods :

- The first is, ARGONet's estimate is the ARGO estimate if ARGO model gives the lowest mean error in the previous K estimates compared to Net model. Otherwise, ARGONet's estimate is the Net estimate. The value of K can be 1, 2, 3 or 4.
- The second is, ARGONet's estimate is the mean between ARGO's estimate and Net's estimate.
- The third is, ARGONet's estimate is the result of a linear regression between ARGO's estimate and Net's estimate. We trained the linear regression model on a period of two years.

The Autoregressive model

To assess the importance of external data sources, we built an autoregressive model of order 52 (AR(52)). We used the LASSO regression with the previous 52 weeks of ILI incidence rates to predict the current week and the two weeks after.

- Real time estimates:
  $y_{it} = \sum_{j=1}^{52} \alpha_j y_{it-j} + \varepsilon_{it}$
- One-week ahead forecast:
  $y_{it+1} = \sum_{j=1}^{52} \alpha_j y_{it-j} + \varepsilon_{it+1}$
- Two-week ahead forecast:
  $y_{it+2} = \sum_{j=1}^{52} \alpha_j y_{it-j} + \varepsilon_{it+2}$

where y_{it} corresponding to the flu incidence rate at time t for the region i, $\sum_{j=1}^{52} \alpha_j y_{it-j}$ corresponding to the previous 52 weeks, ε_t corresponding to residuals.
We applied this model for each region. We used a six years' training dataset. The model was dynamically recalibrated every week.

Evaluation

Our test period is from January 2015 to March 2017.

Metrics

To assess the performance of the models, we compared estimates to the official incidence rates from the Sentinelles network by calculating two metrics : the mean squared error (MSE) and the Pearson correlation coefficient (PCC).

- $MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$
- $PCC = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2}}$

where $\hat{y}_i$ is the predicted value for the week i, $\bar{\hat{y}}_i$ is the mean of predicted values, y_i the real value for the week i, $\bar{y}_i$ is the mean of real values.

Comparisons

First, we assessed the importance of adding external data sources by comparing :

- MSE and PCC of the autoregressive model and the ARGO model including historical data plus the 10 most correlated variables from hospital data and Google data. The individual contribution of hospital data and Google data has already been shown in a previous study.
- MSE and PCC of the autoregressive model and the ARGO model including historical data plus climatic data.
- MSE and PCC of the autoregressive model and the ARGO model including historical data plus Twitter data.

Second, we compared the autoregressive model, ARGO model (including all the data sources), Net model and ARGONet model.

Results

Data sources

Here, we show the results obtained for the comparisons between historical data and the other data sources. We can see that all external data sources improve estimates and this improvement is more important to forecast in longer-term.

Hospital Data and Google Data

In this subsection, we show the difference obtained between historical data and Google data in combination with hospital data. The contribution of Google data and hospital data used individually is shown in previous studies. In most cases, hospital data and Google data improve results.

In real time (Table 1), in term of PCC and MSE, estimates obtained with hospital data and Google data allow to improve results for 9 of the 12 regions. The region Pays de la Loire has the best improvement in term of PCC (+0.027) and MSE (-0.053).

	Auv.	Bour.	Bre.	Cen.	Gd Est	Ht Fra.	Ile Fra.	Norm.	Aqui.	Occi.	Loi.	Pro.
MSE												
AR(52)	**0.083**	0.203	0.167	0.161	0.096	**0.169**	**0.129**	0.241	0.161	0.108	0.368	0.141
ARGO	0.100	**0.168**	**0.134**	**0.110**	**0.087**	**0**.242	0.133	**0.213**	**0.141**	**0.097**	**0.315**	**0.117**
PCC												
AR(52)	**0.958**	0.898	0.916	0.919	0.952	**0.915**	**0.935**	0.879	0.919	0.946	0.815	0.929
ARGO	0.950	**0.916**	**0.933**	**0.945**	**0.957**	0.878	0.933	**0.893**	**0.929**	**0.951**	**0.842**	**0.941**

Table 1. Real time estimate: MSE and PCC for ARGO models including only historical data (AR(52)) and only hospital and Google data, for the period starting from January 2015 to March 2017

For One-week estimate (Table 2), in term of PCC and MSE, results are better for 8 of the 12 regions. The region Nouvelle-Aquitaine has the best improvement in term of PCC (+0.102) and MSE (-0.200).

	Auv.	Bour.	Bre.	Cen.	Gd Est	Ht Fra.	Ile Fra.	Norm.	Aqui.	Occi.	Loi.	Pro.
MSE												
AR(52)	**0.262**	**0.611**	0.433	0.438	0.276	**0.433**	**0.372**	0.580	0.480	**0.278**	0.704	0.389
ARGO	0.275	0.691	**0.372**	**0.288**	**0.206**	0.489	0.373	**0.500**	**0.280**	0.350	**0.523**	**0.304**
PCC												
AR(52)	**0.868**	**0.692**	0.782	0.779	0.861	**0.782**	0.813	0.707	0.758	**0.860**	0.645	0.804
ARGO	0.862	0.653	**0.814**	**0.856**	**0.897**	0.755	0.813	**0.749**	**0.860**	0.825	**0.738**	**0.848**

Table 2. One-week ahead forecast : MSE and PCC for ARGO models including only historical data (AR(52)) and only hospital and Google data, for the period starting from January 2015 to March 2017

For Two-week estimate (Table 3), in term of PCC and MSE, results are better for all the regions. The region Nouvelle-Aquitaine has the best improvement in term of PCC (+0.162) and MSE (-0.321).

	Auv.	Bour.	Bre.	Cen.	Gd Est	Ht Fra.	Ile Fra.	Norm.	Aqui.	Occi.	Loi.	Pro.
MSE												
AR(52)	0.534	0.947	0.665	0.776	0.530	0.679	0.618	0.914	0.775	0.533	0.935	0.596
ARGO	**0.516**	**0.724**	**0.408**	**0.539**	**0.361**	**0.657**	**0.493**	**0.632**	**0.454**	**0.286**	**0.643**	**0.269**
PCC												
AR(52)	0.731	0.522	0.665	0.609	0.733	0.658	0.688	0.539	0.610	0.731	0.528	0.699
ARGO	**0.741**	**0.637**	**0.795**	**0.730**	**0.819**	**0.670**	**0.753**	**0.683**	**0.772**	**0.857**	**0.678**	**0.865**

Table 3. Two-week ahead forecast : MSE and PCC for ARGO models including only historical data (AR(52)) and only hospital and Google data, for the period starting from January 2015 to March 2017

Climatic Data

In this subsection, we show the difference obtained between historical data and climatic data. In most cases, climatic data improve results.

In real time (Table 4), in term of PCC and MSE, estimates obtained with climatic data allow to improve results for all the regions. However, unlike for hospital and Google data, the best improvement in term of PCC is +0.013 and is -0.026 for MSE.

	Auv.	Bour.	Bre.	Cen.	Gd Est	Ht Fra.	Ile Fra.	Norm.	Aqui.	Occi.	Loi.	Pro.
MSE												
AR(52)	0.083	0.203	0.167	0.161	0.096	0.169	0.129	0.241	0.161	0.108	0.368	0.141
ARGO	**0.076**	**0.196**	**0.154**	**0.160**	**0.091**	**0.156**	**0.122**	**0.230**	**0.149**	**0.096**	**0.342**	**0.131**
PCC												
AR(52)	0.958	0.898	0.916	0.919	0.952	0.915	0.935	0.879	0.919	0.946	0.815	0.929
ARGO	**0.962**	**0.901**	**0.922**	**0.919**	**0.954**	**0.921**	**0.939**	**0.884**	**0.925**	**0.951**	**0.828**	**0.934**

Table 4. Real time estimate: MSE and PCC for ARGO models including only historical data (AR(52)) and only climatic data, for the period starting from January 2015 to March 2017

For One-week estimate (Table 5), in term of PCC and MSE, results are better for 10 of the 12 regions. The region Bourgogne-Franche-Comté has the best improvement in term of PCC (+0.034) and MSE (-0.200).

	Auv.	Bour.	Bre.	Cen.	Gd Est	Ht Fra.	Ile Fra.	Norm.	Aqui.	Occi.	Loi.	Pro.
MSE												
AR(52)	**0.262**	0.611	0.433	0.438	0.276	0.433	0.372	0.580	0.480	**0.278**	0.704	0.389
ARGO	0.264	**0.542**	**0.379**	**0.421**	**0.264**	**0.381**	**0.346**	**0.516**	**0.422**	0.294	**0.657**	**0.357**
PCC												
AR(52)	**0.868**	0.692	0.782	0.779	0.861	0.782	0.813	0.707	0.758	**0.860**	0.645	0.804
ARGO	0.867	**0.726**	**0.809**	**0.788**	**0.867**	**0.808**	**0.825**	**0.740**	**0.787**	0.852	**0.669**	**0.820**

Table 5. One-week ahead forecast : MSE and PCC for ARGO models including only historical data (AR(52)) and only climatic data, for the period starting from January 2015 to March 2017

For Two-week estimate (Table 6), results are better for all the regions. The region Bourgogne-Franche-Comté has the best improvement in term of PCC (+0.129) and MSE (-0.256).

	Auv.	Bour.	Bre.	Cen.	Gd Est	Ht Fra.	Ile Fra.	Norm.	Aqui.	Occi.	Loi.	Pro.
MSE												
AR(52)	0.534	0.947	0.665	0.776	0.530	0.679	0.618	0.914	0.775	0.533	0.935	0.596
ARGO	**0.510**	**0.691**	**0.577**	**0.700**	**0.477**	**0.557**	**0.593**	**0.773**	**0.685**	**0.481**	**0.913**	**0.543**
PCC												
AR(52)	0.731	0.522	0.665	0.609	0.733	0.658	0.688	0.539	0.610	0.731	0.528	0.699
ARGO	**0.743**	**0.651**	**0.709**	**0.647**	**0.759**	**0.719**	**0.701**	**0.610**	**0.655**	**0.758**	**0.590**	**0.726**

Table 6. Two-week ahead forecast : MSE and PCC for ARGO models including only historical data (AR(52)) and only climatic data, for the period starting from January 2015 to March 2017

Twitter Data

In this subsection, we show the difference obtained between historical data and Twitter data. In most cases, Twitter data improve results.

In real time (Table 7), in term of PCC and MSE, estimates obtained with Twitter data allow to improve results for 7 of the 12 regions. The regions Occitanie and Pays de la Loire have the best improvements in term of PCC (+0.015) and MSE (-0.030).

	Auv.	Bour.	Bre.	Cen.	Gd Est	Ht Fra.	Ile Fra.	Norm.	Aqui.	Occi.	Loi.	Pro.
MSE												
AR(52)	0.083	**0.203**	**0.167**	0.161	0.096	**0.169**	0.129	**0.241**	0.161	0.108	0.368	0.141
ARGO	**0.078**	0.212	0.170	**0.137**	**0.091**	0.170	**0.109**	0.257	**0.160**	**0.078**	**0.337**	**0.133**
PCC												
AR(52)	0.958	**0.898**	**0.916**	0.919	0.952	**0.915**	0.935	**0.879**	0.919	0.946	0.815	0.929
ARGO	**0.960**	0.893	0.914	**0.931**	**0.954**	0.914	**0.945**	0.871	**0.919**	**0.961**	**0.830**	**0.933**

Table 7. Real time estimate: MSE and PCC for ARGO models including only historical data (AR(52)) and only Twitter data, for the period starting from January 2015 to March 2017

For One-week estimate (Table 8), results are better for all the regions. The region Pays de la Loire has the best improvement in term of PCC (+0.058) and MSE (-0.115).

	Auv.	Bour.	Bre.	Cen.	Gd Est	Ht Fra.	Ile Fra.	Norm.	Aqui.	Occi.	Loi.	Pro.
MSE												
AR(52)	0.262	0.611	0.433	0.438	0.276	0.433	0.372	0.580	0.480	0.278	0.704	0.389
ARGO	**0.236**	**0.570**	**0.355**	**0.331**	**0.251**	**0.422**	**0.305**	**0.548**	**0.376**	**0.232**	**0.589**	**0.330**
PCC												
AR(52)	0.868	0.692	0.782	0.779	0.861	0.782	0.813	0.707	0.758	0.860	0.645	0.804
ARGO	**0.881**	**0.712**	**0.821**	**0.833**	**0.873**	**0.787**	**0.846**	**0.723**	**0.811**	**0.883**	**0.703**	**0.834**

Table 8. One-week ahead forecast : MSE and PCC for ARGO models including only historical data (AR(52)) and only Twitter data, for the period starting from January 2015 to March 2017

For Two-week estimate (Table 9), results are better for all the regions. The region Bourgogne-Franche-Comté has the best improvement in term of PCC (+0.093) and MSE (-0.183).

	Auv.	Bour.	Bre.	Cen.	Gd Est	Ht Fra.	Ile Fra.	Norm.	Aqui.	Occi.	Loi.	Pro.
MSE												
AR(52)	0.534	0.947	0.665	0.776	0.530	0.679	0.618	0.914	0.775	0.533	0.935	0.596
ARGO	**0.503**	**0.764**	**0.490**	**0.603**	**0.466**	**0.624**	**0.571**	**0.874**	**0.624**	**0.417**	**0.866**	**0.507**
PCC												
AR(52)	0.731	0.522	0.665	0.609	0.733	0.658	0.688	0.539	0.610	0.731	0.528	0.699
ARGO	**0.746**	**0.615**	**0.753**	**0.696**	**0.765**	**0.685**	**0.712**	**0.559**	**0.685**	**0.790**	**0.563**	**0.744**

Table 9. Two-week ahead forecast : MSE and PCC for ARGO models including only historical data (AR(52)) and only Twitter data, for the period starting from January 2015 to March 2017

Statistical models

Here, we show the results obtained with AR(52), ARGO, Net and ARGONet models for real time, one-week and two-week forecasts. ARGO model includes all the external data sources : Hospital, Google, Twitter and Climatic data. As shown above, we can see that external data sources improve estimates. Indeed, Figure 1 shows that in term of PCC and MSE, AR(52) is still on 3rd or 4th rank, for each forecast level and each region.
Moreover, even if the accuracy decreases over time (Figure S2 to Figure S13), ARGONet is the most robust model. In real-time, in term of PCC and MSE, ARGONet is on 1st rank for 7 of 12 regions and on 2nd rank for 5 of 12 regions. For one-week and two-week estimates, is still on 1st rank (Figure 1).

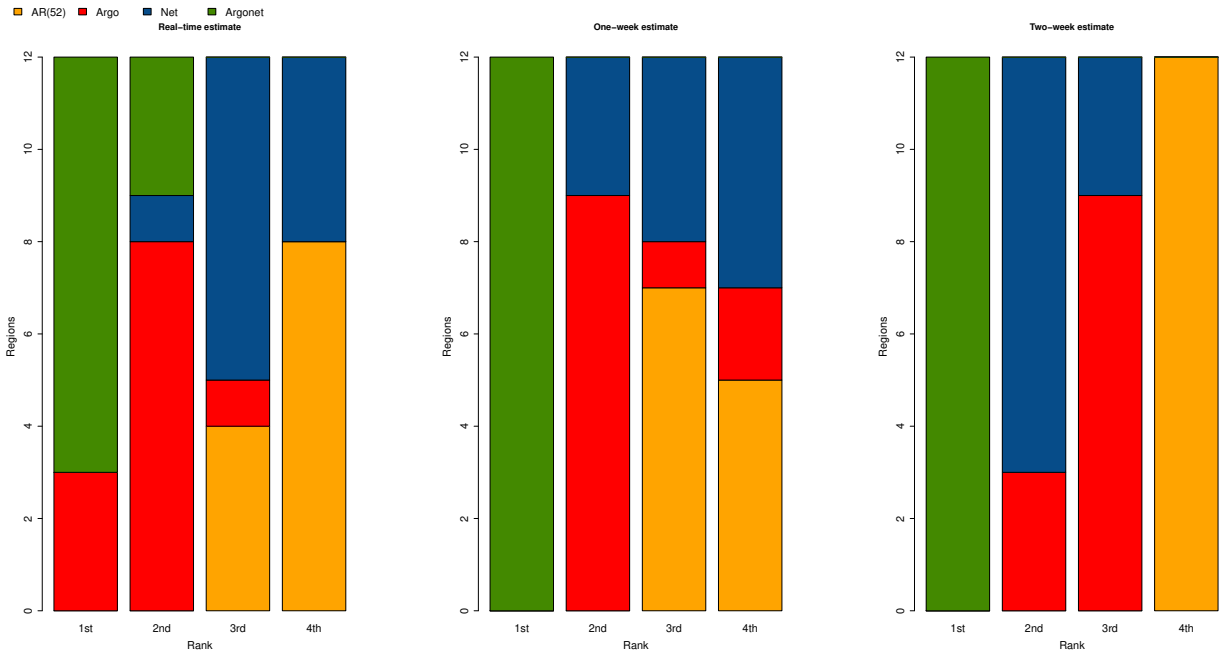


Figure 1. Ranks obtained by each model over the 12 French regions for PCC and MSE

Real time

Table 1 shows results obtained with AR(52), ARGO, Net and ARGONet models for the period starting from January 2015 to March 2017. Over this period, depending on the region, the best PCC varies from 0.852 and 0.971 and the best MSE varies from 0.309 to 0.057. As with Figure 1, barplots in Figure 3 show that the best PCC and MSE are mostly obtained with ARGONet. But ARGO has also good results.

For 3 regions the best model is ARGO with the highest PCC and the lowest MSE. For the other 9 regions, the best model is ARGONet. In the same way, if we compare the distribution of PCC and MSE (Figures 14 and 15), ARGO and ARGONet models are the two best models.

Figure 4 and Figure 5 show a typical example of plot and heatmap obtained for estimates in real time. The heatmap allows to visualize coefficients used for ARGO model. On these plots, we can see that the different models have estimates close to the gold standard. However, for the autoregressive model, there is a time lag more important. On the heatmap, we can see that ARGO model uses mostly 17 variables including 6 variables from Google Data, 9 variables from Hospital Data and 1 variable from Historical Data. One variable from Twitter is also used but only from 2013, the first year we collect twitter data.

	Auv.	Bour.	Bre.	Cen.	Gd Est	Ht Fra.	Ile Fra.	Norm.	Aqui.	Occi.	Loi.	Pro.
MSE												
AR(52)	0.083	0.203	0.167	0.161	0.096	0.169	0.129	0.241	0.161	0.108	0.368	0.141
Argo	0.069	0.129	**0.090**	**0.088**	**0.065**	0.152	0.122	0.219	0.125	**0.060**	**0.309**	0.105
Net	0.071	0.181	0.156	0.104	0.098	0.127	0.141	0.265	0.139	0.105	0.375	0.115
K=1	0.075	**0.120**	0.102	0.098	0.072	0.127	0.116	0.247	**0.118**	0.068	0.357	0.133
K=2	0.066	**0.124**	**0.099**	0.094	0.081	0.126	**0.115**	**0.210**	0.119	0.079	0.311	0.123
K=3	0.081	0.137	0.104	0.093	0.073	0.119	0.117	**0.208**	0.123	0.088	0.314	0.131
K=4	0.068	0.128	0.106	0.101	0.072	0.129	0.116	0.242	0.124	0.069	**0.292**	0.136
Mean	**0.057**	0.141	0.109	**0.090**	0.072	**0.115**	**0.116**	0.224	**0.118**	0.070	0.311	**0.090**
Lm	**0.059**	0.139	0.105	0.097	**0.063**	**0.118**	0.131	0.264	0.129	**0.066**	0.336	**0.100**
PCC												
AR(52)	0.958	0.898	0.916	0.919	0.952	0.915	0.935	0.879	0.919	0.946	0.815	0.929
Argo	0.965	0.935	**0.954**	**0.955**	**0.967**	0.923	0.938	0.890	0.937	**0.970**	**0.844**	0.947
Net	0.964	0.909	0.921	0.948	0.951	0.936	0.929	0.866	0.930	0.947	0.811	0.942
K=1	0.962	**0.956**	0.948	0.951	0.963	0.936	0.941	0.875	**0.941**	0.966	0.820	0.933
K=2	0.967	**0.939**	**0.950**	0.952	0.959	0.936	**0.942**	**0.894**	0.940	0.960	0.843	0.938
K=3	0.959	0.937	0.948	0.953	0.963	0.940	0.942	**0.895**	0.938	0.956	0.842	0.934
K=4	0.966	0.931	0.947	0.949	0.964	0.935	0.941	0.877	0.938	0.965	**0.852**	0.931
Mean	**0.971**	0.929	0.945	**0.955**	0.964	**0.942**	**0.942**	0.887	**0.941**	0.965	0.843	**0.954**
Lm	**0.970**	0.930	0.947	0.951	**0.968**	**0.940**	0.942	0.867	0.935	**0.967**	0.830	**0.949**

Table 10. PCC and MSE for real time estimate for all french regions for the period starting from January 2015 to March 2017

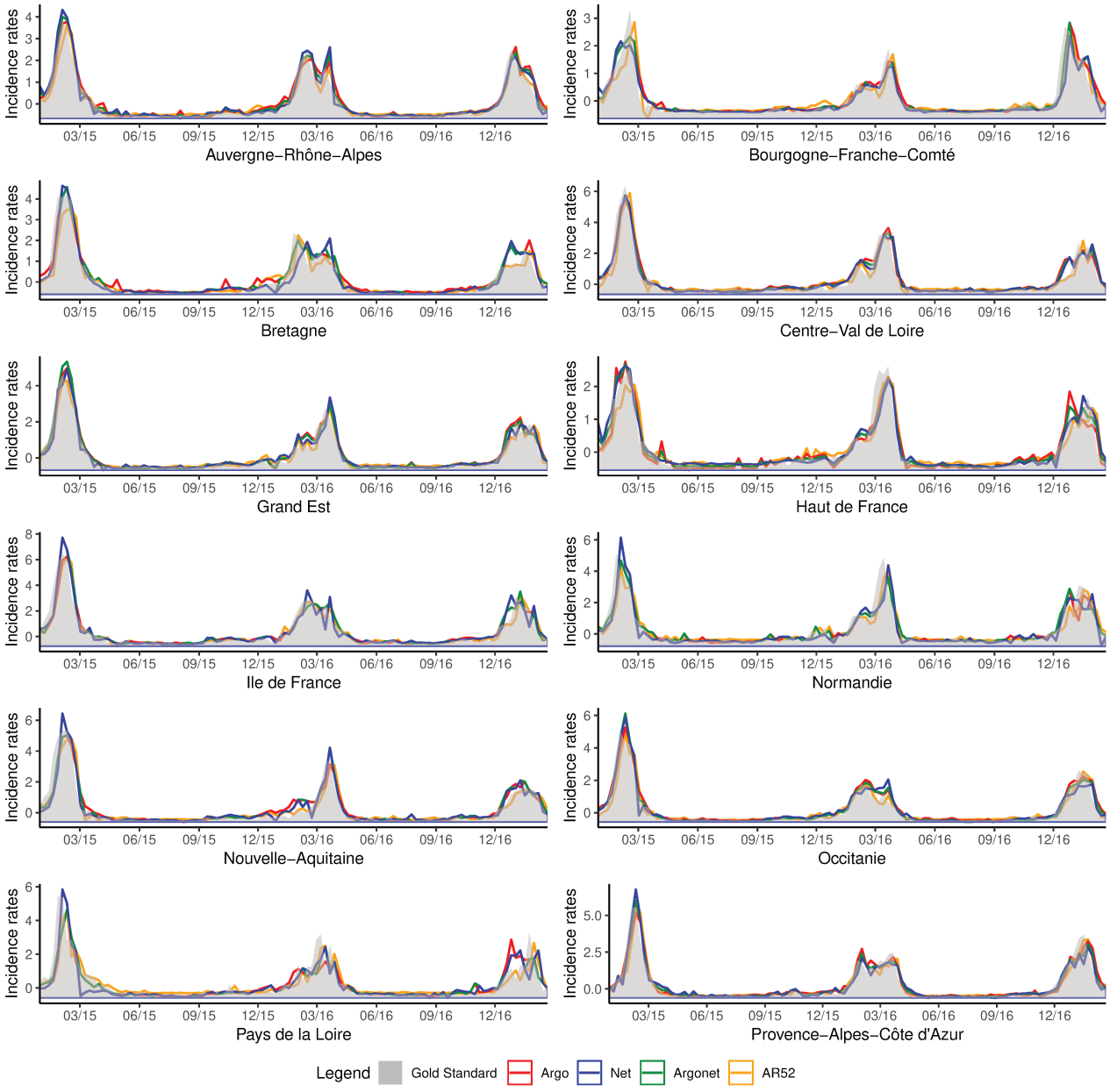


Figure 2. Flu outbreaks from January 2015 to March 2017 and real-time estimates obtained with ARGO, Net and ARGONet models

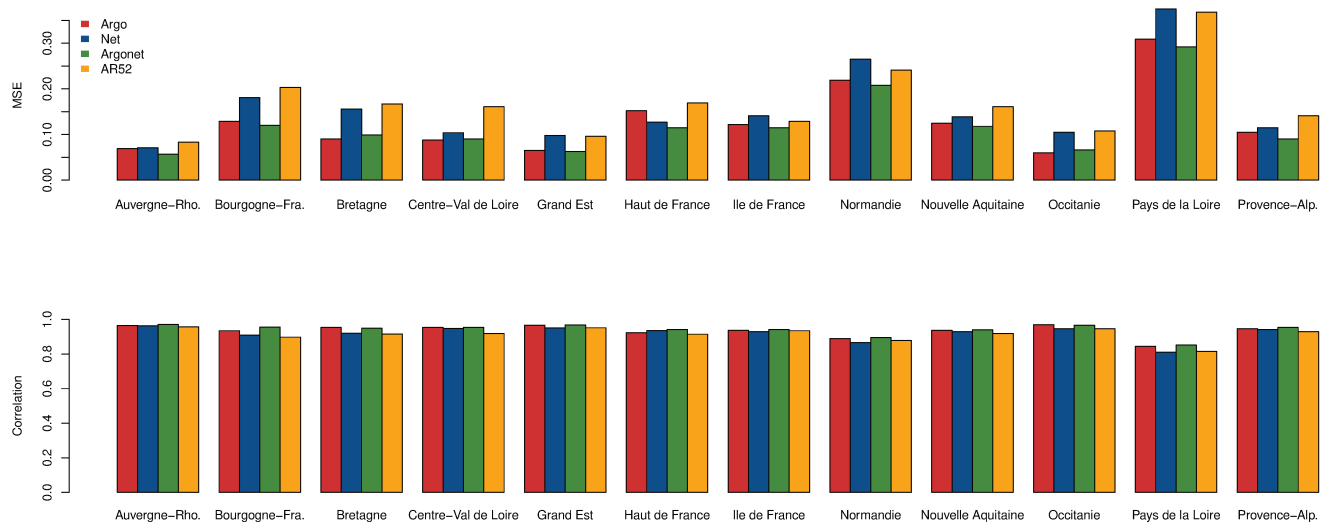


Figure 3. PCC and MSE obtained for predictions in real time with ARGO, Net and ARGONet models

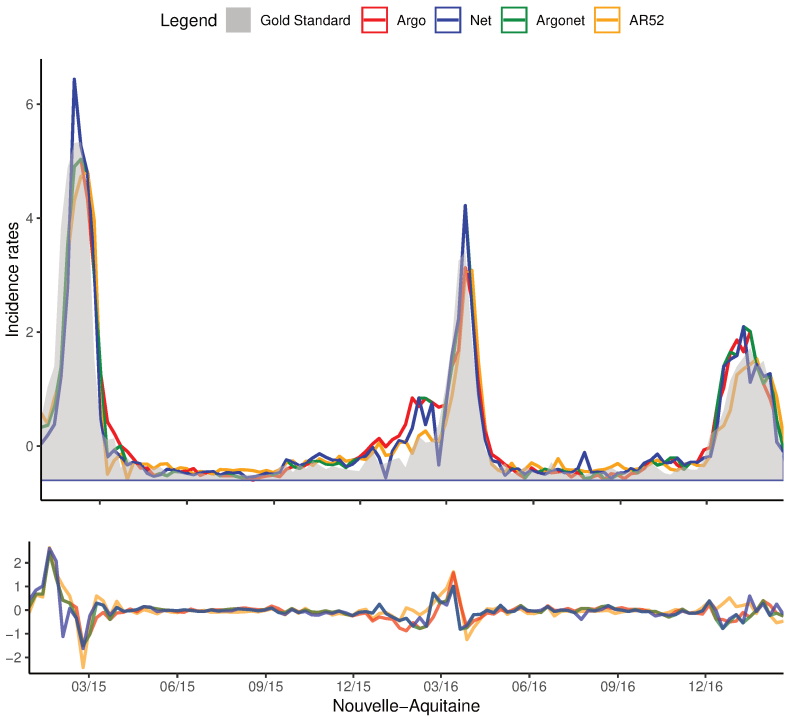


Figure 4. Nouvelle-Aquitaine Real time estimate

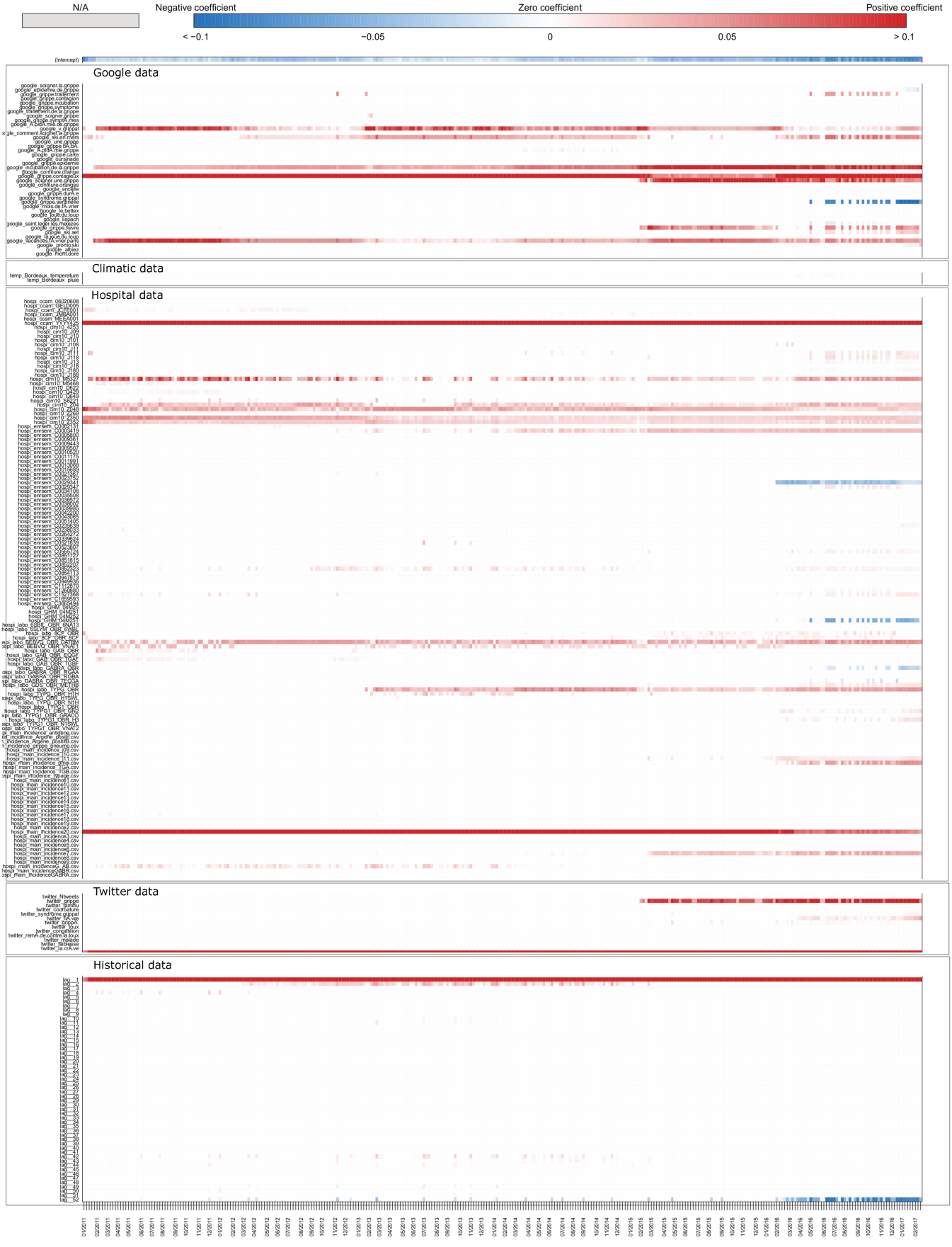


Figure 5. Nouvelle-Aquitaine Real time estimate

One-week forecast

Table 11 shows results for one-week forecast for the time period January 2015-March 2017. Depending on the region, the best PCC is between 0.768 and 0.958 and the best MSE is between 0.460 and 0.083.
For one-week forecast the best model is ARGONet but unlike for real-time estimates, NET model gives better results than ARGO model. AR(52) is the model who gives the weakest results. We can also observe these results on barplots (Figure 7) and for the distribution of PCC and MSE (Figures 14 and 15).

As for Figure 4, Figure 8 shows one-week estimates obtained for the french region Nouvelle-Aquitaine. On this plot, we can see that AR(52) and ARGO models still have a lag of one or two weeks unlike Net and ARGONet models. Net and ARGONet models have very close estimates for the first 2 outbreaks but Net overestimates the third peak. On the heatmap Figure 9, we can see that ARGO model uses mostly 16 variables including 6 variables from Google Data, 9 variables from Hospital Data. One variable from Twitter is used from 2013. We can see also that Climatic data is more used for one-week estimate than for real time.

	Auv.	Bour.	Bre.	Cen.	Gd Est	Ht Fra.	Ile Fra.	Norm.	Aqui.	Occi.	Loi.	Pro.
MSE												
AR(52)	0.262	0.611	0.433	0.438	0.276	0.433	0.372	0.580	0.480	0.278	0.704	0.389
Argo	0.205	0.803	0.334	0.259	0.394	0.289	0.279	0.487	0.300	0.168	0.479	0.254
Net	0.343	**0.269**	0.398	0.282	0.271	0.324	0.439	0.595	0.518	0.235	0.770	0.173
K=1	0.203	0.783	**0.213**	0.237	0.328	0.385	0.238	0.347	**0.196**	0.112	**0.460**	0.159
K=2	0.253	0.766	0.227	0.167	0.320	0.333	0.249	**0.338**	0.212	**0.101**	0.486	0.173
K=3	0.235	**0.205**	0.244	0.173	0.319	0.366	0.245	0.343	**0.209**	0.213	**0.465**	0.156
K=4	0.187	0.765	0.339	0.189	0.314	0.374	0.254	**0.331**	0.227	0.155	0.479	0.180
Mean	**0.101**	0.435	0.217	**0.156**	**0.172**	**0.188**	**0.170**	0.444	0.252	**0.098**	0.523	**0.084**
Lm	**0.118**	0.327	**0.216**	**0.165**	**0.182**	**0.226**	**0.219**	0.484	0.284	0.103	0.536	**0.083**
PCC												
AR(52)	0.868	0.692	0.782	0.779	0.861	0.782	0.813	0.707	0.758	0.860	0.645	0.804
Argo	0.896	0.595	0.832	0.869	0.801	0.854	0.859	0.754	0.849	0.915	0.758	0.876
Net	0.827	**0.864**	0.799	0.858	0.863	0.837	0.778	0.700	0.739	0.881	0.612	0.912
K=1	0.897	0.605	**0.893**	0.881	0.835	0.806	0.880	0.825	**0.901**	0.944	**0.768**	0.920
K=2	0.872	0.614	0.885	0.916	0.839	0.832	0.875	**0.860**	0.893	**0.949**	0.755	0.913
K=3	0.881	**0.897**	0.877	0.913	0.839	0.815	0.877	0.827	**0.895**	0.893	**0.765**	0.921
K=4	0.906	0.614	0.829	0.905	0.842	0.811	0.872	**0.833**	0.885	0.922	0.759	0.909
Mean	**0.949**	0.781	0.890	**0.921**	**0.913**	**0.905**	**0.914**	0.776	0.873	**0.951**	0.736	**0.957**
Lm	**0.940**	0.835	**0.891**	**0.917**	**0.908**	**0.886**	**0.890**	0.756	0.857	0.948	0.729	**0.958**

Table 11. PCC and MSE for one-week forecast for all french regions for the period starting from January 2015 to March 2017

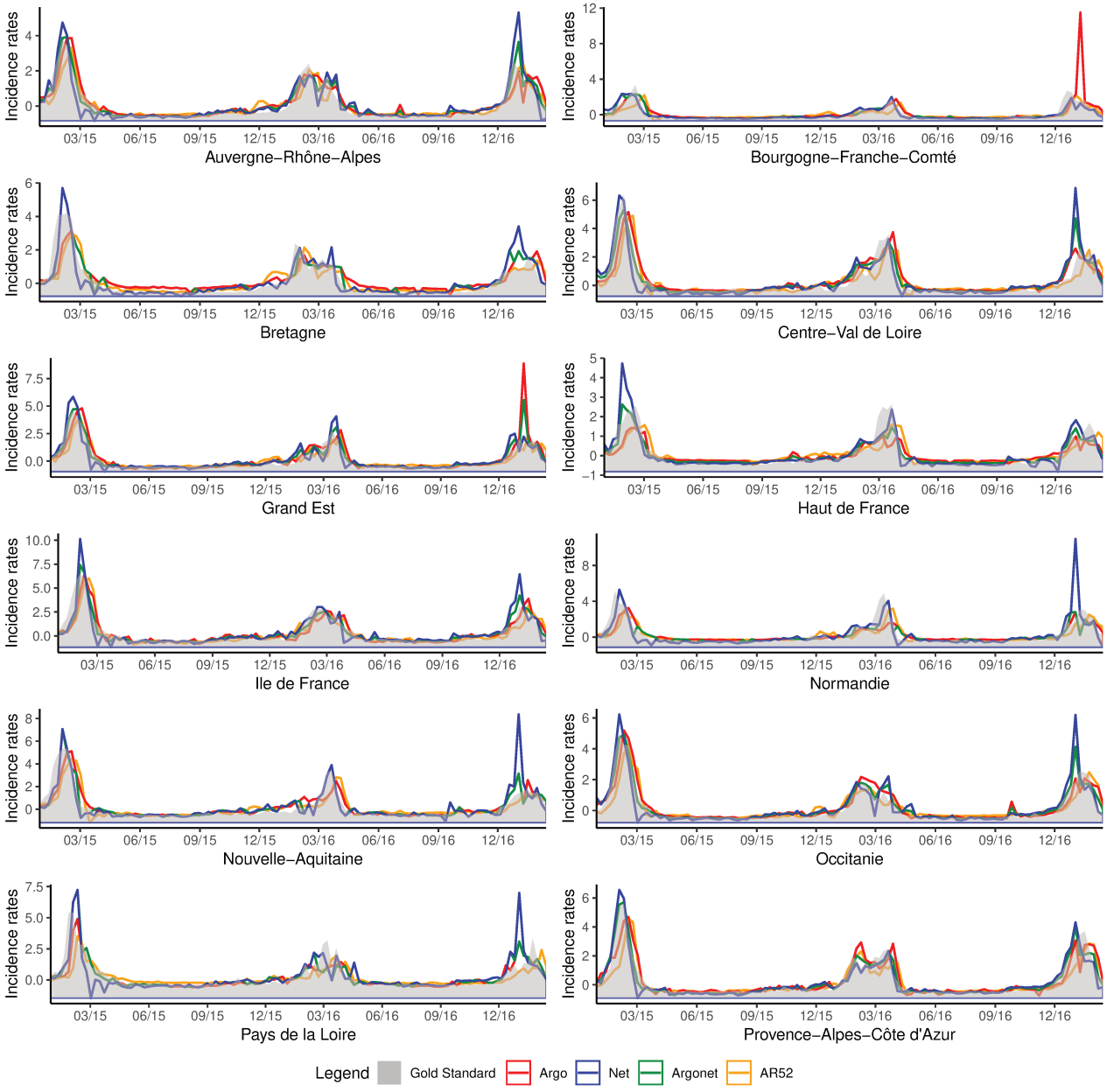


Figure 6. Flu outbreaks from January 2015 to March 2017 and one-week forecast obtained with ARGO, Net and ARGONet models

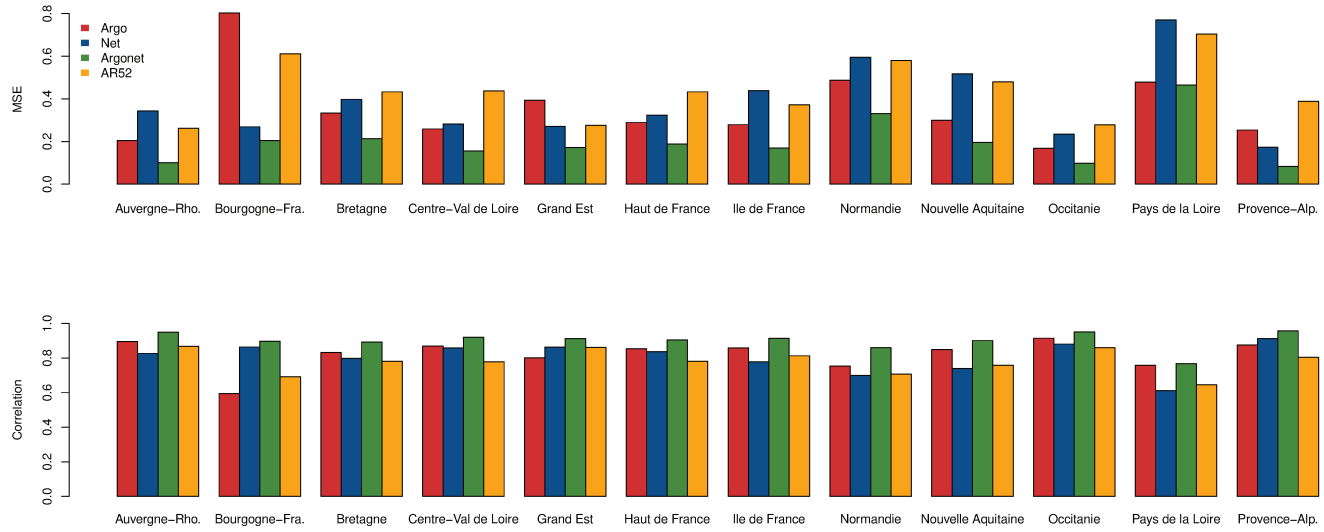


Figure 7. PCC and MSE obtained for one-week forecast with ARGO, Net and ARGONet models

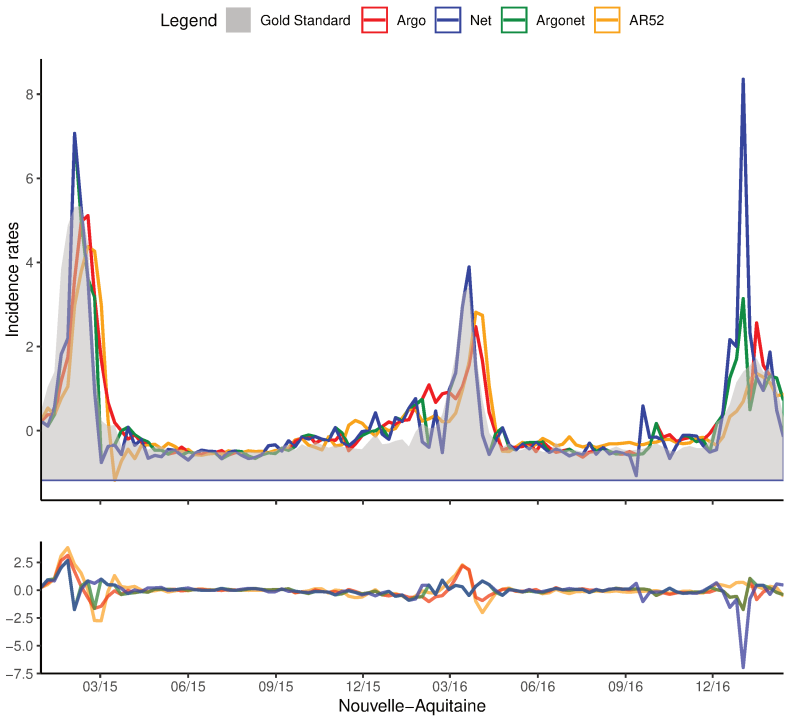


Figure 8. Nouvelle-Aquitaine One-week estimate

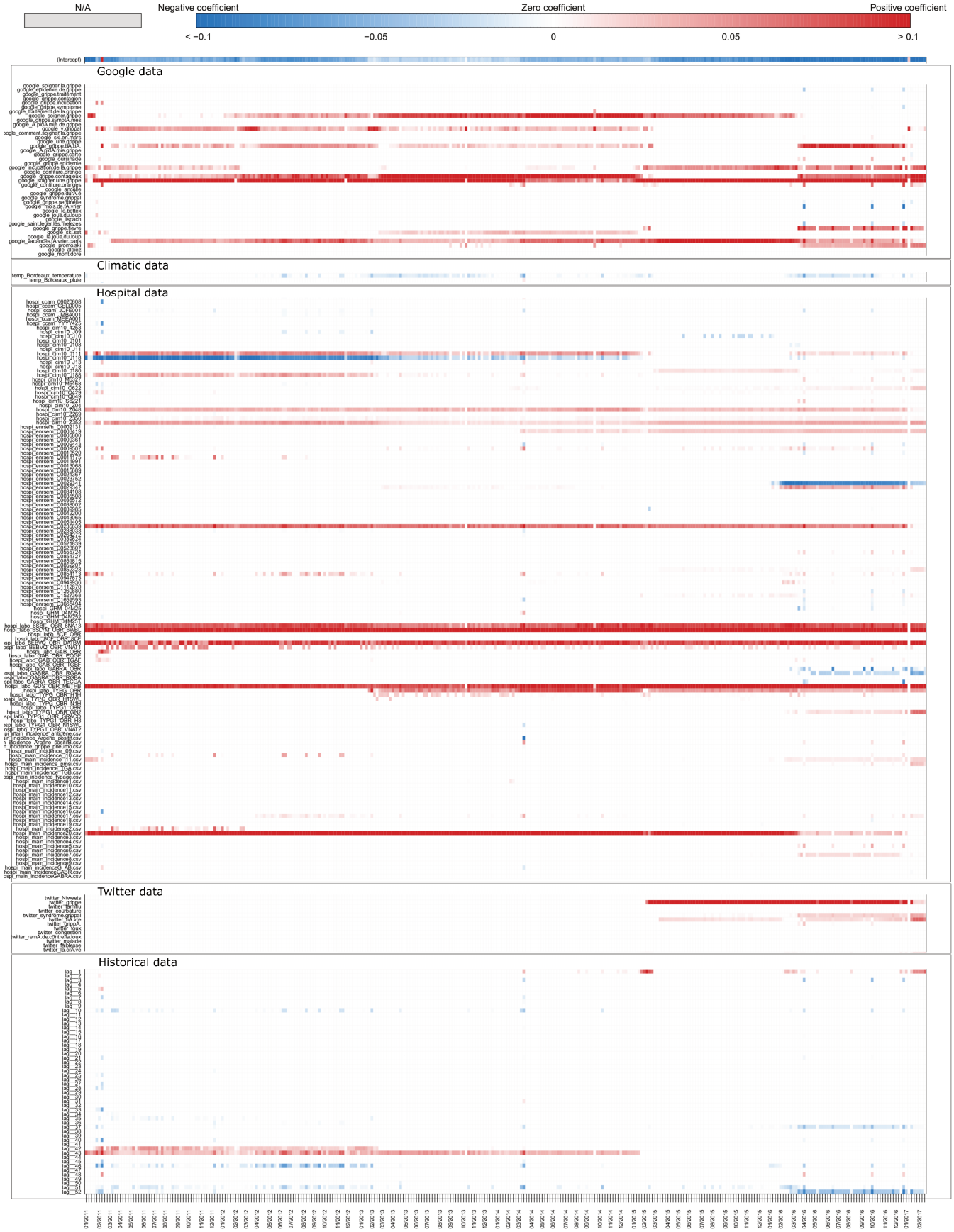


Figure 9. Coefficients Nouvelle-Aquitaine One-week estimate

Two-week forecast

Table 5 shows results for two-week forecast for the time period January 2015-March 2017. Depending on the region, the best PCC varies from 0.779 to 0.957 and the best MSE varies from 0.438 to 0.086. Like for real-time and one-week forecast, AR(52) is the weakest model. The best model is ARGONet for all french regions. The best method on this period is the method using the mean.
Barplot (Figure 11) confirms that ARGONet is the best model for all regions in term of PCC and MSE. On the same way, for the distribution of PCC and MSE (Figure 14 and 15), ARGONet is the best model.

Figure 12 shows two-week estimates for the region Nouvelle-Aquitaine. As for one-week estimate, we can see that AR(52) and ARGO models still have a lag unlike Net and ARGONet models. Nevertheless, Net overestimates the outbreaks. On the heatmap Figure 13, we can see that ARGO model uses mostly 17 variables including 5 variables from Google Data, 9 variables from Hospital Data. Twos variables from Twitter are used from 2013. One variable from Historical data is used before 2013. We can see also that Climatic data is used.

	Auv.	Bour.	Bre.	Cen.	Gd Est	Ht Fra.	Ile Fra.	Norm.	Aqui.	Occi.	Loi.	Pro.
MSE												
AR(52)	0.534	0.947	0.665	0.776	0.530	0.679	0.618	0.914	0.775	0.533	0.935	0.596
Argo	0.424	0.505	0.439	0.386	0.357	0.378	0.416	0.621	0.490	0.222	0.548	0.252
Net	0.351	0.365	0.374	0.282	0.352	0.346	0.573	0.472	0.457	0.265	0.802	0.333
K=1	0.209	0.362	0.263	0.175	0.236	0.377	0.365	0.407	**0.229**	0.132	0.695	0.239
K=2	0.257	0.324	0.283	0.193	0.290	0.371	0.280	0.499	0.281	0.178	**0.438**	0.181
K=3	0.240	0.290	0.301	0.315	0.284	0.390	0.302	0.471	0.276	0.196	0.497	0.143
K=4	0.300	0.280	0.298	0.326	0.282	0.411	0.397	0.477	0.270	0.149	0.636	0.231
Mean	**0.167**	**0.244**	**0.222**	**0.127**	**0.145**	**0.213**	**0.242**	**0.337**	**0.225**	**0.086**	**0.452**	**0.129**
Lm	**0.197**	**0.277**	**0.238**	**0.144**	**0.195**	**0.253**	**0.250**	**0.361**	0.247	**0.099**	0.511	**0.135**
PCC												
AR(52)	0.731	0.522	0.665	0.609	0.733	0.658	0.688	0.539	0.610	0.731	0.528	0.699
Argo	0.786	0.745	0.778	0.805	0.820	0.809	0.790	0.687	0.753	0.888	0.723	0.873
Net	0.823	0.816	0.811	0.858	0.822	0.825	0.841	0.762	0.770	0.867	0.596	0.832
K=1	0.895	0.817	0.867	0.912	0.881	0.810	0.816	0.794	**0.885**	0.934	0.649	0.879
K=2	0.871	0.837	0.857	0.902	0.854	0.813	0.859	0.748	0.858	0.910	**0.779**	0.909
K=3	0.879	0.854	0.848	0.841	0.857	0.803	0.847	0.762	0.861	0.901	0.749	0.928
K=4	0.849	0.859	0.849	0.836	0.858	0.793	0.850	0.759	0.864	0.925	0.679	0.888
Mean	**0.916**	**0.877**	**0.888**	**0.936**	**0.927**	**0.892**	**0.878**	**0.830**	**0.886**	**0.957**	**0.772**	**0.935**
Lm	**0.901**	**0.860**	**0.880**	**0.927**	**0.902**	**0.872**	**0.874**	**0.818**	0.875	**0.950**	0.742	**0.932**

Table 12. PCC and MSE for two-week forecast for all french regions for the period starting from January 2015 to March 2017

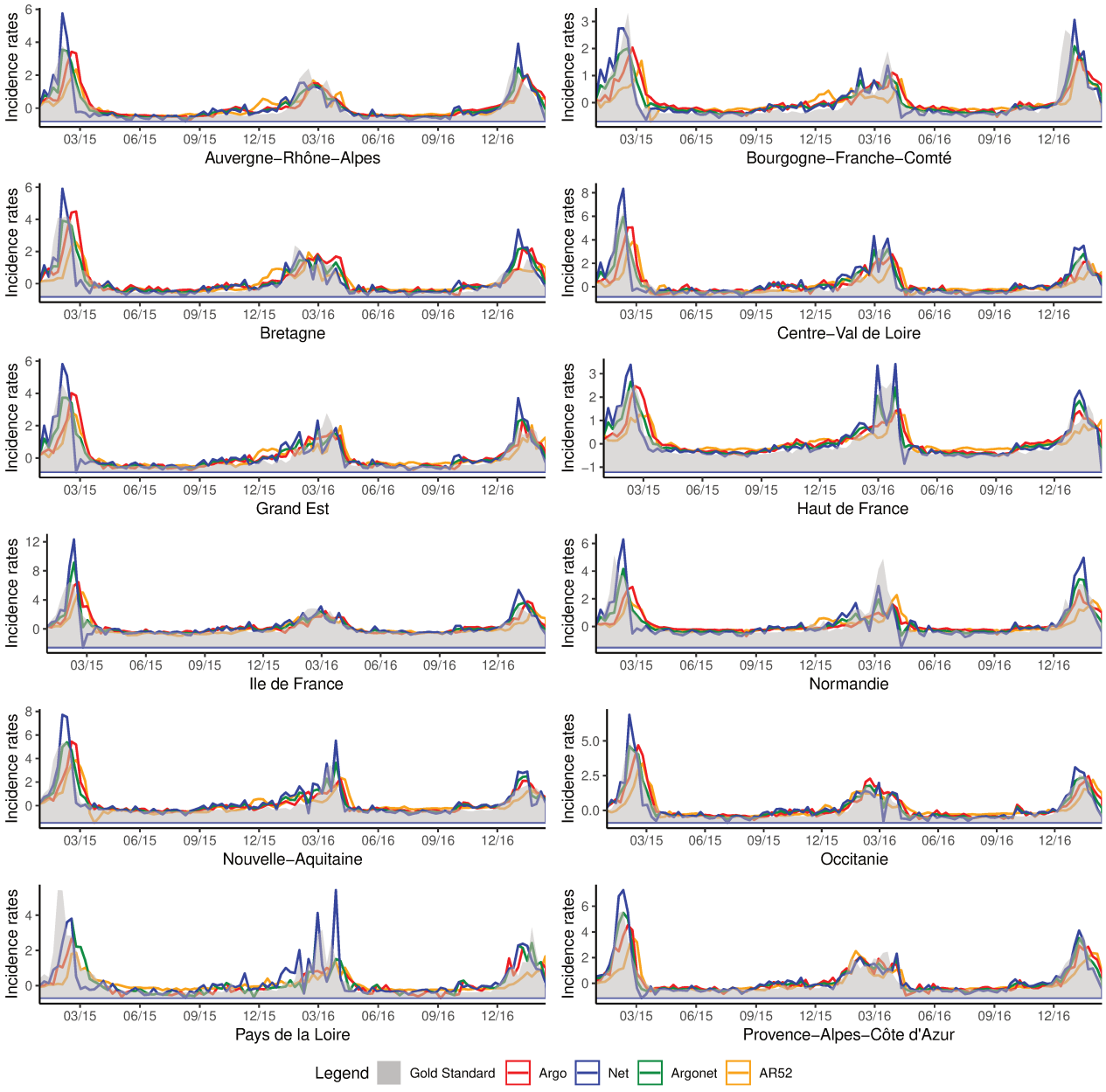


Figure 10. Flu outbreaks from January 2015 to March 2017 and two-week forecast obtained with ARGO, Net and ARGONet models

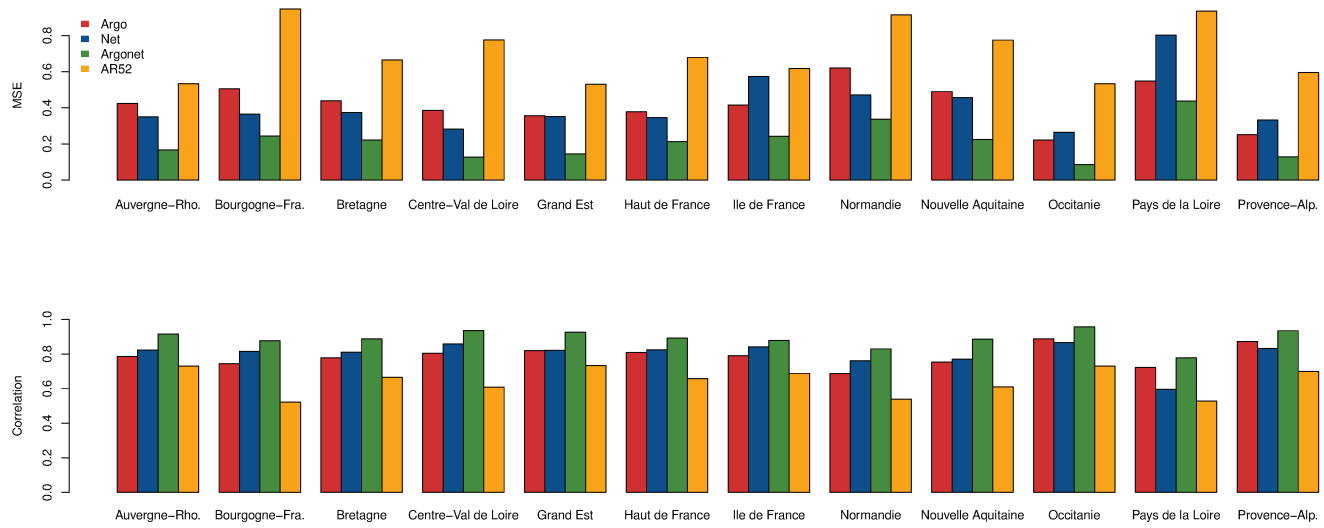


Figure 11. PCC and MSE obtained for two-week forecast with ARGO, Net and ARGONet models

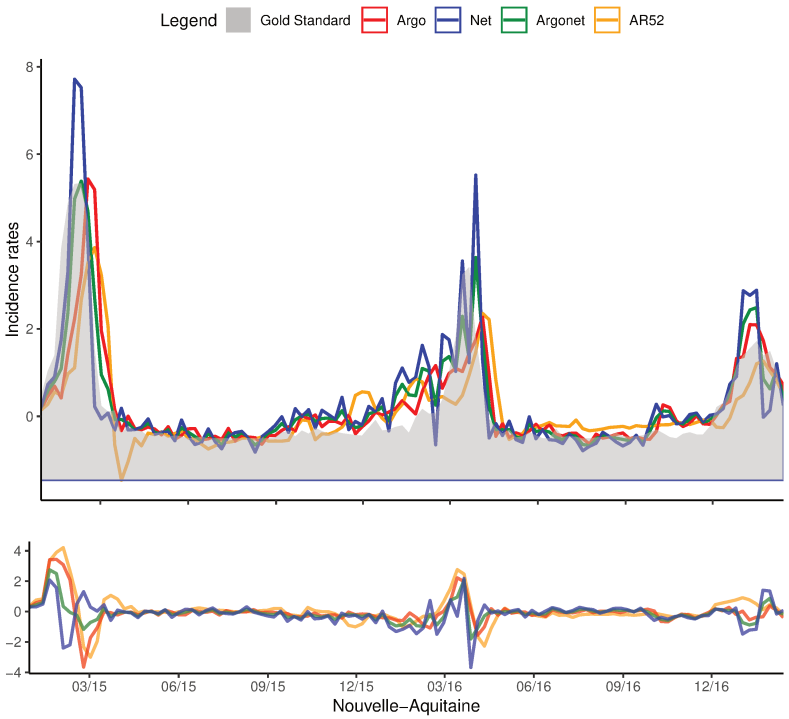


Figure 12. Nouvelle-Aquitaine Two-week estimate

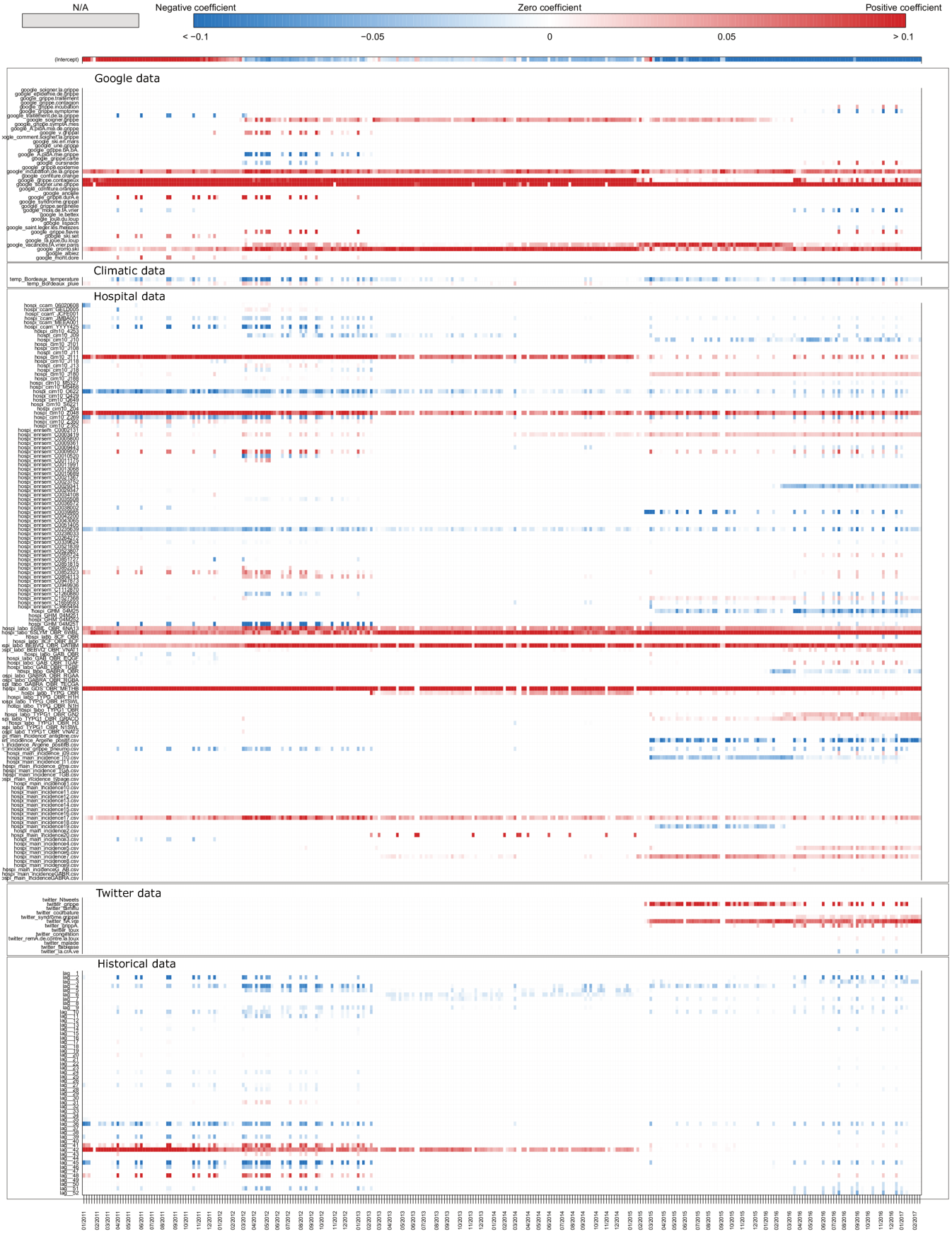


Figure 13. Coefficients Nouvelle-Aquitaine Two-week estimate

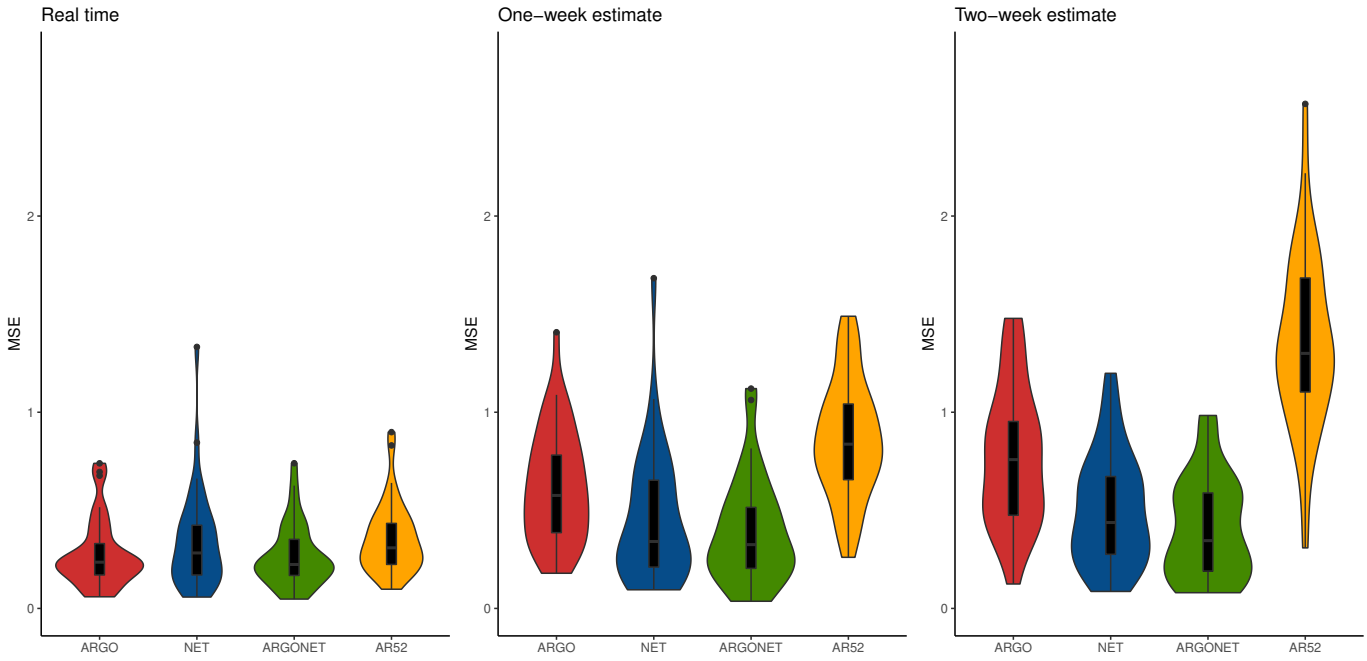


Figure 14. The distribution of the MSE

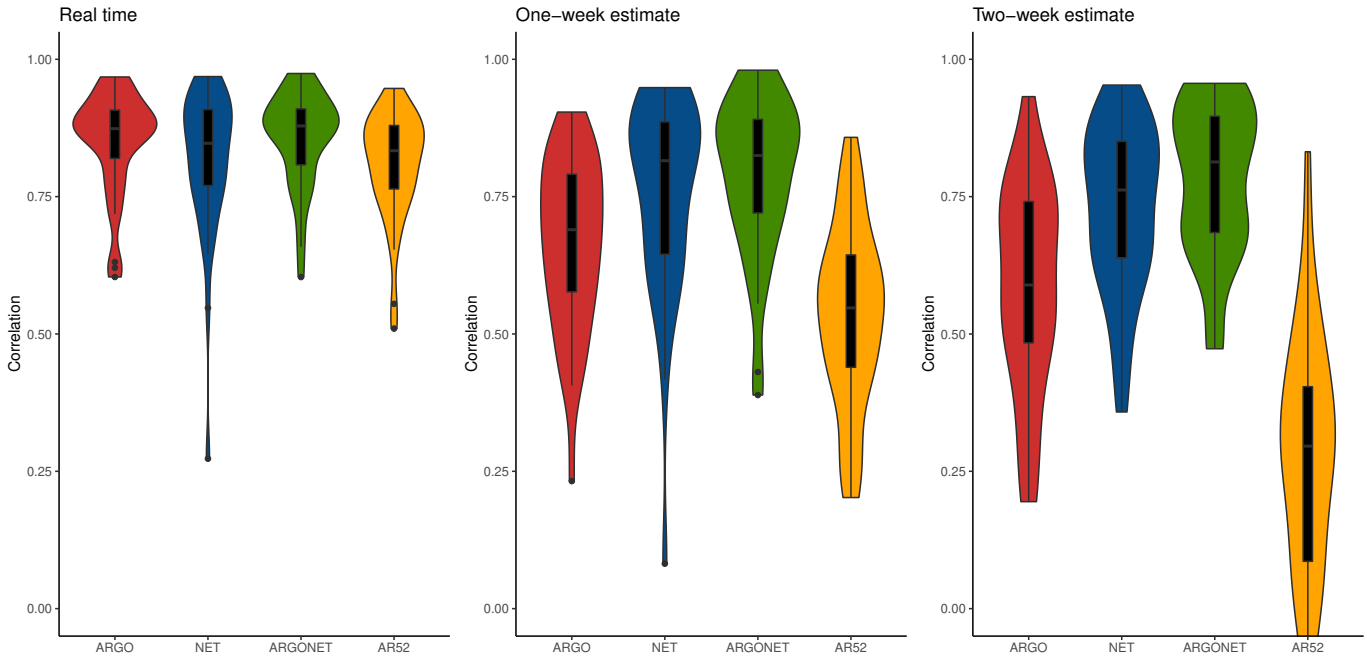


Figure 15. The distribution of the PCC

Discussion

Here we show that methods applied in United States could be used in France. The ARGO model has good results for real time estimates but not for longer-term forecasts. It has a lack of robustness. Conversely, the Net model has better results for longer-term forecasting but tends to overestimate epidemics. The ensemble approach, ARGONet, is the method allowing to produce forecasts with the highest correlation and the lowest errors. This is the most robust model to estimate ILI activity up to two-weeks at the french regional level. Figures S2 to S13 show that the autoregressive model (AR(52)) have a lag increasing and an amplitude decreasing with the forecast period whereas ARGONet keeps closer estimates. Moreover, the method using the mean between estimates with ARGO model and Net model stands out by being better than other methods for the two-week forecast.

In this study, thanks to the comparison with the model using only historical data (AR(52)), we show that external data sources allow to improve estimates and more particularly for longer-term forecasts. Indeed, for the two-week forecast, the combination of Hospital data and Google data allows an improvement of the PCC up to 0.102 and a decrease of MSE up to 0.321. For Climatic data, this improvment reaches (+0.129) for the PCC and (-0.256) for the MSE. For Twitter data, we have (+0.093) for the PCC and (-0.183) for the MSE. If we combine all the data sources for the two-week forecast, we obtaine an improvment of (+0.223) for the PCC and (-0.442) for the MSE.
By analyzing heatmaps obtained for the ARGO models, we can see that the most used data sources are Hospital data and Google data. For the prevision in real-time, the ILI previous incidence rate from historical data is still used, but for longer-term forecasting, historical data are less used. On the other hand, Climatic data and Twitter data are more used for longer-term forecasting than for real time estimate.

One limitation of our study is about the data. Indeed, we have only Hospital data from Rennes University Hospital from the Brittany region. If we could have Hospital data from all the french regions maybe we could have better results. In the same way, Twitter data collected are at the scale of the country and not the region.

To conclude, the ensemble approach ARGONet developed in United States could be used to complete traditional influenza surveillance method in France. The results in real-time and up to two-week forecast could allow to anticipate ILI outbreaks. It could be helpful for decision-making. Thanks to the spatial resolution at the regional level, it could be possible to manage the patients' flow in general practitioners' offices and in hospitals, particularly emergency departments.
Moreover, in addition to hospital data and Google data, we know that Twitter data and Climatic data can be useful for influenza monitoring.

Supplementary Material

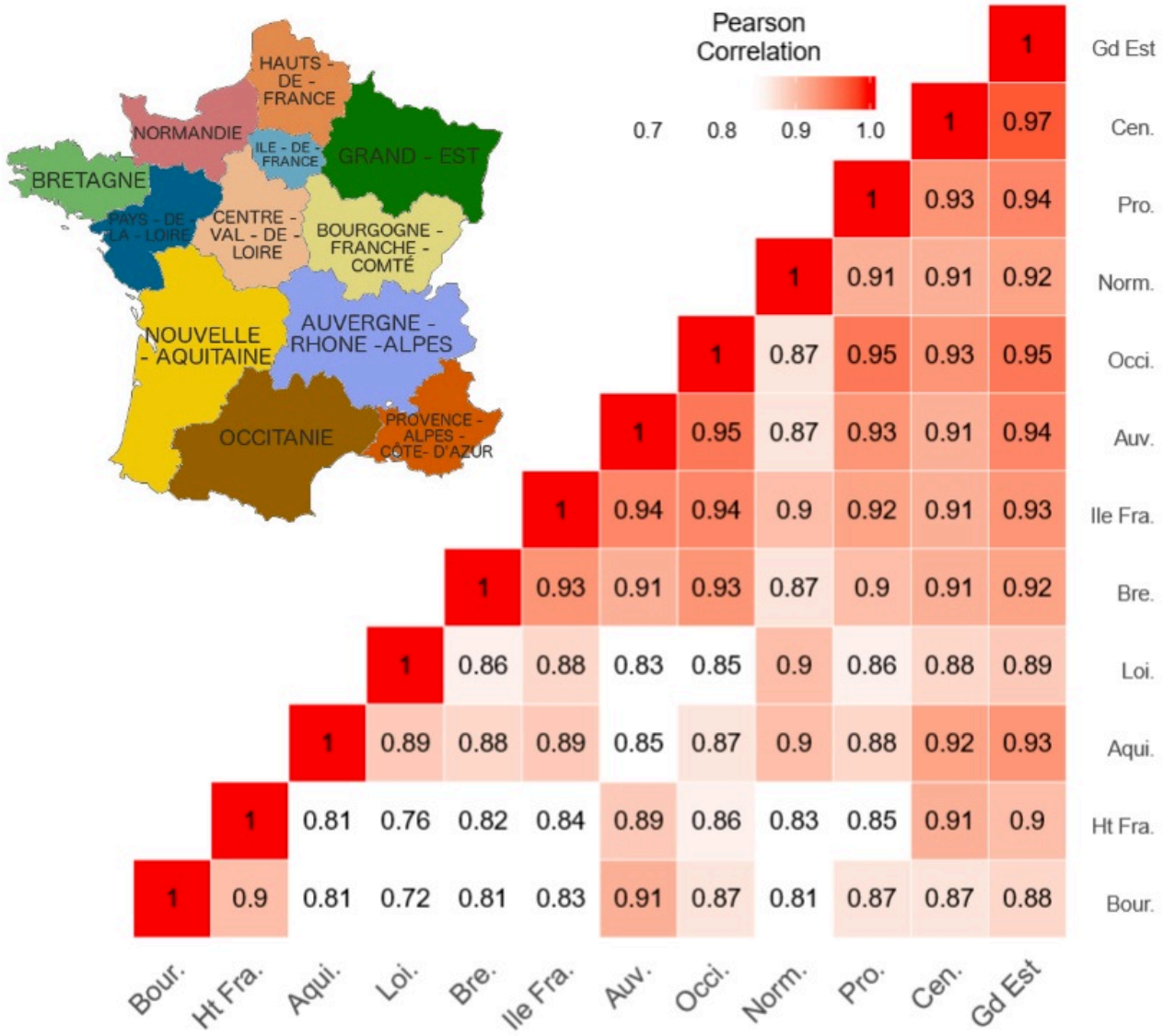


Figure S1. Correlation between French regions on the period starting from January 2013 to March 2017

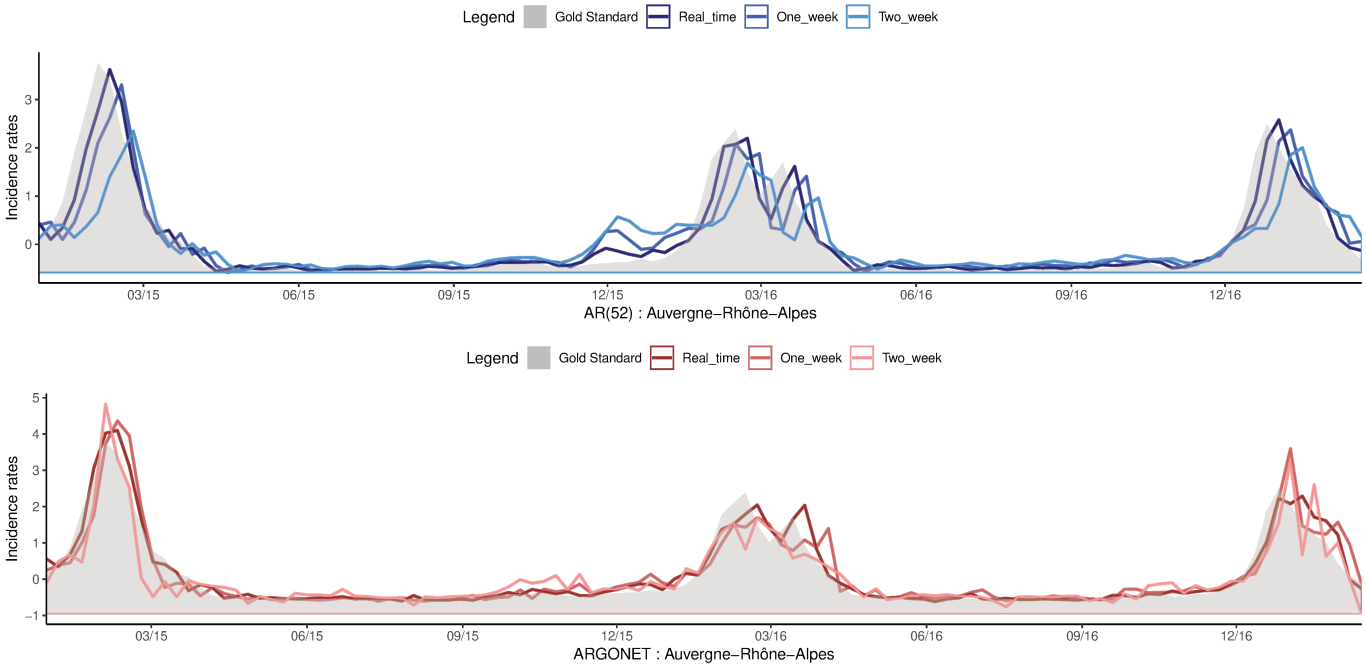


Figure S2. Evolution of Auvergne-Rhône-Alpes estimates over time for AR(52) and ARGONET models

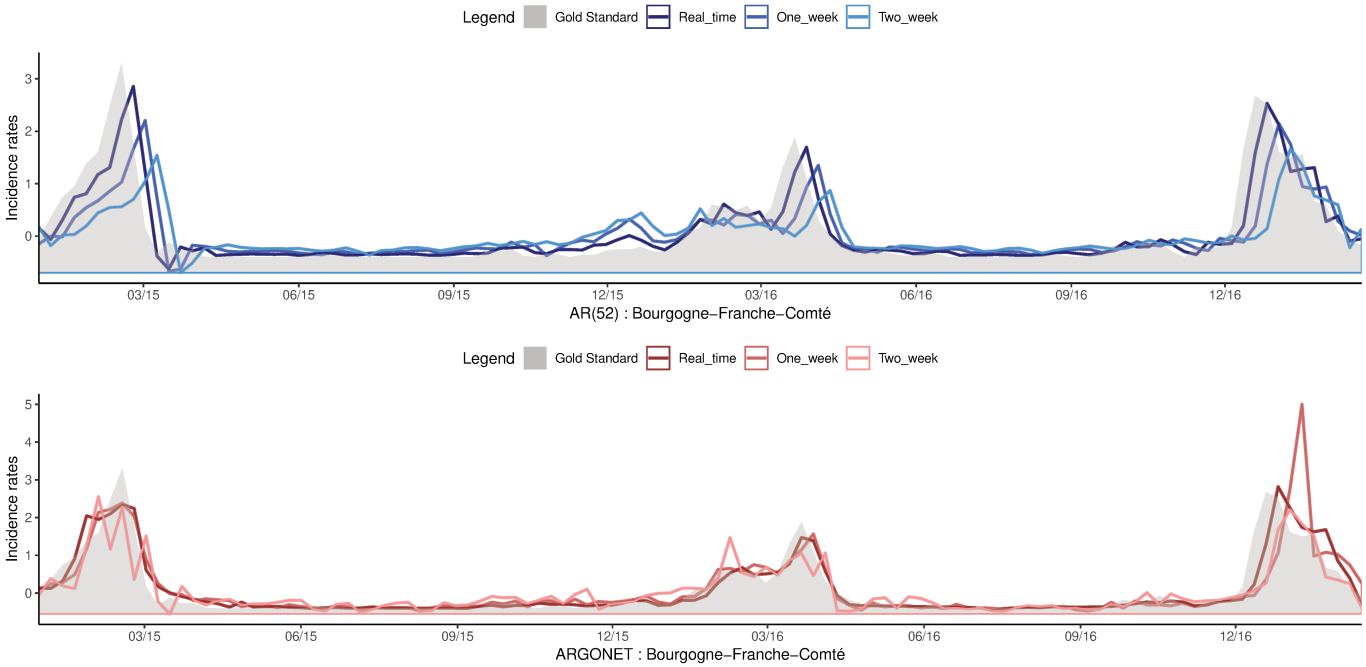


Figure S3. Evolution of Bourgogne-Franche-Comté estimates over time for AR(52) and ARGONET models

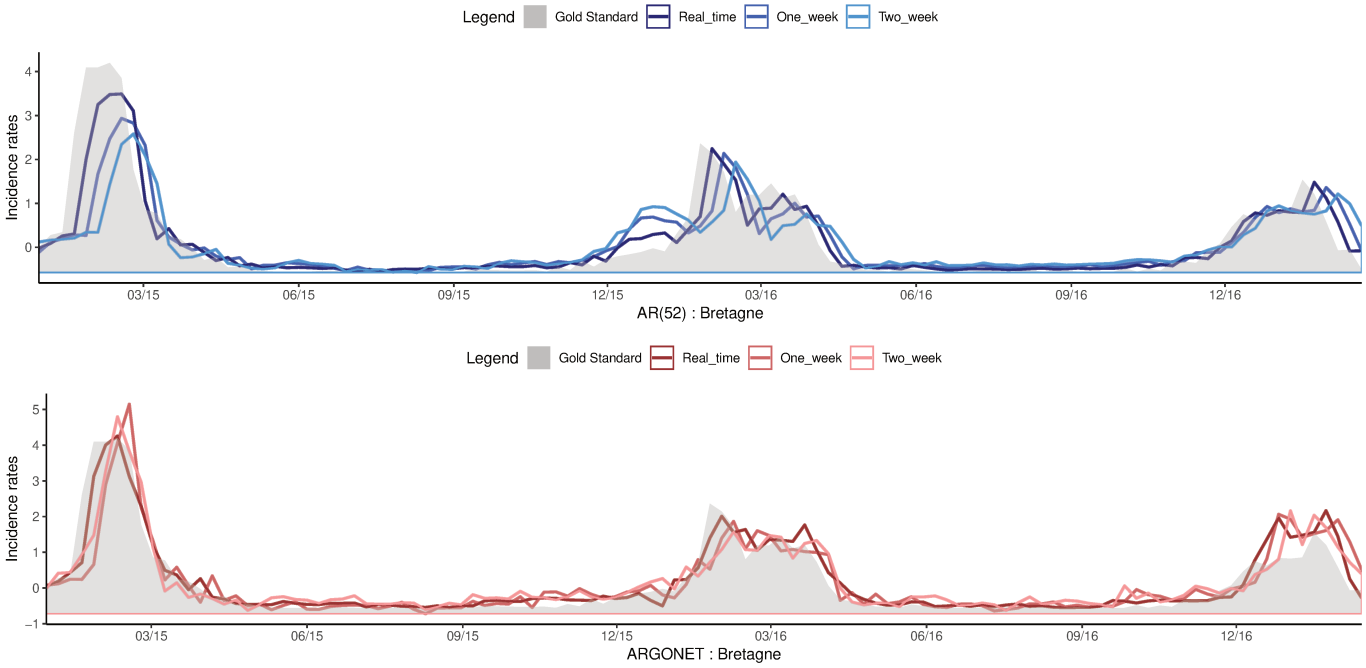


Figure S4. Evolution of Bretagne estimates over time for AR(52) and ARGONET models

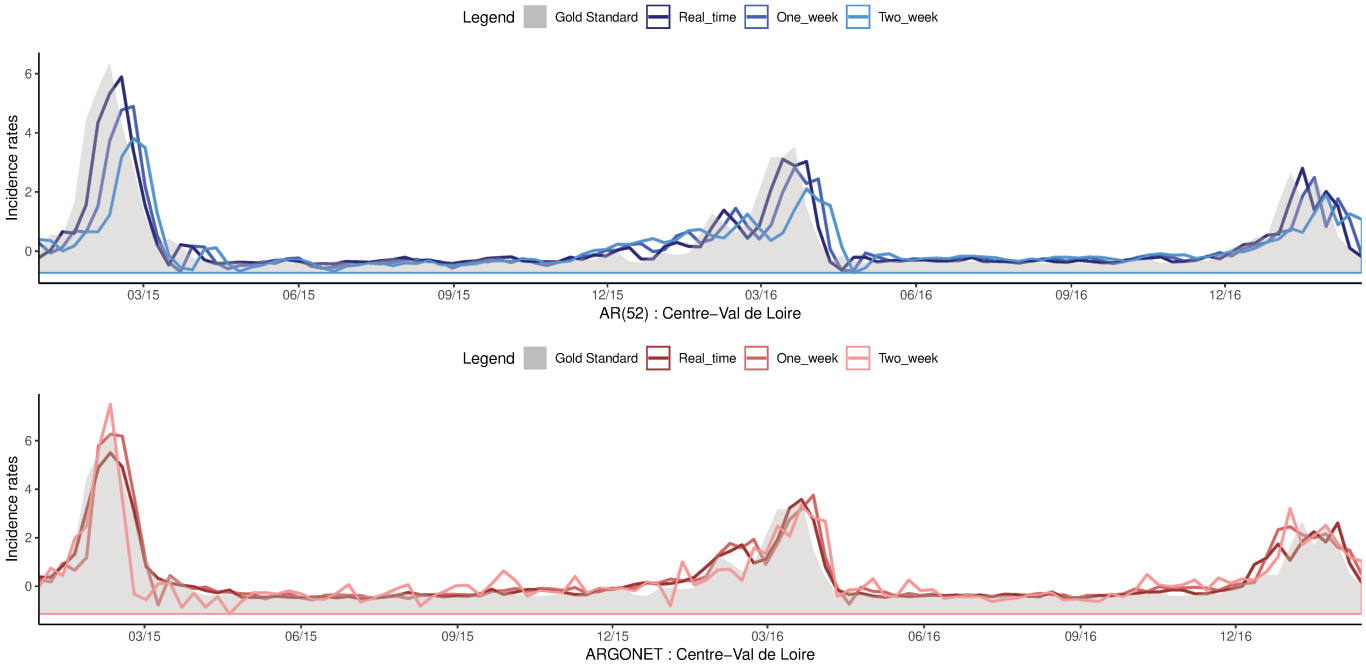


Figure S5. Evolution of Centre-Val de Loire estimates over time for AR(52) and ARGONET models

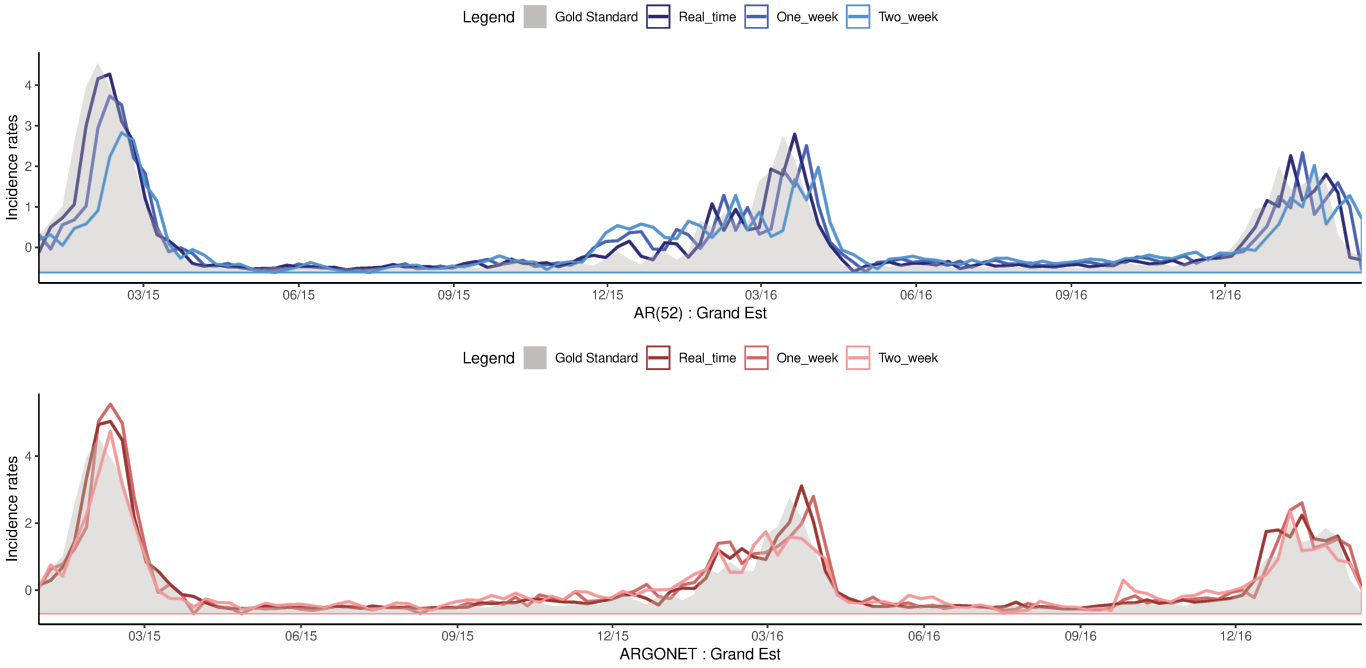


Figure S6. Evolution of Grand Est estimates over time for AR(52) and ARGONET models

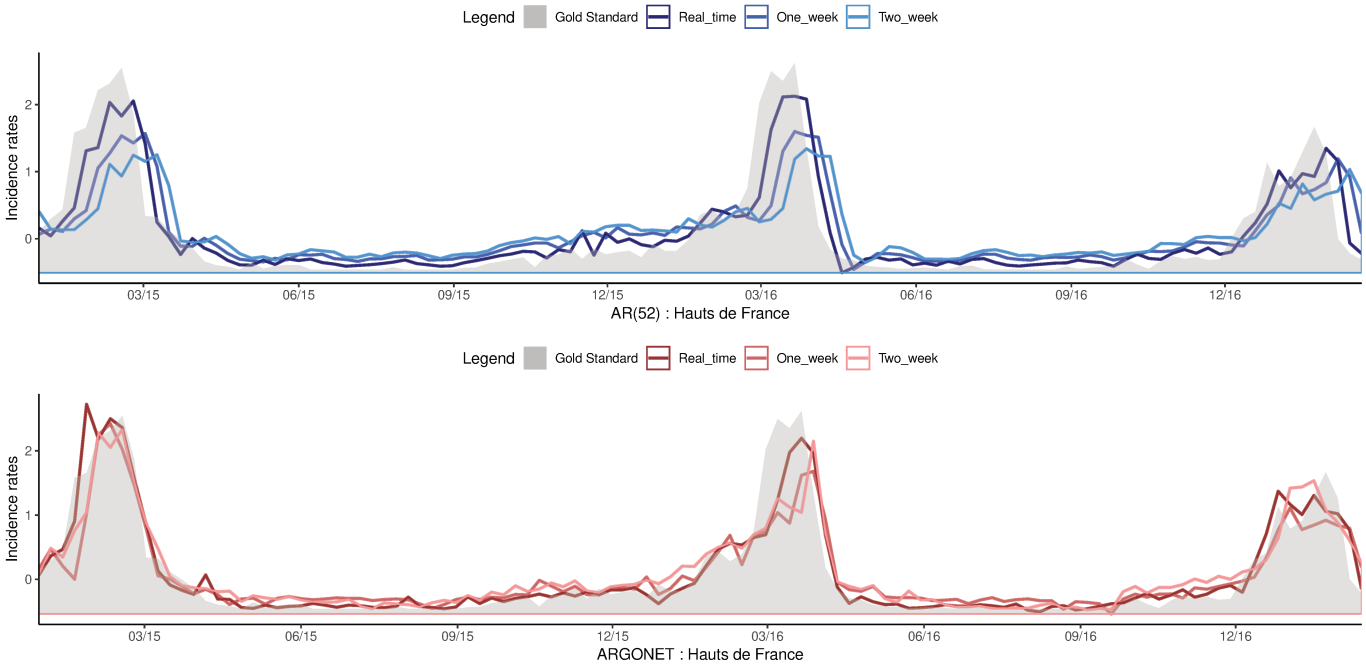


Figure S7. Evolution of Hauts de France estimates over time for AR(52) and ARGONET models

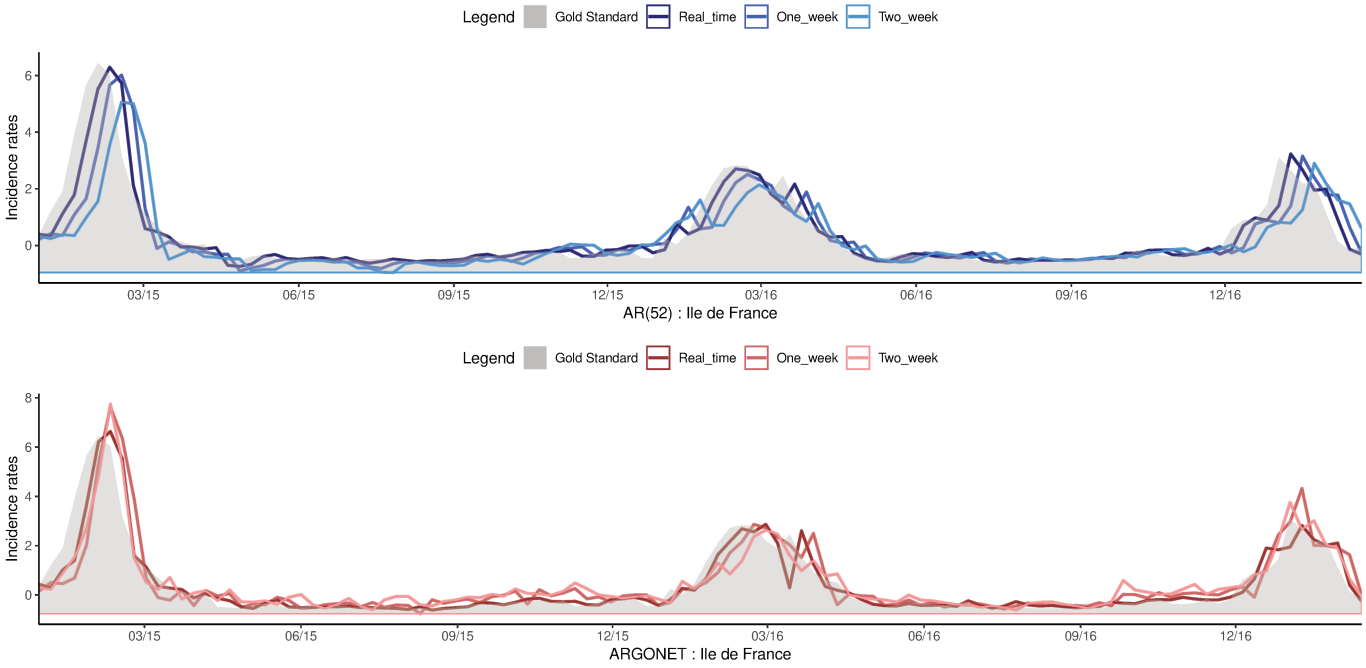


Figure S8. Evolution of Ile de France estimates over time for AR(52) and ARGONET models

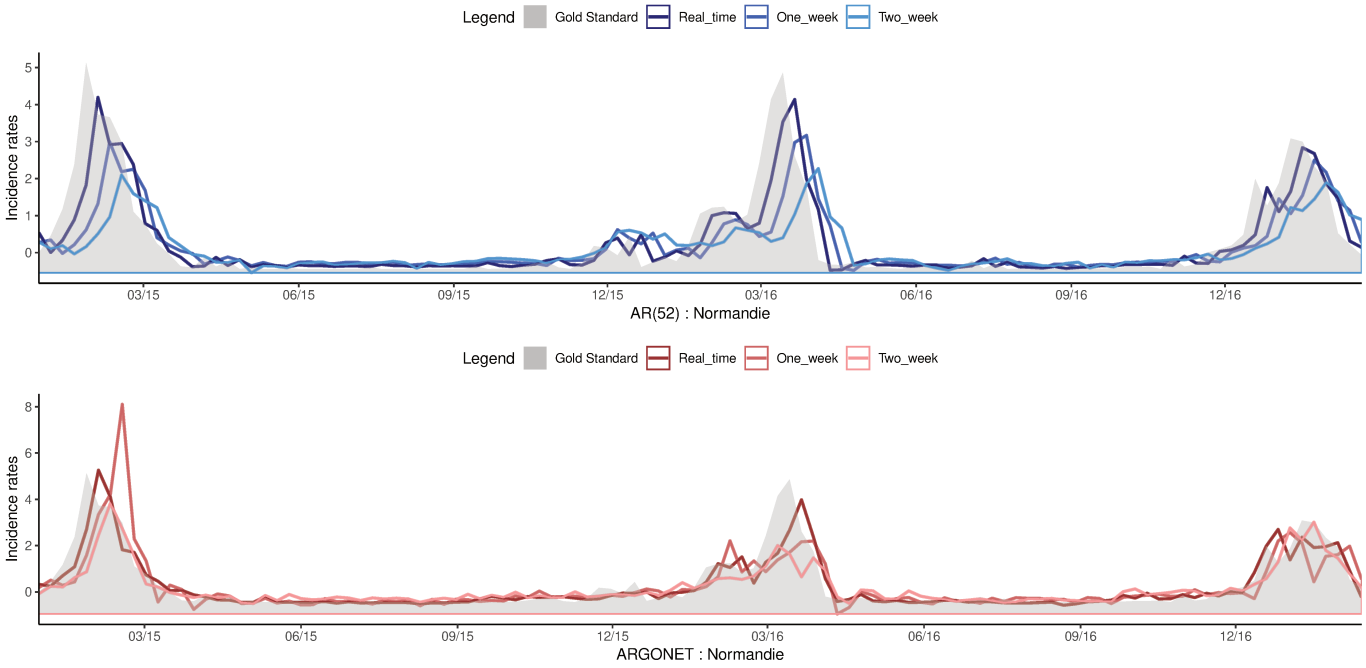


Figure S9. Evolution of Normandie estimates over time for AR(52) and ARGONET models

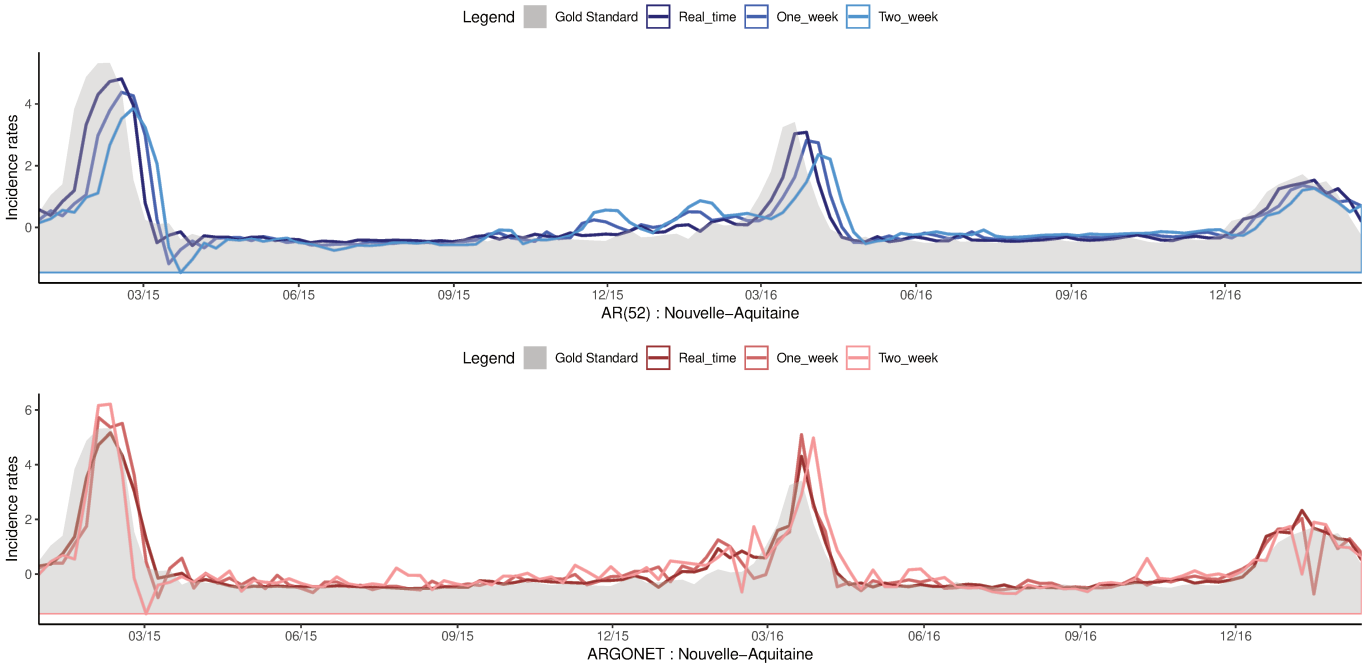


Figure S10. Evolution of Nouvelle-Aquitaine estimates over time for AR(52) and ARGONET models

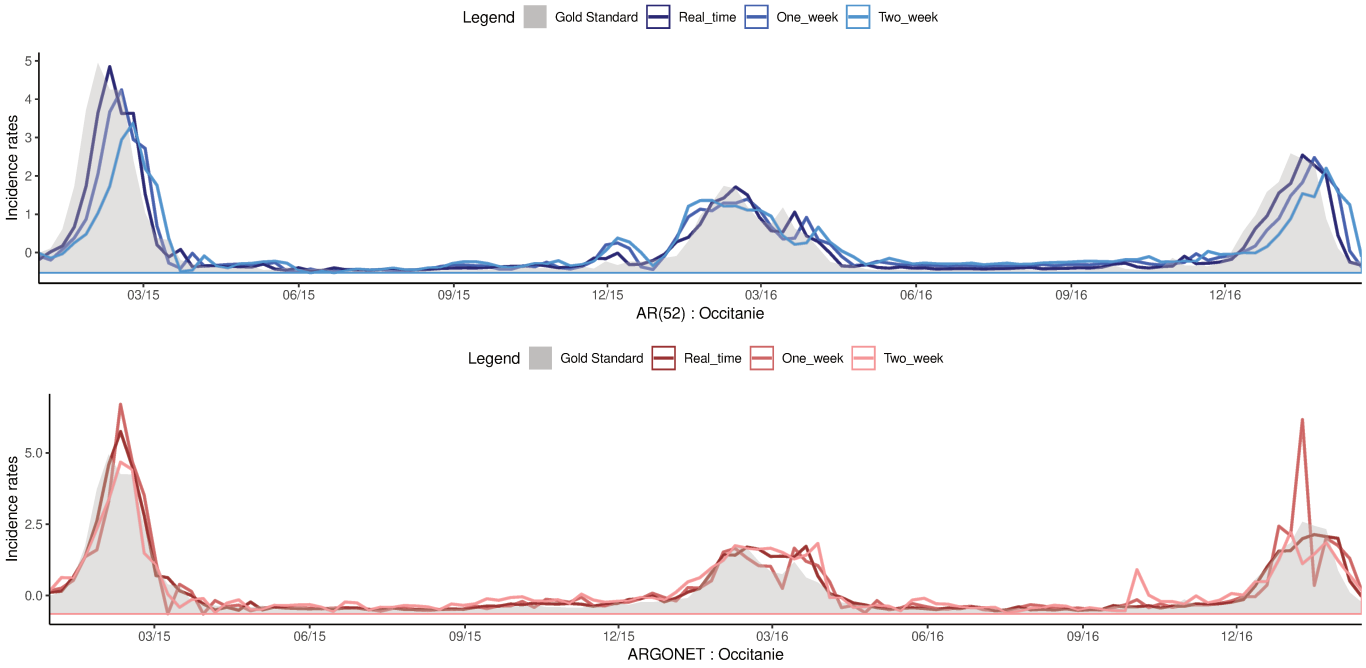


Figure S11. Evolution of Occitanie estimates over time for AR(52) and ARGONET models

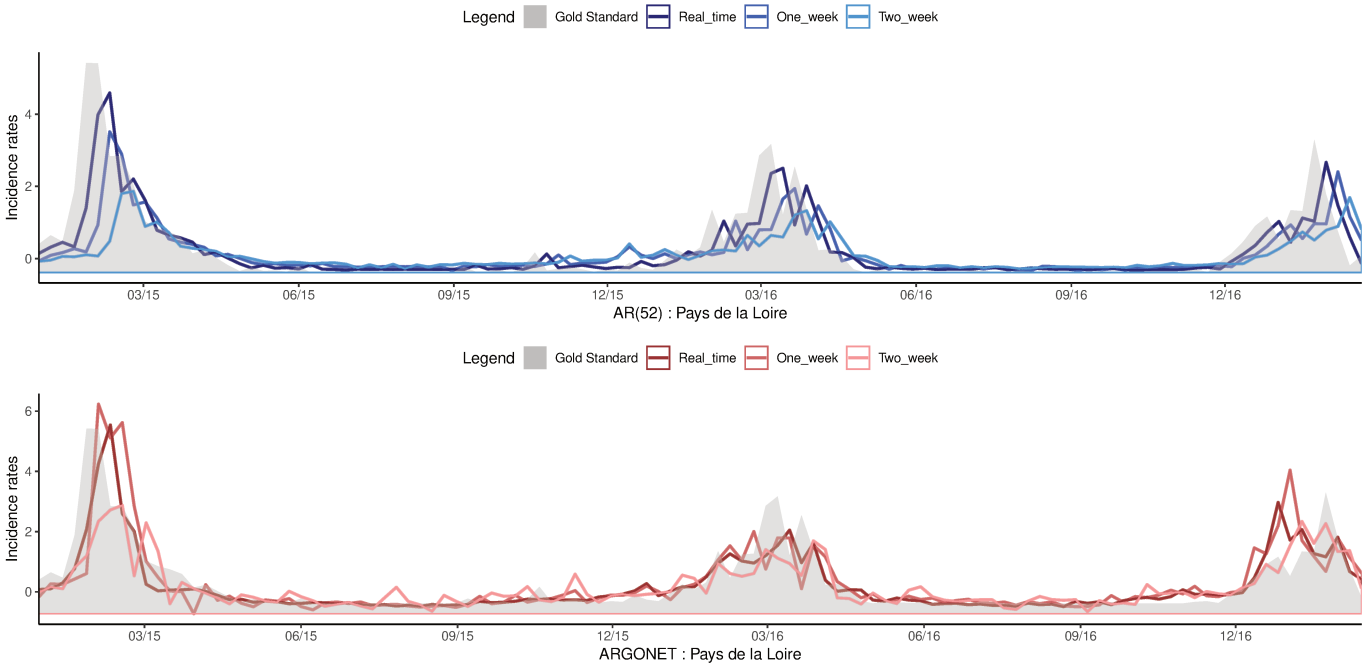


Figure S12. Evolution of Pays de la Loire estimates over time for AR(52) and ARGONET models

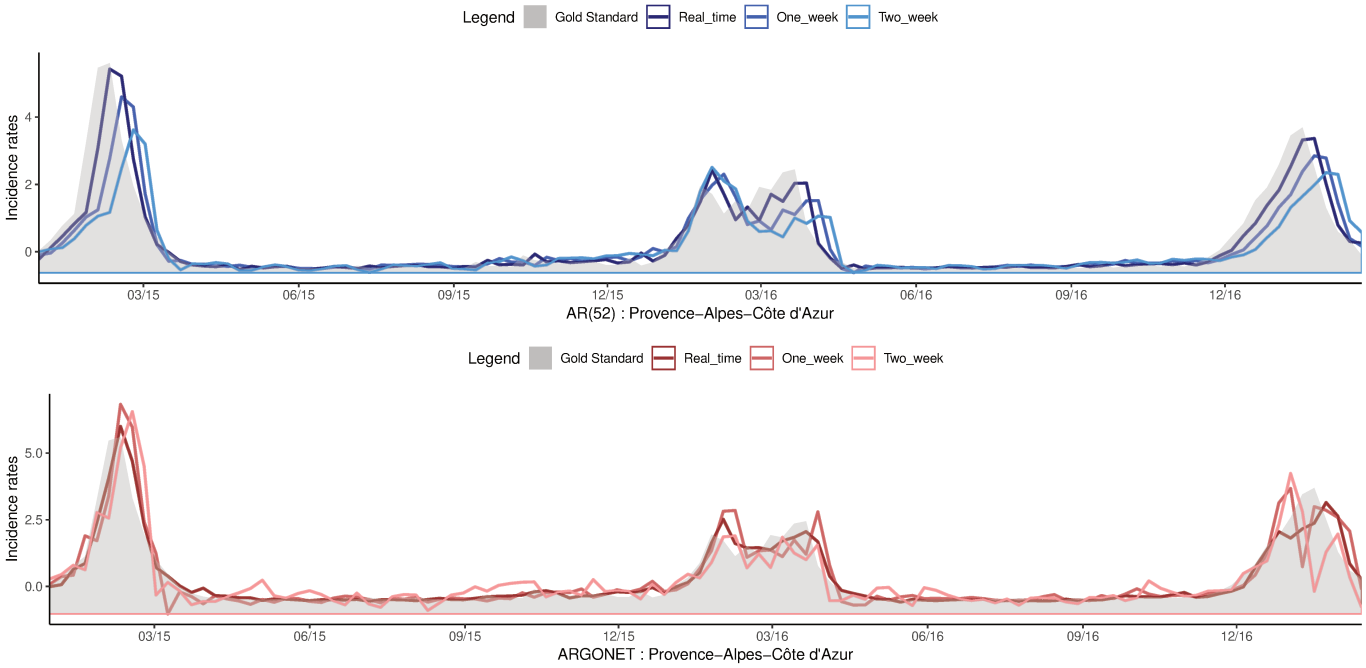


Figure S13. Evolution of Provence-Alpes-Côte d'Azur estimates over time for AR(52) and ARGONET models

References

1. Ferguson, N. M. *et al.* Strategies for mitigating an influenza pandemic. **442**, 448–452, DOI: 10.1038/nature04795.
2. Yang, S., Santillana, M. & Kou, S. C. Accurate estimation of influenza epidemics using google search data via ARGO. **112**, 14473–14478, DOI: 10.1038/srep25732.
3. Yang, W., Lipsitch, M. & Shaman, J. Inference of seasonal and pandemic influenza transmission dynamics. **112**, 2723–2728, DOI: 10.1073/pnas.1415012112.
4. Kalimeri, K. *et al.* Unsupervised extraction of epidemic syndromes from participatory influenza surveillance self-reported symptoms. **15**, e1006173, DOI: 10.1371/journal.pcbi.1006173.
5. D.M. Fleming, W. P., J. van der Velden. The evolution of influenza surveillance in europe and prospects for the next 10 years. **21**, 1749–1753.
6. Santillana, M. *et al.* Cloud-based electronic health records for real-time, region-specific influenza surveillance. **6**, 25732, DOI: 10.1038/srep25732.
7. Nsoesie, E. O., Brownstein, J. S., Ramakrishnan, N. & Marathe, M. V. A systematic review of studies on forecasting the dynamics of influenza outbreaks. **8**, 309–316.
8. Chretien, J.-P., George, D., Shaman, J., Chitale, R. A. & McKenzie, F. E. Influenza forecasting in human populations: A scoping review. **9**, e94130, DOI: 10.1371/journal.pone.0094130.
9. Olson, D. R., Konty, K. J., Paladini, M., Viboud, C. & Simonsen, L. Reassessing google flu trends data for detection of seasonal and pandemic influenza: A comparative epidemiological study at three geographic scales. **9**, e1003256, DOI: 10.1371/journal.pcbi.1003256.
10. Zhang, Y., Bambrick, H., Mengersen, K., Tong, S. & Hu, W. Using google trends and ambient temperature to predict seasonal influenza outbreaks. **117**, 284–291, DOI: 10.1016/j.envint.2018.05.016.
11. Paul, M. J., Dredze, M. & Broniatowski, D. Twitter improves influenza forecasting. **6**, DOI: 10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117.
12. Santillana, M. *et al.* Combining search, social media, and traditional data sources to improve influenza surveillance. **11**, DOI: 10.1371/journal.pcbi.1004513.
13. Mowery, J. Twitter influenza surveillance: Quantifying seasonal misdiagnosis patterns and their impact on surveillance estimates. **8**, DOI: 10.5210/ojphi.v8i3.7011.
14. Sharpe, J. D., Hopkins, R. S., Cook, R. L. & Striley, C. W. Evaluating google, twitter, and wikipedia as tools for influenza surveillance using bayesian change point analysis: A comparative analysis. **2**, DOI: 10.2196/publichealth.5901.
15. McIver, D. J. & Brownstein, J. S. Wikipedia usage estimates prevalence of influenza-like illness in the united states in near real-time. **10**, e1003581, DOI: 10.1371/journal.pcbi.1003581.
16. Global disease monitoring and forecasting with wikipedia. **10**, DOI: 10.1371/journal.pcbi.1003892.
17. Hickmann, K. S. *et al.* Forecasting the 2013–2014 influenza season using wikipedia. **11**, e1004239, DOI: 10.1371/journal.pcbi.1004239.
18. Carneiro, H. A. & Mylonakis, E. Google trends: A web-based tool for real-time surveillance of disease outbreaks. **49**, 1557–1564, DOI: 10.1086/630200.
19. Butler, D. When google got flu wrong. **494**, 155–156, DOI: 10.1038/494155a.
20. Bouzillé, G. *et al.* Leveraging hospital big data to monitor flu epidemics. **154**, 153–160, DOI: 10.1016/j.cmpb.2017.11.012.
21. Poirier, C. *et al.* Real time influenza monitoring using hospital big data in combination with machine learning methods: Comparison study. **4**, e11361, DOI: 10.2196/11361.

22. Lu, F. S., Hattab, M. W., Clemente, L. & Santillana, M. Improved state-level influenza activity nowcasting in the united states leveraging internet-based data sources and network approaches via ARGONet. 344580, DOI: 10.1101/344580.

23. Tibshirani, R. Regression shrinkage and selection via the lasso. **58**, 267–288.

24. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing).

25. from Jed Wing, M. K. C. *et al. caret: Classification and Regression Training* (2018). R package version 6.0-80.

26. Trapletti, A. & Hornik, K. *tseries: Time Series Analysis and Computational Finance*.

Acknowledgements

We would like to thank the French National Research Agency for funding this work inside the Integrating and Sharing Health Data for Research Project (Grant No. ANR-15-CE19-0024). We also thank the French Sentinelles network and Google and Twitter services for making their data publicly available.

Author contributions statement

C.P. and M.S. conceived of the presented idea. C.P. wrote the manuscript with support from M.S.. G.B extracted hospital data. Y.H and T.B. extracted Twitter data. All authors discussed the results and contributed to the final manuscript.

Conflicts of Interest

None declared.

4.5 Discussion des principaux résultats

Dans cet article, nous avons pu montrer que les méthodes développées aux États-Unis pouvaient être appliquées en France. Le modèle ARGO a de bonnes performances pour l'estimation en temps réel mais est moins robuste pour les prévisions à plus long terme. A l'inverse, le modèle Net a de meilleures performances que le modèle ARGO pour les prévisions à plus long terme, mais tend à surestimer les épidémies. La méthode d'ensemble ARGONet, est donc la méthode permettant d'obtenir les meilleures prévisions en termes de PCC et d'EQM. C'est le modèle le plus robuste pour prévoir la grippe jusqu'à 2 semaines pour toutes les régions de France.

Grâce à la comparaison avec le modèle autorégressif, nous avons également pu voir que toutes les sources de données externes permettaient d'améliorer les prévisions et en particulier les prévisions à plus long terme. En analysant les heatmaps représentant les valeurs des coefficients des modèles ARGO, nous avons mis en évidence que les sources de données les plus utilisées étaient les données hospitalières et les données de Google. Pour la prévision en temps réel, une variable du réseau Sentinelles est toujours utilisée, celle correspondant à la semaine précédant celle que l'on cherche à prédire. Cependant, pour la prévision à plus long terme les données historiques sont moins utilisées. A l'inverse, les données climatiques et les données de Twitter sont plus utilisées pour la prévision à plus long terme que pour la prévision en temps réel.

Comme pour l'étude précédente, une des limites de notre étude, est le fait d'utiliser seulement les données hospitalières du CHU de Rennes. En effet, ces données ne sont peut être pas représentatives de la région Bretagne entière et nous pourrions espérer avoir de meilleurs résultats si nous avions des données provenant des CHU de chaque région. De la même façon, les données Twitter sont collectées à l'échelle de la France et non de la région.

Pour conclure, la méthode d'ensemble ARGONet développée aux États-Unis pourrait permettre de compléter les méthodes de surveillance traditionnelles en France. Nous avons également pu voir que les données de Twitter et les données climatiques avaient un apport pour la prévision, tout comme les données hospitalières et les données de Google. Les résultats en temps réel et les prévisions jusqu'à 2 semaines pourraient permettre d'anticiper les épidémies de grippe. Cela pourrait être un apport pour les prises de décision. En effet, grâce à la prévision au niveau régional, cela pourrait être possible d'organiser les flux de patients chez les médecins généralistes et dans les hôpitaux, notamment dans le service des urgences. Cependant, la grippe n'est pas la seule épidémie qui nécessite des moyens de surveillance en France. C'est pour cette raison qu'il est important de savoir si les méthodes que nous avons développées pour la grippe peuvent être généralisées à d'autres maladies.

5 Application à un autre cas d'usage

Figure 5.1: Chapitre 4 - Application à un autre cas d'usage

5.1 Problématique

Les études que nous avons présentées jusqu'ici se sont essentiellement basées sur la grippe. Cependant, ce n'est pas la seule épidémie nécessitant une surveillance. C'est pour cette raison que nous avons souhaité savoir si les modèles développés précédemment, pouvaient être généralisés à d'autres maladies telles que la gastro entérite. En effet, la gastro entérite est également un enjeu de santé publique majeur à travers le monde [39]. Même s'il s'agit généralement d'une maladie bénigne, elle a une morbidité et un poids économique élevés. La diarrhée est l'une des principales causes de mortalité chez les jeunes enfants. Chaque année, à travers le monde, la gastro entérite est responsable d'environ 2.5 millions de décès chez les enfants de moins de 5 ans [40]. En France, chaque année, il y a plus de 21 millions de cas déclarés [41]. Tout comme la grippe, pendant les périodes épidémiques, il y a une forte augmentation de visites chez les médecins généralistes et au service des urgences, ce qui pose problème pour l'organisation des systèmes de santé.

En France, c'est également le réseau Sentinelles qui est en charge de la surveillance des cas de diarrhée aiguë. Les méthodes d'estimation des taux d'incidence sont identiques

à celles de la grippe, le problème de délai dû au traitement et à l'agrégation des données est donc également présent. A l'heure actuelle, des études se sont intéressées aux caractéristiques des épidémies de gastro entérite, comme le virus en circulation, l'impact du vaccin, les changements climatiques, mais peu d'études se sont intéressées à la prévision.

5.2 Objectif

Notre objectif est donc d'évaluer la possibilité d'utiliser les données du web et les données massives hospitalières pour la prévision à 3 semaines des épidémies de gastro entérite. Nous souhaitons effectuer ces prévisions aux niveaux national et régional mais également aux niveaux du service des urgences et des hospitalisations, très impactés par l'augmentation du flux de patients.

5.3 Considérations méthodologiques

5.3.1 Les variables à prédire

Dans cette étude, nous avons 4 niveaux de prédiction :

Le niveau national : Nous avons obtenu les taux d'incidence hebdomadaires français (pour 100 000 habitants) de la diarrhée aiguë, sur le réseau Sentinelles. Nous avons récupéré ces données en avril 2018, pour la période allant de janvier 2008 à mars 2018.

Le niveau régional : De la même manière, nous avons obtenu les taux d'incidence hebdomadaires bretons de la diarrhée aiguë. Nous avons récupéré ces données en avril 2018, pour la période allant de janvier 2008 à mars 2018.

Le niveau des urgences : A partir de l'entrepôt de données cliniques eHOP, nous avons calculé le nombre de patients par semaine se rendant aux urgences pour la gastro entérite. Nous avons récupéré ces données en avril 2018, pour la période allant de janvier 2008 à mars 2018.

Le niveau des hospitalisations : De la même façon, à partir de l'entrepôt eHOP, nous avons calculé le nombre de patients par semaine hospitalisés après avoir été admis aux urgences pour la gastro entérite. Nous avons récupéré ces données en avril 2018, pour la période allant de janvier 2008 à mars 2018.

5.3.2 Les variables prédictives

Ici, nous avons 3 sources de données différentes :

Les données hospitalières : Tout comme pour la grippe, nous avons effectué des recherches sur les données textuelles grâce à des mots-clés en lien avec la gastro entérite : ses symptômes, ses virus et ses traitements. De cette manière, nous avons obtenu 19 signaux extraits de l'entrepôt eHOP. Nous avons également calculé, pour chaque niveau de prédiction, les 100 signaux les plus corrélés. Ces signaux ont été extraits d'une base de données contenant les séries temporelles construites à partir des données structurées de l'entrepôt de données eHOP. La corrélation a été calculée entre janvier 2008 et avril 2014. Au total, pour chaque niveau de prévisions, nous obtenons 119 variables explicatives. 19 variables au minimum sont communes à chaque niveau de prédiction. Ces données ont été extraites de l'entrepôt en avril 2018 pour la période allant de janvier 2008 à mars 2018.

Les données du web : Comme pour les études précédentes, grâce au service Google Correlate, nous avons obtenu la fréquence par semaine des 100 requêtes les plus corrélées pour chaque signal à prédire (national, régional, urgences et hospitalisations). Comme pour les données hospitalières, la corrélation a été calculée entre janvier 2008

et avril 2014 et nous avons récupéré les données pour la période allant de janvier 2008 à mars 2018.

Les données historiques : Pour chaque niveau de prévision, nous utilisons comme variables explicatives les 52 semaines précédant la semaine que l'on cherche à prédire.

5.3.3 Le modèle statistique

Dans cette étude, nous avons choisi d'utiliser le modèle de régression linéaire pénalisée Elastic Net afin de prendre en compte la corrélation et le grand nombre de variables explicatives. Le modèle s'écrit alors :

— Estimation en temps réel :

$y_t = \sum_{i=1}^{52} \alpha_i y_{t-i} + \sum_{j=1}^{100} \beta_j x_{jt} + \sum_{k=1}^{119} \gamma_k z_{kt} + \epsilon_t$

— Prévision à 1 semaine :

$y_{t+1} = \delta_1 \hat{y}_t + \sum_{i=1}^{52} \alpha_i y_{t-i} + \sum_{j=1}^{100} \beta_j x_{jt} + \sum_{k=1}^{119} \gamma_k z_{kt} + \epsilon_{t+1}$

— Prévision à 2 semaines :

$y_{t+2} = \delta_2 \hat{y}_{t+1} + \delta_1 \hat{y}_t + \sum_{i=1}^{52} \alpha_i y_{t-i} + \sum_{j=1}^{100} \beta_j x_{jt} + \sum_{k=1}^{119} \gamma_k z_{kt} + \epsilon_{t+2}$

— Prévision à 3 semaines :

$y_{t+3} = \delta_3 \hat{y}_{t+2} + \delta_2 \hat{y}_{t+1} + \delta_1 \hat{y}_t + \sum_{i=1}^{52} \alpha_i y_{i-1} + \sum_{j=1}^{100} \beta_j x_{jt} + \sum_{k=1}^{119} \gamma_k z_{kt} + \epsilon_{t+3}$

où y_t est le taux d'incidence de la diarrhée aiguë pour la semaine t, $\hat{y}_t$ est l'estimation obtenue par le modèle pour la semaine t, $\sum_{i=1}^{52} \alpha_i y_{t-i}$ les données historiques, $\sum_{j=1}^{100} \beta_j x_{jt}$ les données de Google, $\sum_{k=1}^{119} \gamma_k z_{kt}$ les données hospitalières, $\delta = (\delta_1, ..\delta_3), \alpha = (\alpha_1, ..\alpha_{52}), \beta = (\beta_1, ..\beta_{100}), \gamma = (\gamma_1, ..\gamma_{119})$ les paramètres de la régression et ϵ_t les résidus pouvant être modélisés par un processus ARIMA.

Ces modèles ont été utilisés pour les 4 niveaux à prédire. Cependant, pour les urgences et les hospitalisations, les données sont disponibles en temps réel, il n'est donc pas nécessaire de prédire y_t. Nos modèles ont été optimisés grâce à un jeu d'apprentissage de 6 ans et recalibrés chaque semaine pour incorporer les

nouvelles informations à disposition. Pour une semaine donnée, les paramètres $\alpha = (\alpha_1, ..\alpha_{52}), \beta = (\beta_1, ..\beta_{100}), \gamma = (\gamma_1, ..\gamma_{119})$ de la régression ont été estimés en minimisant le critère :

$$\sum_t \left(y_t - \sum_{i=1}^{52} \alpha_i y_{t-i} - \sum_{j=1}^{100} \beta_j x_{jt} - \sum_{k=1}^{119} \gamma_k z_{kt}\right)^2 + \lambda_\alpha ||\alpha||_1 + \eta_\alpha ||\alpha||_2^2 + \lambda_\beta ||\beta||_1 + \eta_\beta ||\beta||_2^2 + \lambda_\gamma ||\gamma||_1 + \eta_\gamma ||\gamma||_2^2$$

où $\lambda_\alpha, \lambda_\beta, \lambda_\gamma, \eta_\alpha, \eta_\beta, \eta_\gamma$ sont les hyper-paramètres de la régression que nous avons optimisé grâce à une validation croisée 10 blocs. Ces modèles ont été réalisés grâce au package caret du logiciel R et la méthode glmnet associée. Notre période de test débute au mois de mai 2014 et se termine au mois de février 2018.

5.3.4 Évaluation

Afin d'évaluer l'apport des sources de données externes pour la prévision des taux d'incidence de la gastro entérite, nous avons construit 2 modèles autorégressifs. Un premier modèle basé seulement sur la semaine précédant celle que l'on cherche à prédire (AR(1)) et un second modèle basé sur les 52 semaines précédentes (AR(52)). Les signaux de la gastro entérite étant assez bruités, nous avons également testé l'usage d'un lisseur sur les variables à prédire. Nous avons utilisé le lisseur de John Tukey. Ce lisseur fonctionne sur le principe de médianes mobiles, la valeur y_t lissée correspond à la médiane entre y_{t-1}, y_t et y_{t+1}. Ces modèles, ainsi que notre modèle Elastic Net ont été comparés pour chaque niveau de prédiction en calculant le PCC et l'EQM.

5.4 Article

5.4.1 Résumé de l'article

Objectif : La gastro entérite est un enjeu de santé publique majeur. Afin de réduire son impact et d'organiser des mesures sanitaires adaptées, les réseaux de surveillance

produisent des estimations des taux d'incidence. Cependant, ces estimations ont un délai de 1 à 3 semaines. Le principal objectif est de produire des estimations en temps réel et si possible à plus long terme. Pour cela, nous souhaitons évaluer une des méthodes déjà appliquée pour prédire les épidémies de grippe, en effectuant des prédictions jusqu'à 3 semaines à différentes échelles (nationale, régionale, urgences et hospitalisations).

Méthodes : Afin de prédire les épidémies de gastro entérite, nous avons appliqué un modèle Elastic Net, réutilisant les données du web, les données massives hospitalières et les données historiques. Une méthode de lissage a également été testée sur nos signaux à prédire. Nous avons comparé ce modèle à des modèles naïfs utilisant seulement les données historiques. Pour cela, nous avons utilisé 2 indicateurs le PCC et le MSE.

Résultats : Nous observons alors que, jusqu'à 3 semaines de prévision, en fonction de la zone que nous cherchons à prédire, les meilleurs indicateurs sont compris entre 0.73 et 0.87 pour le PCC et 0.533 et 0.257 pour le MSE. Pour l'échelle nationale et régionale, et pour le service des urgences, c'est le modèle Elastic Net avec l'utilisation d'un lisseur sur la variable à prédire, qui nous donne les meilleurs résultats. A l'échelle des hospitalisations, c'est également le modèle Elastic Net, mais sans l'usage du lisseur.

Conclusion : Nous avons pu montrer que le modèle Elastic Net développé pour la grippe, pouvait nous permettre d'obtenir des prévisions précises jusqu'à 3 semaines pour les épidémies de gastro entérite. De plus, l'usage de données externes comme les données du web ou les données hospitalières est important. Cette méthode pourrait venir compléter les méthodes de surveillance traditionnelles.

Big Data to Predict Gastroenteritis Outbreaks

Canelle Poirier[1,2,*], Guillaume Bouzillé[1,2,3], Valérie Bertaud[1,2], Marc Cuggia [1,2,3], Mauricio Santillana[4,5], Audrey Lavenu[6,7,8]

[1]INSERM, U1099, Rennes, F-35000, France;
[2]Université de Rennes 1, LTSI, Rennes, F-35000, France;
[3]CHU Rennes, Centre de Données Cliniques, Rennes, F-35000, France;
[4]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA;
[5]Department of Pediatrics, Harvard Medical School, Boston, MA, USA;
[6]Université de Rennes 1, Faculté de médecine, Rennes, F-35043, France;
[7]INSERM CIC 1414, Université de Rennes 1, Rennes, F-35043, France;
[8]IRMAR, Institut de Recherche Mathématique de Rennes, UMR CNRS 6625, Rennes, France
[*]Laboratoire d'informatique médicale, 2 rue Henri le Guilloux, F-35033 Rennes, France , canelle.poirier@univ-rennes1.fr

Abstract

Background Accute gastroenteritis (AG) is a major public health issue. To reduce impact and to organize adapted sanitary responses, traditional surveillance produce estimates but with 1- to 3-week delay. The main challenge is to produce near real-time and longer term estimates.
Objective For the flu, alternative modeling strategies have been proposed to avoid this delay. We assess one of these methods to predict AG up to 3 weeks at different levels.
Methods We used Web data, Hospital data and Historical data in combination with an Elastic Net model and a smoother.
Results We observe that up to three weeks forecasts, we still obtain PCC between 0.73 and 0.87 and MSE between 0.533 and 0.257 depending to the level of prediction.
Conclusions We found that external data sources in combination with Elastic Net give accurate estimates. It could complement traditional surveillance.

Introduction

Accute gastroenteritis (AG) is a major public health issue worldwide.[1] It is commonly defined as diarrhoea (three or more loose stools) or vomiting in the past 24 hours.[2] Even if it is generally a mild disease, its morbidity and economic burden are high.[3] Indeed, diarrhoea is one of the principal causes of morbidity and mortality among young people. It causes up to 2.5 million deaths in children under 5 years per year worldwide.[4] In France, there are more than 21 million episodes of AG each year.[5] There are AG all year-round but there is a winter peak, usually between December and March, which is mainly due to norovirus and rotavirus. During AG peaks, the large increase of visits to general practitioners and to emergency departments causes healthcare system disruption.
To reduce its impact and to help organizing adapted sanitary responses, it is necessary to monitor AG activity. In US, participating laboratories report weekly to the Centers for Disease Control and Prevention (CDC), the total number of rotavirus tests performed, and the number of those tests that were positive as well as waterborne, foodborne, and enteric disease outbreaks of all etiologies.[6,7] In France, each week, volunteer outpatient healthcare providers report all accute diarrhoea cases seen during consultation. An estimation of the incidence rate is then computed, at the national or regional scale, by taking into account the number of sentinel physicians and the medical density of the area of interest. Surveillance networks produce estimates of incidence rates, but with a one to three-week delay due to the time needed for data processing and aggregation. This time lag is an issue for public health decision-making. Therefore, there is a growing interest in finding ways to avoid this information gap.[7,11]

Related work on real-time disease surveillance. In the past decade, multiple research teams have shown that it is possible to produce accurate and reliable disease activity estimates ahead of traditional healthcare-based systems. However, in case of AG, several studies are looking at the clinical characteristics, impact of the vaccine, the climate change or the type of virus responsible for the epidemics[8–10,13–15], but very few studies are interested in forecasting. Nevertheless, some studies assessed correlation between data sources as drug reimbursement data, or emergency department visits and AG general practitioner visits.[3,12] Moreover, other studies showed a significant correlation between search query trends on

Internet and AG incidence rates.[16,17]
In the case of influenza epidemics, alternative modeling strategies have been proposed to avoid the delay of traditional disease surveillance networks, in particular by developing statistical models leveraging Internet search data that is available in near real-time. A well-known implementation on the use of internet data for detecting influenza epidemics is the now-discontinued Google Flu Trends,[18] a web service operated by Google. They showed that internet users' searches are strongly correlated with influenza epidemics. But internet users' searches are strongly correlated with other diseases including AG.[17,19] Some authors updated the Google Flu Trends algorithm to predict influenza incidence rates in real time by including data from other sources, such as historical flu information for instance.[11]
Internet is not the only data source that can be used to produce information in real time. With the widespread adoption of electronic health records (EHRs), a digital version of a patient's paper chart, hospitals also produce a huge amount of data during hospitalization. This is commonly called Hospital Big Data (HBD). Bouzillé et al.[27] showed that HBD are strongly correlated to influenza incidence rates. Some authors[20,26] proposed statistical models using HBD to predict influenza incidence rates in real time.

Our contribution. We evaluate the feasibility of using Internet search query trends data and HBD to predict AG incidence rates, in near real time, at the national and regional scales and for up to 3 weeks. Moreover, hospitals and particularly emergency department are impacted during AG peak, for this reason we wanted to predict AG incidence rates for people going to emergency departments and those hospitalized.

Methods

Variables to be predicted

National level We obtained the national (Metropolitan France) accute diarrhea weekly incidence rates (per 100000 inhabitants) from the French Sentinel network (websenti.u707.jussieu.fr/sentiweb) from January 2008 to March 2018. We retrieved these data in April 2018.

Regional level In the same way, We obtained the regional (Brittany region) accute diarrhea incidence rates (per 100000 inhabitants) from the French Sentinel network (websenti.u707.jussieu.fr/sentiweb) from January 2008 to March 2018. We choose the Brittany region because HBD data that we used are from this region. We retrieved these data in April 2018.

Emergency level We had access to data from the clinical data warehouse (CDW) of Rennes University Hospital (France). This CDW, called eHOP, integrates structured (laboratory test results, prescriptions, ICD-10 diagnoses) and unstructured (discharge letter, pathology reports, operative reports) patients' data. It includes data from 1.2 million inpatients and outpatients and 45 million documents that correspond to 510 million structured elements. eHOP consists of a powerful search engine system that can identify patients with specific criteria by querying unstructured data with keywords, or structured data with querying codes based on terminologies. We calculated from eHOP data, the number of patients per week coming to the emergency for AG. We retrieved this signal from January 2008 to March 2018 in April 2018.

Hospital level In the same way, we calculated from eHOP data, the number of patients per week hospitalized after they have been admitted to the emergency for AG. We retrieved this signal from January 2008 to March 2018 in April 2018.

Predictive variables

Web Data We obtained the frequency per week of the 100 most correlated internet queries (if correlation ≥ 0.60) by French users from Google Correlate (https://www.google.com/trends/correlate). Because our test period is from May 2014 to February 2018, we inserted each signal to be predicted from January 2008 to April 2014 into Google Correlate. In this way, we obtain the most correlated internet queries that can be different depending on the signal to be predicted. The internet queries obtained correspond to requests performed by French users at the national level. We retrieved Google Correlate data for the period going from from January 2008 to March 2018.

Clinical data The first approach to obtain eHOP data connected with AG was to perform different full-text queries to retrieve patients who had at least one document in their EHR that matched the search criteria related with gastroenteritis, symptoms, virus or treatments. For each query, the eHOP search engine returned all documents containing the chosen keywords (often, several documents for one patient and one stay). For query aggregation, we kept the oldest document for

one patient and one stay and then calculated, for each week, the number of stays with at least one document mentioning the keyword contained in the query. In this way, we obtained 19 variables from the CDW eHOP.

From the CDW eHOP, we built a database containing the time series constructed from the structured data. In all, we have 1 335 347 time series. As Google Correlate, the Pearson correlation between each signal to be predicted and the time series from the database was calculated. In this way, for each signal to be predicted, the second approach was to retrieve the 100 most correlated signals to AG signal. Because our test period is from May 2014 to February 2018, we calculated the correlation between January 2008 and April 2014.
In all, we had 119 variables that can be different depending on the signal to be predicted excepted manual queries common to all. We retrieved retrospective data for the period going from January 2008 to March 2018 in April 2018.
This study was approved by the local Ethics Committee of Rennes Academic Hospital (approval number 16.69).

Historical Data For each signal to be predicted, in order to have a history of AG incidence rates in our models, we used the 52 weeks preceding the week we were predicting.

Statistical model

In order to minimize the negative effects of using a large number of input variables, potentially including redundant (or highly co-linear) information, we used Elastic Net, a regularized multi-variate regression methodology, capable of identifying parsimonious models.[22] With datasets that may have up to 271 potentially correlated variables, we performed the Elastic Net regression analysis using the R package caret and the associated function fit with the method glmnet.[21,23] We fixed a coefficient alpha equal to 0.5. We optimized the shrinkage parameter lambda via a 10-fold cross-validation.
As for residuals of a linear regression, we checked means, variances and correlation to know if we obtained white noises. To test the stationarity and whiteness, we used Dickey Fuller's and Box-Pierce's tests available from the R packages tseries and stats.[23,24] If we didn't have white noises, we fitted residuals with a model of temporal series, called ARIMA.
The statistical formulation of our models is :

- Prevision for week t (real time) :
  $y_t = \sum_{i=1}^{52} \alpha_i y_{t-i} + \sum_{j=1}^{100} \beta_j x_{jt} + \sum_{k=1}^{119} \gamma_k z_{kt} + \epsilon_t$
- Prevision for week $t+1$:
  $y_{t+1} = \delta_1 \hat{y}_t + \sum_{i=1}^{52} \alpha_i y_{t-i} + \sum_{j=1}^{100} \beta_j x_{jt} + \sum_{k=1}^{119} \gamma_k z_{kt} + \epsilon_{t+1}$
- Prevision for week $t+2$:
  $y_{t+2} = \delta_2 \hat{y}_{t+1} + \delta_1 \hat{y}_t + \sum_{i=1}^{52} \alpha_i y_{t-i} + \sum_{j=1}^{100} \beta_j x_{jt} + \sum_{k=1}^{119} \gamma_k z_{kt} + \epsilon_{t+2}$
- Prevision for week $t+3$:
  $y_{t+3} = \delta_3 \hat{y}_{t+2} + \delta_2 \hat{y}_{t+1} + \delta_1 \hat{y}_t + \sum_{i=1}^{52} \alpha_i y_{i-1} + \sum_{j=1}^{100} \beta_j x_{jt} + \sum_{k=1}^{119} \gamma_k z_{kt} + \epsilon_{t+3}$
  - y_t : Accute diarrhea incidence rate at time t
  - $\hat{y_t}$: Estimate obtained by the model for the week t.
  - y_{t-i} for i from 1 to 52 : Historical variables
  - x_{jt} for j from 1 to 100 : Google data
  - z_{kt} for k from 1 to 119 : Hospital data
  - $\delta = (\delta_1, ..\delta_3), \alpha = (\alpha_1, ..\alpha_{52}), \beta = (\beta_1, ..\beta_{100}), \gamma = (\gamma_1, ..\gamma_{119})$: Regression parameters
  - ϵ_t : Residuals

These models are used for all variables to be predicted. Nevertheless, for emergency level and hospitalization level we didn't have to predict the week t because we have access to hospital data in near real time. In this way, to predict up to 3 weeks for emergency and hospitalization levels, we used the real value y_t and not the predicted value $\hat{y_t}$.
Our test period started on May 2014 and finished on February 2018. We fitted our model using a training dataset that corresponded to the data for the previous six years. In a given week, we had to find parameters $\alpha = (\alpha_1, ..\alpha_{52}), \beta = (\beta_1, ..\beta_{100}), \gamma = (\gamma_1, ..\gamma_{119})$ that minimize :

$$\sum_t \left(y_t - \sum_{i=1}^{52} \alpha_i y_{t-i} - \sum_{j=1}^{100} \beta_j x_{jt} - \sum_{k=1}^{119} \gamma_k z_{kt}\right)^2 + \lambda_\alpha ||\alpha||_1 + \eta_\alpha ||\alpha||_2^2 + \lambda_\beta ||\beta||_1 + \eta_\beta ||\beta||_2^2 + \lambda_\gamma ||\gamma||_1 + \eta_\gamma ||\gamma||_2^2$$

where $\lambda_\alpha, \lambda_\beta, \lambda_\gamma, \eta_\alpha, \eta_\beta, \eta_\gamma$ were hyper-parameters of the Elastic Net regression. We used a 10 blocks cross-validation to optimize these hyper-parameters. For previsions for week $t+1$ to week $t+3$ we had to estimate $\delta = (\delta_1, ..\delta_3)$ which implies 2 additional hyper-parameters : $\lambda_\delta, \eta_\delta$. The model was dynamically recalibrated every week to incorporate new information.

Evaluation

We compared our estimates with the AG signals from the Sentinel network and CDW eHOP (Variables to be predicted) by calculating the mean squared error (MSE) and the Pearson's correlation coefficient (PCC) over our entire test period. Indeed, clinicians want to know when an epidemic starts and finishes as well as its amplitude and severity. PCC is an indicator for the trend of the signal, if the end and the start of epidemics are well estimated, and MSE is an indicator for peaks' amplitude. To assess the contribution of hospital data and web data to predict AG incidence rates, we built a naive model (AR(52)) based only on the 52 previous AG incidence rates as well as a second naive model based only on the previous week. This model was built at national scale, regional scale and also at emergency and hospitalization scales. The signals being quite noisy, we also smoothed variables to be predicted with the Tukey's smoother. We compared estimates obtained with smoothed variables to the true unsmoothed signals from the Sentinel network and CDW eHOP.

Results

Here, we show the Pearson's correlation coefficient (PCC) and mean square error (MSE) obtained at the different levels of prediction : National, Regional, Emergency and Hospitalization with the different models.All results are presented in table 1. We can see that in most cases, external data sources allow to give better results in term of PCC and MSE as well as the use of the Tukey's smoother excepted for the hospitalization level.

	real-time		forecast 1 week		forecast 2 week		forecast 3 week	
	PCC	MSE	PCC	MSE	PCC	MSE	PCC	MSE
National								
Elastic Net	0.91	0.189	**0.91**	**0.187**	0.87	0.260	0.84	0.314
Elastic Net+Smoother	**0.95**	**0.107**	0.90	0.196	**0.89**	**0.227**	**0.87**	**0.257**
AR(1)	0.93	0.131	0.86	0.274	0.80	0.404	0.72	0.549
AR(52)	0.94	0.114	0.90	0.202	0.88	0.246	0.86	0.284
Region								
Elastic Net	0.72	0.556	0.695	0.608	0.70	0.598	0.66	0.683
Elastic Net+Smoother	**0.80**	**0.402**	**0.81**	**0.381**	**0.77**	**0.453**	**0.75**	**0.505**
AR(1)	0.69	0.618	0.64	0.722	0.54	0.911	0.53	0.933
AR(52)	0.73	0.528	0.69	0.618	0.65	0.695	0.65	0.704
Emergency								
Elastic Net			0.78	0.438	0.77	0.464	0.78	0.445
Elastic Net+Smoother			**0.86**	**0.277**	**0.86**	**0.273**	**0.86**	**0.280**
AR(1)			0.65	0.701	0.60	0.800	0.52	0.963
AR(52)			0.77	0.454	0.72	0.561	0.70	0.602
Hospitalization								
Elastic Net			**0.75**	**0.504**	**0.73**	**0.535**	**0.73**	**0.533**
Elastic Net+Smoother			0.67	0.668	0.60	0.795	0.66	0.668
AR(1)			0.54	0.911	0.57	0.854	0.46	1.067
AR(52)			0.66	0.669	0.66	0.681	0.65	0.690

Table 1. PCC and MSE obtained over all prediction period (May 2014 to March 2018) for all levels and models and predictions up to three weeks.

National analysis

For the current week's prediction, PCC ranges from 0.95 to 0.91 and MSE ranges from 0.107 to 0.189. Elastic Net model with the use of Tukey's smoother on the variable to explain gives the best results. For the one-week prediction, PCC ranges from 0.91 to 0.86 and MSE ranges from 0.187 to 0.274. Elastic Net model without smoother gives the best results. For the two-week prediction, PCC ranges from 0.89 to 0.80 and MSE ranges from 0.227 to 0.404. Elastic Net model with

Tukey's smoother gives the best results. In the same way, for the three-week prediction, PCC ranges from 0.87 to 0.72 and PCC ranges from 0.257 to 0.549 and Elastic Net model with Tukey's smoother gives the best results. Figure 1 illustrates predictions obtained at the national scale with the different models

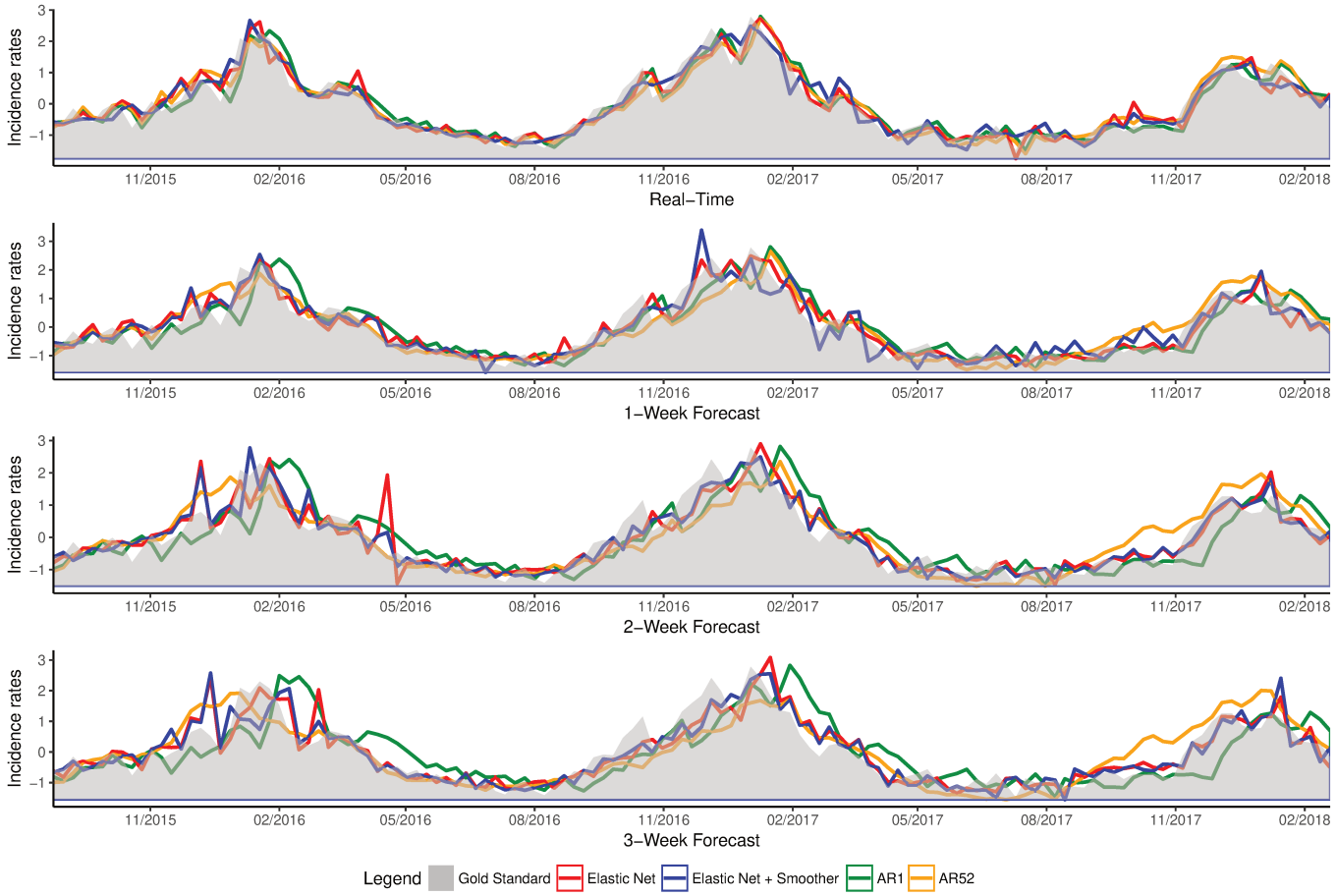


Figure 1. National level. Predictions up to three weeks obtained at national level with all models

Regional analysis

For the current week's prediction, PCC ranges from 0.80 to 0.69 and MSE ranges from 0.402 to 0.618. Elastic Net model with Tukey's smoother gives the best results. In the same way, for the one-week prediction, PCC ranges from 0.81 to 0.64 and MSE ranges from 0.381 to 0.772 and Elastic Net model with Tukey's smoother gives the best results. For the two-week prediction, PCC ranges from 0.77 to 0.54 and MSE ranges from 0.453 to 0.911 and Elastic Net model with Tukey's smoother gives the best results. For the three-week prediction, PCC ranges from 0.75 to 0.53 and PCC ranges from 0.555 to 0.933 and Elastic Net model with Tukey's smoother gives the best results. Figure 2 illustrates predictions obtained for the regional scale.

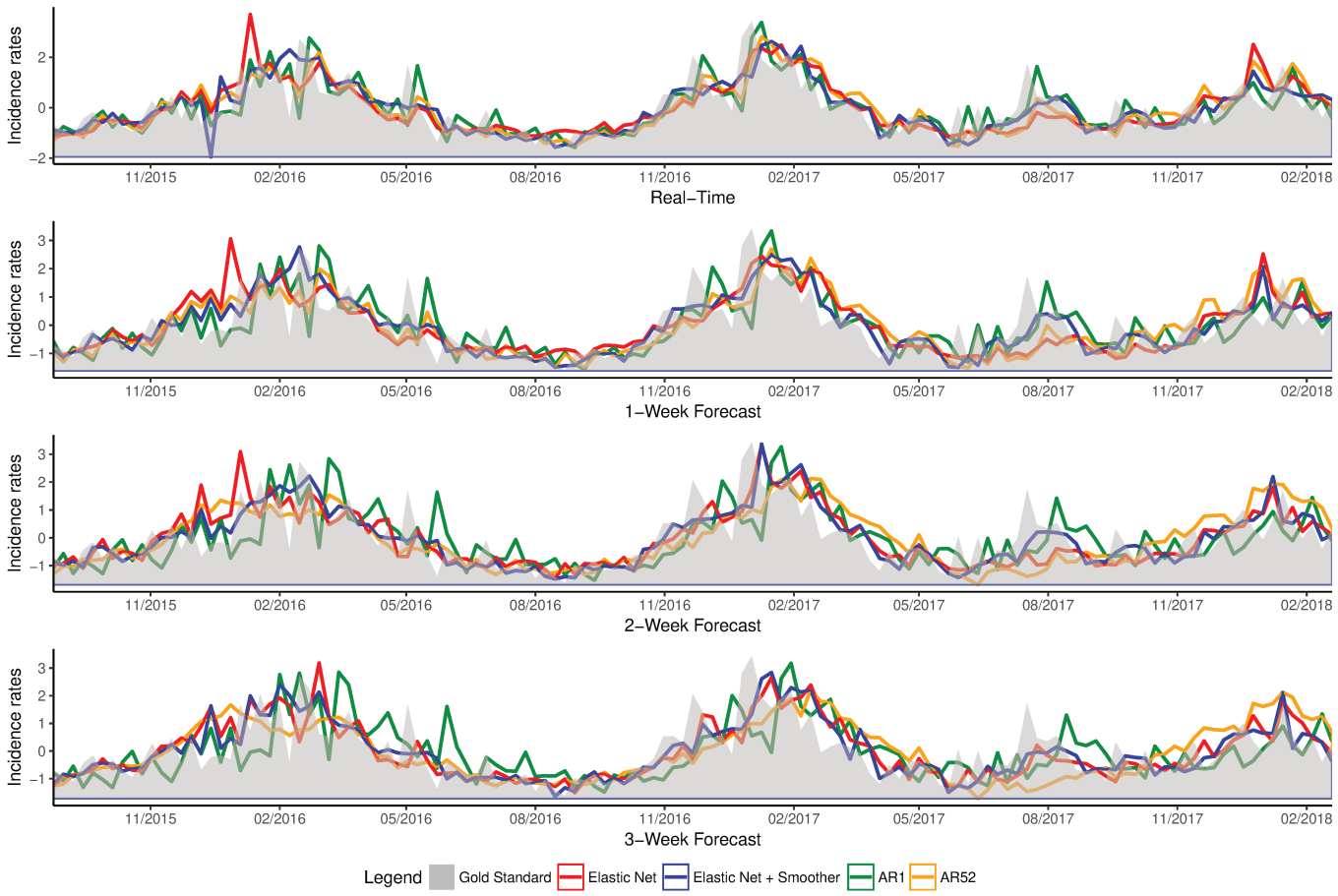


Figure 2. Regional level. Predictions up to three weeks obtained at regional level with all models

Emergency analysis

For the one-week prediction, PCC ranges from 0.86 to 0.65 and MSE ranges from 0.277 to 0.701 and Elastic Net model with Tukey's smoother gives the best results. In the same way, for the two-week prediction, PCC ranges from 0.86 to 0.60 and MSE ranges from 0.273 to 0.800 and Elastic Net model with Tukey's smoother gives the best results. For the three-week prediction, PCC ranges from 0.86 to 0.52 and PCC ranges from 0.280 to 0.963 and Elastic Net model with Tukey's smoother gives the best results. Figure 3 illustrates predictions obtained for the emergency scale with Elastic Net model using Tukey's smoother and the naive model.

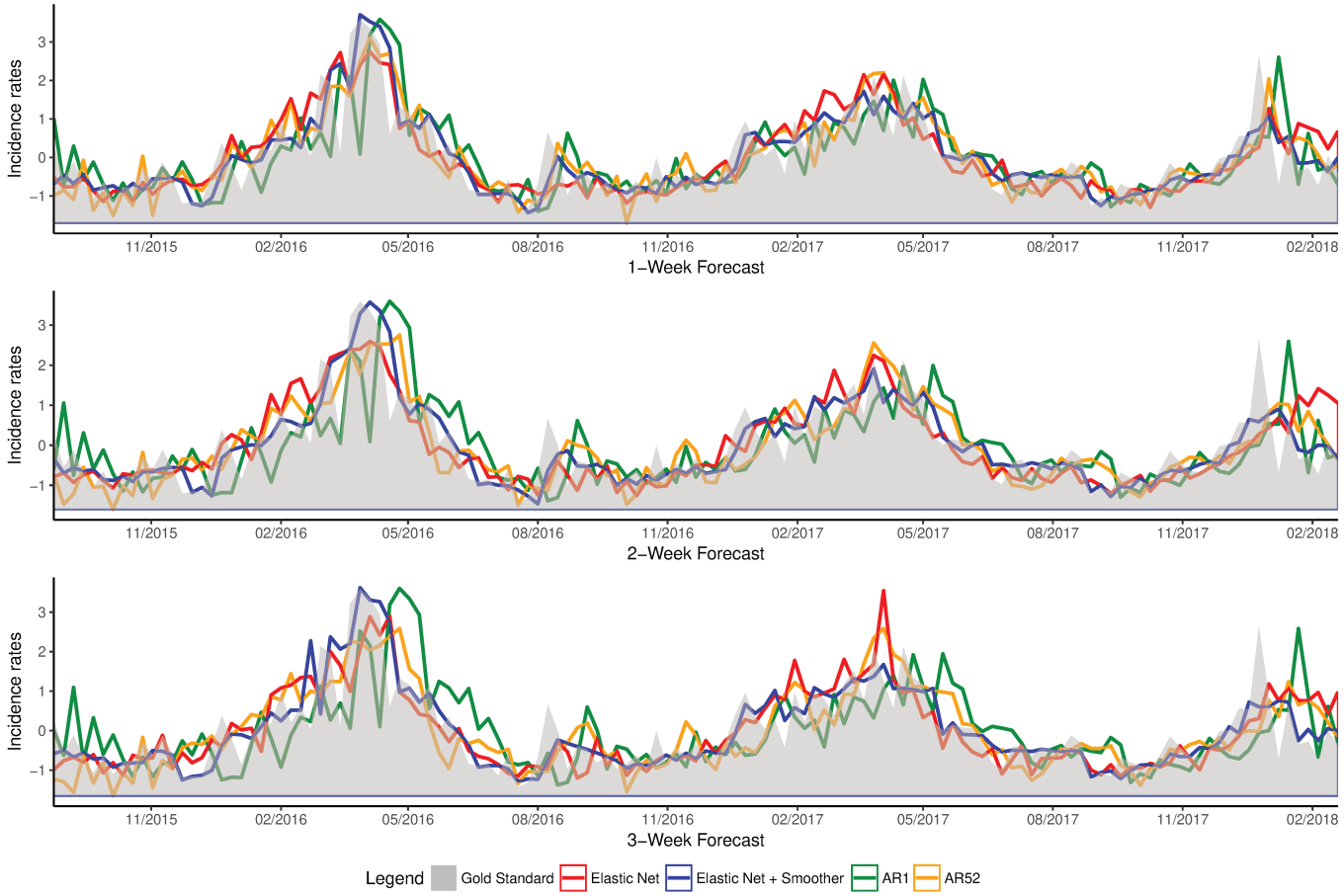


Figure 3. Emergency level. Predictions up to three weeks obtained at emergency level with all models

Hospitalization analysis

For the one-week prediction, PCC ranges from 0.75 to 0.54 and MSE ranges from 0.911 to 0.504 and Elastic Net model without Tukey's smoother gives the best results. In the same way, for the two-week prediction, PCC ranges from 0.73 to 0.57 and MSE ranges from 0.535 to 0.854 and Elastic Net model without Tukey's smoother gives the best results. For the three-week prediction, PCC ranges from 0.73 to 0.46 and PCC ranges from 0.533 to 1.067 and Elastic Net model without Tukey's smoother gives the best results. Figure 4 illustrates predictions obtained for the emergency scale with Elastic Net model without Tukey's smoother and the naive model.

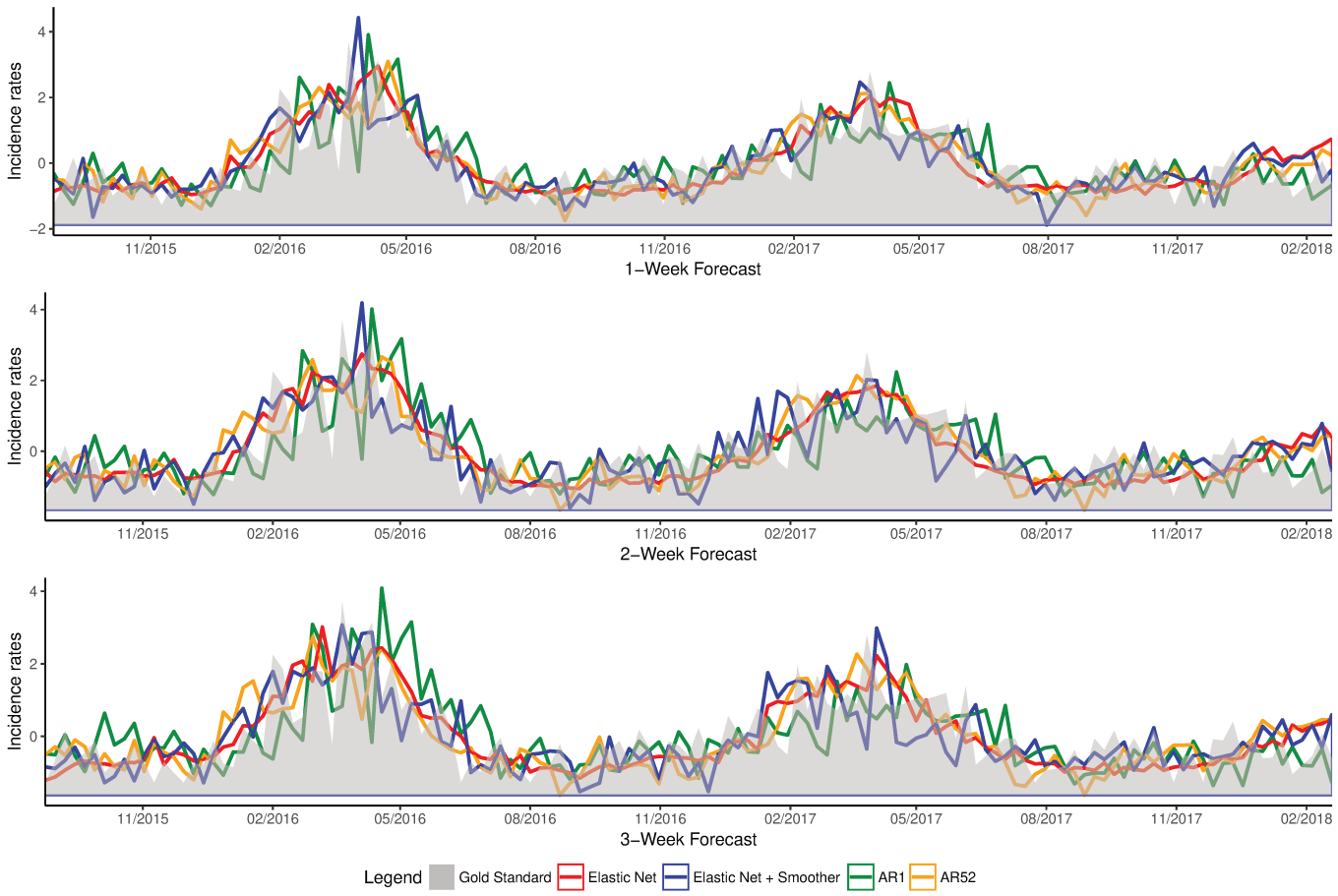


Figure 4. Hospitalization level. Predictions up to three weeks obtained at hospitalization level with all models

Discussion

Here we show that the use of web data and HBD in combination with historical data is a good way to predict AG at national, regional, emergency and hospitalization levels. Indeed, if we compare the naive model AR(52) with the Elastic Net model without Tukey's smoother, we can see that the results are better in term of PCC and MSE excepted at the national level where the results are similar. One possible cause could be that the national signal is less noisy than the other signal. It may therefore be easier to obtain more information from historical data than for other prediction levels. In the same way, if we compare the naive model AR(52) to the naive model AR(1), we can observe that the model AR(52) still gives better results. It is thus important to integrate historical data to improve our estimates.

In term of statistical methods used, the signals for the AG being noisier than those for influenza on which the methods have been developed, we made the choice to apply a smoother. In all cases, excepted at the hospitalization level, we get better results if the signal to be predicted is smoothed. For the hospitalization level, data are from the CDW eHOP from Rennes University Hospital and in terms of volume we do not have a large number of patients hospitalized for AG. This implies that the difference between the number of cases during the epidemic peak and the non-epidemic period is less important than for other prediction levels. In this case, using a smoother can cause too much information loss. The Tukey's smoother has a real contribution but it could be interesting to test other smoother to know if one of them could give better results.

Finally, even if indicators decrease with the number of weeks to predict, with the best model, we observe that up to three weeks forecasts, we obtain PCC between 0.73 and 0.87 and MSE between 0.533 and 0.257 depending to the level of prediction. These results compared to the results obtained to predict flu outbreaks at united states level[25], seem good, which is why the Elastic Net model with web data, hospital data and historical data could complement traditional surveillance methods.

However, one major limitation of our work is to use previous estimates to predict the next weeks. Indeed, there is a risk to multiply errors. Indeed, if the prediction at time t is wrong, prediction time $t+1$ could be worse. It could be interesting to test an other approach, with good results for influenza, for example an ensemble method combining the power of different statistical models.[28]

Conclusion

Regarding the results, we show that hospital data, web data and historical data have a real contribution to predict AG outbreaks. The use of these external data sources with Elastic Net model and also a smoother could be a good way to complete traditionnal surveillance. Indeed, it could be a way to organize and to reduce the impact of AG peak in particular in hospitals by anticipating epidemics up to 3 weeks.
Nevertheless, maybe we could improve the accuracy of estimates with an other smoother or a method avoiding to use predictions made by previous models.

Acknowledgements

We would like to thank the French National Research Agency (ANR), for funding this work inside the INSHARE (INtegrating and Sharing Health dAta for Research) project (grant no. ANR-15-CE19-0024). We also thank the French Sentinel network and Google search engine for making their data publicly available.

References

1. Farthing MJ. Diarrhoea: a significant worldwide problem. Int J Antimicrob Agents. 2000;14(1):65?9.
2. Majowicz SE, Hall G, Scallan E, Adak GK, Gauci C, Jones TF, et al. A common, symptom-based case definition for gastroenteritis. Epidemiol Infect. 2008;136(7):886?94.
3. Rivière M, Baroux N, Bousquet V, Ambert-Balay K, Beaudeau P, Jourdan-Da Silva N, et al. Secular trends in incidence of acute gastroenteritis in general practice, France, 1991 to 2015. Euro Surveill. 2017 22(50).
4. Kosek M, Bern C, Guerrant RL. The global burden of diarrhoeal disease, as estimated from studies published between 1992 and 2000. Bull World Health Organ. 2003;81(3):197?204.
5. Cauteren DV, Valk HD, Vaux S, Strat YL, Vaillant V. Burden of acute gastroenteritis and healthcare-seeking behaviour in France: a population-based study. Epidemiology Infection. 2012;140(4):697?705.
6. Laboratory-Based Surveillance for Rotavirus – United States, July 1997-June 1998 [Internet]. Disponible sur: https://www.cdc.gov/mmwr/preview/mmwrhtml/00055737.htm
7. Shah MP. Near Real-Time Surveillance of U.S. Norovirus Outbreaks by the Norovirus Sentinel Testing and Tracking Network - United States, August 2009-July 2015. MMWR Morb Mortal Wkly Rep. 2017;66.
8. Arena C, Amoros JP, Vaillant V, Ambert-Balay K, Chikhi-Brachet R, Jourdan-Da Silva N, et al. Acute diarrhea in adults consulting a general practitioner in France during winter: incidence, clinical characteristics, management and risk factors. BMC Infect Dis. 2014 14.
9. Charles MD, Holman RC, Curns AT, Parashar UD, Glass RI, Bresee JS. Hospitalizations associated with rotavirus gastroenteritis in the United States, 1993-2002. Pediatr Infect Dis J. 2006;25(6):489?93.
10. . Al AJH et. Acute Gastroenteritis Surveillance through the National Outbreak Reporting System, United States - Volume 19, Number 8, August 2013 - Emerging Infectious Disease journal - CDC.
11. Yang S, Santillana M, Kou SC. Accurate estimation of influenza epidemics using Google search data via ARGO. Proc Natl Acad Sci USA. 2015;112(47):14473?8.
12. Kirian ML, Weintraub JM. Prediction of gastrointestinal disease with over-the-counter diarrheal remedy sales records in the San Francisco Bay Area. BMC Med Inform Decis Mak. 2010;10:39.
13. Rohayem J. Norovirus seasonality and the potential impact of climate change. Clin Microbiol Infect. 2009;15(6):524?7.
14. Amador JJ, Vicari A, Turcios-Ruiz RM, Melendez D AC, Malek M, Michel F, et al. Outbreak of rotavirus gastroenteritis with high mortality, Nicaragua, 2005. Rev Panam Salud Publica. 2008;23(4):277?84.

15. Greer AL, Drews SJ, Fisman DN. Why « winter » vomiting disease? Seasonality, hydrology, and Norovirus epidemiology in Toronto, Canada. Ecohealth. 2009;6(2):192?9.

16. Shah MP, Lopman BA, Tate JE, Harris J, Esparza-Aguilar M, Sanchez-Uribe E, et al. Use of Internet Search Data to Monitor Rotavirus Vaccine Impact in the United States, United Kingdom, and Mexico. J Pediatric Infect Dis Soc. 2018;

17. Pelat C, Turbelin C, Bar-Hen A, Flahault A, Valleron A-J. More Diseases Tracked by Using Google Trends. Emerg Infect Dis. 2009;15(8):1327?8.

18. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457(7232):1012?4.

19. Carneiro HA, Mylonakis E. Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks. Clin Infect Dis. 2009;49(10):1557?64.

20. Santillana M, Nguyen AT, Louie T, et al. Cloud-based Electronic Health Records for Real-time, Region-specific Influenza Surveillance. Scientific Reports 2016;6:25732. doi:10.1038/srep25732

21. Kuhn M. caret: Classification and Regression Training. R package version 6.0-77.9000. https://github.com/topepo/caret/

22. Zou H, Hastie T. Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society, Series B 2005;67:301–320.

23. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: : R Foundation for Statistical Computing 2015. https://www.R-project.org/

24. Trapletti A, Hornik K. tseries: Time Series Analysis and Computational Finance. 2015. http://CRAN.R-project.org/package=tseries

25. Yang S, Santillana M, Brownstein JS, Gray J, Richardson S, Kou SC. Using electronic health records and Internet search information for accurate influenza forecasting. BMC Infectious Diseases. 2017

26. Poirier C, Lavenu A, Bertaud V, Campillo-Gimenez B, Chazard E, Cuggia M, et al. Real Time Influenza Monitoring Using Hospital Big Data in Combination with Machine Learning Methods: Comparison Study. JMIR Public Health Surveill. 2018;4(4):e11361.

27. Bouzillé G, Poirier C, Campillo-Gimenez B, Aubert M-L, Chabot M, Chazard E, et al. Leveraging hospital big data to monitor flu epidemics. Computer Methods and Programs in Biomedicine. 2018;154:153-60.

28. Lu FS, Hattab MW, Clemente L, Santillana M. Improved state-level influenza activity nowcasting in the United States leveraging Internet-based data sources and network approaches via ARGONet. bioRxiv. 2018;344580.

5.5 Discussion des principaux résultats

Ici, nous avons pu montrer que l'utilisation d'un modèle Elastic Net réutilisant les données du web et les données hospitalières était un bon moyen de prédire les épidémies de gastro entérite aux niveaux national, régional mais aussi aux niveaux des urgences et des hospitalisations. En effet, lorsque nous comparons le modèle autorégressif d'ordre 52 et le modèle Elastic Net utilisant toutes les sources de données externes, nous pouvons voir que dans tous les cas, au niveau régional et aux niveaux des urgences et des hospitalisations, nous obtenons des estimations plus précises. Ce n'est pas toujours le cas au niveau national, ce qui peut être dû au fait que nous avons un signal beaucoup moins bruité et ainsi que les données historiques apportent plus d'information.

Les signaux de la gastro entérite étant plus bruités que les signaux des syndromes grippaux sur lesquels la méthode a été développée, nous avons choisi de tester l'usage d'un lisseur sur nos variables à prédire. Dans tous les cas, excepté au niveau des hospitalisations, l'usage du lisseur améliore nos estimations. Au niveau des hospitalisations, les données ont été extraites de l'entrepôt eHOP du CHU de Rennes. Le volume de patients hospitalisés après qu'ils soient passés aux urgences pour gastro entérite n'est pas très important. Cela implique que la différence du nombre de cas pendant les périodes épidémiques et non épidémiques est moins importante que pour les autres niveaux de prédiction. L'usage d'un lisseur n'est donc pas forcément adapté dans ce cas là et peut causer la perte de trop d'informations. Pour tous les autres niveaux, nous avons vu que le lisseur de Tukey avait une réelle contribution, mais il pourrait être intéressant de tester une autre méthode.
Une limitation importante de notre travail est d'utiliser les estimations obtenues par notre modèle pour prédire les semaines suivantes. Ceci est un risque de multiplication des erreurs si la prévision effectuée pour la semaine t ou $t+1$ est fausse. La méthode d'ensemble ARGONet présentée précédemment a été développée en parallèle de

cette étude. Il serait important de voir si cette méthode pourrait nous permettre d'améliorer nos estimations et ainsi éviter ce risque de multiplication. De plus, nous pourrions tester l'usage des données de Twitter et des données climatiques.

Pour conclure, au vu des résultats, les données du web, les données hospitalières et les données historiques ont une réelle contribution pour prédire les épidémies de gastro entérite. L'utilisation de ces données externes avec un modèle Elastic Net ainsi que l'usage d'un lisseur pourraient permettre de compléter les moyens de surveillance traditionnels. Cela pourrait être une alternative utilisée pour organiser et réduire l'impact des épidémies, en particulier dans les hôpitaux, en anticipant les épidémies jusqu'à 3 semaines. Néanmoins, nous pourrions peut être améliorer la précision de nos estimations en utilisant d'autres sources de données ou un autre modèle statistique tel que la méthode d'ensemble ARGONet.

6 Discussion

Ces dernières années, l'exploitation des données massives à l'aide de nouvelles méthodes de data mining a ouvert de nouvelles perspectives dans de nombreux domaines. C'est le cas du domaine de la santé, où dorénavant, un grand volume de données est produit, notamment grâce à l'adoption du dossier patient électronique. Le Big Data en santé est très important dans le domaine de la recherche, c'est un véritable enjeu pour la surveillance épidémique, la prévention des maladies, ou encore la pharmacovigilance [42]. Dans ce travail de thèse, l'objectif était de montrer qu'à l'aide de modèles statistiques adaptés, la réutilisation des données massives hospitalières pouvait être un réel apport pour la surveillance syndromique.

Dans une étape préliminaire, nous avons cherché à exploiter les données présentes dans l'entrepôt de données biomédicales du CHU de Rennes et nous avons voulu montrer que ces données pouvaient être pertinentes pour la surveillance syndromique. Pour cela, nous avons choisi comme cas d'usage la grippe, qui est un enjeu de santé publique majeur. Nous avons réussi à montrer que les signaux extraits de l'entrepôt clinique, en lien avec le syndrome grippal, étaient très corrélés au signal estimé par les moyens de surveillance traditionnels.

La seconde étape portait sur le développement d'un modèle statistique, permettant la prévision en temps réel, afin d'éviter le délai engendré par les réseaux de surveillance et en particulier le réseau Sentinelles. Trois modèles statistiques ont été comparés, les forêts aléatoires, les machines à vecteurs supports et un modèle de régression linéaire pénalisée Elastic Net. Nous avons également testé l'apport d'une modélisation ARIMA des résidus. Ces modèles permettaient la réutilisation des données massives hospitalières ou des données du web. Dans cette étude, nous avons montré que 2 modèles avaient des performances comparables : le modèle SVM et le modèle Elastic Net et que les données hospitalières permettaient des estimations en temps réel plus

précises que les données du web. Cependant, une des limites de cette étude est qu'elle ne teste pas l'apport des données hospitalières et des données du web utilisées en simultané. De plus, nous effectuons des estimations en temps réel à l'échelle nationale et à l'échelle de la Bretagne, mais il parait important de prévoir à plus long terme et à une échelle plus fine. Enfin, une SVM linéaire a été appliquée à nos données. La méthode à noyaux n'a pas été plus performante, la perspective d'une étude de simulation pourrait être intéressante pour comprendre pourquoi

La troisième partie de ce travail a été réalisée lors de ma mobilité de 6 mois aux États-Unis, financée grâce à l'obtention de la bourse Fulbright. Ainsi, j'ai pu bénéficier de l'expertise statistique de chercheurs travaillant également sur l'utilisation de modèles prédictifs réutilisant les données massives pour la surveillance syndromique. L'objectif était ici la réutilisation de différentes sources de données : les données hospitalières, les données du web incluant Google et Twitter et les données climatiques à l'aide d'un modèle statistique adapté afin de prévoir à plus long terme les épidémies de grippe. Ces prévisions ont été faites au niveau des régions françaises. 3 méthodes statistiques appliquées aux États-Unis et développées dans l'équipe où j'ai séjourné ont été testées sur nos données. Cependant, les modèles ont été modifiés afin de ne pas prédire seulement en temps réel mais jusqu'à 2 semaines. De plus, les données de Twitter et les données climatiques ont été ajoutées en tant que variables explicatives et 2 méthodes supplémentaires ont été implémentées pour le modèle d'ensemble ARGONet : la moyenne et la régression linéaire des prédictions obtenues par les modèles ARGO et Net. Dans cette étude, nous avons montré que toutes les sources de données étaient importantes pour la prévision, même si les données hospitalières et les données de Google était les plus utilisées dans nos modèles. Le modèle statistique qui a permis d'obtenir les meilleurs résultats en terme d'estimation est la méthode d'ensemble ARGONet, utilisant le pouvoir prédictif des deux autres modèles développés, ARGO et Net. Une des principales limites de cette étude est liée aux données utilisées. En effet, il est dommage de n'avoir à disposition que les données hospitalières du CHU de Rennes qui ne sont pas représentatives de toutes les

régions. De plus, les données du web et en particulier les données Twitter ont été extraites pour la France entière.

Enfin, la dernière partie a été consacrée à l'application de nos modèles statistiques à un autre cas d'usage : la gastro-entérite. Ici, nous avons choisi de développer un modèle de régression linéaire pénalisée Elastic Net réutilisant les données massives en santé et les données du web afin de prévoir les épidémies jusqu'à 3 semaines. Ces prévisions ont été faites à différentes échelles : au niveau national, au niveau régional, au niveau des urgences et au niveau des hospitalisations. Tout comme pour l'étude précédente, une des limites de cette étude est la zone géographique couverte par ces données. De plus, nous avons appliqué ici un modèle de régression linéaire pénalisée Elastic Net. Il serait intéressant de pouvoir comparer nos résultats avec la méthode d'ensemble développée précédemment.

7 Conclusion

Les méthodes statistiques développées dans cette thèse et l'analyse de la réutilisation de différentes sources de données dont les données massives hospitalières montrent un réel apport pour la surveillance syndromique. Une méthode d'ensemble combinant le pouvoir prédictif de différents modèles ainsi que l'apport des différentes sources de données pourrait permettre de compléter les méthodes de surveillance traditionnelles en France. En effet, cela pourrait être un soutien pour l'aide à la décision et également un moyen de réduire l'impact des épidémies et d'organiser le flux de patients chez les médecins généralistes et dans le service des urgences. Une des perspectives majeure à notre travail est l'utilisation de données hospitalières provenant de différents CHU du grand Ouest : Nantes, Brest, Angers, Poitiers et Tours ainsi que l'utilisation des données du Système national d'information inter-régimes de l'Assurance maladie. De plus, nous avons choisi de privilégier les modèles de régression linéaire pénalisée permettant de pouvoir interpréter facilement les sources de données importantes pour la surveillance syndromique. Cependant les méthodes de machine learning telles que la SVM montrent également de bonnes performances. Il pourrait alors être intéressant de tester d'autres méthodes comme les réseaux de neurones ou de combiner les résultats de ces différents modèles en développant une nouvelle méthode ensembliste.

Références

[1] What Is Big Data ? - Gartner IT Glossary - Big Data ;. Available from : https://www.gartner.com/it-glossary/big-data/.

[2] Tuffery S. Data Mining et statistique décisionnelle : L'intelligence des données. Editions TECHNIP ; 2012.

[3] Besse P, Le Gall C, Raimbault N, Sarpy S. Data mining et statistique. Journal de la société française de statistique. 2001 ;142(1) :5–36.

[4] Formenty P, Roth C, Gonzalez-Martin F, Grein T, Ryan M, Drury P, et al. Les pathogènes émergents, la veille internationale et le Règlement sanitaire international (2005). Médecine et Maladies Infectieuses. 2006 ;36(1) :9–15.

[5] Santé publique France - Qui sommes-nous ? ;. Available from : https://www.santepubliquefrance.fr/Sante-publique-France/Qui-sommes-nous.

[6] Réseau Sentinelles > France > Accueil ;. Available from : https://websenti.u707.jussieu.fr/sentiweb/.

[7] Yang S, Santillana M, Kou SC. Accurate estimation of influenza epidemics using Google search data via ARGO. Proceedings of the National Academy of Sciences of the United States of America. 2015 ;112 :14473–14478.

[8] Santillana M, Nguyen AT, Louie T, Zink A, Gray J, Sung I, et al. Cloud-based Electronic Health Records for Real-time, Region-specific Influenza Surveillance. Scientific Reports. 2016 ;6 :25732.

[9] Josseran L, Fouillet A. La surveillance syndromique : bilan et perspective d'un concept prometteur. Revue d'Épidémiologie et de Santé Publique. 2013-04-01 ;61(2) :163–170.

[10] Bulletin épidémiologique grippe, semaine 14. Saison 2018-2019 ;. Available from : http://invs.santepubliquefrance.fr/Dossiers-thematiques/Maladies-infectieuses/Maladies-a-prevention-vaccinale/Grippe/Grippe-generalites/Donnees-de-surveillance/Bulletin-epidemiologique-grippe-semaine-14.-Saison-2018-2019.

[11] Nsoesie EO, Brownstein JS, Ramakrishnan N, Marathe MV. A systematic review of studies on forecasting the dynamics of influenza outbreaks. Influenza and Other Respiratory Viruses. 2014 ;8 :309–316.

[12] Carneiro HA, Mylonakis E. Google Trends : A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks. Clinical Infectious Diseases. 2009 ;49(10) :1557–1564.

[13] Butler D. When Google got flu wrong. Nature. 2013 ;494 :155–156.

[14] Broniatowski DA, Paul MJ, Dredze M. National and Local Influenza Surveillance through Twitter : An Analysis of the 2012-2013 Influenza Epidemic. PLOS ONE. 2013 ;8 :e83672.

[15] Paul MJ, Dredze M, Broniatowski D. Twitter Improves Influenza Forecasting. PLoS Currents. 2014 ;6.

[16] Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JS. Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. PLoS Computational Biology. 2015-10-29 ;11(10).

[17] Sharpe JD, Hopkins RS, Cook RL, Striley CW. Evaluating Google, Twitter, and Wikipedia as Tools for Influenza Surveillance Using Bayesian Change Point Analysis : A Comparative Analysis. JMIR Public Health and Surveillance. 2016-10-20 ;2(2).

[18] Generous N, Fairchild G, Deshpande A, Valle SYD, Priedhorsky R. Global Disease Monitoring and Forecasting with Wikipedia. PLOS Comput Biol. 2014 ;10 :e1003892.

[19] McIver DJ, Brownstein JS. Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time. PLOS Comput Biol. 2014 ;10 :e1003581.

[20] Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). Journal of the American Medical Informatics Association. 2010 ;17 :124–130.

[21] Murphy S, Wilcox A. Mission and Sustainability of Informatics for Integrating Biology and the Bedside (i2b2). eGEMs (Generating Evidence & Methods to improve patient outcomes). 2014-09-11 ;2(2).

[22] Cuggia M, Garcelon N, Campillo-Gimenez B, Bernicot T, Laurent JF, Garin E, et al. Roogle : an information retrieval engine for clinical data warehouse. Studies in Health Technology and Informatics. 2011 ;169 :584–588.

[23] Costagliola D, Flahault A, Galinec D, Garnerin P, Menares J, Valleron AJ. A routine tool for detection and assessment of epidemics of influenza-like syndromes in France. American Journal of Public Health. 1991-01 ;81(1) :97–99.

[24] Google Correlate ;. Available from : https://www.google.com/trends/correlate/.

[25] Zou H, Hastie T. Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society, Series B. 2005 ;67 :301–320.

[26] Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. 1996 ;58 :267–288.

[27] Kennard EHe. Ridge regression : biaised estimation for nonorthogonal problems. Technometrics. 1970 ;1.

[28] R Core Team. R : A Language and Environment for Statistical Computing. R Foundation for Statistical Computing ; 2015.

[29] J Friedman RT T Hastie. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010 ;33(1) :1–22.

[30] Breiman L. Random Forests. Machine Learning. 2001 ;45 :5–32.

[31] Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002 ;2(3) :18–22.

[32] Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995 ;20 :273–297.

[33] Drakos G. Support Vector Machine vs Logistic Regression ; 2018-08-12. Available from : https://towardsdatascience.com/support-vector-machine-vs-logistic-regression-94cc2975433f.

[34] Lahiri S, Ghanta K. The Support Vector Regression with the parameter tuning assisted by a differential evolution technique : Study of the critical velocity of a slurry flow in a pipeline. Chemical Industry and Chemical Engineering Quarterly. 2008 07 ;14.

[35] Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071 : Misc Functions of the Department of Statistics, Probability Theory Group (Formerly : E1071), TU Wien ; 2015.

[36] Improved state-level influenza nowcasting in the United States leveraging Internet-based data and network approaches | Nature Communications ;. Available from : https://www.nature.com/articles/s41467-018-08082-0.

[37] Infoclimat - la météo en temps réel : observations météo en direct, prévisions, archives climatologiques, photos et vidéos... ;. Available from : https://www.infoclimat.fr/.

[38] Kuhn M. caret : Classification and Regression Training ; 2018. R package version 6.0-80.

[39] Farthing MJ. Diarrhoea : a significant worldwide problem. International Journal of Antimicrobial Agents. 2000-02 ;14(1) :65–69.

[40] Kosek M, Bern C, Guerrant RL. The global burden of diarrhoeal disease, as estimated from studies published between 1992 and 2000. Bulletin of the World Health Organization. 2003 ;81(3) :197–204.

[41] Cauteren DV, Valk HD, Vaux S, Strat YL, Vaillant V. Burden of acute gastroenteritis and healthcare-seeking behaviour in France : a population-based study. Epidemiology Infection. 2012-04 ;140(4) :697–705.

[42] Big data en santé ;. Available from : https://www.inserm.fr/information-en-sante/dossiers-information/big-data-en-sante.

Annexes

Les annexes présentées ici correspondent aux annexes de l'article "Real Time Influenza Monitoring Using Hospital Big Data in Combination with Machine Learning Methods : Comparison Study" du chapitre 3.

Multimedia Appendix 1 : eHOP queries (with the number of concerned hospital stays from 2003 to 2016)

Full text queries :

1. flu (20 135)
2. flu without vaccine (16 893)
3. flu vaccine (3 317)
4. fever (118 671)
5. aches (8 584)
6. fever or aches (123 332)
7. fever and aches (3 923)
8. fever or aches or flu (132 504)
9. fever and aches or flu (22 405)
10. fever or aches or flu without vaccine (130 254)
11. fever and aches or flu without vaccine (19 275)
12. fever or flu (128 124)
13. aches or flu (26 785)
14. fever or flu without vaccine (125 842)
15. aches or flu without vaccine (23 687)
16. flu in the emergency room (9 886)
17. Tamiflu (1 899)

Appropriate terminologies :

18-23. ICD-10 queries (16 504)
24-34. Test laboratories (6 805)

Figure .1: Multimedia Appendix 1

1. grippe incubation
2. grippe contagion
3. incubation grippe
4. epidemie de grippe
5. grippe symptome
6. grippe symptomes
7. epidemie grippe
8. épidémie de grippe
9. contagion grippe
10. épidémie grippe
11. grippe traitement
12. traitement de la grippe
13. comment soigner la grippe
14. soigner grippe
15. symptome de la grippe
16. soigner une grippe
17. grippe en France
18. symptomes de la grippe
19. ski en mars
20. calendrier fevrier
21. grippe carte
22. symptome grippe
23. mois de février
24. la grippe en france
25. mois de fevrier
26. skiset
27. lispach
28. oursinade
29. made in angers
30. date mardi gras
31. chaine thermale
32. ancelle
33. chaine thermale du soleil
34. costume de carnaval
35. fête du mimosa
36. syndrome grippal
37. salon peche
38. fete du mimosa
39. calendrier février
40. joue du loup
41. banh chung
42. la joue
43. le mois de février
44. chastreix
45. confiture oranges
46. la joue du loup
47. grippe epidemie
48. greolières
49. fete des citrons menton
50. orange amere
51. les jouvencelles
52. salon de la peche
53. location ski lyon
54. masque de carnaval
55. esf
56. météo serre chevalier
57. ski set
58. grippe bébé
59. recette de bugnes
60. recette des bugnes
61. oursinades
62. la bresse lispach
63. etat grippal
64. saint jean montclar
65. recette bugne
66. meteo les angles
67. vacance de fevrier
68. sentinelle grippe
69. albiez
70. fete des citrons
71. date de la saint valentin
72. fête des citrons à menton
73. minable le pingouin
74. melezes
75. gresse
76. coloriage carnaval
77. tarif location de ski
78. les melezes
79. oranges ameres
80. carnaval de cadiz
81. dhg
82. printemps ete
83. costume carnaval
84. saint leger les melezes
85. le mont dore
86. gresse en vercors
87. date carnaval
88. le mourtis
89. masque carnaval
90. oranges amères
91. maternelle carnaval
92. vacances de fevrier
93. de carnaval
94. carnaval de limoux
95. ski laguiole
96. fevrier
97. chabanon
98. grippe enceinte
99. montclar
100. bettex

Figure .2: Multimedia Appendix 2

1. soigner la grippe
2. grippe contagion
3. grippe symptomes
4. epidemie de grippe
5. grippe incubation
6. grippe symptome
7. traitement de la grippe
8. epidemie grippe
9. grippe traitement
10. état grippal
11. symptome de la grippe
12. grossesse et grippe
13. symptomes de la grippe
14. soigner une grippe
15. épidémie grippe
16. grippe en france
17. comment soigner la grippe
18. symptomes grippe
19. soigner grippe
20. symptome grippe
21. la grippe en france
22. chaine thermale
23. chaine thermale du soleil
24. made in angers
25. incubation de la grippe
26. traitement grippe
27. chaines thermales du soleil
28. costume de carnaval
29. salon peche
30. oursinade
31. lispach
32. salon de la peche
33. skiset
34. confiture oranges
35. oranges ameres
36. confiture orange
37. ski en mars
38. mois de février
39. chastreix
40. texte faire part mariage
41. mois de fevrier
42. fête du mimosa
43. fete des citrons menton
44. location ski lyon
45. ancelle
46. calendrier fevrier
47. fetes des citrons
48. mont dore
49. sentinelle grippe
50. la bresse lispach
51. www.vacaf.org
52. costume carnaval
53. etat grippal
54. les jouvencelles
55. ski esf
56. confiture orange amere
57. ski laguiole
58. minable le pingouin
59. esf
60. joue du loup
61. la joue du loup
62. recette de bugne
63. grippe france
64. la joue
65. printemps ete
66. infosup clermont
67. chaines thermales
68. météo risoul
69. bal de carnaval
70. ski mont dore
71. les mourtis
72. fete du mimosa
73. fete du citron a menton
74. orange amere
75. masque de carnaval
76. fete des citrons
77. collection printemps ete
78. jour de la st valentin
79. orange amère
80. saint jean montclar
81. meteo egypte
82. le mont dore
83. carnaval de limoux
84. recette des bugnes
85. grippal
86. coloriage carnaval
87. meteo metabief
88. nuit de neige
89. météo chatel
90. meteo orcieres
91. date mardi gras
92. vacances février paris
93. grippe intestinale
94. superdevoluy
95. affiche carnaval
96. ski manigod
97. chabanon
98. esf la bresse
99. fête des citrons à menton
100. sentiweb

Figure .3: Multimedia Appendix 3

NATIONAL	2010-2011				2011-2012				2012-2013				2013-2014			
	PCC	MSE	ΔH	ΔL	PCC	MSE	ΔH	ΔL	PCC	MSE	ΔH	ΔL	PCC	MSE	ΔH	ΔL
eHOP Custom																
RF	0.95	4119	50	2	0.88	6447	68	0	0.90	10904	-63	3	0.86	3664	30	1
RF+Arima	0.98	1174	42	1	0.96	1928	79	0	0.95	6128	105	0	0.91	2212	75	1
SVM	0.97	1932	-8	1	0.95	1877	35	0	0.96	4930	-21	1	0.95	996	19	1
SVM+Arima	0.97	1436	23	1	0.97	1450	53	0	0.96	4876	39	0	0.94	1217	43	1
ElasticNet	0.97	1855	4	1	0.92	3056	12	0	0.95	6343	-12	1	0.92	1447	-6	1
Elastic+Arima	0.98	1222	23	1	0.96	1735	56	0	0.96	5102	48	0	0.95	1145	27	1
Google Custom																
RF	0.89	6476	-112	5	0.86	4849	-44	1	0.88	15642	-152	1	0.87	2651	27	1
RF+Arima	0.96	1966	-35	0	0.96	1376	-9	0	0.94	7171	83	1	0.93	1347	67	1
SVM	0.96	2815	16	1	0.95	1971	7	0	0.96	5803	12	1	0.91	1711	43	1
SVM+Arima	0.96	2311	41	1	0.96	1653	42	0	0.96	5194	36	1	0.90	1958	-36	1
ElasticNet	0.94	3606	-27	1	0.93	2654	-8	0	0.94	7658	12	1	0.88	1981	2	1
Elastic+Arima	0.96	2394	35	1	0.95	1801	45	0	0.95	6108	65	1	0.92	1664	55	1
eHOP Complete																
RF	0.96	3121	54	1	0.89	4797	-2	0	0.94	7832	-112	2	0.86	10518	83	-1
RF+Arima	0.97	1917	67	1	0.96	1636	-9	0	0.97	3148	-15	2	0.85	5285	93	1
SVM	0.95	3046	66	4	0.94	2366	22	0	0.96	4646	-33	1	0.94	2234	49	1
SVM+Arima	0.95	3087	66	4	0.94	2158	28	0	0.96	5174	-32	1	0.94	2267	49	1
ElasticNet	0.94	4807	113	0	0.96	1685	17	0	0.94	7762	92	0	0.88	2164	5	1
Elastic+Arima	0.94	7387	254	0	0.97	1732	30	-1	0.94	7807	136	0	0.94	1730	48	1
Google Complete																
RF	0.95	2743	23	0	0.94	2674	8	0	0.96	5711	-112	1	0.93	4931	84	1
RF+Arima	0.96	2392	49	2	0.98	1188	58	0	0.97	3391	27	1	0.97	1249	57	1
SVM	0.95	2671	12	3	0.97	3893	-27	0	0.97	3421	-8	0	0.92	1564	56	1
SVM+Arima	0.94	3107	32	1	0.97	931	-28	0	0.97	3608	-30	0	0.92	1641	56	1
ElasticNet	0.90	5392	-49	4	0.95	1975	-11	0	0.95	6142	6	1	0.94	3360	93	1
Elastic+Arima	0.96	2153	6	1	0.94	2260	11	0	0.96	4081	36	1	0.95	2511	85	1
Historical Variables																
RF	0.95	4118	52	1	0.87	5796	45	0	0.92	9265	-72	3	0.85	3588	13	1
RF+Arima	0.97	1307	102	1	0.96	1427	7	0	0.97	4116	-45	0	0.93	2052	85	1
SVM	0.96	2202	50	1	0.94	1902	19	0	0.95	5336	1	1	0.94	1068	37	1
SVM+Arima	0.96	2135	49	1	0.95	1739	18	0	0.96	5301	1	1	0.94	1095	37	1
ElasticNet	0.96	2090	29	1	0.94	1909	2	0	0.96	4957	4	1	0.93	1178	19	1
ElasticNet+Arima	0.97	2060	17	1	0.95	1908	2	0	0.96	4957	4	1	0.93	1178	19	1

NATIONAL	2014-2015				2015-2016				Global		Means					
	PCC	MSE	ΔH	ΔL	PCC	MSE	ΔH	ΔL	PCC	MSE	PCC	MSE	ΔH	|ΔH|	ΔL	|ΔL|
eHOP Custom																
RF	0.94	11606	-167	0	0.87	4754	-51	2	0.947	2292	0.9	6916	-22	72	1.33	1.33
RF+Arima	0.97	4546	-55	0	0.85	6202	58	1	0.974	1227	0.94	3698	51	71	0.5	0.5
SVM	0.98	3750	-11	1	0.94	2809	24	1	0.980	866	0.96	2716	6	19	0.83	0.83
SVM+Arima	0.98	3482	-8	1	0.94	2703	25	1	0.981	819	0.96	2527	29	33	0.66	0.66
ElasticNet	0.96	5638	-43	0	0.90	4125	-34	1	0.974	1133	0.94	3744	-13	18	0.66	0.66
Elastic+Arima	0.98	3333	-13	1	0.91	3448	15	1	0.980	872	0.96	2664	26	30	0.66	0.66
Google Custom																
RF	0.86	18706	-133	2	0.87	6507	-40	2	0.937	2607	0.87	9139	-76	94	2	2
RF+Arima	0.95	8316	-15	1	0.92	3730	44	1	0.972	1171	0.94	3984	22	42	0.66	0.66
SVM	0.97	5325	36	0	0.94	2461	10	1	0.977	968	0.95	3348	21	23	0.66	0.66
SVM+Arima	0.97	5135	36	0	0.95	2189	-72	1	0.979	899	0.95	3073	8	38	0.66	0.66
ElasticNet	0.94	9513	-10	1	0.89	4275	-7	1	0.968	1382	0.92	4948	-6	11	0.83	0.83
Elastic+Arima	0.97	5076	27	1	0.93	3069	38	1	0.977	988	0.95	3352	44	44	0.83	0.83
eHOP Complete																
RF	0.92	13468	-123	2	0.85	5843	-45	-5	0.954	2148	0.90	7597	-24	75	-0.2	1.5
RF+Arima	0.96	6706	-134	1	0.89	4612	37	1	0.974	1172	0.93	3884	7	64	1	1
SVM	0.96	6146	114	0	0.95	2379	59	1	0.972	1173	0.95	3469	46	57	1.2	1.2
SVM+Arima	0.96	7044	112	0	0.95	2440	60	1	0.971	1247	0.95	3495	47	58	1.2	1.2
ElasticNet	0.94	9899	72	1	0.89	4311	-74	1	0.962	1693	0.93	5105	38	62	0.5	0.5
Elastic+Arima	0.96	6194	120	1	0.93	2951	-19	1	0.970	1427	0.95	4634	95	101	0.33	0.66
Google Complete																
RF	0.92	11601	-121	1	0.94	6922	36	1	0.963	1706	0.94	5764	-14	70	0.66	0.66
RF+Arima	0.94	8529	-33	2	0.96	2780	24	1	0.979	917	0.96	3255	30	45	1.2	1.2
SVM	0.98	4148	81	0	0.94	4130	110	1	0.974	1192	0.96	2805	37	49	0.8	0.8
SVM+Arima	0.99	3430	61	0	0.95	3117	64	0	0.976	11310	0.96	2639	26	45	0.3	0.3
ElasticNet	0.96	7128	36	1	0.96	1920	28	1	0.965	1604	0.94	4320	17	37	1.33	0.33
Elastic+Arima	0.97	5047	67	1	0.97	1751	26	0	0.978	1057	0.96	2967	39	38	0.66	0.66
Historical Variables																
RF	0.95	10845	-172	0	0.81	7132	-46	-5	0.945	2320	0.89	6790	-30	67	0	1.66
RF+Arima	0.96	6033	-86	0	0.93	2648	3	1	0.971	959	0.95	3030	11	55	0.5	0.5
SVM	0.96	5427	29	0	0.95	1772	18	1	0.978	920	0.95	2935	26	26	0.66	0.66
SVM+Arima	0.97	4818	15	0	0.95	1724	9	1	0.979	891	0.95	2802	22	22	0.66	0.66
ElasticNet	0.97	5686	27	1	0.94	2180	24	1	0.974	991	0.95	3000	18	18	0.83	0.83
Elastic+Arima	0.96	6459	27	1	0.95	1809	24	1	0.978	980	0.95	3062	16	16	0.83	0.83

REGIONAL	2010-2011				2011-2012				2012-2013				2013-2014			
	PCC	MSE	ΔH	ΔL	PCC	MSE	ΔH	ΔL	PCC	MSE	ΔH	ΔL	PCC	MSE	ΔH	ΔL
eHOP Custom																
RF	0.91	5796	-29	-1	0.90	1665	4	-4	0.87	7209	-60	1	0.65	4577	-4	1
RF+Arima	0.92	5524	-24	1	0.85	2029	-6	-4	0.83	9785	-59	1	0.58	5039	3	1
SVM	0.92	5502	-53	1	0.89	1533	-53	-1	0.85	7874	-80	1	0.76	2477	-28	1
SVM+Arima	0.92	4721	-40	1	0.87	1548	-39	-1	0.83	8451	-80	1	0.73	2926	-77	1
ElasticNet	0.88	7029	-99	1	0.90	2092	-93	-1	0.82	12232	-226	1	0.76	2154	-81	1
Elastic+Arima	0.92	4689	-28	0	0.82	2146	-26	-1	0.81	9584	-90	1	0.71	2855	-27	1
Google Custom																
RF	0.86	7706	-9	0	0.89	1657	17	-3	0.82	10031	3	1	0.70	2887	-46	1
RF+Arima	0.90	6570	37	0	0.78	3141	57	-3	0.73	15320	54	1	0.68	3277	-26	1
SVM	0.91	6010	-71	0	0.84	2102	-54	1	0.70	14528	1	1	0.65	3247	-36	1
SVM+Arima	0.90	6041	-67	0	0.84	2134	-54	1	0.70	14600	4	1	0.64	3270	-36	1
ElasticNet	0.88	9682	-137	1	0.76	4208	-93	1	0.68	17815	-206	1	0.65	3445	-102	1
Elastic+Arima	0.91	5494	-58	0	0.82	2189	-40	1	0.71	14566	-46	1	0.74	2637	-31	1
eHOP Complete																
RF	0.92	4263	-40	0	0.89	2222	18	-1	0.88	8525	-46	1	0.65	9735	26	1
RF+Arima	0.93	4036	-30	0	0.90	1984	20	-1	0.89	8218	-51	1	0.53	8530	13	-2
SVM	0.89	7682	69	3	0.87	1637	-7	-1	0.85	7249	-104	3	0.74	4851	24	1
SVM+Arima	0.89	7607	55	3	0.86	1833	-7	-1	0.85	7293	-104	3	0.72	4552	26	1
ElasticNet	0.86	11416	-178	1	0.87	2755	-107	-1	0.81	12443	-220	1	0.71	2771	-100	1
Elastic+Arima	0.90	5740	-2	1	0.80	2393	-27	-1	0.83	8635	-61	3	0.66	3677	7	1
Google Complete																
RF	0.92	4650	-80	0	0.90	1423	-8	-3	0.88	6824	-54	1	0.63	5955	-9	2
RF+Arima	0.92	4512	-28	0	0.80	2659	3	-3	0.82	9973	-47	1	0.60	5248	19	1
SVM	0.84	8664	-97	1	0.33	7233	24	-4	0.83	7955	-67	1	0.56	4735	-43	-1
SVM+Arima	0.85	8231	-98	1	0.28	8047	24	-4	0.83	8262	-68	1	0.55	4871	-1	1
ElasticNet	0.85	10735	-148	1	0.60	4090	-97	-3	0.74	15086	-192	1	0.67	3764	-134	1
Elastic+Arima	0.89	6455	-63	1	0.64	3907	-44	-1	0.75	12229	-28	1	0.78	2113	-26	1
Historical variables																
RF	0.92	4520	-45	0	0.77	3120	-3	-1	0.82	8352	-90	1	0.44	7277	4	1
RF+Arima	0.93	3574	-28	0	0.73	3701	10	1	0.75	11561	-84	1	0.45	7522	21	2
SVM	0.89	6392	-46	0	0.71	3406	-19	1	0.73	12857	-173	1	0.60	3934	-42	1
SVM+Arima	0.90	6132	-46	0	0.71	3490	-19	1	0.73	12590	-173	1	0.60	3728	-42	1
ElasticNet	0.91	6211	-81	1	0.68	3702	-63	1	0.71	13385	-89	1	0.58	3369	-62	1
Elastic+Arima	0.91	6193	-84	1	0.68	3637	-55	1	0.74	12423	-86	1	0.53	3673	-58	1

REGIONAL	2014-2015				2015-2016				Global		Means					
	PCC	MSE	ΔH	ΔL	PCC	MSE	ΔH	ΔL	PCC	MSE	PCC	MSE	ΔH	|ΔH|	ΔL	|ΔL|
eHOP Custom																
RF	0.88	18111	-97	1	0.85	4218	-56	1	0.911	2777	0.84	6929	-40	42	-0.2	1.5
RF+Arima	0.85	18236	-15	0	0.90	3181	-2	1	0.910	2807	0.82	7299	-17	18	0	1.3
SVM	0.91	12865	-58	-1	0.83	6052	-90	1	0.923	2364	0.86	6050	-60	60	0.3	1
SVM+Arima	0.91	12718	-30	-1	0.84	5899	-60	1	0.916	2491	0.85	6044	-54	54	0.3	1
ElasticNet	0.90	21569	-215	-1	0.87	6152	-115	1	0.907	3283	0.86	8538	-138	138	0.3	1
Elastic+Arima	0.90	11949	19	0	0.85	4770	-38	1	0.918	2451	0.84	5999	-32	38	0.3	0.7
Google Custom																
RF	0.76	27094	-85	1	0.79	8211	-32	7	0.897	3221	0.80	9598	-25	32	1.2	2.2
RF+Arima	0.77	26400	-87	1	0.80	7422	-21	1	0.880	3780	0.78	10355	2	47	1.2	1.2
SVM	0.88	15569	27	0	0.82	5610	-71	1	0.902	2903	0.8	7844	-34	43	0.7	0.7
SVM+Arima	0.88	15529	26	0	0.81	5680	-72	1	0.903	2894	0.8	7876	-33	43	0.7	0.7
ElasticNet	0.86	20276	-125	0	0.76	8064	-148	1	0.891	3887	0.77	10582	-135	135	0.8	0.8
Elastic+Arima	0.90	12748	113	0	0.78	7041	-36	2	0.900	3021	0.81	7446	-16	54	0.8	0.8
eHOP Complete																
RF	0.89	15654	-132	0	0.90	4141	-15	1	0.914	2906	0.86	7423	-32	46	0.3	0.7
RF+Arima	0.88	16625	-134	0	0.89	3601	10	1	0.916	2758	0.84	7166	-29	43	-0.2	0.8
SVM	0.88	15132	92	0	0.86	4363	-35	1	0.911	2756	0.85	6819	7	55	1.2	1.5
SVM+Arima	0.88	16229	131	0	0.86	4250	-36	1	0.909	2816	0.84	6961	11	60	1.2	1.5
ElasticNet	0.89	20134	-166	0	0.86	6992	-124	1	0.905	3745	0.83	9419	-149	149	0.5	0.8
Elastic+Arima	0.92	10032	48	0	0.84	4880	-76	1	0.917	2483	0.83	5893	-19	37	0.8	1.2
Google Complete																
RF	0.88	17301	-131	1	0.80	6581	-91	3	0.912	2736	0.83	7122	-62	62	0.7	1.7
RF+Arima	0.88	17178	-131	1	0.79	6666	-84	3	0.909	2767	0.80	7706	-45	52	0.5	1.5
SVM	0.87	17624	-83	0	0.76	8609	-45	3	0.89	3348	0.70	9137	-52	59	0	1.7
SVM+Arima	0.87	15552	7	0	0.76	8522	-48	3	0.89	3265	0.69	8914	-31	41	0.3	1.7
ElasticNet	0.87	21226	-168	-1	0.75	7302	-120	3	0.893	3815	0.75	10367	-143	143	0.3	1.7
Elastic+Arima	0.91	11243	6	0	0.77	7488	-28	2	0.903	2967	0.79	7239	-31	32	0.7	1
Historical variables																
RF	0.79	31289	-249	2	0.79	8259	79	2	0.876	3627	0.76	10470	-51	78	0.8	1.2
RF+Arima	0.85	20251	-60	0	0.78	10314	129	1	0.885	3432	0.75	9487	-2	55	0.8	0.8
SVM	0.82	25670	-8	0	0.74	8516	7	2	0.876	3578	0.75	10021	-47	49	0.8	0.8
SVM+Arima	0.82	24632	-15	0	0.74	8479	15	2	0.885	3431	0.75	9950	-47	52	0.8	0.8
ElasticNet	0.85	20298	-13	0	0.77	7321	8	1	0.884	3378	0.75	9048	-50	52	0.8	0.8
Elastic+Arima	0.85	20038	4	0	0.77	7361	8	1	0.887	3287	0.75	8887	-45	49	0.8	0.8

Figure .4: Multimedia Appendix 6

REGIONAL	2010-2011				2011-2012				2012-2013				2013-2014			
	PCC	MSE	ΔH	ΔL	PCC	MSE	ΔH	ΔL	PCC	MSE	ΔH	ΔL	PCC	MSE	ΔH	ΔL
eHOP Custom																
Dataset 1	0.92	4689	-28	0	0.82	2146	-26	-1	0.81	9584	-90	1	0.71	2855	-27	1
Dataset 2	0.91	6298	50	1	0.73	3357	14	-4	0.84	8291	-34	3	0.74	3584	7	1

REGIONAL	2014-2015				2015-2016				Global		Means					
	PCC	MSE	ΔH	ΔL	PCC	MSE	ΔH	ΔL	PCC	MSE	PCC	MSE	ΔH	|ΔH|	ΔL	|ΔL|
eHOP Custom																
Dataset 1	0.90	11949	19	0	0.85	4770	-38	1	0.92	2451	0.84	5999	-32	38	0.3	0.7
Dataset 2	0.93	8638	14	0	0.90	3066	40	1	0.92	2347	0.84	5539	16	27	0.3	1.7

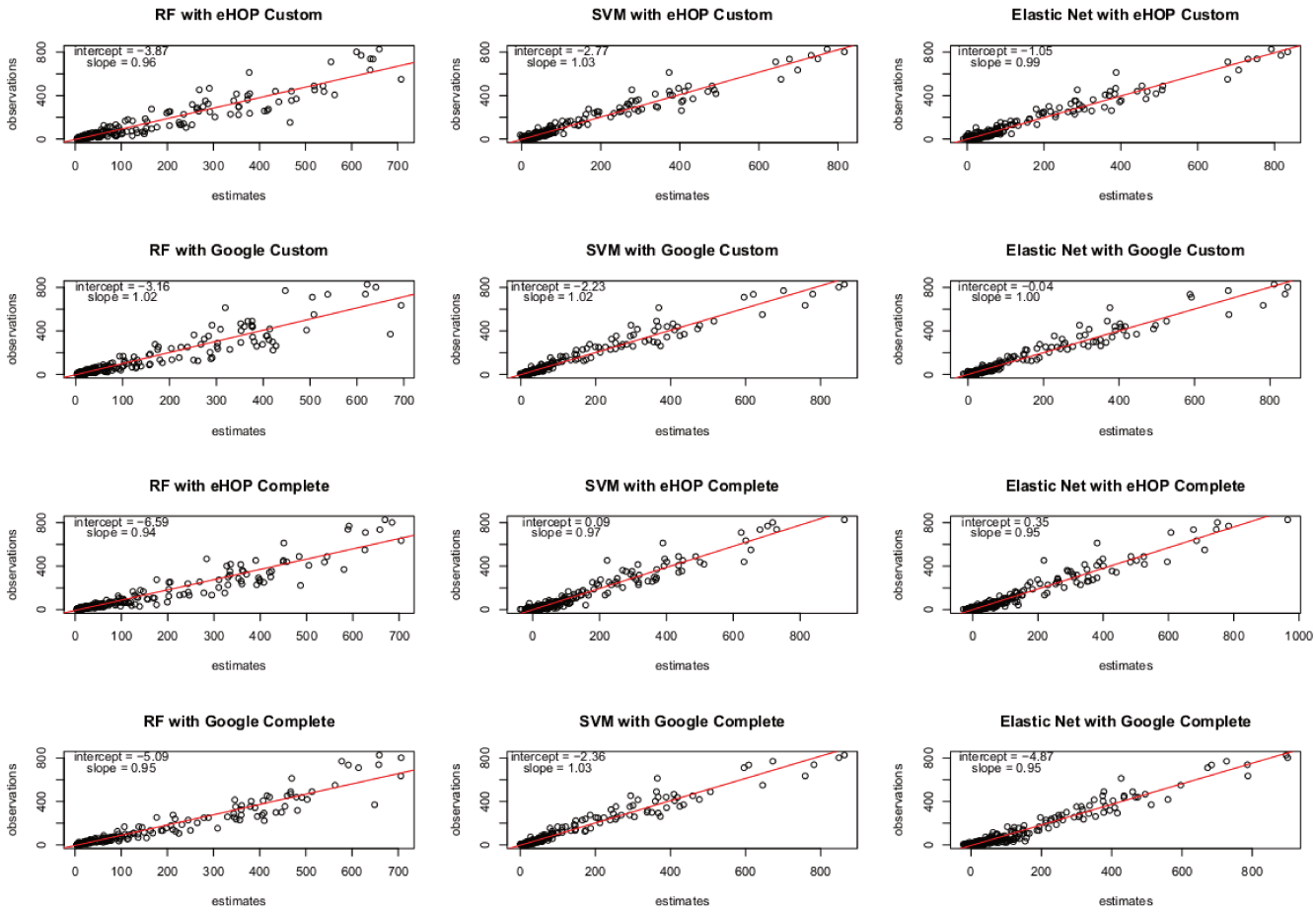


Figure .5: Multimedia Appendix 7

Multimedia Appendix 8. Accuracy metrics for the 2010-2011 (flu outbreak period for which the best estimates were obtained with all models) and 2013-2014 (flu outbreak period for which the worst estimates were obtained with all models) seasons. PCC and MSE for the global period (Global) and mean values (Means) of all indicators for each model during the epidemic periods. In bold, the best results for each dataset. a. Data for the whole France. b. Data for the Brittany region.

a. National	2010-2011				2013-2014				Global		Means					
	PCC	MSE	ΔH	ΔL	PCC	MSE	ΔH	ΔL	PCC	MSE	PCC	MSE	ΔH	|ΔH|	ΔL	|ΔL|
eHOP Custom																
RF	0.95	4119	50	2	0.86	3664	30	1	0.947	2292	0.9	6916	–22	72	1.33	1.33
SVM	0.97	1932	–8	1	**0.95**	**996**	19	1	**0.98**	**866**	0.96	2716	6	19	0.83	0.83
Elastic+Arima	**0.98**	**1222**	23	1	0.95	1145	27	1	0.98	872	**0.96**	**2664**	26	30	0.66	0.66
Google Custom																
RF	0.89	6476	–112	5	0.87	2651	27	1	0.937	2607	0.87	9139	–76	94	2	2
SVM	0.96	2815	16	1	0.91	1711	43	1	**0.977**	**968**	**0.95**	**3348**	21	23	0.66	0.66
Elastic+Arima	**0.96**	**2394**	35	1	**0.92**	**1664**	55	1	0.977	988	0.95	3352	44	44	0.83	0.83
eHOP Complete																
RF	0.96	3121	54	1	0.86	10518	83	–1	0.954	2148	0.9	7597	–24	75	–0.2	1.5
SVM	**0.95**	**3046**	66	4	0.94	2234	49	1	**0.972**	**1173**	**0.95**	**3469**	46	57	1.2	1.2
Elastic+Arima	0.94	7387	254	0	**0.94**	**1730**	48	1	0.97	1427	0.95	4634	95	101	0.33	0.66
Google Complete																
RF	0.95	2743	23	0	0.93	4931	84	1	0.963	1706	0.94	5764	–14	70	0.66	0.66
SVM	0.95	2671	12	3	0.92	**1564**	56	1	0.974	1192	**0.96**	**2805**	37	49	0.8	0.8
Elastic+Arima	**0.96**	**2153**	6	1	**0.95**	2511	85	1	**0.978**	**1057**	0.96	2967	39	38	0.66	0.66

b. Regional	2010-2011				2013-2014				Global		Means					
	PCC	MSE	ΔH	ΔL	PCC	MSE	ΔH	ΔL	PCC	MSE	PCC	MSE	ΔH	|ΔH|	ΔL	|ΔL|
eHOP Custom																
RF	0.91	5796	–29	–1	0.65	4577	–4	1	0.911	2777	0.84	6929	–40	42	–0.2	1.5
SVM	0.92	5502	–53	1	**0.76**	**2477**	–28	1	**0.923**	**2364**	**0.86**	**6050**	–60	60	0.3	1
Elastic+Arima	**0.92**	**4689**	–28	0	0.71	2855	–27	1	0.918	2451	0.84	5999	–32	38	0.3	0.7
Google Custom																
RF	0.86	7706	–9	0	0.7	2887	–46	1	0.897	3221	0.80	9598	–25	32	1.2	2.2
SVM	0.91	6010	–71	0	0.65	3247	–36	1	**0.902**	**2903**	0.80	7844	–34	43	0.7	0.7
Elastic+Arima	**0.91**	**5494**	–58	0	**0.74**	**2637**	–31	1	0.9	3021	**0.81**	**7446**	–16	54	0.8	0.8
eHOP Complete																
RF	**0.92**	**4263**	–40	0	0.65	9735	26	1	0.914	2906	**0.86**	7423	–32	46	0.3	0.7
SVM	0.89	7682	69	3	**0.74**	**4851**	24	1	0.911	2756	0.85	6819	7	55	1.2	1.5
Elastic+Arima	0.9	5740	–2	1	0.66	3677	7	1	**0.917**	**2483**	0.83	**5893**	–19	37	0.8	1.2
Google Complete																
RF	**0.92**	**4650**	–80	0	0.63	5955	–9	2	**0.912**	**2736**	**0.83**	**7122**	–62	62	0.7	1.7
SVM	0.84	8664	–97	1	0.56	4735	–43	–1	0.890	3348	0.70	9137	–52	59	0	1.67
Elastic+Arima	0.89	6455	–63	1	**0.78**	**2113**	–26	1	0.903	2967	0.79	7239	–31	32	0.7	1

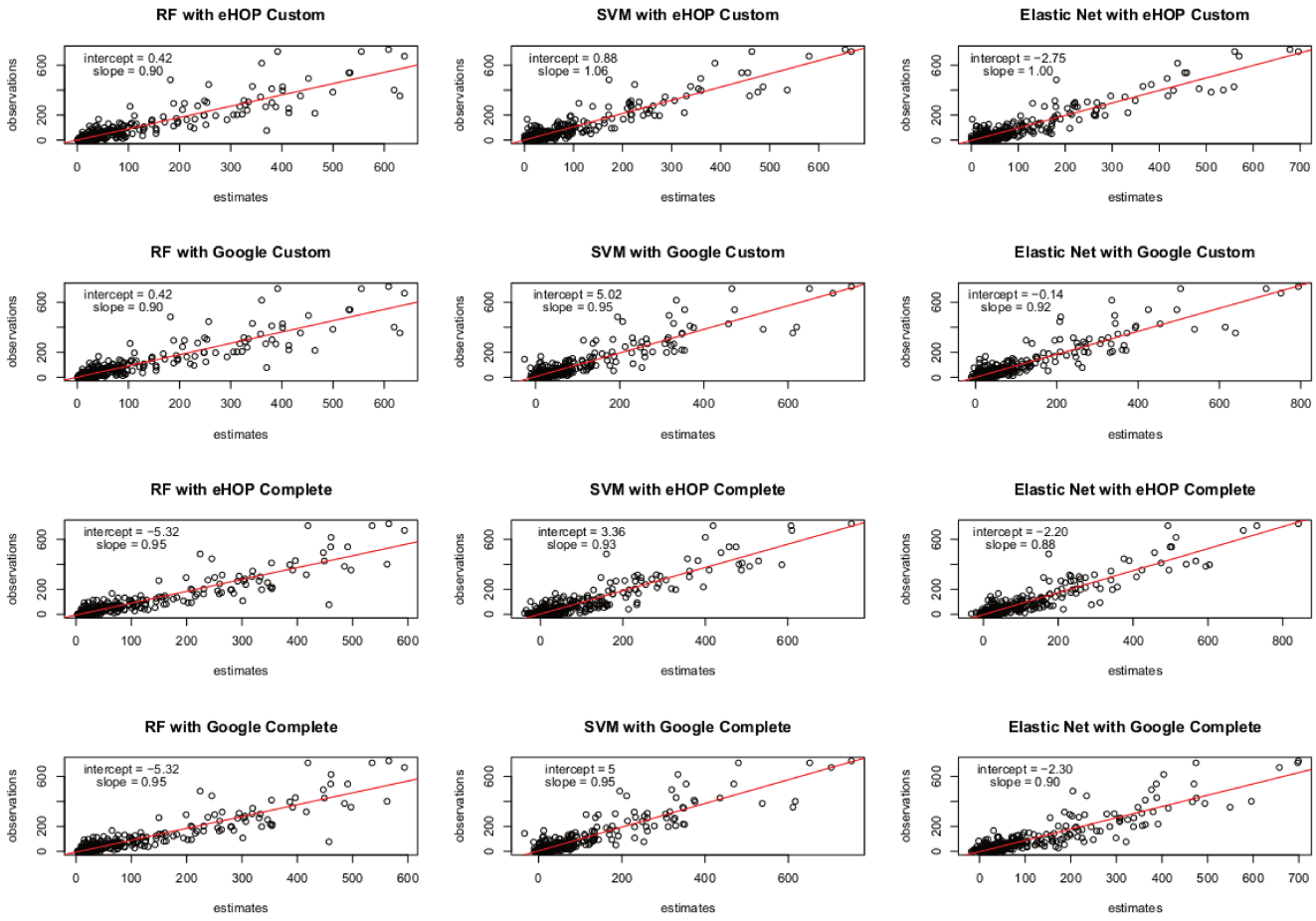


Figure .6: Multimedia Appendix 9

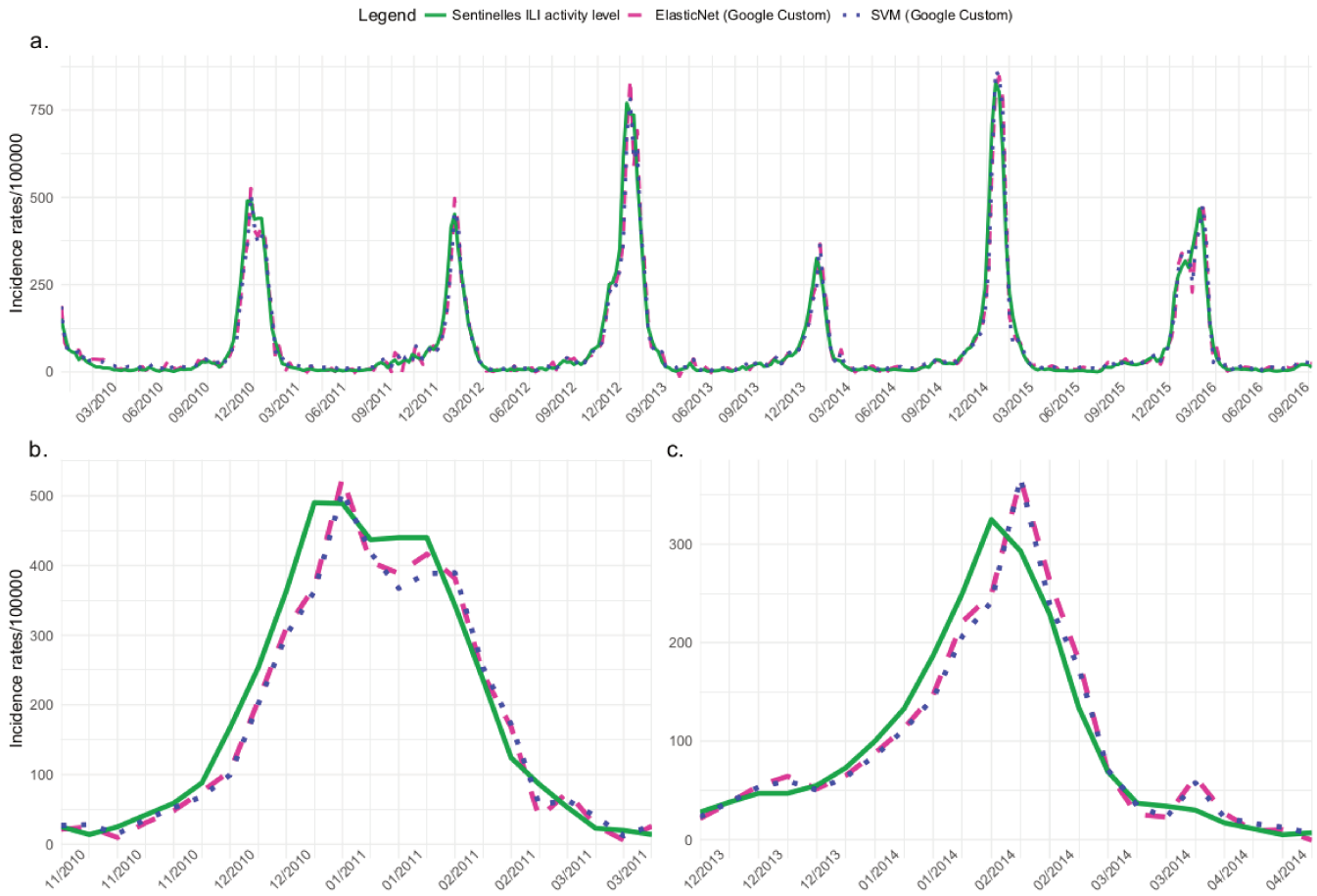


Figure .7: Multimedia Appendix 10

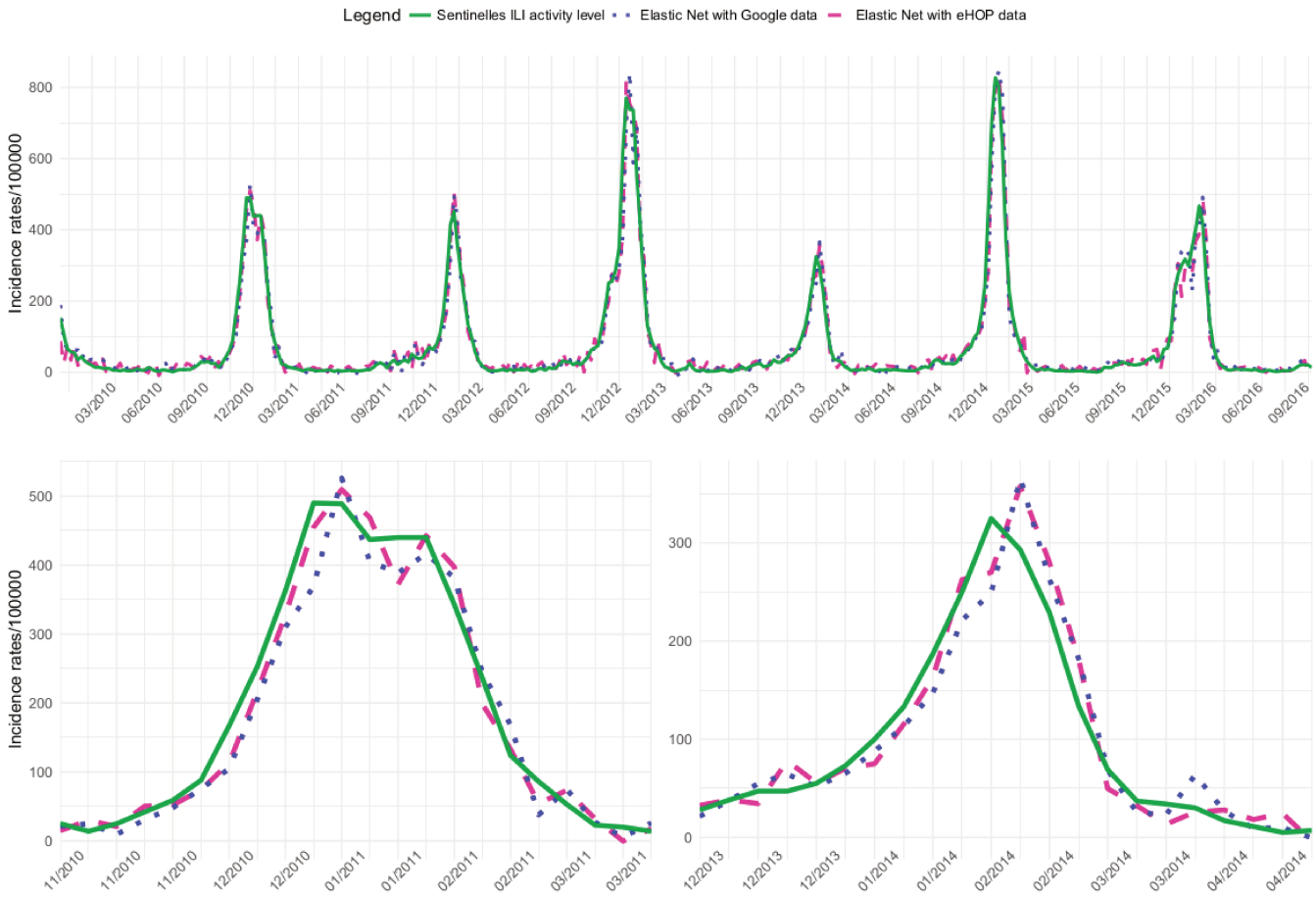


Figure .8: Multimedia Appendix 11

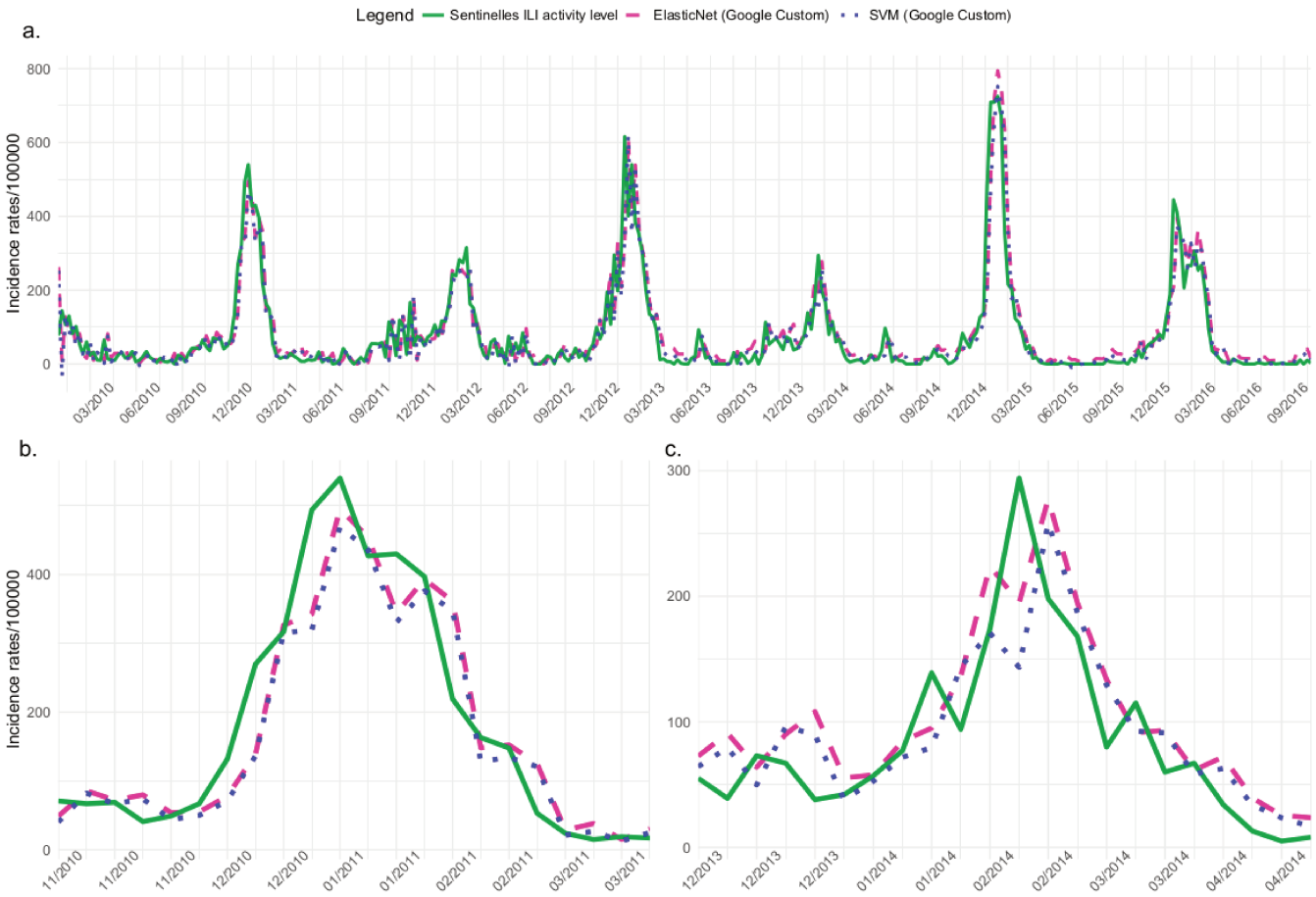


Figure .9: Multimedia Appendix 12

UNIVERSITE BIOLOGIE BRETAGNE SANTE LOIRE

UNIVERSITÉ DE RENNES 1

Titre : Modèles statistiques pour les systèmes d'aide à la décision basés sur le réutilisation des données massives en santé : Application à la surveillance syndromique en santé publique.

Mots clés : Données massives, Machine Learning, Modélisation statistique, Aide à la décision, Santé publique, Surveillance syndromique

Résumé : Depuis plusieurs années, la notion de Big Data s'est largement développée. Afin d'analyser et explorer toutes ces données, il a été nécessaire de concevoir de nouvelles méthodes et de nouvelles technologies. Aujourd'hui, le Big Data existe également dans le domaine de la santé. Les hôpitaux en particulier, participent à la production de données grâce à l'adoption du dossier patient électronique. L'objectif de cette thèse a été de développer des méthodes statistiques réutilisant ces données afin de participer à la surveillance syndromique et d'apporter une aide à la décision.
Cette étude comporte 4 axes majeurs. Tout d'abord, nous avons montré que les données massives hospitalières étaient très corrélées aux signaux des réseaux de surveillance traditionnels.
Dans un second temps, nous avons établi que les données hospitalières permettaient d'obtenir des estimations en temps réel plus précises que les données du web, et que les modèles SVM et Elastic Net avaient des performances comparables.
Puis, nous avons appliqué des méthodes développées aux Etats-Unis réutilisant les données hospitalières, les données du web (Google et Twitter) et les données climatiques afin de prévoir à 2 semaines les taux d'incidence grippaux de toutes les régions françaises.
Enfin, les méthodes développées ont été appliquées à la prévision à 3 semaines des cas de gastro-entérite au niveau national, régional et hospitalier.

Title : Statistical models for decision support systems based on the reuse of Health Big Data : Application to syndromic surveillance in public health.

Keywords : Big Data, Machine Learning, Statistical modelling, Decision support, Public Health, Syndromic surveillance

Abstract : Over the past few years, the Big Data concept has been widely developed. In order to analyse and explore all this data, it was necessary to develop new methods and technologies. Today, Big Data also exists in the health sector. Hospitals in particular are involved in data production through the adoption of electronic health records. The objective of this thesis was to develop statistical methods reusing these data in order to participate in syndromic surveillance and to provide decision-making support.
This study has 4 major axes. First, we showed that hospital Big Data were highly correlated with signals from traditional surveillance networks.
Secondly, we showed that hospital data allowed to obtain more accurate estimates in real time than web data, and SVM and Elastic Net models had similar performances.
Then, we applied methods developed in United States reusing hospital data, web data (Google and Twitter) and climatic data to predict influenza incidence rates for all French regions up to 2 weeks.
Finally, methods developed were applied to the 3-week forecast for cases of gastroenteritis at the national, regional and hospital levels.